

Tests worth teaching to

incentivising quality in qualifications
and accountability

Edited by

Gabriel Heller Sahlgren

With contributions from

Dale Bassett

Robert Coe

Gabriel Heller Sahlgren

Geoffrey Holden

Tim Oates

J. R. Shackleton



The Centre for Market
Reform of Education

First edition published 2014
by The Centre for Market Reform of Education Ltd

ISBN 978-1-63068-737-2

The moral right of the authors has been asserted.

All rights reserved. No part of this publication may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form, or by any means (electronic, mechanical, photocopy, recording, or otherwise) without the prior permission of the publisher. Any person who does any unauthorised act in relation to this publication may be liable to criminal prosecution and civil claims for damages.

A CIP catalogue record for this book is available from the British Library.

Every care has been taken that all information was correct at the time of going to press. The publisher accepts no responsibility for any error in detail, inaccuracy or judgement whatsoever.

Typeset by Great White Designs, Stroud, Gloucestershire
Printed and bound in Great Britain by Intype Libra Ltd, Wimbledon 2014

ABOUT THE AUTHORS

Dale Bassett is Head of Public Policy at AQA, the awarding body and education charity. A specialist in education policy, he has written extensively on structural reform of the schools system, as well as on issues related to curricula, qualifications, and teaching.

Robert Coe is Professor in the School of Education and Director of the Centre for Evaluation and Monitoring (CEM) at Durham University. His research interests include evaluation methodology, evidence-based education policy, and the statistical comparability of examinations in different subjects and over time. Before embarking on an academic career, he was a teacher of mathematics, with experience in a range of secondary schools and colleges.

Gabriel Heller Sahlgren is Director of Research at the Centre for Market Reform of Education and Affiliated Researcher with the Research Institute of Industrial Economics in Stockholm, Sweden. He is the author of *Incentivising Excellence: School Choice and Education Quality* (CMRE and IEA 2013), among other publications on issues relating to applied microeconomics.

Geoffrey Holden is Senior Policy Advisor at City & Guilds, the leading vocational awarding body. Among his broad interests in education policy, he has a particular interest in 14-19 education and ensuring that there is a high quality vocational offer for all young people.

Tim Oates was Director of Research at the Qualifications and Curriculum Authority from 1997 until he joined Cambridge Assessment in 2006 as Group Director of Assessment Research and Development. He recently chaired the expert panel providing advice to the Secretary of State on revisions to the National Curriculum in England. He is also Visiting Professor at the University of Leeds and Fellow of Churchill College, Cambridge.

J. R. Shackleton is Professor of Economics at the University of Buckingham and editor of *Economic Affairs*. A specialist in labour market economics, he has been a Dean of two business schools and worked as an economist in the Civil Service. He has also been involved with schools examinations for many years.

CONTENTS

Introduction: the other education market and how to improve it	1
<i>Gabriel Heller Sahlgren</i>	
1 The history of qualifications and the role of competition	6
<i>J. R. Shackleton</i>	
Introduction	6
The early history of school exams	7
From GCE and CSE to GCSE, and associated administrative and regulatory changes	10
How A Levels have changed	15
Have standards fallen?	17
Is competition good or bad?	20
Conclusion	24
References	26
2 The ‘qualifications sledgehammer’: why assessment-led reform has dominated the education landscape	28
<i>Tim Oates</i>	
Introduction: assessment and qualifications as drivers for education reform	28
Taking the brunt of it: the unmanageably contradictory pressures imposed on qualifications and assessment	29
Modularisation: a case in point	35
The importance of a coherent incentive structure: what we can learn from Finland	38
Conclusion: authentic piloting and responsiveness to unintended consequences	42
References	43
3 Regulatory overkill: school accountability, qualifications, and the future	46
<i>Dale Bassett</i>	
Introduction: view from the back seat	46
A bang and a whimper	48
Homogenisation: or, how ‘standards’ have trumped quality	49
Oligopoly: or, how high-stakes accountability stifles the benefits that might accrue in a functioning qualifications market	52
Slaying the accountability behemoth	54
Conclusion: producing qualifications that support education	55
References	57

4 The vocational question: in pursuit of quality rather than equivalence	58
<i>Geoffrey Holden</i>	
Introduction	58
Qualification-based reform and the growing role of the state	60
A broad education system	63
Levels and the myth of equivalence	66
The problem of competence	67
Knowledge- and curriculum-based reform	70
Conclusion and policy solutions	71
References	73
5 Incentives and ignorance in qualifications, assessment, and accountability	74
<i>Robert Coe and Gabriel Heller Sahlgren</i>	
Introduction	74
Quality and purposes of qualifications and assessment	76
The issue of accountability	82
Potential advantages and problems of accountability systems	82
Evidence on the impact of accountability	84
Features of accountability systems	87
Reconciling educational goals with demands of accountability	91
Conclusion	94
References	95

INTRODUCTION: THE OTHER EDUCATION MARKET AND HOW TO IMPROVE IT

GABRIEL HELLER SAHLGREN

THE DEBATE ABOUT CHOICE and competition in education is in most countries confined to the issue of school choice. School choice produces competition between schools, which forms the essence of what is normally described as an education quasi-market, characterised by public funding combined with consumer choice among different providers. Whether or not this market should be promoted has become one of the biggest controversies of education policy of recent decades.

Yet in England, Wales, and Northern Ireland, there is another education quasi-market, the existence and dynamics of which have increasingly fuelled debate. For many years, schools in these nations have had the right to decide which qualifications their pupils take from a range of options offered by multiple independent providers, rather than a single, government board, as is the international norm. This allows a measure of diversity in assessment and qualifications, stimulating competition among different exam boards, and providing the essential features of ‘the other education market’ referenced in the title of this introduction. Whereas the debate about school choice focuses heavily on competition in regard to where pupils will study, the debate about qualification and assessment choice is fundamentally about competition in regard to what pupils will study and how they will be assessed.

However, in the last couple of years, competition in qualifications and assessment has been the subject of increasing criticism. Because of perverse

incentives to maximise the number of customers, critics argue that it induces exam boards to dumb down their examinations and inflate grades. Furthermore, as consecutive governments have sought to expand parental choice of schools, while simultaneously increasing school accountability, the pressure to raise achievement is higher than ever. Instead of working to raise quality, it is argued, exam board competition leads to a ‘race to the bottom’.

At the same time, proponents of markets generally point to the oft-observed failures of monopolies in producing innovation and serving their customers well. In a monopoly situation, in which all schools sit exactly the same examinations set by a single national board, it is not clear what would incentivise higher quality. The only way to do this is by government diktat. But even if politicians wanted to improve quality unconditionally, which is not clear given the political incentives at work, it is still unlikely that government is equipped for the task. Instead of market failures, we might just end up with government failures – which can be much more far-reaching and difficult to correct. As in any other field, many would therefore find it difficult to see how the qualifications and assessment system would benefit from monopolisation.

Yet it is clear that all competition is not equal: the structure and design of the market matters immensely for whether boards have incentives to raise quality. At present, the strength of accountability and regulation in the qualifications and assessment market means that incentives are currently produced by a mishmash of government diktats and market forces.

This monograph discusses the current arrangements in regard to qualifications, assessment, and accountability in England, and how we can improve the incentives at play in order to raise quality. The contributions differ in their specific prescription as to how to improve the system, but they all agree that choice and competition in qualifications and examinations are not the fundamental problem. Instead, all highlight the problems arising as a result of detailed and yet ill-informed, government intervention in the market. Clearly, politicians of every hue need to take more care and work to develop a healthy overall framework within which competition might work more constructively.

In the first chapter, Professor J.R. Shackleton of the University of Buckingham sets the present arrangements in their historical context, showing how the current diversity of qualifications and assessments on offer has its roots in a time when state

intervention in education was minimal. Spontaneous developments early on led to diverse qualifications, and were gradually influenced by government policy to steer these developments towards specific goals. Shackleton argues that the problem today lies not with exam board diversity and competition, but rather with pervasive government meddling in the market and constant flip-flopping policy changes. The solution is therefore not to curtail or remove competition, but rather to extend it, while leaving the broad structure of the system alone and resisting further intervention.

Advancing this argument further, Cambridge Assessment's Tim Oates argues that problems have arisen as a result of the government co-opting the qualifications and assessment system to drive education reform more broadly. Reforms have often created new problems instead of solving old ones, so further tinkering is not the way forward. Instead, the government must pay greater attention to the overall incentive structure across all parts of the education system, to ensure that it works to encourage all actors to pull in the same direction and improve quality. In sharp contrast to the lessons generally drawn from the Finnish experience, Oates argues that the country's rise in international league tables was due precisely to policymakers' attention to the coherence between different parts of the system. Due to the dangers of unintended consequences, he also argues in favour of using pilot programmes before reforms are scaled up to the national level.

In the third chapter, Dale Bassett of AQA argues that detailed government regulation of content and grading has ensured that innovation and diversity have been strangled at the altar of school accountability to uniform standards. While the emphasis on regulation and standardisation has ensured that fewer young people today fail outright in academic terms, it has produced a situation in which schools do not have meaningful and effective choices in respect of the qualifications on offer. It has also prevented competition on quality in respect of content and level of difficulty. The accountability system places too heavy an emphasis on high-stakes examinations and comparability. It must be reformed to focus more on the long-term success of pupils beyond school. Changing the focus in this way and reducing micro-regulation, Bassett argues, could produce a qualifications market that better serves pupils' needs, rather than the preoccupations of government.

The dangers of equivalency are also emphasised in the fourth chapter by City & Guild's Geoffrey Holden. Addressing the issue of vocational education, he shows how, due to the perceived need for the two to be comparable, consecutive governments have

attempted to force vocational qualifications into a framework of standards designed for their academic counterparts. This is a futile quest, since vocational and academic knowledge and skills cannot be made equivalent. The question should rather be whether pupils acquire the right knowledge and skills to do their specific jobs well. Yet Holden also argues that the distinction between academic and vocational has been made worse by policymakers, who have focused too much on competency and skills in vocational education, rather than on the broader academic knowledge that underpins them. Such knowledge is crucial for pupils to be able to progress, both within and beyond the sector for which they are educated. The flurry of changes in government policy on vocational education, Holden argues, displays the futility of a top-down approach in general. Only by refraining from such interventions can the government induce healthy incentives in the private sector to improve quality.

In the final chapter, Professor Robert Coe of Durham University and I advocate an evidence-based and experimental approach to reform. In order to empirically evaluate whether a qualification is fit for purpose, it is important first of all to define the different purposes that qualifications serve. Having outlined suggestions in this respect, we argue that exam boards should be required to state which purposes their qualifications and assessments are supposed to fulfil, and show evidence of how well they do. Assessments used for accountability purposes should also be designed to meet clear quality criteria, and exam boards should again be asked to provide evidence regarding the extent to which they meet those criteria. Because of our ignorance regarding the optimal design of accountability structures, we also call for randomised pilot programmes in which schools are exposed to different accountability features in order to find out which work best. Similarly, squaring educationally desirable practices with the demands for high-stakes accountability will in many cases require similar experimentation. We discuss the example of teacher assessment, and offer suggestions as to how it can be reconciled with the demands of accountability.

Overall, it is clear that the current state of affairs is far from the unregulated market that opponents sometimes conjure up when arguing for monopolisation and increased regulation. Instead, the authors show various examples of how continuously increasing regulation and tinkering with the system at the central level have produced more problems than they have solved. Government is clearly trapped in a vicious circle of continuous reform. The solution is therefore not to dismantle

or curtail choice and competition further, but to improve the incentives at work. In other words, rather than more intrusive regulation, we clearly need leaner and smarter regulation.

For example, it seems increasingly clear that the conceptual framework maintaining that all qualifications are equivalent and the government apparatus designed to ensure this equivalency prevents exam boards from differentiating themselves on quality. For strong quality competition to occur, it therefore seems highly likely that this framework will have to be dismantled.

Yet, as noted by the monograph's contributors, the idea that we can design the perfect system via a top-down approach is foolish. Instead of promoting yet another hubristic Great Leap Forward in qualifications, assessment, and accountability reform, the solution is to take a more experimental approach to education policymaking overall. Taking a broadly light-touch regulatory approach, we can trial different measures on a local scale and avoid the mistakes of the past. By using pilot schemes we can road-test different ideas by subjecting them to competition with each other, and thereby advance our knowledge about what works.

Some argue that we should not experiment with children's education. The problem with this argument is not only that we will never be able to improve quality significantly without it, but also that it is a fallacy to believe that the status quo or nation-wide reforms do not represent experiments. The status quo has of course never been rigorously tested, meaning that support for it rests on belief rather than evidence. National reforms, meanwhile, tend to be large-scale experiments that are very difficult (sometimes impossible) to properly evaluate. This situation might suit politicians, since it means that they cannot be held accountable for failure, but it is not in the best interests of children.

An evidence-based education policy requires good evidence, and the only way to amass that evidence is to allow more experimentation. This means that politicians' ideas regarding what they think works in education should always be put to the test in carefully designed trials before they guide national education policy. In other words, in order to ensure quality in assessment and qualifications, and square these with demands for accountability, we need a market approach to education policymaking too.

I THE HISTORY OF QUALIFICATIONS AND THE ROLE OF COMPETITION

J. R. SHACKLETON¹

Introduction

THIS CHAPTER LOOKS IN detail at the roots of today's schools examinations system, and assesses the arguments regarding a continuing role for competition in the delivery of qualifications and assessment. Examinations in England, Wales, and Northern Ireland – though not Scotland – are unique in the sense that schools can choose between different awarding bodies offering their own versions of qualifications, which are accepted for state funding and as entry tickets to universities, professions, and a wide variety of employment. According to the House of Commons Education Committee (2012), no other country offers such choice.

This is not the result of a deliberate neo-liberal policy. It is rather the consequence of the historical development of education in England, and the early initiatives of universities and other independent institutions at a time before heavy state involvement in education became the norm.

But is it sensible that we should continue to be an outlier in this respect? Critics argue that competition between boards has been at least partly responsible for perceived 'grade inflation'. In this view, standards have dropped as awarding bodies have competed for customers by offering narrow and dumbed-down syllabi, coupled with generous marking. The solution, according to the critics,

¹ The author thanks Gabriel Heller Sahlgren, Quintin Brewer and Peter Maunder for helpful comments.

is therefore to eliminate competition and have a single awarding body for each qualification.

However, a case can be made that awarding bodies are dynamic innovators and that competition between them benefits pupils, schools, employers, and other parties. If there is a problem with grade inflation, and (more importantly) with insufficiently high levels of educational achievement, it may rather have more to do with constant government meddling, contradictory and short-lived policy initiatives, and over-fussy regulation that have produced perverse incentives in the system. The solution is not to suppress competition and impose a qualifications and assessment monopoly, but instead to relax micro-regulation and obsessive short-term tinkering with the exam system. Instead of continuing the perpetual reforms in qualifications and examinations, it would be wise to focus our attention on arguably more important issues, including how to increase the quality of teachers and expand parental choice further.

The early history of school exams

In order to understand the current system of qualifications and examinations in England, Wales, and Northern Ireland, and what we should do to improve it, it is useful briefly to survey its modern history. The current system emerged out of a combination of spontaneous developments and increased government intervention to steer those developments. Consequently, there have been many changes in the system over the years.

Beginning in the mid-19th century, there was a determined move towards modernisation of many English institutions, and to a more meritocratic way of allocating jobs and positions of influence. The Army, the Civil Service,² and the universities were all affected by a drive for greater openness and promotion on merit and achievement rather than family- or patron-based influence.

In this climate, schools looked to raise educational standards. This seemed to involve a role for external examinations – schools sought guidance primarily from the universities, which responded with enthusiasm. In 1857, the

2 The Northcote-Trevelyan reforms, from 1855, required recruitment to the Civil Service to be ‘entirely on the basis of competitive examinations’ (Civil Service 2014). Within a few years, the reforms were said to have ‘eliminated all dunces’ – perhaps a rather optimistic conclusion.

University of Oxford Delegacy of Local Examinations commenced operations. It was quickly followed by the University of Cambridge Local Examinations Syndicate in 1858,³ and the University of Durham Matriculation and School Examination Board (also in 1858). Other bodies followed in the late 19th and early 20th centuries.⁴ In 1902, the University of London Extension Board was founded, to be followed by the Joint Matriculation Board (by Manchester, Leeds, and Liverpool universities) in 1903.

The early boards were usually small, with plenty of direct involvement by quite senior academics who set and marked papers. Administrators played a limited role and schoolteachers were hardly involved at all, although in some cases headmasters of leading schools sat on councils or committees.

The system seems to have worked reasonably well, but to the bureaucratic mind it was untidy and there were no common standards. In 1911, the Consultative Committee on Examinations in Secondary Schools called for national qualifications to be offered, albeit mainly by university-run bodies. Accordingly, the School Certificate (taken at 16) and the Higher School Certificate (taken at 18) were introduced in 1918.⁵ These awards were overseen by the Secondary Schools Examinations Council, the first of many regulatory bodies (see Box 1).

The inter-war years were relatively stable for school examinations in comparison to what would be the case later on. The pre-existing boards offered the School Certificate and remained largely unchanged, although some were renamed.⁶ Over this period, the number of pupils taking the School Certificate and Higher School Certificate rose slowly, but as compulsory schooling only

3 This was made up of thirteen university academics who set regulations, wrote question papers, marked scripts and made awards. Examinations were held in December (to avoid conflict with summer university exams, initially at venues in Birmingham, Brighton, Bristol, Cambridge, Grantham, Liverpool, London and Norwich. Incidentally, Cambridge is the only university still involved in schools awards, through its role in OCR (Cambridge Assessment 2014).

4 Most of these were university spin-offs, although the Central Welsh Board was founded in 1896 by Welsh local authorities. Some pupils also sat examinations set by professional institutes.

5 The Certificates covered a range of subjects, and resembled the continental baccalaureate model rather than the stand-alone subject awards (O and A levels) developed after World War II.

6 The University of London Extension Board became the University of London Matriculation and Schools Examination Council, while the University of Durham Matriculation and School Examination Board became the marginally snappier Durham University Examinations Board.

Box 1: Regulatory bodies

The Secondary Schools Examination Council, set up in 1917 to oversee the new School Certificate qualification, operated for over 45 years until its responsibilities were taken over by the Schools Council for Curriculum and Examinations in 1964. In 1982, the Schools Council was split into two new bodies: the Secondary Examination Council and the School Curriculum Development Committee. In 1988, these bodies were replaced by, respectively, the Schools Examination and Assessment Council and the National Curriculum Council. Only five years later, in 1993, the examination and curriculum functions were combined into the Schools Curriculum and Assessment Authority (SCAA). This, in turn, was merged in 1997 with the National Council for Vocational Qualifications to form the Qualifications and Curriculum Authority (QCA). The QCA lasted until 2009, when functions were divided up again with the introduction of the Office of Qualifications and Examination Regulation (Ofqual) and the Qualifications and Curriculum Development Agency (QCDA). The QCDA absorbed another body, the National Assessment Agency, which had been set up in 2004 to develop National Curriculum tests. However, it did not last long, being abolished under the Coalition in 2012. National Curriculum tests are now the responsibility of the Standards and Testing Agency while part of the examination administration responsibility has gone to the Teaching Agency.

went up to age 14, most young people had no opportunity to gain these academic qualifications.⁷ Some were able to obtain various commercial or technical qualifications. But most inter-war pupils left school without any certification of their abilities and achievements.

In 1941, a Secondary Schools Examination Council committee investigated secondary-school examinations and curricula as part of the early thinking about the post-war period. Its approach tied in with proposals embodied in the 1944

⁷ Before H. A. L. Fisher's Education Act of 1918 enforced compulsory education to 14, the school leaving age was 12. Intriguingly, Fisher also proposed compulsory part-time education from 14 to 18. This was scrapped because of public spending cuts in the aftermath of the First World War.

Education Act, introducing the tripartite structure of grammar, technical, and secondary modern schools. A case was made for abandoning external examinations in favour of schools experimenting with internal assessment, but this was not widely supported and was eventually vetoed by the Headmasters Association.

From GCE and CSE to GCSE, and associated administrative and regulatory changes

In 1947, the government agreed to introduce a new General Certificate of Education (GCE) at Ordinary, Advanced, and Scholarship levels.⁸ The papers differed significantly from the school certificates, as candidates would sit discrete subject-based papers that allowed for greater choice and depth. The standard was also to be pitched differently: an O-level pass, for example, was to be made equivalent to 'credit' level in the old School Certificate. O levels were clearly aimed at grammar and independent school pupils with a strong academic bent, intending to go on to university and/or professional and management careers.

In 1951, pupils sat GCE papers for the first time, awarded by the same exam boards as the school certificates. These were primarily the university-based English boards plus the Welsh Joint Education Committee, which had replaced the Central Welsh Board in 1948, and the Northern Ireland Schools Examination Council. However, over the next few years, there were some changes: the Associated Examining Board (AEB) was set up by City and Guilds⁹ in 1953; the Southern Universities' Joint Board was founded in 1954 as a successor to the University of Bristol School Examinations Council; and ten years later the Durham University Examinations Board closed down.

The GCE system was never intended for the 75 per cent of the age group in secondary modern schools overwhelmingly taught by non-graduates.¹⁰ The

8 The Scholarship 'S' level was introduced in a limited number of subjects for the top A-level pupils, initially as the basis for awarding state scholarships at university. Papers were graded as Distinction, Merit, or Unclassified. The state scholarship awards ceased in 1962 with reform of pupil funding, and the papers were afterwards known as Special (again 'S') level. The last S levels were sat in 2001.

9 The City and Guilds of London Institute was founded in 1878, to promote technical and vocational education. It has been involved in many of the most important developments in UK education and still has over two million pupils.

10 Fewer than 20 per cent of secondary modern teachers were graduates, compared with almost four out of five grammar school teachers (Brooks 2008).

Norwood Report (1943), which had laid out the basis for the 1944 Education Act, argued that secondary modern schools should be unfettered by external examinations and conduct bold experiments with internal assessment instead. However, this was not what young people and their parents wanted. Many who had failed the 11-plus selection process tried, some successfully, to transfer to grammar schools at 12 or 13. Of those remaining in secondary modern schools, a good number sought widely recognised qualifications to improve their job prospects. External vocational certificates – such as Pitman awards and nursing certificates – were one possibility. Some areas developed their own school leaving awards that had some regional currency.

But many pupils wanted to take O levels. There were considerable hurdles, since secondary modern pupils normally left school at 15 – a year before O levels were normally taken – and there was only limited financial support available for them. Yet by 1954, from very low initial levels, more than 5,500 candidates from 357 secondary moderns entered for O levels. By 1958, the number of candidates had risen to nearly 17,000, and by 1960 almost 40 per cent of pupils were attempting some O level awards (Brooks 2008).

The Beloe Committee (1960) – set up by the Secondary Schools Examinations Council to consider the ‘problem’ of the proliferation of examinations in secondary modern schools – called for a new external examination system below GCE. In the Committee’s view, GCEs were suitable for the top 20 per cent of the age group, and new subject-based awards were to focus on the next 40 per cent of the cohort. The Committee argued that the new qualifications ‘should largely be in the hands of teachers serving in the schools which will use them’ (p. 47). At the same time, regional examinations boards, independent of the existing GCE awarders, were advocated. On the governing bodies of these new boards, there would be representatives of teachers, local education authorities, further education institutions, training organisations, and employers. Each regional body, serving a defined area, were in effect to have a local monopoly on sub-GCE awards.

The Certificate of Secondary Education (CSE) duly came into being in 1965, offered by a large number of local boards. These were completely new

organisations, with the exception of those in Northern Ireland and Wales.¹¹ There were five pass grades in the CSE. The top grade in common subjects like mathematics and English was considered equivalent to an O-level pass. It was very much a teacher-led qualification at all stages: teachers sat on subject committees, approved papers, and marked them (Tattersall 2008). The range of courses for CSE was wider than that for O levels, and included many vocational subjects. In a significant development, individual schools were able to set their own ('Mode 3') syllabi, subject to approval from their regional exam board. This also meant that assessment methods moved away from heavy reliance on unseen examination papers.

Over the next twenty years, comprehensive schools increasingly replaced grammar and secondary modern schools, and in 1972 the school leaving age was raised to 16. The distinction between the 'O level candidate' and the 'CSE candidate' thus became more blurred, with many pupils taking a combination of CSEs and O levels. Accordingly, the qualifications were merged from 1987, when the award of General Certificate of Secondary Education (GCSE) was introduced. In the following year, the Education Reform Act introduced the National Curriculum, which had significant implications for 16–18 school examinations, such as much greater centralisation of syllabus approval and prescribed learning outcomes at both GCSE and A level. The Act also led to a new set of assessments (SATs) for younger children at Key Stage 1 (age 7), Key Stage 2 (age 11), and Key Stage 3 (age 14).¹²

The introduction of the GCSE led to further institutional change, as CSE and GCE boards joined together in new examining groups. For example, the

11 In Wales and Northern Ireland, the universities had not been involved in school examinations and the boards were already much closer to schools. There were thirteen new CSE boards. Their numbers were reduced slightly in 1979 when the Metropolitan and Middlesex boards merged in 1979 to form the London Regional Examinations Board, while in 1982 the West Yorkshire and Lindsey and Yorkshire and Humberside boards coalesced into the Yorkshire Regional Examinations Board.

12 SATs (Standard Assessment Tasks), now National Curriculum Tests, were set and organised by central government agencies rather than the existing exam boards. These have gone through several changes and are now the responsibility of the Standards and Testing Agency. Tests at Key Stage 3 have been abandoned.

The other examining groups were the Midland Examining Group, the Northern Examining Association, and the Southern Examining Group. Each was made up of a mixture of regional CSE boards and older GCE boards. The Northern Ireland Schools Examination Council and the Welsh Joint Education Committee, which already offered both GCE and CSE, continued as before.

University of London School Examinations Board (responsible for GCEs) linked with two regional CSE boards, the London Regional Examinations Board and the East Anglian Examinations Board, to form the London and East Anglian Examining Group.¹³

The merger of GCE and CSE lent impetus to the move towards assessed coursework. Although the Joint Matriculation Board had experimented with some O-level coursework in the early 1960s, it was standard in CSE. In the 1990s, it became the norm for GCSE assessment to include coursework, and the practice spread to A levels as well. This presented considerable logistical and moderation challenges to examinations boards.

The late 1980s and 1990s also saw major changes in GCE A levels, detailed in the next section, with the introduction of AS awards and, later on, the development of modular A and AS qualifications. A-level entries have risen enormously since the early 1950. At that time, just over 100,000 candidates sat these exams; the figure today is in excess of 850,000 pupils.

Another element was the perceived need to give secondary education a more vocational flavour. In 1986, the National Council for Vocational Qualifications was set up. It developed a framework for locating all qualifications within five levels of 'competence' or 'standards of performance'. GCSEs were at level 2 on a five-point scale, alongside basic craft certificates, with A levels at level 3 alongside long-established qualifications like Ordinary National Diplomas. In addition, new A level-equivalent qualifications were later developed for schools and colleges.

This rapidly changing environment led to yet more changes in exam boards as they reshaped to face the challenges of new markets and new regulatory requirements. One important common feature was the gradual disengagement of universities from boards. For example, the University of London School Examinations Board formally merged in 1991 with the London and East Anglian Group to form the University of London Examinations and Assessment Council (ULEAC). In 1996, ULEAC merged with the vocational Business and

13 The other examining groups were the Midland Examining Group, the Northern Examining Association, and the Southern Examining Group. Each was made up of a mixture of regional CSE boards and older GCE boards. The Northern Ireland Schools Examination Council and the Welsh Joint Education Committee, which already offered both GCE and CSE, continued as before.

Technology Educational Council to form a new educational charity, the Edexcel Foundation, with a much wider remit and no connection to the University of London. In turn, Edexcel was taken over in 2003 by Pearson PLC, the British-based multinational publishing and education company which publishes the Financial Times. It was renamed Pearson Edexcel, but from April 2013 is officially known simply as Pearson. It no longer has any connection with the University of London.

No other exam board is run as a profit-making institution, but most of them have also undergone considerable change. They all prefer short titles telling little about their origins. The organisation with the largest share of the UK schools market is AQA (the Assessment and Qualifications Alliance), which in 2011 had 45 per cent of GCSE and 42 per cent of A-level candidates. AQA is an amalgam of several older boards such as the Joint Matriculation Board, the Associated Examining Board, and the Oxford Delegacy, and has been closely associated with City and Guilds. The other major English awarding body is OCR (Oxford, Cambridge and RSA Examinations), which grew out of the University of Cambridge Local Examinations Syndicate, the Southern Universities' Joint Board, the A level section of the Oxford Delegacy, and the Royal Society of Arts Examination Board. It is owned by the University of Cambridge, but managed at arm's length through its Cambridge Assessment division, which also runs Cambridge International Examinations.

Another smaller private body is the International Council for the Accreditation of Academic Evaluation, owned by an accountancy body. The Northern Ireland Schools Examination Council eventually transmogrified into the Council for the Curriculum, Examinations and Assessment, now a non-departmental government body which dominates schools assessment in the province. Finally, the Welsh board, the WJEC, has changed the least and is still owned by the Welsh local authorities, although it has operational independence. Between them, these awarding bodies now handle about 15 million scripts a year.

Clearly, therefore, the current framework of qualifications and examinations has been produced by a combination of spontaneous developments and government intervention to direct these towards specific goals. Overall, the processes have meant that England, Wales, and Northern Ireland have a market for qualifications and curricula, which is unique in the developed world.

How A levels have changed

If the awarding institutions and their regulators have changed, the qualifications they offer have changed even more. For example, when GCE A levels were introduced, they were only graded as 'pass' or 'fail'. In 1953, a further grade of 'distinction' was introduced. In 1963, a five-grade scheme was introduced, with pass grades of A-E and an Ordinary-level award below the passing grades.

These grades were 'norm-referenced', meaning that fixed quotas of candidates received particular grades. The top 10 per cent received an A, the next 15 per cent a B, the next 10 per cent a C, the next 15 per cent a D and the next 20 per cent an E. This meant that 70 per cent passed the examination. A further 20 per cent were awarded the O-level consolation prize, and the last 10 per cent received no award.

In the 1980s, this system came under sustained attack. Far more candidates were taking A levels, and it seemed demotivating and unfair to condemn 30 per cent of them automatically to failure. In 1987, this complaint was addressed by the introduction of a hybrid system, which combined 'criterion-referencing' with statistical boundaries. Criteria were introduced to distinguish the qualities dividing an A from a B, and an E from a fail. Examining teams were to use their academic judgment on a sample of scripts to determine where these boundaries lay. The other grades were then parcelled out on the basis of fixed percentages. At the same time, the substitution of GCSEs for O levels led to A-level 'near-misses' being labelled N rather than O.

In another development, Advanced Supplementary (AS) awards were introduced in 1989. These qualifications were intended to take two years of study alongside A levels. The plan was to give pupils a broader post-compulsory education by letting them take one or two AS levels in addition to a relatively narrow diet of A levels. Ideally, for example, pupils who were taking three science A levels would also take an arts subject at AS.¹⁴ The content of the courses was intended to be half that for an A level, but at the same level of difficulty.

However, the advent of 'Curriculum 2000' changed things dramatically. Modularity was introduced across the post-compulsory examination system. Year-12 pupils now took three AS-level modules in each subject. They could exit

¹⁴ One of many attempts to get pupils to broaden their sixth-form studies, it was not notably successful.

with an AS based on these modules, or take three more 'A2' modules in Year 13 to achieve a full A-level award. Clearly, A2 modules were now at a different level from AS modules. Moreover, it was possible to re-sit modules to improve marks. Overall grades were aggregated across modules, but the best attempt at a module would count when grades were 'cashed in'. Grades were now supposed to be fully criterion-referenced, though in practice statistical considerations still played an important role in determining boundaries. The N grade was abolished. The first new-style AS awards were made in 2001, with A2s awarded in 2002.

The system proved popular with universities, as it enabled offers to be made on the basis of candidates' year-12 AS grades. However, it also involved an excessive amount of assessment, and so after 2008 most subjects involved two AS modules and two A2 modules, with the previous six-module content hurriedly shoehorned into four modules.

A further change was the introduction of Advanced Extension Awards (AEAs) in 2002. These were intended to replace the previous S-level papers for top pupils. The AEAs were in turn scrapped in 2010, when the A* grade was introduced at A level to enable universities to differentiate between the large numbers now achieving grade A. Around 8 per cent of all candidates now achieve A*, which is roughly the same proportion that achieved the original A grade back in the 1950s. As Smithers (2012, p. 2) puts it, 'A* has, in effect, become the new A'.

Finally, Secretary of State Michael Gove has decided to end the modular system and return – with effect from 2015 or 2016 depending on subject – to a 'linear' A level. This involves decoupling the AS from the A level award. Coursework, which came into vogue in the 1980s, is also being phased out.

There are yet more complications, such as the development of various vocational qualifications, which were rebadged as 'vocational A levels', and the provision of international A-level awards by Cambridge International Examinations and Pearson. But enough has been said to show that the A level has changed very considerably over the sixty-odd years of its life.

A similar, and even more convoluted, story could be told about O levels and GCSEs. For example, until 1975 different GCE boards had different grading systems for O level. The new GCSE awards after 1987 involved tiered papers,

with weaker candidates taking the lower tiers and answering easier questions and only able to obtain a C grade. These convolutions continue: Mr Gove has insisted that the current GCSE alphabetical grades are going to be replaced by a larger number of numerical grades.

Have standards fallen?

Have all these changes worked? Critical comment about the schools examination system has a number of dimensions. Employers complain that pupils with apparently good GCSE and GCE A-level qualifications have problems with writing and numeracy. Top universities complain that even pupils with excellent A levels no longer have the depth of knowledge and understanding necessary to undertake university study, and need remedial support and a ‘dumbed down’ first year in order to progress. Meanwhile, teachers complain that assessments are too narrow and encourage ‘teaching to the test’ rather than enabling them to develop pupils’ interests and wider abilities.¹⁵

Figure 1 shows the proportions of pupils achieving particular GCE A-level grades over a 23-year period: it shows an almost continuous improvement in pass rates (which mostly applies to GCSEs as well). Yet this was achieved despite a substantially larger proportion of the age group taking these qualifications. How is this possible?

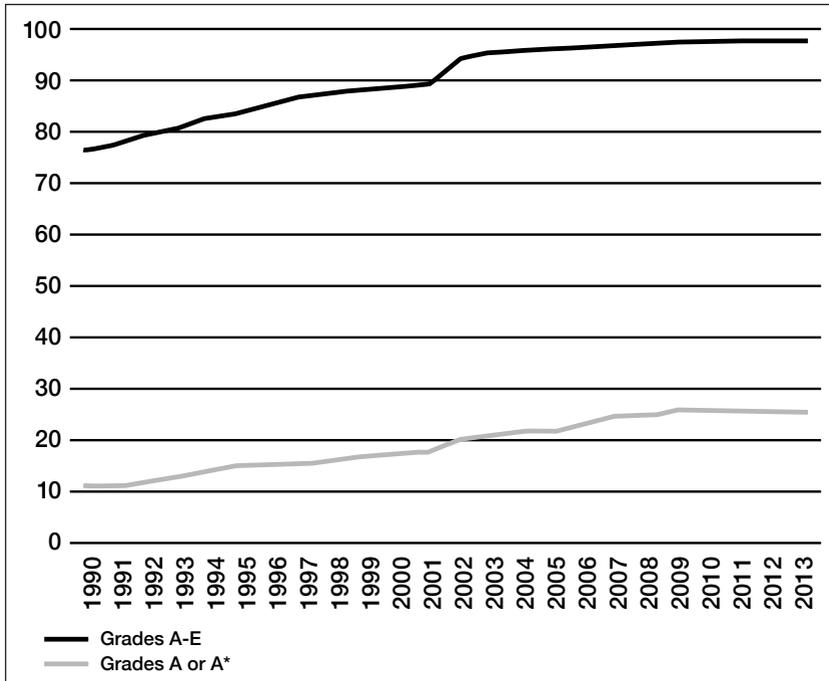
Educationalists sometimes claim that results have improved because of better teaching, better resources – such as smaller classes, more appropriate textbooks, and online support – and more effort by pupils who increasingly realise the importance of hard work if they are to progress to higher education or employment.

These factors may have had some impact, but it seems unlikely that they can account for improved pass rates and grades on the scale illustrated in Figure

15 State schools are judged on league tables showing, for example, the proportion of pupils getting A*-C results at GCSE. This forces teachers into an instrumental approach to ensure that as many pupils as possible achieve these grades, possibly neglecting brighter and lower-achieving pupils as a consequence. Independent and selective state schools, by contrast, have talented pupils who need a real challenge. This has led a growing number of such schools to opt out of the standard system, taking IGCSEs (International awards, without coursework, which resemble old O levels) at 16 and the International Baccalaureate at 18.

1. The reality is that standards have changed, and comparisons over anything other than the short term are well-nigh impossible.¹⁶

Figure 1
Proportion of A-level candidates achieving particular grades 1990–2013



Sources: Smithers (2012) and Department for Education.

Whereas A levels were originally conceived with the demands of university entrance selection in mind, their function today is much wider. The approach has increasingly been to enable pupils to ‘access the curriculum’.¹⁷ This means tightly prescribed syllabi and methods of assessment, which split marks according

¹⁶ Coe (2007) shows how standards have changed with respect to various different criteria, though he is agnostic about how these changes should be interpreted.

¹⁷ Universities and employers are consulted from time to time over specifications, but have little impact on assessment.

to a number of objectives that carry predetermined proportions of the marks.¹⁸ This makes assessment more predictable and limits the type of questions set and the type of responses rewarded. Open-ended questions, especially longer essay-type ones, have gradually disappeared and shorter questions, including multiple-choice and structured responses, are favoured. This enables candidates to build up marks in a predictable manner, and is increasingly emphasised in classroom teaching as pupils are prepared for examinations.¹⁹

Changes in the structure of awards can have noticeable effects on results. In Figure 1, for example, there is a visible upward shift in the overall A-level pass rate between 2001 and 2002, reflecting the impact modularisation.

There may also be impacts from the increasing use of statistical controls to regulate the proportions receiving grades. Examiners are required to set grade boundaries by examining a sample of borderline scripts and comparing them with papers at the boundary in the previous session. These decisions have to be related to published qualitative criteria. In practice, this is very difficult since papers awarded the same mark may have very different characteristics. For example, two strong answers and a very weak one must be compared with three moderate answers. Attention is focused on statistics showing what proportion of candidates would achieve a specific grade if mark X was set as the boundary. This is then compared with the proportion at the last examination session, while other statistics are introduced, such as teacher predictions, the proportion of candidates from examination centres with a strong previous record, and the GCSE results achieved previously by A-level candidates. The use of such data has increasingly influenced boundary decisions and may be implicated in the upward drift of grades.

The plus side of all this is that young people have been given confidence to progress in their studies. Far more school-leavers now have some qualifications to show for their long years of education. This is no mean achievement, and it

18 There are three, four, or five objectives depending on subject. For instance, Economics assesses Knowledge and Understanding, Application, Analysis and Evaluation.

19 As Sheldon (2011) notes, 'In order to standardise results and to ensure consistency of marking, mark schemes are now very detailed ... There is no examiners' discretion to reward a talented but unconventional response. In a recent case, an Oxford tutor reported on a piece of history work submitted by a candidate. It was some AS-level written work – large parts of excellent work had been crossed out by the student's A level teacher with the comment "not required for AS level".'

is not unreasonable to downplay older generations' complaints that their A level results have been devalued. Times change, and 50-year olds are not in the same labour market as today's young people. The fact that older people's A grade in History indicated that they were well inside the top 5-10 per cent of their age cohort, while a nominally similar grade might cover 15-20 per cent of today's candidates, is a curiosity but does not have much practical significance.

However, it may still be a problem if employers and others do not understand what is being measured, and how the current qualifications are organised and assessed. While the continuity of the GCE and GCSE brands is useful, it has obscured the significant changes catalogued here. Few people not directly involved can possibly be expected to follow the ins and outs of curricular and assessment changes. Constantly changing qualifications and examinations may therefore produce confusion, and in the end create more problems than the changes are solving.

It can also be a problem if politicians and teachers believe their own rhetoric about rising levels of achievement and become complacent. Many countries have been improving their domestic qualification results at least as fast as we have. Clearly, improved GCSE and GCE results should not be taken as evidence of progress towards a 'world class' educational system. It is salutary, for example, to note that while from 2006 to 2012 the proportion of candidates achieving A*-C in GCSE mathematics rose by around 5 percentage points, PISA data suggest that there was no significant change in average scores of UK pupils, while their world ranking slipped from 24th to 26th. Similarly, in the equivalent TIMSS tests, English pupils did not improve between 2007 and 2011 either (CEER 2013; IEA 2013; OECD 2013).

Is competition good or bad?

Because of concern about changing standards, people seek scapegoats. The awarding bodies are one target. It is asserted that unnecessary competition between these bodies contributed to changing standards, each pursuing increased market share and offering easier qualifications to achieve this.

To economists, suspicion of competition and markets is only too familiar. Is there any substance in this case? The Sykes Review for the Conservative Party

before the last election, the Walport Report (2010) and the House of Commons Education Committee (2012) have given voice to such suspicion. But in reality there is little evidence to support it. As the Walport Report (paragraph 101) says, '[W]e should emphasise that we do not have proof that competition between awarding bodies has affected grading standards'.²⁰ Indeed, available research does not support the view that exam boards lower their standards to compete for pupils (Malacova and Bell 2006).

The main indictment against competition between boards is that they put commercial interests before education, and generate undemanding syllabi with softer assessment. But this ignores the fact that all awards have to be approved by the regulator, nowadays Ofqual, and must conform to tight subject criteria. As noted above, the determination of grades is also tightly controlled. Other charges relate to information problems, which schools are said to face when making choices between boards; the possibility of boards pushing 'tie-in sales' of teacher seminars and learning materials; and excessive charges for examination entries. Yet such concerns are certainly not unique to the examination and assessment market, and regulatory interventions are possible – in some cases under general competition law – if there is genuine evidence of malpractice.²¹

The idea that commercial imperatives lead to a 'race to the bottom' is common amongst supporters of state control in many areas. But if a board was perceived to lower standards and give weak pupils generous grades, there could well be a reaction from universities and employers.²² Concern about this possibility is a deterrent to boards tempted to move downmarket, as is fear of losing international pupils, who are a very important and profitable section of customers, particularly in the case of OCR and Pearson.

Furthermore, the alternative of having a single assessment body for all schools examinations would not be without problems. There would be difficulties in

20 This lack of evidence did not stop Michael Gove from deciding that in future there should only be one syllabus for each GCSE core subject, and that awarding bodies should compete for franchises to run these qualifications – a proposal now abandoned because it may have been in breach of European rules.

21 For example, Ofqual reacted quickly to revelations in the *Daily Telegraph* that examiners were giving too much away in teacher seminars (Richardson 2012).

22 For example, Business Studies A level, which was initially highly popular, has suffered decline as top universities regard it as insufficiently demanding. Of course, this does not mean that the subject was deliberately dumbed down, but merely that reputation is important in the qualifications market.

moving quickly to such a monolithic institution, and it is not certain that there would be economies of scale sufficient to offset the costs of disruption. There is no reason to suppose that the provision of public examinations is a natural monopoly. There might also be legal challenges to such a move, as the awarding bodies are in most cases organisations independent of the state, and could not be expected to submit to effective nationalisation without opposition.

Assuming it could be accomplished, however, problems would remain. If perceived grade inflation is the issue, single-board systems under state control, such as that in Scotland, have had similar concerns (Kerevan 2011). And a single exam board would be even more at risk from political interference than the current arrangements.

There would also be greater operational risks involved as examination paper, marking, and grading errors would have a much wider impact. Such problems would also be more obviously the responsibility of ministers, who can currently pass the buck to the particular exam board at fault.

Monopoly suppliers of any description can neglect consumers and pursue their own agendas, and there is no reason to expect a unitary assessment body to be any different. Furthermore, preferences and experiences differ from individual to individual, and from school to school. Dissatisfaction with a board's services, which currently can be accommodated by switching boards, might in a monopoly environment lead schools to opt for alternative qualifications that are outside of government control, such as the International Baccalaureate.²³ Then pressures might arise to forbid such exit, or make it prohibitively expensive for state schools by refusing funding for exam entries outside the government-run board. Such measures would be distinctly illiberal and counter-productive to producing a better qualification system.

There is a more positive case to be made for continuing and indeed expanding competition. Competing boards have an incentive to engage in innovation, which benefits schools and pupils. This can take several forms. There has been assessment innovation: in the past, particular boards have pioneered different forms of assessment, including various forms of data response that have enlivened

23 There are significant numbers of schools that switch boards from time to time. The House of Commons Education Committee (2012, paragraphs 132-135) found no evidence that switches were related to the perceived easiness of the assessments.

previously dry subjects. Individual boards have also driven process innovation. For example, the assessment of large numbers of pupils has become more cost-effective through on-line marking and the employment of different markers for different types of question. Feedback to schools has also been greatly improved.

Then there is the perhaps more controversial issue of subject innovation. Since the introduction of a National Curriculum in the 1980s, the determination of what should be taught in a particular discipline has been increasingly seen as the prerogative of government agencies, albeit advised occasionally by outside bodies.²⁴ Indeed, in its conclusions, the House of Commons Education Committee (2012) says, ‘We see no benefit to competition on syllabus content’.

But over the long period of post-war schools examinations, there have been many valuable innovations developed by individual boards, with initiatives from schools, universities, and other organisations. One particularly fecund area has been the work pioneered by the Nuffield Foundation (2014), which since 1962 has funded 60 or so curriculum projects in science, mathematics, languages, history, and other subjects. Typically, such projects have involved teachers, academics, and others coming together in partnership with individual boards to develop new curricula and learning materials.²⁵ The freedom to work with individual boards to develop new ideas has been very valuable. It is increasingly narrowly circumscribed and would disappear completely with just one centrally determined specification for each subject area, administered by a single awarding body.

Of course, innovation could still occur under a single assessment board, but it is less likely as there would be little competitive pressure, and experiments might have to be ‘all or nothing’ because the single board’s remit would probably exclude the possibility of running alternative versions of the same

24 Occasional and somewhat random input from universities into syllabus redesign has been the pattern for many years, but has had little obvious impact. Perhaps the newly-formed A Level Content Advisory Board (ALCAB), a spin-off from the Russell Group of leading universities, will have more influence. But ALCAB is unlikely to replicate the influence of hands-on university-based examiners, who have largely disappeared from assessment bodies. At one time, they were quite common, until the relentless pressure of the Research Assessment Exercise/Excellence Framework drove them out.

25 An example with which I am familiar is the Nuffield Economics and Business, which combined disciplines and involved an active team with a strong emphasis on ‘investigation, progression and integration’. The option was offered in partnership with ULEAC (later Edexcel) alongside its more conventional single-subject awards. It ran as an A/AS level from 1994–2008, and a GCSE version ran from 1994 to 2009 (Nuffield Foundation 2009).

award side-by-side. Moreover, a nationalised board would have no incentive to expand its overseas operations, and might well be discouraged from overseas activity altogether, thus losing significant invisible exports and international influence. At the same time, a system of franchised awards might encourage innovation prior to a contract being awarded, but limit it for the duration of the contract, therefore limiting its potential.

Instead, a case could be made that new competition should be positively encouraged in various ways. Examples include drawing subject specifications more loosely, which would allow for a wide range of options to be developed in most subjects; promoting new entry from higher education and professional bodies to offer at least some new syllabi; and allowing more aggressive marketing of competing qualifications – with, for example, the endorsement of leading universities and employers.²⁶

Conclusion

This chapter has surveyed the history of external schools examinations over the last century and a half. Unlike other countries, where examinations have been centrally controlled by the state, the system in England, and to a lesser extent in Northern Ireland and Wales, has evolved from voluntarist foundations. This has left us a legacy of competing awarding bodies between which schools can choose.

I do not doubt that there has been a change in the standards applied in assessing pupils at 16 and 18. As documented here, changes in awarding bodies, regulatory systems, and methods of assessment have been plentiful. These changes have been inevitable given the policy objectives of school comprehensivisation and the expansion of higher education, together with a changing public mood that would no longer accept the exclusion of the vast majority of young people from formal qualifications. There is little point in trying to recreate a past that had a distinct downside for a large proportion of pupils.

²⁶ At the moment, the official position is that universities have no preferences between competing awards – although teachers suspect this is misleading. Making endorsements open would improve information to candidates and also encourage universities to once again get involved properly in the business of schools examinations.

Nevertheless, change could have been handled much better. The insistence by successive generations of politicians and the teaching profession that standards have been maintained virtually unchanged does not bear close examination. But nor does the more recent attempt to blame lapses and failings in the system on the awarding bodies.

If an A level is no longer quite the ‘gold standard’ of fifty years ago, it is politicians with their ‘here today, gone tomorrow’ policies who are largely responsible. While we may all have pet changes we would like to introduce, perhaps we should just be a little more laid back about the way it works. On the whole, it functions better than a lot of other systems, as recognised by the many thousands of international pupils who seek our awards. It is certainly free from the corruption of systems in less-developed countries, and offers a wider range of choice between and within subjects than is available in most other jurisdictions.

In retrospect, explicitly acknowledging that the function of the examination system had changed would probably have been preferable. One way might have been to set a clearly defined floor with a basic pass standard that everyone was expected to reach, and then simply rank all pass candidates in terms of their performance, allowing universities and employers to choose the decile or percentile at which they would aim recruitment – rather than trying to maintain arbitrary grade standards.

Yet I have doubts whether further tinkering would achieve very much. Indeed, the Secretary of State who *did not* want to make major changes to the exam system on his or her watch would in this case get my vote. There has been reform frenzy in the qualifications and examinations market, which has to stop. Of course, smaller reforms to improve the exam system are always possible, but we need to get away from constant tinkering as successive Secretaries of State try to remake the system according to some new template devised by their advisers.

In fact, arguably, examinations are a side issue. More fundamental imperatives are to improve the quality of our teachers, too many of whom are poorly qualified; to increase the range and scope of parental choice; to reduce the power and influence of backward-looking teaching unions; and to improve discipline, behaviour, and safety in some of our poorer schools. If we get these things right, there would be no need to worry so much about the form of our qualifications and examinations system.

References

- Beloe Committee (1960), *Secondary School Examinations other than the GCE* London: HM Stationery Office. <http://www.educationengland.org.uk/documents/beloe/beloe.html> (accessed 7th May 2014).
- Brooks, V. (2008), 'The Role of External Examinations in the Making of Secondary Modern Schools in England 1945-1965', *History of Education* 37(3):447–467.
- Cambridge Assessment (2014), 'Our Heritage', <http://www.cambridgeassessment.org.uk/about-us/who-we-are/our-heritage/> (accessed 1st May 2014).
- CEER (2013), 'GCSE 2012'. Report, Centre for Education and Employment Research, University of Buckingham, <http://www.buckingham.ac.uk/wp-content/uploads/2010/11/GCSE12.pdf> (accessed 14th June 2014).
- Civil Service (2014), 'The Origins of the Modern Civil Service: the 1850s', <http://www.civilservice.gov.uk/about/a-partial-history-of-the-civil-service/the-origins-of-the-modern-civil-service-the-1850s> (accessed 14th June 2014).
- Coe, R. (2007), 'Changes in Standards at GCSE and A level: Evidence from ALIS and YELLIS'. Report, Centre for Curriculum, Evaluation, and Management, Durham University, <http://www.cem.org/attachments/ONS%20report%20on%20changes%20at%20GCSE%20and%20A-level.pdf> (accessed 29th May 2014).
- House of Commons Education Committee (2012), 'The Administration of Examinations for 15–19 Year Olds in England'. First Report of Session 2012–13 Volume One. <http://www.publications.parliament.uk/pa/cm201213/cmselect/meduc/141/141.pdf> (accessed 30th April 2014).
- Kerevan, G. (2011), 'Could Do Better at Stopping Exam Grade Inflation', *The Scotsman*, 5th August 2011, <http://www.scotsman.com/news/george-kerevan-could-do-better-at-stopping-exam-grade-inflation-1-1778110> (accessed 13th May 2014).
- Malacova, E. and J. Bell (2006), 'Changing boards: investigating the effects of centres changing their specifications for English GCSE', *The Curriculum Journal* 17(1):27–35.
- NCES (2012), 'Figure 3. Change in average mathematics scores of 8th-grade students, by education system: 2007–2011 and 1995–2011'. Data retrieved from the National Center for Education Statistics: http://nces.ed.gov/timss/figure11_3.asp (accessed 14th June 2014).
- Norwood Report (2003). *Curriculum and Examinations in Secondary Schools* London: HM Stationery Office. Available at <http://www.educationengland.org.uk/documents/norwood/norwood1943.html#05> (accessed 3rd May 2014).

- Nuffield Foundation (2014), 'Curriculum Projects', <http://www.nuffieldfoundation.org/curriculum-projects> (accessed 13th May 2014).
- Nuffield Foundation (2009), 'Nuffield Economics & Business: A Short History', http://www.nuffieldfoundation.org/sites/default/files/files/NEB_History_of_2008.pdf (accessed 13th May 2014).
- OECD (2013), 'PISA 2012 Results'. Data retrieved from: <http://www.oecd.org/pisa/keyfindings/pisa-2012-results.htm> (accessed 14th June 2014).
- Richardson, H. (2012), 'Rules Tightened on GCSE and A-level Exam Seminars', BBC, 27 April 2012, <http://www.bbc.co.uk/news/education-17853507> (accessed 13th May 2014).
- Sheldon, N. (2011), 'History Examinations from the 1960s to the Present Day'. <http://www.history.ac.uk/history-in-education/project-papers/topics> (accessed 9th May 2014).
- Smithers, A. (2012), 'A-Levels 2012'. Report, Centre for Education and Employment Research, University of Buckingham.
- Tattersall, K. (2008), 'The Relationship of Examination Boards with Schools and Colleges: A Historical Perspective'. Speech at Cambridge Assessment seminar, <http://www.cambridgeassessment.org.uk/Images/126056-kathleen-tattersall-s-speech.pdf> (accessed 30th April 2014).
- Walport Report (2010), *Science and Mathematics Secondary Education for the 21st Century: Report of the Science and Learning Expert Group*, <http://webarchive.nationalarchives.gov.uk/+/http://www.bis.gov.uk/wp-content/uploads/2010/02/Science-Learning-Group-Report.pdf> (accessed 11th May 2014).

2 THE ‘QUALIFICATIONS SLEDGEHAMMER’: WHY ASSESSMENT-LED REFORM HAS DOMINATED THE EDUCATION LANDSCAPE

TIM OATES

Introduction: assessment and qualifications as drivers for education reform

IN ENGLAND, ASSESSMENT AND qualifications are at the heart of education reform. Indeed, review after review has focused on those features much more often than other parts of the system. In recent decades, this has led to a swathe of new policies in these areas, which subsequently have been changed or abolished in the light of unintended consequences that policymakers failed to foresee. Examples include the modularisation and later de-modularisation of A levels and GCSEs; the introduction of the AS-level award and its subsequent downgrading in relevance; and the development and subsequent abolition of Diploma qualifications. Furthermore, these changes have occurred against the backdrop of the rise of GCSE and A-level equivalents, the development of vocational awards, and a legion of changes to GCSE content, such as the constantly shifting policy regarding the use of calculators in mathematics and the form and function of teacher-assessed components.

This chapter focuses on the tendency to use assessment and qualifications as drivers for education reform. It argues that this tendency has produced a constantly changing and untenable system, which does not, and cannot, address the deficiencies policymakers hope to solve. Naturally, qualifications need to be

evaluated and updated when circumstances and evidence indicate that this is the right thing to do. For example, in the mid-1980s, the reform of O Levels and CSEs into GCSE enhanced both equity and efficiency in the system. But contemporary education reforms in England have depended far too heavily on changes in assessment and qualifications policy, which have produced a mishmash of incentives in the education system, pulling its actors in different directions.

In order to provide incentives for improvements, therefore, policymakers must examine the structure of the system as a whole in order to ensure that all parts are working to pull actors towards the same goal. Such coherence is indeed an essential feature of high-performing systems worldwide.

The government must also be much more agile when it comes to recognising the emergence of perverse incentives, such as responding to the adverse impacts of the grade C threshold target. A history of insufficiently prompt action to remedy such problems is evidence of clear government failure over the past decade and more. Granted that it is not always possible, the failure first to model the effects in smaller-scale pilot schemes has further added to instability and sub-optimal performance in the system.

Taking the brunt of it: the unmanageably contradictory pressures imposed by government on qualifications and assessment

While Tony Blair and his education secretaries were ultimately responsible for the education policies under New Labour, it is clear that Michael Barber was a key figure behind the scenes. Indeed, his book *The Learning Game*, published in 1996, essentially became a handbook for the New Labour education ministers. Barber's emphasis on 'standards, not structures' fuelled assessment-focused change. His strict focus on failure in the education system led to a legitimate interest in high expectations (Wilby 2011) but undue emphasis, in reform measures, on thresholds and targets dominated by assessment and qualifications.

Barber was right to emphasise the huge disparities in attainment across the education system. For example, in 1989, only 30 per cent of pupils attained 5+ GCSE grades with A–C grades (Payne 2001), with significant variation depending on school type, ethnic group, and social background (Gillborn and

Mirza 2000). But the idea of using qualification outcomes as an indicator of educational quality was perverted into an exaggerated focus on assessment and qualifications as the key instruments of improvement in the education system.

The use of qualifications and assessment as major policy instruments is not new. This approach intensified in the 1950s, with the introduction of A levels and O levels, following the Education Act of the mid-1940s. Although examinations have continued mainly to be produced by independent assessment bodies, successive governments have increased the levels of state regulation of the form and content of these examinations, through codes, criteria, and the development of increasingly elaborate national regulatory organisations.

But despite this escalation of central control, it would be quite wrong to cast assessment and qualifications simply as crude tools of education policy. The reality is far more complex. Much of the complexity derives from the multiple functions that assessment and qualifications fulfil: Newton (2007) outlines twenty functions of national assessments, while Mike Coles and I trace forty functions of general and vocational qualifications (CEDEFOP 2010). Some of these functions relate to curriculum intent, since assessment and qualifications embody and convey certain curriculum intentions, for example to focus on certain knowledge and skills. Other functions relate to standards, for example whether pupils and the education system at large are improving.

Qualifications and assessment are likely to continue to carry multiple functions, but it is time to recognise this over-dependence on assessment and qualifications in respect of efforts to drive education reform, and the relative neglect of other factors.

The importance of balanced policy is highlighted in Schmidt's work on 'curriculum coherence' (see Schmidt and Prawat 2006). Drawing from empirical work on features shared by high-performing education systems, curriculum coherence is seen as key, which in turn displays two dimensions. The first is that all elements of the system together pull in the same direction, and the second is that curriculum content is based on well-grounded progression in learning. To this idea of 'curriculum coherence', I have earlier added fourteen 'control factors' – amenable to policy intervention – across which coherence should be established and maintained (Oates 2010). These factors are:

- Curriculum content (specifications, support materials, etc.)
- Assessment and qualifications
- A national framework for qualifications
- Inspection
- Pedagogy
- Professional development
- Institutional development
- Institutional forms and structures (e.g. school size and education phases)
- Allied social measures (linking social care and health care with education)
- Funding
- Governance (autonomy versus direct control)
- Accountability arrangements
- Labour market/professional licensing
- Allied market regulation (e.g. health and safety legislation and insurance regulation)

Different education systems place a different emphasis on different factors – but there are important complementarities and dependencies between them, which requires joined-up thinking about how to generate improvements. In high-performing systems, the concerted management of these factors encourages curriculum coherence. This suggests that attention to the relations between policy areas within the system is as important as the form of policies in one specific area, such as accountability.

My concern is that English policymakers have emphasised qualifications and assessment to the neglect of other factors, therefore inhibiting policy movement towards coherence and a step-change in system performance. As principal instruments in the accountability agenda, qualifications and assessment have carried an overblown policy burden.

The use of public examinations in target setting and for measuring teacher, school, and national performance is obvious. This is a classic assessment-led strategy. But the assessment-focused change strategy is more prevalent than one might at first suppose. A less obvious example of assessment-led reform

is the National Curriculum. Technically, the National Curriculum is not a curriculum at all, and this is no trivial matter. The term ‘curriculum’ refers to the totality of the experience of learning, encompassing aims, content, methods, assessment, and evaluation (Eraut 1976), and curriculum theory explains the distinctions between intended curriculum, enacted curriculum, and actual learning outcomes (Valverde et al. 2002). It encompasses ‘taught curriculum’ and ‘untaught curriculum’ as elements of the schooling experience. Understanding these elements and the interaction between them is vital for understanding school performance and national arrangements. Where the National Curriculum masquerades as a curriculum is where it states content – things that should be taught – and it does determine to a degree, and in certain areas, the pedagogical approach. For example, both the requirement for experimentation in science and for development of phonological awareness in English carry strong implications for pedagogy. Yet it is more accurate to describe the National Curriculum as a framework of standards, which outlines the goals in terms of outcomes. Although it determines aspects of curriculum, it is far more assessment-oriented than curriculum-oriented.

Only with the Literacy and Numeracy Strategies did government action around the National Curriculum introduce substantial school intervention in terms of pedagogy. The Numeracy Strategy appears responsible for a minor increase in mathematics attainment in TIMSS (Hodgen et al. 2010), but remains controversial in respect of curriculum control. John Bangs, the then NUT head of education, regarded the strategies as invaluable professional development support to teachers, while other educationalists regarded it as inappropriate subversion of school autonomy (Whitty 2006). As a non-statutory part of government policy, the Numeracy Strategies do not detract from the fact that the government’s main legislative instrument – the National Curriculum – remains an assessment-oriented, standards-focused one.

Scrutinising the research on the advantages of having a National Curriculum, many cite a general culture of high expectations, which intensified as a result of the New Labour focus on standards. But the impact of a general concern for high standards in England has been moderated by the specific impact of detailed accountability measures and the focus on examination standards as the key metric for judging whether the high expectations are being met.

And the emphasis on high expectations can be distorted, leading to a heavy focus on specific pupils and/or a very instrumental focus on a restricted set of outcomes. England has indeed seen the emergence of such problems – some of which are well documented but subject to an exceptionally slow policy response.

A well-known problem has been the tendency among schools to focus on GCSE C/D borderline candidates, in order to push a higher percentage of pupils over the 5 A*–C pass threshold. This led to relative neglect of those well above the threshold and well below it. Indeed, one large metropolitan authority relentlessly targeted such candidates, even sending letters home to households with children in this category. At one time, the then Department for Education and Skills was actually advocating this focus on C/D borderline candidates as a key improvement strategy, despite its known adverse impact on equity (Gillborn and Youdell 2000; Marx 2012).

It is clear that previous governments were extraordinarily slow to respond to the distortions associated with the threshold measures. This allowed a highly non-egalitarian principle to become embedded in teachers' thinking and practice for a protracted time. Although the new, more balanced Progress 8 measure seeks to drive teachers towards a concern for all pupils and remove undue focus on grade D candidates, given the absence of a pilot programme the precise effects of the new Progress 8 measure will be difficult to anticipate. Careful monitoring and necessary fine-tuning are likely to be required to ensure that the combination of high standards and equity is driven into educational practice.

The second strong moderation of a general culture of high expectations was the distorting effect of 'teaching to the test'. Again well documented, the impact has been wide ranging, including (1) a general narrowing of the curriculum (Gilbert 2012); (2) a dramatic rise in strategic retakes in both GCSE and A levels (Vidal Rodeiro 2014); (3) narrow assessment-driven instruction; and (4) a deleterious effect on both the quality of learning resources, and the relation between those resources and qualifications (Education Select Committee 2012).

The form of accountability adopted from 1997 onwards created a dominant focus on qualifications outcomes. This is combined with recognition among pupils that high grades are of increasing importance for entry to higher education and in the labour market (Sissons and Jones 2012), which makes

them believe they should narrow their focus on the part of the curriculum that is important to pass the test. Indeed, in the 2010 National Curriculum review, evidence cited a strong lack of pupil motivation for uncertificated components of the Key Stage 4 curriculum (Oates 2010).¹

The narrow assessment-driven instruction combined with highly strategic entry and retaking behaviour has put extraordinary pressure on exam boards. Such strategic behaviour includes entering pupils for more than one exam board in the same subject, a dramatically increased rate of retaking of modules/units, and intense focus on the specifics of the examination.

But one pressure is the most serious and elementary: the issue regarding the extent to which a rise in grades reflects a genuine rise in underlying attainment, versus the extent to which it reflects ‘gaming’. This is increasingly well documented, with the Cambridge Assessment ‘standards debate’ of 2010 and Durham University’s excellent triangulation of the rise in grades being the two clearest examples (Coe 2007). Indeed, during the last decade, one of the most serious failures in education policy was the inability, or unwillingness, to develop an independent metric for measuring underlying educational standards. England’s participation in TIMSS, PIRLS, and PISA did provide independent data, notwithstanding questions regarding the extent to which the content in these tests matched the English curriculum. Nonetheless, the results from these surveys did indeed show a discrepancy with domestic assessments, with the latter showing potential signs of inflation, confirmed in a range of well-designed domestic studies (Coe 2007; Hodgen et al 2010; Massey et al. 2003).

The problem was simple: the assessments used for judging whether standards were improving were also what pupils, teachers, and schools were targeting. In such circumstances, one would expect the data to become distorted, as gaming emerges. This is an elementary problem, well known in performance-measurement theory (Elton 2004). Many other countries use independent tests to sample national attainment in order to gain an independent yardstick against which policy and improvement strategy can be measured. Despite clear signs of systemic problems, the English government of the time was remarkably slow to respond to the emerging problems of the accountability agenda.

1 Ironically, the narrowing of focus is taking place despite evidence that teaching beyond the syllabus enhances the chances of higher grades (Suto et al. 2012).

Modularisation: a case in point

A relevant example of these problems is the modularisation of qualifications, which is worth exploring further. Indeed, the (mis)management of modular qualifications now looks like yet another failure to manage relations between different elements in the education system. Modularisation has been both a cause and a victim of the problems outlined. In an accountability system with over-reliance on qualifications, modularisation allows for more gaming behaviour, and it creates more problems for the maintenance of standards. In addition, gaming has increased the assessment burden on schools and pupils, while driving up overall public expenditure on examinations.

These are serious problems, and an apparently easy solution immediately presents itself: banning modular examinations. But the solution is not that easy after all, since some positive aspects of modularisation are lost by such a ban. At the same time, careful cost-benefit analyses of modular examinations have been clouded by the deleterious impact of accountability policy on qualifications.

When modular A levels were developed in the 1980s and 1990s, candidates seldom improved their grades when retaking modules. The simple reason for this was that schools did not encourage cramming for retakes in order to maximise final grades. Modularisation at advanced level, applied only in subjects that seemed suited to the approach, was seen to have considerable merit since it reduced the assessment burden at the end of the course. It also meant that the amount of assessment could be increased, which improved each qualification's measurement accuracy. Modularisation encourages pupils to work consistently throughout the course rather than coast and cram for the exam – which helped raise attainment in various groups of pupils. Early assessment in the first modules gave pupils feedback on strengths and weaknesses, as well as familiarity with the form of A-level assessment. Again, this had a beneficial impact on underlying attainment.

The early modular A-level pilots added some important, innovative qualifications to the total catalogue in the country. In the mid-1990s, awarding bodies were responding to the extraordinary diversity that existed, and remains, in the English education system. This diversity is rarely recognised. For example, we have one of the most varied systems in the world in terms of school types,

LEAs (some very active, others less active), school sizes, and school structures. This extends to acute diversity in ideas about assessment. The overall structure gives rise to a variety of local eco-systems of schools.

A subject of research and policy discussion in its own right, the issue is raised here simply to make the point that during the 1990s, awarding bodies responded to the diversity of the system by maintaining high diversity in the qualifications catalogue. Although awarding bodies were consolidating during this period due to various pressures, development and provision of modular qualifications helped to ensure that diverse curriculum approaches were supported by suitable qualifications.

Nobody knows exactly what might have happened without Dearing's (1996) across-the-board application of modularisation in A levels, but it is likely that modular A levels would have become established as part of the provision in some subjects and in some segments of the system. A mix of modular and non-modular qualifications allows awarding bodies to monitor and maintain standards more effectively, and promotes better understanding and control of inflationary elements as well as for better identification of genuine changes in educational attainment. All of this was rendered far more difficult with universal implementation.

Still, it is possible to compensate for the losses of removing modular qualifications. For example, schools can use 'staged' assessments, developed in-house or provided by awarding bodies. Not contributing to the final grade, these assessments can help pupils with feedback, while increasing their motivation to work from the outset of the course. In this way, schools can ensure that all pupils understand the need for engagement throughout the course – and that 'coasting' could have negative consequences. Awarding bodies could accept the reduction in assessment time, and focus on maximising the measurement qualities of the terminal examinations. But all this needs to be executed well in order to establish in schools the practices that were encouraged and applied by the early pilot modular programmes. From this perspective, removal of modular A levels does not necessarily detract from the quality of upper-secondary level assessments or compromise attainment.

Despite the problems with wholesale modularisation, AS levels did provide a specific benefit to the English system. The 1996 Dearing Report intended

to broaden 16-19 programmes through the 'four AS, three A levels' model (Dearing 1996), which allowed pupils to gain feedback and experience from four subjects in the first year, and afterwards focus on the three optimum subjects in the second year.

In other words, the AS-level system takes into account that pupil preferences can and often do change. It is highly efficient to encourage means by which pupils study the subjects that most motivate and engage them; in this, the 'four AS, three A levels' model clearly has considerable intrinsic merit. Its deficit is in the extent to which – in a universally modularised system with qualifications playing such an important role for accountability – the credit from AS qualifications is used to contribute to the final grades in the A levels.

Again, a solution could be to let pupils take four AS levels, but move towards entirely 'staged' assessment that does not contribute to the final A-level grades. Pupils would still gain three A levels in the second year, while also obtaining a non-graded certificate in the fourth subject studied to AS level only.

The main impediment to this solution would be the reaction by schools, one produced by the accountability culture dominated by qualifications and assessment. In essence, schools would argue that if funding is tied to qualifications, and the key outcomes are three A-level grades, then AS levels not contributing to those grades should not be taught. The result is a reduced academic diet from four AS and three A levels to just three A levels.

This would not appear to be a problem in a system where higher education still focuses on three A-level grades. Yet the removal of the choice at the end of the first year of post-GCSE study is likely to have hidden but significant negative consequences for the reasons described above. Indeed, viewing curriculum solely through the lens of assessment and qualifications is unwise and yet pervasive. Consider the removal of practical science coursework from science qualifications, which has precipitated a storm of protests from those who claim that unless something explicitly counts for an examination, it will not get taught.

But this notion represents an extraordinarily reductionist position. First, coursework undertaken under strong accountability pressure places both teachers and awarding bodies in a desperately conflicted position. Indeed, qualifications with a substantial teacher-assessed coursework component put intolerable pressure

on teachers, pulling them in very different directions. On the one hand, teachers are pressured to act on behalf of the school to drive continual improvement in the exam performance of their pupils. On the other hand, they must act on behalf of awarding bodies to be impartial and reliable assessors, ensuring that their judgments and marking practices are in line with awarding body marking schemes and national standards. This leads to a highly conflicted professional role regarding internal assessment. For the majority of teachers, it does not lead to maladministration of assessment, but it appears to drive bunching and upwards tilting of marking, and may include a strong element of ‘benefit of the doubt’ for borderline pupils (Cambridge Assessment 2014).

Second, exam boards are also conflicted. The boards design qualifications to national criteria, elements of which lead to highly compromised qualification structures. One example is the controversial change of grade boundaries in GCSE English part way through the 2011-12 academic year – a qualification with coursework components worth 60 per cent of the final grade. The judicial review into the problems which arose cited poor design criteria, emanating from the government, as a principal contributing factor to the issues surrounding the award. Exam boards are under pressure from subject organisations and teachers to include coursework, while at the same time having to ensure dependability, which is both hugely costly and perceived as draconian by schools.

In the current context of drivers and incentives, coursework assessment puts unmanageably contradictory pressures on teachers, and different but equally unmanageable pressures on awarding bodies. Yet coursework remains desirable from an educational perspective. Within the current incentive framework, coursework not contributing to final grades is not worth teaching. This collapse of ‘curriculum thinking’ into ‘qualifications thinking’ also conditions both research and policy regarding curricula – and it is clear that the principal culprit is the view of qualifications as key drivers of education reform.

The importance of a coherent incentive structure: what we can learn from Finland

It is worth considering systems with alternative approaches to the lopsided focus on assessment and qualifications. Finland is an important example, but not for

the factors commonly assumed to be in operation there. Much of the discussion about the Finnish 'education miracle', rising from a low achiever to one of the top performers in the world, has focused on the degree of autonomy enjoyed by Finnish schools. Low levels of inspection and the absence of high-stakes national tests in primary- and lower-secondary education have been heralded by British educationalists as proof that school autonomy with low accountability is the key to ensuring high quality.²

But this overlooks three vital features of the Finnish system: (1) the nature of the system in the 1970s and 1980s, when Finland dramatically transformed its education system; (2) the locus of control that continues to exist in the Finnish system; and (3) the importance of the rigorous matriculation examinations at the end of upper-secondary schooling. Schools may appear more autonomous than schools in England, but the system demonstrably is not free of restriction and high-stakes assessment.

Finland leapt to international attention following its performance in PISA 2000, and prominent commentators have focused on elements of the *current* Finnish system in explaining the country's educational success (e.g. Hancock 2011; Partanen 2011; Guardian 2014). But this is not a sufficient approach. As argued in the 2010 English curriculum review, in order to understand the underlying reasons behind high-achieving countries' success, one has to analyse the arrangements in place *prior to and during* the period of improvement. Merely looking at the system as it is now, when it already has achieved high performance, is insufficient to unveil causation (Oates 2010).

From the late 1990s to the present day, Finland's education system has been characterised by relatively high school autonomy, with low levels of central inspection and low levels of external testing (Sahlberg 2011). The system is also noteworthy for its 'front-end restriction', associated with a highly selective, and long duration, teaching training. This contrasts with systems focusing on

2 It is worth noting that Finland fell out of the top ten countries in mathematics in PISA 2012, confirming a decline since 2006 (OECD 2013). Similarly, while Finland came in eighth place in mathematics in TIMSS 2011, it was also revealed that Finnish seventh graders had fallen radically since TIMSS 1999 (IEA 2012, p. 56). So at the same time as the 'miracle' was discovered, the country's pupils were beginning to slip. This further suggests that Finland's ascendancy is far more complicated than what many commentators suggest.

‘back-end restriction’, characterised by a strong emphasis on inspection and target-based accountability arrangements.

A key question is whether the current characteristics of the system were also present when Finland’s transformation from a relatively low-performing to a high-performing country occurred. The historical record suggests that the answer is a resounding ‘no’. A historical analysis of the system’s characteristics and the nature of policy preceding and during the transformation suggests that high control from the centre – including high-intensity inspections, state-approved textbooks, and national benchmark tests – played an important role.

Indeed, key Finnish educational analysts, such as Hautamäki (2014), emphasise that the system between 1972 and 1985 was strongly state controlled, with all teachers having to go through extensive in-service training in which the mandatory content was delivered. At the same time, school inspections were extensive, and all teaching material had to be approved to ensure that it was aligned with a very detailed, national curriculum, spanning over 600 pages. While there were no national assessments in any subject in compulsory education, the detailed curriculum, intensive in-service teacher training, and standardised tests in some school subjects – which were used by educational researchers – ensured comparability of school marks.

Thus, there were two major phases in the development of Finland’s contemporary education system. The first phase involved the enactment of fundamental reform from 1968 onwards, which created a fully comprehensive system and the foundation that gave rise to high performance in the late 1990s. At this time, implementation at the school level was ensured by heavy centralised state involvement.

The second phase, on the other hand, involved a strategic move towards more school autonomy and low levels of centralised inspections. In the decentralising spirit of the late 1980s, the office responsible for approving textbooks was closed in 1990. In my own interviews with current Finnish teachers and educationalists, they emphasise high-quality teachers and high-quality materials as the key ingredient of Finnish success. And, of course, it is important to note that the rigorous matriculation exams at the end of upper-secondary school remain a key part of the system.

It is clear that most international commentators have inaccurately focused on the second phase, frequently associating the current system's characteristics with the previous period of transformation and substantial improvement – during which arrangements were very different. The first phase tends to be ignored outside Finland, and the highly centralised change strategy may indeed be the 'inconvenient truth', at odds with the oft-desired and appealing narrative regarding autonomy (Alexander 2012; Benton 2014).

Once the system had been established, central control was relaxed, but it is vital to recognise that the quality criteria established in the first phase were vital for the transformation – and continue to be the basis of contemporary system performance. One of the factors existing in both phases is the high-quality teacher training, which is highly selective. Only 10 per cent of applicants are accepted on demanding criteria relating to both command of the subject discipline and disposition towards teaching. All teachers are expected to have master's degrees, with research and evaluation playing an integral role in the training curriculum.

With such demanding criteria and content, teacher training can certainly be characterised as a key control mechanism in the system. It ensures that all teacher practice embodies the values and practices of the system. This 'front end' type of control explains the lack of need for 'back end' type of control in the form of a strong inspection system and national assessment, which characterise the English system.

While it is important to acknowledge the problems involved in drawing causal conclusions from narratives of this type – since there could be other, unrecognised changes, not necessarily in education, contributing to Finland's rise – it is still important to have an accurate picture of what policies the country pursued during its transformative stage. Clearly, at the very least, the reasons behind Finland's improvements are not as clear-cut as commonly assumed in the debate.

The first lesson from Finland is that ideas about issues such as equity and ability played a vital role in the transformation of its system. The social and political discussion prior to the adoption of fully comprehensive compulsory education was important for the concerted and coherent implementation of the new system, and for its continued success. Rather than focusing too much

on assessment and qualifications as drivers for change, the Finnish discussion concentrated on ensuring coherence of all elements in the proposed system (in line with Schmidt's notion of 'curriculum coherence'). In fact, the structure and assessment approaches of the main high-stakes matriculation examinations at the end of upper-secondary education were pretty much left alone, continuing in much the same form as it had for about a century. The examinations have not remained static, but changes have reflected changes in the curriculum, rather than vice versa. Thus, curriculum drove assessment and qualifications – which contrasts sharply with the English situation.

The second lesson from Finland is consequently that coherence is vitally important. In the first phase of transformation, this was ensured via strong central control. Once the new system was established, this control was replaced by the 'front end' restriction in the form of a highly selective teacher training that became the bedrock for ensuring continued coherence between the different elements of the system.

Conclusion: authentic piloting and responsiveness to unintended consequences

This chapter has argued that English policymakers place far too much emphasis on assessment and qualifications in their attempts to reform education. This has created a situation in which qualifications and assessment are in constant motion – with the direction reflecting the zeitgeist of the time – which is untenable. In review after review, earlier reforms have been targeted as key problems that have to be reversed, thereby ignoring the precise reasons why the reforms were implemented in the first place. The back-and-forth reforms of qualifications and assessment merely highlight the lack of joined-up thinking among education analysts and policymakers, which amounts to a significant barrier to sustained improvement in outcomes.

The first policy conclusion is thus to end the over-dependence on assessment and qualifications as drivers of education reform. Instead, the government must ensure coherence between the different elements of the education system in order to ensure that they all pull in the same direction. Careful attention must be paid to the incentive structure produced by all elements in the system – such

as inspections and curriculum aims – to ensure that they join up and exert pressure on professionals to generate higher quality.

The second policy conclusion is that the government must be more responsive and agile when it is clear that there are perverse and contradicting incentives in the system. For example, the delay in acting on gaming behaviour that arose because of strong incentives to focus on 'marginal pupils' on the boundary of obtaining a C grade was a clear government failure. Such delays must be avoided.

While the perverse incentives arising from this specific policy might have been prevented by a clear analysis prior to implementation, it is difficult to foresee what unintended consequences could arise from system-wide reforms. The third policy conclusion is therefore to pursue pilot programmes before implementing policies nationally. This relates both to education reforms generally as well as qualifications and assessment reforms more specifically. A role model in this case is the Singaporean government, which trials policies extensively before deciding whether to scale them up nationally. In England, policies have too often lacked an authentic trial phase.

In short, therefore, reforming qualifications and assessment further is unlikely to produce better outcomes. Instead, policymakers must be prepared to reform other parts of the education system to ensure coherence. Given the over-zealous reforms that have characterised qualifications and assessment in England over many years now, it is time for politicians to stop tinkering with these and instead focus their attention on balanced innovation and management of the arrangements in the system.

References

- Alexander, R. (2012), 'Neither National nor a Curriculum?' *Forum* 54(3):369–384.
- Benton, T. (2014), 'A Re-evaluation of the Link between Autonomy, Accountability and Achievement in PISA 2009'. Report, Cambridge Assessment.
- Cambridge Assessment (2014), 'Radical Solutions in Demanding Times: Alternative Approaches for Appropriate Placing of "Coursework Components" in GCSE Examinations'. Report, Cambridge Assessment.
- CEDEFOP (2010), 'Changing Qualifications – A Review of Qualification Policies and Practices'. Report, European Centre for the Development of Vocational Training, Luxembourg.

- Coe, R. (2007), 'Changes in Standards at GCSE and A level: Evidence from ALIS and YELLIS'. Report, Centre for Curriculum, Evaluation, and Management, Durham University.
- Dearing, R. (1996), 'Review of Qualifications for 16–19 Year Olds', Qualifications and Curriculum Authority, London.
- Education Select Committee (2012), 'The Administration of Examinations for 15–19 Year Olds in England', First Report of Session 2012–13, House of Commons, London.
- Elton, L. (2004), 'Goodhart's Law and Performance Indicators in Higher Education', *Evaluation & Research in Education* 18(1-2):120–8.
- Eraut, M. (1976), 'The Analysis of Curriculum Materials'. Occasional Paper, University of Sussex.
- Gilbert, R. (2012), 'Curriculum planning in a context of change – a literature review'. Report, Department of Education and Early Childhood, Victoria, Melbourne.
- Gillborn, D. & D. Youdell (2000), *Rationing Education: Policy Practice Reform and Equality*. Philadelphia: Open University Press.
- Gillborn, D. & H. Mirza (2000), 'Educational Inequality – Mapping Race Class and Gender – A Synthesis of Research Evidence'. Report, Office for Standards in Education, London.
- Guardian (2014), 'How Finnish Schools Shine', *The Guardian*, <http://www.theguardian.com/teacher-network/teacher-blog/2012/apr/09/finish-school-system> (accessed 11th June 2014).
- Hancock, L. (2011) 'Why are Finland's Schools Successful?' *Smithsonian Magazine*, September.
- Hautamäki, J. (2014), 'How do Finns Know? To Trust or Not to Trust Teachers' Assigned Grades'. Draft document kindly supplied by commissioning editor Prof. Mary James, Cambridge.
- Hodgen, J., D. Kuechemann, M. Brown, R. Coe (2010), 'Multiplicative Reasoning, Ration and Decimals: A 30-year Comparison of Lower Secondary Students' Understandings' in Pinto M.F. & T.F. Kawasaki (eds.), *Proceedings of the 34th conference in the International Group of the Psychology of Mathematics Education*, volume 3, pp. 89–96 Belo Horizonte, Brazil.
- IEA (2012), 'TIMSS 2011 International Results in Mathematics', Report, Boston College, Chestnut Hill, MA.
- Marx, R. (2012), "'I Get the Feeling that it is Really Unfair": Educational Triage in Primary Mathematics' in Smith C. (ed.) *Proceedings of the British Society for research into Learning Mathematics* 32(2):58–63.

- Massey, A., S. Green, T. Dexter, and L. Hamnett (2003), 'Comparability of National Tests over Time: Key Stage Test Standards between 1996 and 2011'. Report, Research and Evaluation Division, University of Cambridge Local Examinations Syndicate.
- Newton, P. (2007), 'Clarifying the Purposes of Educational Assessment', *Assessment in Education* 14(2):149–70.
- Oates, T. (2010), 'Could do Better – Using International Comparisons to Refine the National Curriculum', Report, Cambridge Assessment.
- OECD (2013), 'PISA 2012 Results in Focus: What 15-year-olds Know and What They Can Do with What They Know: Key results from PISA 2012'. Report, OECD, Paris.
- Partanen, A. (2011), 'What Americans Keep Ignoring About Finland's School Success', *The Atlantic* 29(12), <http://www.theatlantic.com/national/archive/2011/12/what-americans-keep-ignoring-about-finlands-school-success/250564/> (accessed 11th June 2014).
- Payne, J. (2001), 'Patterns of Participation in Full-time Education After 16: An Analysis of the England and Wales Youth Cohort Study'. Report, Policy Studies Institute, London.
- Sahlberg, P. (2011), *Finnish Lessons*. New York: Teachers College Press.
- Schmidt, W. and R. Prawat (2006) 'Curriculum Coherence and National Control of Education: Issue or Non-issue?' *Journal of Curriculum Studies* 38(6):641–658
- Sissons, P. and K. Jones (2012), 'Lost in Transition? The Changing Labour Market and Young People not in Employment, Education or Training'. Report, The Work Foundation, Lancaster University, London.
- Suto, I., G. Elliot, N. Rushton, and S. Mahta (2012), 'Going Beyond the Syllabus: A Study of A Level Mathematics Teachers and Students', *Educational Studies* 38(4):479–483.
- Valverde, G., L. J. Bianchi, R. G., Wolfe, W. H Schmidt, and R. T. Houang (2002), *According to the Book: Using TIMSS to Investigate the Translation of Policy into Practice through the World of Textbooks*. Dordrecht: Kluwer Academic Publishers.
- Vidal Rodeiro, C. L. (2014), 'Multiple Entries in GCSE/IGCSE Qualifications'. Internal Report, Cambridge Assessment.
- Wilby, P. (2011), 'Mad Professor Goes Global', *The Guardian*, 14th June 2011, <http://www.theguardian.com/education/2011/jun/14/michael-barber-education-guru> (accessed 11th June 2014).
- Whitty, G. (2006), 'Teacher Professionalism in a New Era'. Paper presented at the first General Teaching Council for Northern Ireland Annual Lecture, Belfast.

3 REGULATORY OVERKILL: SCHOOL ACCOUNTABILITY, QUALIFICATIONS, AND THE FUTURE

DALE BASSETT¹

Introduction: view from the back seat

THINGS WERE SO DIFFERENT 30 years ago. There were a couple of dozen exam boards, regionally based and often set up by universities, offering different suites of qualifications for 14–16 year olds (what we now call Key Stage 4) to meet different educational needs. Largely unencumbered by the whims of government, the boards had huge discretion over the content of their qualifications and how they were assessed and awarded, within a loose framework agreed by the Schools Council for the Curriculum and Examinations, a body established by central and local government in partnership with teachers and comprising representatives of all the main stakeholders in education.

Today, things are very different. Four exam boards dominate provision of Key Stage 4 qualifications, one of which is this author's employer. Operating nationally, only one has even a passable claim to a meaningful link with a university. The number of qualifications they offer is now relatively small and decreasing, their content is prescribed in detail by government, and their assessment and grading strictly controlled by a regulator, Ofqual. In this

¹ The author is writing in a personal capacity, and the views expressed do not represent those of his employer. He thanks Daniel Acquah and Ali Wood for reviewing drafts, and to Gemma O'Brien for assisting with research. Any errors or omissions are the author's alone.

regime, innovation, heterogeneity, and stakeholder ownership take a back seat to consistency, uniformity, and central prescription.

This change has had a very significant positive impact on the education system, helping to ensure that the vast majority of young people leave education with at least some useful, respected qualifications, giving them a chance to progress and to succeed.

But it has also had a negative effect, which is becoming increasingly damaging to the education of young people. It has restricted the effective functioning of the qualifications market, rendering schools almost powerless to exercise meaningful influence or choice over the qualifications their pupils take. And it has driven the creation of more homogeneous qualifications that are, of necessity, often designed to prioritise regulatory compliance over educational utility.

This matters because, in the 14–19 phase, qualifications essentially dictate the curriculum young people follow and, increasingly, the pedagogical approach teachers take. If teachers really are professionals, trusted to use their expertise and experience to inform their practice, then imposing a uniform curriculum and a single approach to teaching and learning denies them the discretion that is essential to that professionalism. Many would argue that one curriculum and one approach to teaching and learning cannot possibly be right for all pupils; some might take the opposite view, but if we believe in teachers as professionals, surely it is their call to make. There are several arguments for and against having a choice of exam boards providing a choice of different qualifications, but one of the most compelling in favour must be that by exercising choice where they can, schools have demonstrated just how highly they value diversity and the ability to tailor their curriculum. Many have flocked to the International GCSE (IGCSE) in recent years for precisely this reason.

But in a world of increasingly similar qualifications, teachers have less and less freedom to exercise meaningful choice. The result of this burgeoning homogeneity can be, in the worst cases, exam-centric curricula that fuel student disengagement, militate against the development of a love of learning, and perhaps most damningly, fail to secure the high-quality education that will prepare young people to progress and to succeed.

This chapter argues that the solution to this dilemma is to decrease the regulatory burden imposed by the accountability framework, which currently stifles innovation and heterogeneity in the system. A more light-touch approach will enable a genuine, functioning qualifications market to develop, which would better serve the needs of young people, schools, universities, and employers alike.

A bang and a whimper

The regulation of qualifications that drives the stultifying developments outlined is the direct result of a major paradigm shift, in which government took control of the school system by introducing centralised accountability. Much has been written on the effects of this accountability system, good and bad. However, nobody has sufficiently explored the way in which its design encourages (in fact, requires) qualifications and exam boards to be regulated to such an extreme degree, and the way in which this damages the very thing the school accountability system is intended to improve: the education of our young people.

In 1986, the Thatcher government introduced sweeping reforms to financial markets that quickly became known as the ‘Big Bang’. This burst of deregulation changed the face of finance overnight, fuelling innovation and causing money to pour into the City from across the globe.

At the same time, another series of sweeping reforms was underway in the schools system. Just as revolutionary and transformative, and also driven by government, there was one major difference with what was going on in the financial sector: whereas the financial big bang was deregulatory, its educational counterpart heralded an unprecedented, and in the end stifling, level of regulation of schools.

The Thatcher government’s schools reform package included the introduction of a statutory National Curriculum, of statutory national tests at the end of each Key Stage, and of GCSEs – for the first time, a universal, end-of-school qualification taken by all. These reforms served two purposes. One was about the curriculum itself, with the goal being to ensure a minimum standard or level of provision for every child. The second was accountability. If you want

to have standardised national testing to facilitate school accountability, a standardised curriculum is a pre-requisite – you have to teach the same thing in every school so that you can test the same thing in every school, which is necessary to compare schools for accountability purposes.²

With the accountability floodgates now wide open, initiatives flowed thick and fast. Thatcher's government introduced performance tables; Major's established Ofsted; and Blair's set the first 'floor targets' for school performance, which gained increasing influence under Brown's government and now Cameron's.

Arguably, the accountability agenda has succeeded on its own terms. It has achieved, in large part, what it was meant to achieve. Attainment has increased significantly; the vast majority of young people now leave school with a recognised qualification; and these days badly underperforming schools do not stay underperforming for long.³ It is hard to imagine that anyone would want to reverse the educational gains that have happened over the past quarter-century. The big bang was probably the right policy at the right time, as influential work by Mourshed, Chijioke, and Barber (2010) implies.⁴

But, as anyone who has worked in, or for that matter gone through, the education system in the last decade or so knows, that success has come at a considerable cost.

Homogenisation: or, how 'standards' have trumped quality

All this accountability – which for secondary schools is largely based on GCSE results – has driven changes in qualifications, which in turn have driven changes in curriculum and teaching in many schools.

For the accountability system to work, we declare equivalence between different qualifications – between GCSEs and BTECs, between GCSEs of different subjects (so biology is equivalent to history, for example), and between GCSEs of the same subject, both between and within exam boards (which

2 It is not necessarily the case that standardised testing requires a standardised curriculum. Some large-scale, credible standardised tests do not rely on a common curriculum, with the OECD's PISA survey being a high-profile example. There is certainly a discussion to be had, which is beyond the scope of this paper, on the wider question of how and what we test uniformly at a national level.

3 Indeed, the number of failing schools has fallen from 400 to 150 in the last four years (Adams 2014).

4 Of course, Sir Michael was the driving force behind the data-fuelled accountability explosion in the 2000s.

offer multiple versions of some GCSEs). When we declare this equivalence in a high-stakes accountability system, we create incentives that may result in the so-called ‘race to the bottom’, where schools and pupils are tempted to choose the qualifications that are the easiest (or *perceived* to be the easiest) to pass.⁵ These, of course, are not necessarily the same ones that support the best teaching and learning or the best student progression.

The response of the Department for Education and Ofqual has been to increase prescription. The accountability system assumes that GCSEs are equivalent in terms of content, assessment, grading, size, and the time they require for teaching and learning. This requirement of comparability necessitates a regulatory approach that specifies in detail the structure and content of qualifications, and the way in which they are assessed and awarded – minimising any difference between, say, different mathematics GCSEs.

This is achieved in three ways. First, subject criteria published by DfE determine the content of the qualification (i.e. the curriculum), and specify what skills and knowledge exam boards must assess. Second, assessment objectives and technical guidance published by Ofqual prescribe how the content must be assessed, for example to what extent boards must use exams or teacher-assessed practical work, and even the minimum time each exam must last. Finally, grading or awarding is the most heavily regulated part of the process, with Ofqual detailing precisely the approach exam boards must use when setting grade boundaries. Awarding is very heavily guided by statistical predictions: in the most extreme case, for major qualifications like GCSE English and mathematics, a tolerance of just 1 per cent is set for deviation from the statistics, beyond which it is very difficult for boards to go without overwhelming evidence of a change in pupils’ performance compared with previous cohorts.

Exam boards do try to create qualifications that are as varied as possible to support different educational needs, for example offering different English qualifications that are either based around individual set texts or have a more thematic, genre-based approach – but it is becoming increasingly difficult to innovate in this way as content and assessment become ever more regulated.

5 Whether the ‘race to the bottom’ is actually reflected in the demand of qualifications is a matter of some debate. There is a body of research suggesting that this may not be the case, and particularly so after the strengthening of the regulatory regime in recent years (AQA 2011).

Of course, this is done to ensure that standards are maintained following years of outcry over ‘grade inflation’ – repeated year-on-year increases in pass rates that seemed implausible, were they solely to reflect improvements in learning. The intention, then, is to try to avoid undermining quality. But since regulation only works when it is followed, there is an obvious risk that the bar could be lowered in order to facilitate compliance. Why allow exam boards flexibility over, say, the methods of assessment they use, if there is a risk that doing so could undermine standards – especially in the context of the pressure schools face from the accountability regime? So flexibility is curtailed, resulting in qualifications that may be robust in terms of grading standards, but at the same time allow little room for inspiring teaching and learning.

Why do exam boards not just ignore all this prescription and innovate in the way they wish? Why do schools not shun regulated qualifications in favour of more innovative fare from outside the mainstream? There is one simple reason: if exam boards do not follow the rules, Ofqual will not accredit their qualifications. And if schools do not teach accredited qualifications, they will not count in the all-important performance tables. So, yet again, accountability is to blame.

It is already clear that the need for equivalence is beginning to squeeze out diversity from the system. The role of non-exam assessment, not just in creative subjects but also in English and science, is radically changing and curtailed – a direct regulatory response to the pressures of the accountability system. Vocational and applied qualifications at GCSE level are undergoing a major transformation, and their importance in school performance tables has been significantly reduced. Even the IGCSE – brought into performance tables by the current government specifically because of the flexibility and challenge it provided teachers and learners – is by no means certain to retain its prominence once new performance tables are introduced in 2016.

It is notable that where exam boards do have room to innovate, they use it. AQA, for example, offers a GCSE-style qualification in further mathematics, which schools are using to stretch their brightest pupils and better prepare them for A-levels. This is a challenging, rigorous, and innovative qualification, lauded by pupils and teachers alike, and it works because it sits outside the main accountability system and the dominating need for comparability that requires

prescription and regulation. Similarly, the Extended Project Qualification – deliberately designed not to fit the A-level or GCSE moulds – has proved excellent at giving pupils the space to develop their independent research and writing skills, and is all the more popular with universities because of this.

Nevertheless, overall, the drive towards equivalence and standardisation has stifled innovation and decreased diversity in qualifications and assessment. While the most immediate effect of the accountability-driven regulatory regime is on the substance of qualifications themselves, there is another impact that over the longer term could also have a pernicious effect.

Oligopoly: or, how high-stakes accountability stifles the benefits that might accrue in a functioning qualifications market

The market for GCSEs is almost exclusively supplied by four exam boards. For most of the core subjects in the English Baccalaureate,⁶ just two boards account for over 80 per cent of qualifications awarded. It is true that there are historical reasons for this, notably mergers between boards encouraged by previous Secretaries of State. But today, it is extremely difficult for this market to diversify further.

The regulatory burden imposed on exam boards offering high-volume, high-stakes qualifications such as GCSEs is simply huge. There are, for example, 54 ‘general conditions of recognition’ with which any approved board must comply. Although Ofqual theoretically takes a risk-based approach to regulation, boards that offer major qualifications like GCSEs are subject to detailed ‘close and continuous’ monitoring of almost every aspect of their operational activities.

This burden of regulation constitutes an extremely high barrier to entry in the GCSE market. If a new exam board (or for that matter an existing one offering, say, specialist vocational qualifications) wanted to begin providing GCSEs, it would have to apply to Ofqual for recognition to do so. Ofqual requires evidence that the board will be able to deliver high-quality qualifications that will be developed, assessed, and awarded robustly; that its operations will be reliable and scalable; and that its financial position is secure. For a new entrant

6 The English Baccalaureate comprises English, mathematics, the sciences (including computer science), history or geography, and a language.

– without a track record to draw on – to demonstrate that it could meet these criteria would be a challenge, to say the least.

Why does this matter? Because it does not take an economist to see that a small number of providers and high barriers to market entry are unlikely to be in schools' best interests. A well-functioning market needs strong competition to ensure that providers are responsive to their customers' demands, providing products and services of quality that meet schools' needs. This is by definition harder to ensure in a closed market with a very small number of competitors.

And we know that new entrants to a market are often the disruptive innovators, developing radical new products that depart from conventional approaches to meet customers' needs in completely new ways – and in the process shaking up entrenched provision and forcing those players to up their game. If we want innovation that will improve the quality of what schools are buying, and potentially have a positive impact on teaching and learning itself, we need energetic, nothing-to-lose start-ups to challenge the risk-averse, traditional thinking of the existing giants.

Without a driving need for homogeneity in the qualifications market, and the regulation that assures it, barriers to entry could be lowered, schools could enjoy a properly-functioning market that put their needs first, and pupils could benefit from heterogeneity and innovation that would give teachers the flexibility to make learning more engaging, challenging, and inspiring than a one-size-fits-all approach to qualifications ever could.

It is important to acknowledge that these innovative new qualifications will not always be good. Exam boards will sometimes get it wrong or fail to persuade universities/employers of a new qualification's value. And schools will sometimes make a poor choice for a particular pupil. But this is a risk that is equally present in the existing centrally managed system: countless examples from GNVQs to the Diploma have failed to gain traction for a variety of reasons. Arguably, wholesale failure may even be a greater risk in a centralised system, especially where government mandates the creation of qualifications for which there is no demand from employers or higher education (Key Skills being a prime example).

But if schools' incentives are centred on ensuring that pupils progress, rather than acquiring as many certificates as possible, they will want qualifications

that support high-quality teaching and learning, and develop the skills young people really need – and they will demand that those qualifications have currency in the real world. Exam boards will have to exemplify to business and higher education what a particular qualification means its holder can actually do. Getting those incentives right is the essential pre-requisite to allow this to happen.

Slaying the accountability behemoth

We are faced with a dilemma. On the one hand, we have a school accountability system that, in addition to its well-documented impact on schools, demands a regulatory regime for qualifications so prescriptive that it is gradually depriving schools of a functioning market – geared towards their needs rather than the regulator’s requirements – and homogenising the curriculum for every one of the 600,000 or so young people per year who sit GCSEs. It goes far beyond ensuring a minimum standard, depriving teachers of the freedom to teach the way they want and to tailor the curriculum to context or a given pupil’s needs. On the other hand, it is undeniable that the same accountability system has driven a vast improvement in average school performance and almost eliminated the plague of chronically underperforming schools, while ensuring that nearly every young person leaves school holding a piece of paper that has currency with employers, colleges, and universities. The challenge to policymakers is to retain these positives while ameliorating the negatives.

The school accountability system as we know it has played a vital role, but its usefulness in its current form is arguably nearing the end. The tipping point at which it does more harm than good must come soon, if it has not already. It dictates curriculum to too great an extent, and puts far too much emphasis on summative assessment. As long as the system defines school success in terms of performance in high-stakes summative assessment, it will drive a reductionist approach to curriculum and demand a qualifications market so tightly regulated that it severely inhibits innovation, whether from existing providers or would-be new entrants that are unable to scale the regulatory barrier standing between them and the market.

A centralised accountability regime can be a way of guaranteeing a minimum standard or incentivising improvement, but it needs to be much broader and much more sophisticated in order to do that in a way that provides room for innovation. The move from a threshold-based floor standard to one based on pupils' progress is clearly a significant improvement: by removing the laser focus on a single borderline between 'pass' and 'fail', schools will at least be rewarded (or penalised) for how all their pupils perform. But any approach that relies solely on exam results remains extremely limited. By continuing to pile more pressure onto qualifications than they can bear, severe regulation will be just as necessary to maintain standards under the new system as it is now – and diversity, innovation, and education quality will continue to be threatened as a result.

Policy-makers must therefore start working towards the next evolution of the accountability system. School accountability must become far more about where young people end up than how they get there. And in an era where teaching is becoming increasingly professionalised, outstanding school leaders have a wider reach than ever before, and peer support and challenge is ever more commonplace, we should have the vision to recognise that government accountability is not the only accountability – and perhaps, in the future, not even the best. A subtler and more diverse approach would yield many benefits for the education of young people. One could undoubtedly be a diverse qualifications market focused on delivering rich, rigorous curricula and assessments that support the best teaching and learning and open doors for young people, instead of prioritising regulatory compliance above all else.

Conclusion: producing qualifications that support education

An accountability system that places significant emphasis on other outcomes, in addition to exam results, might allow for qualifications that do not need to be absolutely comparable or 'equivalent' at the expense of being as good as possible educationally for pupils and teachers.⁷ Of course, it is true that many good teachers try, as far as possible, to ignore exams and focus instead on making sure

7 The question of what such an accountability system might look like is outside the scope of this chapter, but interesting contributions have been made by the likes of ASCL (2012), the Headteachers' Roundtable (2013), and my colleagues at AQA (2013).

their pupils benefit from excellent teaching, rooted in a rounded, stimulating curriculum. But it should not have to be like this. Why is it impossible to have fantastic qualifications that actually support good teaching and learning, rather than hindering it?

To be clear, Ofqual cannot fairly bear the responsibility for this situation. It has, as a new regulator, done a good job in difficult circumstances, and has through regulation tackled some of the biggest problems in qualifications that emerged as a consequence of an ever more pressuring accountability system. Nor is it the fault of any one government or minister. The accountability system and the regulation that flowed from it may have begun with Keith Joseph's reform initiative, but it has been its steady expansion over the past quarter-century that has resulted in the overbearing system we have today.

With the pressure on individual qualifications greatly reduced, the opportunity could arise for a move to genuinely risk-based, exam board-focused regulation, rather than today's heavily qualification-centric approach. Ofqual would rightly still ensure that boards have the capacity and capability to deliver quality and reliability, but could put greater trust in boards' assessment expertise and customer-responsiveness to ensure that schools had a diverse range of high-quality qualifications to choose from in each subject.

Reduced prescription would give exam boards the freedom to innovate. Removing the need for strict comparability between qualifications, and reducing the weight of the accountability system on summative assessment, would also have benefits for validity, for example by allowing an increased role for teacher or other non-exam assessment where it would help to assess skills or knowledge in ways in which traditional exams cannot. A lower regulatory burden could encourage new entrants to the market, and also help to ensure that breadth of provision is maintained – the demands of compliance and monitoring have certainly been a challenge for DfE, Ofqual, and the exam boards during the current wave of GCSE and A-level reform.

Moving towards this system may seem impossible in an age of oppressive accountability and prescriptive regulation, but the prize is great: qualifications that help teachers to fascinate, stimulate, and challenge their pupils to prepare them for a future world that we can barely imagine.

References

- Adams, R. (2014), 'More Schools in England Meeting Minimum Standards', *The Guardian*, 23th January 2014, <http://www.theguardian.com/education/2014/jan/23/schools-improve-performance-dfe-data> (accessed 28th May 2014).
- AQA (2011), 'Response to the Education Select Committee Inquiry Into the Administration of Examinations for 15–19 Year Olds in England', Centre for Education Research and Practice, https://cerp.aqa.org.uk/sites/default/files/pdf_uploads/CERP-IP-CERP-07112011_0.pdf (accessed 20th May 2014).
- AQA (2013), 'Secondary School Accountability'. Consultation Response and Discussion Paper, <http://filestore.aqa.org.uk/news/pdf/AQA-NEWS-ACCOUNTABILITY-CONSULTATION-DISCUSSION-PAPER-130424.PDF> (accessed 1st April 2014).
- ASCL (2012), 'Intelligent Accountability', Policy Paper No. 86, <http://www.ascl.org.uk/download.77099223-2E0D-4343-A9C423C1D3EFF021.html> (accessed 10th May 2014).
- Headteachers' Roundtable (2013), 'Accountability Consultation', <http://headteachersroundtable.wordpress.com/accountability-measures-consultation/> (accessed 20th April 2014).
- Mourshed, M., C. Chijioke, and M. Barber (2010), *How the World's Most Improved School Systems Keep Getting Better*. McKinsey & Co: London.

4 THE VOCATIONAL QUESTION: IN PURSUIT OF QUALITY RATHER THAN EQUIVALENCE

GEOFFREY HOLDEN

Introduction

FEW WOULD QUESTION THE value of a good education. Yet there are different forms of education, and this heterogeneity requires a multi-pronged approach to qualifications and assessment. This is most clearly displayed in the case of vocational schooling, which remains highly controversial. Proponents claim that it motivates young people, and facilitates re-engagement with education, while others argue that it bifurcates schooling into a two-track education system in which some will inevitably lose. Partly reflecting these different views, consecutive governments have introduced swaths of reviews, reforms, and proposals to improve the system. And yet, we still do not have a stable, well-understood, and respected vocational pathway.

The reasons for this are complex, but at its heart, in the context of general widespread confusion over incentives, accountability, and performance, are two key policy fallacies:

1. the enduring belief that the problem of vocational qualifications is one of supply and that the 'right design' will solve the problem; and
2. the simplistic view that 'vocational' equals 'occupational' and that they can be measured in the same way.

Wider deficiencies in the system underlie these issues. First, there is no alternative subject-based education pathway for lower or later achievers after the age of 16. The vast majority of young people currently sit their GCSEs at the end of year 11. Those who fail to achieve the ‘benchmark’ 5 A*–C, including English and mathematics are offered little alternative than to attempt re-sits or to pursue vocational courses. This means they are effectively cut off from any further subject-based study and have no real opportunity to continue any learning in literature, languages, or the sciences. The vast majority of school sixth forms are selective, requiring a minimum set of GCSEs for entry, and tend to offer mainly A-level provision. Vocational education is therefore offered as the default route for those deemed in this context to be low-achievers, which clearly reinforces negative perceptions of the framework.

Second, reforms to vocational qualifications have been driven by a particular approach to ‘competency’, introduced in the National Vocational Qualification system and set up following the De Ville Review (Manpower Services Commission 1986). In this approach, skills linked to particular occupations became the sole focus of vocational qualifications. Consequently, it led to the dominance of a skills-based approach, while the necessary acquisition of wider knowledge and experience were neglected. The focus on competence, while important, has meant that the processes by which skills are developed, and the specialist pedagogy behind successful vocational education, do not get the attention they deserve.

While incentives are intended to guide behaviour in a particular direction, there is always a danger that they deliver perverse consequences that run counter to the original intentions. This is particularly likely to happen when an incentive to behave in a particular way is also used as an accountability measure. As Wolf (2011) noted in her report on vocational education, the incentives in place, based on the performance and funding systems, encourage the teaching of qualifications which attract the most performance points or the most funding – not the qualifications that help young people to progress. Consequently, young people often take qualifications that are not going to help them succeed in the labour market, a clear unintended consequence of the current framework.

The truth is that vocational pedagogy and assessment are quite distinctive to those in the academic field. By viewing vocational reforms solely through the

academic lens, while concentrating on skills and downplaying the crucial role of knowledge, policymakers have ended up repeating the errors of the past. Only by setting basic parameters and then standing back can the government offer the right incentive for lasting reform – one that is not driven by a centralised top-down approach, but instead ensures space for innovation to develop organically within the market.

Qualification-based reform and the growing role of the state

Before the mid-1980s, government played a minor role in curricula and qualifications in England. Since then, however, it has expanded its remit considerably, introducing, reforming, and abolishing a huge range of governmental or quasi-governmental bodies, qualifications, and frameworks. Significant disruption has been caused as an unintended consequence of a flurry of Education Acts, as consecutive governments extended their role in the field. The School Examinations Council (SEC) was established in 1984, in a move designed to reduce the individual influence of teachers in curriculum development and establish a national approach. Its members were nominated by the Secretary of State, and in 1986 the National Council for Vocational Qualifications (NCVQ) was set up to promote National Vocational Qualifications (NVQs), again staffed with government appointees.

More bodies followed thick and fast. In 1988, the Education Reform Act made provision for two new councils to be established: the National Curriculum Council (NCC) and the School Examinations and Assessment Council (SEAC). 1991 saw the proposals for General National Vocational Qualifications (GNVQ) to be delivered in schools or colleges to complement NVQs awarded in the workplace. GNVQs grew from the initial pilot to become a national offer by 1993. They were expected to both secure a young person entry to higher education and place vocational education and training on equal standing to A levels, and to be clearly related to NVQs so that young people could ‘progress quickly and effectively’ from them to NVQs if that was their choice. Their take up was in part driven by powers introduced by the 1988 Education Act, which gave the Secretary of State control over post-16 provision. This desire for GNVQs to meet two quite distinct purposes led to a questioning of their

credibility by their main external users. In addition, they added another layer to the vocational offering for young people, exacerbating, rather than resolving, the problem of a ‘qualifications jungle’. However, although popular, at advanced level they suffered from the inevitable ‘academic drift’ in a misguided attempt to prove parity with A levels, and they were phased out in the early 2000s.

Shortly after the GNVQ scale up in 1993, the Education Act abolished both the NCC and SEAC and replaced them with the School Curriculum and Assessment Authority (SCAA). This was followed by another Education Act in 1977, which abolished the NCVQ and the SCAA, replacing them with the Qualifications and Curriculum Authority (QCA) and a similar body for Wales. These bodies were responsible for overseeing both academic and vocational qualifications. The Act also brought the first major step in the establishment of direct state control over courses in all state schools, which in turn led to external qualifications. Perhaps most surprising is that this significant extension of state control over qualifications was introduced by a Conservative administration. This approach was retained and expanded under the successive Labour governments, which followed shortly afterwards.¹

The next significant step in vocational qualification reform came in 2004 when the working group, chaired by former chief inspector Mike Tomlinson, published its report ‘14–19 Curriculum and Qualifications Reform’. Tomlinson (2004) identified a number of problems, ranging from the UK’s poor record on keeping teenagers in school and their low skill levels in numeracy, literacy, and ICT, to the poor status of vocational courses and qualifications, and the difficulty of differentiating between thousands of pupils with top grades in their A Levels. This was attributed to the complexity and lack of transparency in the web of academic and vocational qualifications. Tomlinson’s key recommendation was to replace GCSEs, A Levels, and vocational qualifications with a new single modular diploma at four levels. There was political reluctance to announce the abolition of the ‘Gold Standard’ A level, and the government rejected most of Tomlinson’s recommendations. It responded with proposals for vocational 14–19 Diplomas covering 14 occupational sectors, but also decided to keep the existing ‘gold standard’ GCSE and A-Level exams. The 14–19 Diplomas were

1 Dates and bodies are sourced from Education in England: <http://www.educationengland.org.uk/index.html>.

introduced, but take up was low despite significant public funding support, and their demise followed the 2010 general election.

At the same time, the QCA embarked on a major programme for reform of the qualifications system. Taking its remit from the 2003 White Paper ‘21st century skills: realising our potential’, the agency proposed that qualifications reform should be underpinned by a unit-based national system of credits. This ‘qualifications system’ encompassed more than just the National Qualifications Framework (NQF). The outcome was the Qualifications and Credit Framework (QCF) – and meant the conversion of all existing vocational provision into standard unit formats, each given a credit rating, and combined into different sizes of qualifications.

The next significant change was in 2009 when the Apprenticeships, Skills, Children and Learning Act created the Young Person’s Learning Agency, the Skills Funding Agency, the Office of the Qualifications, and Examinations Regulator (Ofqual). It also created a new agency to carry out the non-regulatory functions currently performed by QCA.

This Act gave the Secretary of State further extensive powers to

1. define the content of certificates for apprenticeships;
2. stipulate which courses – other than mathematics, English, and ICT – pupils aged 16 to 19 should be entitled to study;
3. direct a local authority to provide information about accountable resources held, received, or expended by its schools; and
4. stipulate the minimum level of attainment in literacy and numeracy needed for qualifications for young people aged 19 or over. Even Ofqual, which was set up to be independent of ministers and reports to parliament, has to consider the powers given to the Secretary of State and his successors.

Then, in 2010, the Coalition confirmed that it was closing the Qualifications and Curriculum Development Agency (QCDA), which had been created in 2008 when the QCA was split into the QCDA and Ofqual, the watchdog for exam standards. The decision to close the QCDA raised some concerns, since its job was to give independent advice based on its members’ experience as

curriculum developers and former teachers, and sometimes even to challenge politicians. Curriculum planning was now the responsibility of the DfE, with a panel of government-appointed ‘experts’ offering advice.

Qualification-driven reform is an attractive option for politicians since it gives the impression of change and that they ‘get something done’. However, these reforms have been invariably undermined by subsequent links to funding, inspection, and accountability regimes, distorting the original intentions and leading to reviews, revisions, or even abandonment in favour of another set of reforms. If there is any lesson to be taken from the history of government intervention in education, it is that centrally driven, qualification-based reforms have a poor track record. NVQs, GNVQs, Modern Apprenticeships, and the 14–19 Diplomas all required reviews and amendments within a few years of introduction or were scrapped. Indeed, the QCF now appears to be heading the same way, as Ofqual (2014) noted recently.

A broad education system

Providing young people with learning opportunities that will enhance their lives in the fullest sense should lie at the heart of all education. It is also a truism that not all people learn in the same way – some have a greater facility for dealing with concepts and abstract thinking, while others will show an aptitude for solving practical problems or working with their hands (Claxton, Lucas, and Webster 2010; Lucas 2007). A very few multi-talented people can operate at a high level in both domains.

Reflecting the prevailing zeitgeist, societies tend to place greater value on certain talents over others. For many years, it has undoubtedly been true that English pupils who have demonstrated aptitudes in academic subjects have been more highly regarded than those who are more at ease in a vocational environment. Some of these values are at odds with the creation of a successful workforce in a labour market where a constantly evolving set of knowledge, skills, and personal attributes are needed.

There is certainly a case to be made for an assessment at the age of 16 to ensure that all young people have gained a broad general education as a sound basis for progression and further study in more specialist areas. The current weakness

in the English system is rather that it makes no allowances for those who, for whatever reason, have not yet achieved this at 16. Currently, a significant number of 16-year-olds fail to meet the current benchmark of 5 A*–C GCSEs, including mathematics and English; in 2012/13, the figure amounted to just over 40 per cent of the cohort (DfE 2013). For those who have not yet achieved their GCSEs, the lack of a subject-based curriculum after the age of 16 is a serious weakness, reinforcing the perception that the vocational route is for the ‘less capable’ pupils.

There are those who argue for a single ‘unified system’, as envisaged in the original remit of the Tomlinson Report and repeated in the latest paper from Labour’s Skills Task Force (2014). This is in part derived from the view that it is in some way discriminatory to have a tracked system with clear vocational and academic pathways. In this view, the tracked system condemns many young people to a second-class route, while the elite follow the ‘Royal Route’ through A levels to higher education.

But the experience of other countries where tracked systems are well established suggests that this does not need to be the case. European countries with vocational routes ensure that the same elements of general education are included in the vocational schools – but contextualised within the vocational setting in order to maintain engagement. Critical success factors in continental systems are that: (1) selection, advice, and guidance are of high quality; (2) routes are clearly aligned with the labour markets and further education; and (3) that there are bridges between the routes.

In contrast, the current English system fails too many. There is no doubt that a broad education is an essential grounding for life and work, and this is the policy adopted by the current administration following the recommendations in the Wolf Review. Nevertheless, policymakers should not assume that one set of capabilities is worth more respect or greater financial support than others. As a society, we need people who ‘do’ as well as people who ‘think’ – and the best vocational programmes produce people who can do both.

If we do not recognise at an early stage whether or not a young person may be more suited to one pathway over another, we run the risk of wasting (expensive) investments in education. This does not mean that pupils should be locked into one path at a particular age; they may have the right and opportunity to change

their minds in future. The answer may thus be to allow the development of an offer alongside GCSEs, which would enable young people to explore the world of work and get a feel for particular occupations.

For many pupils, the first important educational decision will be their GCSE choices, and it is therefore important that high-quality guidance is available to parents of 12- or 13-year-olds, so that they can navigate a set of complex choices in discussions with their children. A full suite of GCSEs is clearly not right for all, but we still do not provide other high quality pathways for those who are less suited to them. We should be offering imaginative and worthwhile programmes that will tap into these pupils' learning styles and interests.

This is not to say that they should be denied the opportunity to explore and be exposed to the diversity of a traditional liberal education. It is also a mistake to abandon the core curriculum elements of English and mathematics, which can and should be delivered to all children up to the age of 16. Opportunities to start learning a trade can be a valuable part of the overall educational provision, but this experience must be balanced by broader learning.

There is a tendency to use the terms 'vocational education' and 'practical learning' almost interchangeably, but this is not helpful. It sets up a false choice between 'academic' and 'vocational', and therefore fails to take account of the subtle shadings that exist between a curriculum that introduces the world of work at one end, and the development of specific occupational competence at the other. Unpacking 'practical learning', it is easy to see that the term can describe everything from art, music, dance, and sport, on the one hand, to pottery, cabinet making, and metalwork on the other. Either grouping may be vocational, but they may equally just be pursuits in which an individual has developed, or wishes to develop, a talent.

Similarly, there are a number of vocational subjects that are offered in academic institutions, which employ many characteristics of academic teaching. The most obvious examples are medicine, architecture, and accountancy, but even engineering is often delivered in a very academic way. So the line between 'academic' and 'vocational' is clearly not as clear-cut as commonly thought.

Levels and the myth of equivalence

In recent years, attempts to introduce frameworks covering all qualifications have led to the use of ‘levels’, to describe the level of education to which a particular qualification corresponds. The creation of NVQs at levels 1–5 was accompanied by the introduction of the NQF, which includes the general academic options of GCSEs and A Levels. Alongside this were descriptors attempting to define the levels in terms of the various domains. This has led to the perverse consequence of equivalence being read into the system. Vocational qualifications have been designated levels that are said to be broadly equivalent to academic qualifications, but although administratively convenient, it bears little relation to reality.

The imposition of this system assumes that status could be conveyed on vocational education by demonstrating its equivalence with academic education. But this provides nothing but a bureaucratic convenience with little relevance in the real world. To claim that an A level in art is equivalent to an A level in physics is a meaningless comparison. Yes, the two qualifications may offer an equal level of challenge, and they are valuable as progression markers for the next stage of learning, but other than that they are simply not comparable.

Even within the world of work, there is no direct comparison between different occupations just because they are so different. A good hairdresser is not ‘equivalent’ to a good electrician or a butcher. They have different skills, but the crucial thing is if they have the *right* skills and knowledge to do their job well. It is important to have progression ladders within an occupation, and these do not always neatly fit within an accountability framework that emphasises vertical progression. For example, an automotive technician might need further training to undertake specialised repair, but it makes no sense to try to determine the ‘level’ of this training – it is simply training to obtain an additional skill. Unfortunately, such simplicity is undermined by funding rules and accountability measures, which focus on vertical progression rather than recognising and valuing lateral progression.

The problem of competence

A key problem with the current vocational system is its heavy focus on skills over knowledge in its conception of 'competence'. Research by City & Guilds found that ideas about what constitute competence affect the usefulness of a qualification. Indeed, Brockmann, Clarke, and Winch (2008) argue that the English system's 'skills-based' approach lacks a developed notion of citizenship, of broad competence development, and of occupational identity. It neglects general education as well as personal development. In contrast, vocational education in the Netherlands and Germany takes a 'knowledge-based' approach, where content is high in theoretical input, valuing both tacit and explicit forms of knowledge, and ideas of personal development and civic education. Competence, in these countries, is seen as a multi-dimensional concept, whereby individuals integrate theory and practice, bring together resources, and apply the 'whole' by reflecting on a given work situation and upon their own actions. In this way, as Brockmann, Clarke, and Winch (2008, p. 562) put it, 'Students and workers thus become producers of knowledge, central to the success of knowledge-based labour processes'. Unlike the situation in England, 'employability' is not solely focused on the interests of employers-at-large, and qualification and curriculum designers are not operating under centrally driven directives.

Defining competence is indeed a complex endeavour, without a straightforward solution. For this reason, a key problem with the strong assessment focus on easily observable actions, behaviours, and outcomes is clearly that important aspects of competence may effectively be removed from assessment simply because they cannot be easily observed and measured. Consequently, the drive towards accountability has meant that important, broader knowledge has been removed from the concept of competence in vocational education.

It was not always like this. Academic education has of course always been viewed as a coherent programme of learning, taking people along a well-understood continuum from basic principles to the most advanced study. Learners, teachers, and parents understand the progression from say GCSE physics through A level to degree and postgraduate work. It involves gradually learning a body of knowledge and the associated skills. But it is also true that this continuity and coherence were mirrored in the vocational field up

to the latter part of the 20th century. Examples include the City & Guilds suites of Foundation, Craft, Advanced Craft, and Technician qualifications. Achieving 'City & Guilds advanced craft' status was highly valued by those seeking to progress in their chosen field and those who had this qualification were welcomed by employers. These qualifications involved not only learning and practising a set of skills until they could be performed to a pre-determined standard, but also the underlying theory and principles of mechanics, science, mathematics, as well as broader learning.

In addition, City & Guilds used to offer a nationally validated programme of 'Centre Devised' awards. In these, providers could propose revisions to the syllabi and/or local assessment variations. These proposals were reviewed and validated by subject experts and quality assured by City & Guilds. Responsibility was devolved to the professionals, which enabled providers to respond quickly and flexibly to local needs, while still maintaining a national standard. This flexibility was unfortunately also the awards' downfall, since they did not fit into the approved categories devised by QCA for the NQF in the way it developed. Ultimately, the awards were therefore withdrawn, as the lack of funding for providers made their delivery untenable.

Smithers (1993) clearly demonstrated the lack of technical knowledge in NVQs compared to the old City & Guilds qualifications. This is clear when we compare, for example, the contents of the old City & Guilds Process Plant Part 3 qualification with its nearest equivalent, the Level 3 NVQ in Materials Processing and Finishing. The Process Plant qualification had a module on Process Science, which expected that pupils would be able to grasp issues such as atomic structure, chemical bonding, chemical equilibrium, and reaction kinetics. The goal was to ensure 'a generalised mastery of the science background required for a proper understanding of the technology used in practical tasks, so that they may progress to other applications of it in new tasks or new training without re-learning the background studies'.²

In contrast, the only science knowledge demanded in the equivalent NVQ concerns generalities, such as having knowledge and understanding of how to use other information sources – including standard reference charts for limits

2 Content quoted from City & Guilds archive material.

and fits, tapping drill reference charts, and metal specifications – imperial and metric systems of measurements, and the meaning of symbols and abbreviations on the documents used.³ This may adequately describe the knowledge necessary for competent performance in the work place, but does not provide wider domain knowledge or offer any potential for personal advancement.

Unfortunately, the approach required following the introduction of the QCF was arguably even worse for the cause of good vocational education in England than the introduction of NVQs. It was based on the premise that individual ‘units of assessment’ could be placed in a central databank and combined to meet individual needs. These units could be packaged into qualifications, but each unit is assessed alone and there was no longer any holistic approach to a learning programme or any concept of curriculum. Furthermore, by requiring each unit to be derived from the relevant National Occupation Standards, they merely reflect current practice in a particular occupation without adding any wider knowledge.

The methodology underlying the competency model requires candidates to demonstrate achievement of every single component part, which led to some extreme examples of units with endless lists of ‘assessment criteria’, each of which had to be achieved (or ticked off) in order for the candidate to achieve the qualification. This may be reasonable in an on-the-job context, but when delivered in education settings driven by financial incentives based on qualification attainment, it can place downward pressure on standards. In addition, by focusing on endless descriptions of small pieces of learning, any overarching judgment on all-round ability has been lost.

The changes have been introduced in an attempt to improve the allegedly low quality of vocational education and training. They were part of an attempt to align education more closely to the ‘needs’ of industry and commerce, and to rectify some of the knowledge, skill, and attitude deficits of school leavers. This type of instrumental, economic analysis remains important in political debates about education across the main political parties. But there is still relatively little discussion about whether vocational education (or any form of education) should, or can, play the functional role assigned to it by the prevailing instrumental discourse.

3 Content cited from City & Guilds archive material.

Indeed, tensions between vocational education and economic policy remain. At least for full-time qualifications, some progress has been made following the Wolf Review, which accepted that narrow, outcome-led learning does not meet the needs of a knowledge economy. That approach to learning is simply not compatible with the depth and breadth of understanding needed to compete in an increasingly globalised market place.

Knowledge- and curriculum-based reform

In 2001, City & Guilds commissioned a research project to investigate the issue of the role of knowledge in a vocational curriculum. The key conclusion was that ‘the core of a high quality system of vocational education is the knowledge that it enables students to acquire’.⁴ This knowledge must encompass both theoretical and specialist elements applicable to a range of occupations, and help pupils in their chosen careers.

Vocational knowledge is clearly linked to the underlying academic discipline of the profession. For example, physics and mathematics underpin traditional vocational occupations in engineering and construction, while social sciences and the humanities underpin occupations in the service industries. At the same time, vocational knowledge also leads to the skills and knowledge demands of specific occupations. It is this combination that should form the basis of the vocational curriculum. The workplace knowledge enables the student to develop broad occupational skills, while the theoretical element enables the learner to see beyond the workplace and provides a sound basis for further progression both within and beyond the sector.

The point is that all pupils should have the right to learn particular occupational skills of their choice, but there must also be an entitlement to a more general intellectual and critical understanding of the world of work. In this way, academic knowledge would underpin vocational learning. As Dewey (1916, pp. 318–19) argued a century ago:

An education which acknowledges the full intellectual and social meaning of a vocation would include instruction in the historic background of present

⁴ The quote is from the unpublished City & Guilds research project.

conditions; training in dealing with material and agencies of production; and study of economics, civics, and politics, to bring the future worker into touch with the problems of the day and the various methods proposed for its improvement.

Conclusion and policy solutions

The purpose, content, and reliability of examinations have been intensely scrutinised in recent years. But the main focus has been on academic qualifications, such as GCSEs and A levels. The Wolf Review (2011) brought a welcome focus on the vocational offer, but there are still concerns in terms of the implementation of the ideas into reforms. Current government policies are attempting to avoid the errors of the past by taking a less prescriptive approach, while at the same time insisting that ‘employer-led’ developments are the way forward. The hard part is finding the right balance between employer involvement and employer prescription. As argued above, it is the possession of the relevant body of disciplinary knowledge that enables young people to broaden their horizons and look beyond the narrow focus on the workplace. A broad view of work and occupation is therefore necessary.

The approach for 14–19 is indirect. Rather than attempting to design new national qualifications for vocational provision, the DfE has set out certain guidelines and characteristics to which qualifications must conform if they are to receive recognition in the performance league tables. These include not only external assessment, but also a significant element of synoptic assessment. This is a move in the right direction, but still suffers from the underlying assumption that to be of ‘high quality’, vocational qualifications must be judged by the standards of academic provision. This means that they must feature external assessment, involve grading, and be of a minimum size. It is entirely understandable that these features are chosen, since they are well known and it is easy to make sure they are implemented. Yet it does not follow that these features by themselves ensure quality within the vocational context. As argued in this chapter, the essence of a good vocational qualification is the development of both a body of knowledge as well as the associated skills and understanding – one without the other is indeed of little value.

NVQs and GNVQs focused on skills and competencies, with the government trusting that knowledge would follow alongside automatically. Meanwhile, the QCF atomised content down to specific sets of assessment criteria with no recognition of holistic learning and achievement. In turn, the current approach risks a weighting to the more easily externally assessed elements of knowledge. We need carefully designed combinations of education, training, and experience. The critical point in the vocational context is that these will vary from area to area – as noted earlier, the issue should not be about comparability or equivalence, but about what is *right* for any given occupation or vocation.

The place for vocational options within full time education for 14–19 year olds has to reflect the changing world around us, and recognise that the educational experience of a young person has a profound effect on their future careers and lives. Young people need to be prepared for a changing society and for structural changes in the labour market. For vocational education, this issue remains critical. Employers regularly call for employees with wider skills, such as problem solving, the ability to work in teams, resilience, and entrepreneurialism, in addition to high-level functional skills and technical expertise.

A curriculum-based approach to vocational education could go a long way to solve the problems. We need to re-establish accepted and understood progression routes in the vocational area. Given the space and freedom to operate, awarding bodies could create such programmes – indeed, they have to come from the sector, not from government departments or their agencies. If there is anything the experience of educational reforms in the past decades has shown, it is that centrally devised and government-controlled initiatives tend to fail. This contrasts with the continuing popularity with learners and employers of qualifications where the development has come from the awarding sector and practitioners (e.g. Stanton 2008).

The most effective and healthy incentive any government could offer would therefore be to stand back from any further direction and interference in the development of qualifications, frameworks, or equivalences. Awarding bodies have a long history of working closely with employers, academics, as well as school and college providers. They are already subject to an external regulator and accountable for the quality of their offer. More importantly, they are accountable to those who make use of their qualifications, whether as providers

of learning, learners seeking employment and progression, or employers looking for respected qualifications that meet their needs.

References

- Brockmann, M., L. Clarke, and C. Winch (2011), 'Knowledge, Skills, Competence: European Divergences in Vocational Education and Training (VET) – The English, German and Dutch Cases', *Oxford Review of Education* 34(5):547–67.
- Claxton, G., B. Lucas, and R. Webster (2010), *Bodies of Knowledge: How the Learning Sciences Could Transform Practical and Vocational Education*. London: Edge Foundation.
- Dewey, J. (1916), *Democracy and Education*. New York: Free Press.
- DfE (2013), 'GCSE and Equivalent Results in England 2012/13', SFR 40/2013, National Statistics, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/251184/SFR40_2013_FINALv2.pdf (accessed 29th May 2014).
- Labour's Skills Task Force (2014), 'Qualifications Matter: Improving the Curriculum and Assessment for All'. The Third Report of the Independent Skills Taskforce, http://www.yourbritain.org.uk/uploads/editor/files/Skills_Taskforce_3rd_report.pdf (accessed 29th May 2014).
- Lucas, B. (ed.) (2007), *New Kinds of Smart: Emerging Thinking About What it is to be Intelligent Today*. London: The Talent Foundation.
- Manpower Services Commission (1986), *Review of Vocational Qualifications*. London: HMSO.
- Ofqual (2014), 'Chief Executive Report', Paper 117/13, <http://www.ofqual.gov.uk/documents/paper-11713-chief-executives-report/> (accessed 29 May 2014).
- Smithers, A. (1993), *All our Futures: Britain's Education Revolution; A Despatches Report on Education*. Channel 4 Books: London.
- Stanton, G. (2008), 'Learning Matters', Report, CfBT, <http://cdn.cfbt.com/~media/cfbtcorporate/files/research/2008/r-learning-matters-report-2008.pdf> (accessed 29th May 2014).
- Tomlinson, M. (2004), '14-19 Curriculum and Qualifications Reform', Final Report of the Working Group on 14–19 Reform, <http://dera.ioe.ac.uk/11961/8/annexes%20to%20final%20report.pdf> (accessed 29th May 2014).
- Wolf, A. (2011), 'Review of Vocational Education: the Wolf Report', Department for Education, <https://www.gov.uk/government/publications/review-of-vocational-education-the-wolf-report> (accessed 29th May 2014).

5 INCENTIVES AND IGNORANCE IN QUALIFICATIONS, ASSESSMENT, AND ACCOUNTABILITY

ROBERT COE AND GABRIEL HELLER SAHLGREN

Introduction

IN RECENT YEARS, IT has become clear that qualifications, assessment, and accountability drive the curriculum. That which is perceived to gain credit on high-stakes assessments is what will get taught in schools. Successive governments have also invested hope in the idea that changes to the assessment, qualifications, and accountability frameworks can leverage improvements in system-wide performance. The design of these structures is therefore crucial in determining young people's educational experiences and outcomes.

Yet we know little about how high-stakes assessments should be designed to optimise outcomes. The potential prize it offers makes the prospect attractive, but empirical evidence worldwide indicates that it is easier said than done. It is extremely difficult to foresee all unintended consequences of policy measures to counter these effectively. This means that grand schemes that change the framework significantly and universally often create more problems than they solve.

This chapter argues in favour of a more experimental approach. First, it clarifies the different goals of qualifications and assessment, and the effects we want them to have on motivations, curriculum, and standards. Current English national assessments are not fit for all purposes we want them to fulfil.

One problem is that exam boards are neither required explicitly to spell out the purposes their assessments attempt to fulfil, nor do they have to show any evidence how successful they are in this respect. Our first recommendation is therefore that exam boards should be required by Ofqual to state explicitly which purposes their assessments intend to support, and that they should be required to provide evidence indicating the extent to which they are successful.

Second, the chapter discusses the theoretical advantages and problems with high-stakes accountability, which inevitably impacts on assessment and qualifications; reviews the empirical evidence on its impact; and outlines requirements for the achievement measures used in accountability systems, as well as a typology of different accountability structures. Our second policy recommendation is that assessments used in accountability systems should be designed to meet specific quality criteria, examples of which are stipulated here, and evidence on whether or not they do indeed meet these criteria should be collected.

However, as noted above, it is important to acknowledge our ignorance in terms of our ability to design the perfect system from scratch. In fact, we know little about how different features of the accountability system interact. For this reason, an experimental approach is preferable. Our third policy recommendation is therefore to undertake a research programme investigating the optimal combination of accountability features. By randomly assigning schools to different features, we would be able to radically increase our knowledge regarding the types of accountability that maximise system-wide improvements.

Finally, we discuss ways in which we can reconcile certain educationally desirable practices with the need for accountability. Again, an experimental approach is preferable. Our fourth and final policy recommendation is therefore to advance pilot programmes trialling a range of strategies to reconcile educationally desirable practices with accountability structures. We discuss one conspicuous example, how teacher assessment can be made safe for accountability, and suggest one approach to be trialled.

In sharp contrast to previous attempts to improve the incentive structure within qualifications and assessment, therefore, the chapter acknowledges our ignorance about optimal system design. It emphasises that theories of how to improve the assessment and qualifications system – and how to square it with

demands of high-stakes accountability – must be put to the test in carefully designed trials before they are scaled up to national level.

Quality and purposes of qualifications and assessment

It may seem obvious that ‘high quality’ in assessment is desirable, but it is less obvious exactly what it means. This is because quality has multiple meanings – there a number of different dimensions along which we might choose to define it. A common approach is to start by clarifying the ‘purpose’ of an assessment, in order to provide a basis for judging whether it is suitable. Of course, as Newton (2007) points out, there are different meanings of purpose too, and most assessments have more than one. Nevertheless, the notion of whether qualifications and assessments are ‘fit for purpose’ is useful for determining their level of quality, and we therefore need to be clear what purposes we want assessments and qualifications to support.¹

Newton (2007, p.150) makes a helpful distinction between purpose as the ‘decision, action or process which it enables’ (the ‘decision level’) and purpose as ‘the intended impacts of running an assessment system’ (the ‘impact level’). While listing eighteen distinct uses of assessments, he points out that these are just a selection – and warns of over-simplification by grouping different purposes together, even if they share particular characteristics. In order to evaluate whether a particular assessment is fit for a certain purpose, we do need to be specific. In practice, this means identifying a particular assessment outcome, such as a C grade in GCSE mathematics, and a specific interpretation, use, or decision that might be applied to it. For example, we might stipulate that only candidates with a certain qualification will be shortlisted for a particular job, or that we will interpret a certain grade to indicate that a candidate is able to solve a specific problem (such as dividing an office coffee bill in proportion to the number of days each worker are in the office). This level of specificity may seem excessive, but it is necessary to avoid discussing generalities that are too vague to be testable. At the very least, such generalities need to be exemplified

1 It is common to invoke the concept of validity as a key element of quality. But as Newton and Shaw (2014) show, validity is itself a concept with multiple meanings, whose definition is unclear and contested. For this reason, it is still essential to define the different aspects of quality in which we are interested.

with specific illustrative instances, which allow us to empirically verify whether the outcome is indeed a good indicator of the intended interpretation or use.

Newton (2007) does not interpret purpose as relating to such specific uses and interpretations of assessment outcomes, or has at least not provided the level of detail called for above. Identifying a comprehensive list of purposes is a challenge, but it is important given the common concern that assessments with too many purposes inevitably lead to compromises of fitness (Pellegrino et al. 2001). In order to limit the scope of this challenge, however, we focus on national assessments in England.²

Table 1 lists the main uses of these assessments for decision-level purposes and Table 2 lists them for impact-level purposes. It is important to note that the lists should be seen as a starting point, intended to start a conversation and to illustrate what we think is required, rather than as a definitive listing. Clearly, there would need to be a more systematic, open, democratic, and market-influenced process of identifying and prioritising different purposes before such lists could be seen as final.

Table 1: Decision-level purposes of national assessments

What interpretations or decisions should the assessment support?	Examples	How well do current assessments do this?
1. Indicate specific areas of skill, knowledge, or competence that individuals would be expected to demonstrate in another context.	a) Ability to write accurate English. b) Ability to converse in French. c) Ability to use a spreadsheet to calculate an average of a set of figures.	General qualifications (GCSEs and A levels) are specifically designed not to do this, since overall grades allow for compensation. Some vocational qualifications may support these kinds of interpretations.
2. Identify gaps in learning that need to be addressed.	a) Achieving Level 3 in KS2 reading indicates the need for a catch-up programme in Year 7. b) Achieving a D or lower in GCSE mathematics indicates that continued study in this study is required.	Diagnostic information is very general, and the deficit model implied may be questioned, but these kinds of interpretations are probably broadly sound.

2 Some of Newton's uses (e.g. pupil monitoring or diagnosis) are not relevant to these assessments, so need not feature here. Lists of uses from the US context (e.g. Baker and Linn 2002, p. 5) also provide examples, although some do not readily transfer to England.

<p>3. Allocate individuals to appropriate teaching groups.</p>	<p>a) Setting in Year 7 based on KS2 performance.</p>	<p>Notwithstanding the lack of evidence about the benefits of setting (Higgins et al. 2013), current assessments probably broadly meet this need.</p>
<p>4. Decide whether an individual is equipped to go on to a further course of study or employment.</p>	<p>a) Requirement of C grades in mathematics and English to qualify as a teacher. b) Requirement of at least 5 C grades at GCSE to start an A-level programme. c) Requirement of a B grade in GCSE mathematics to take A-level mathematics. d) General guidance about what combinations of A levels are appropriate, based on GCSE grades. e) Requirement for an A grade in chemistry A level before applying to read medicine.</p>	<p>It is likely to depend on specific judgements, but feedback loops in these decisions help to make the required level appropriate. The alignment between what is assessed by the prior qualification and what is actually required probably varies according to context. In many cases, the relationship may be quite weak or unknown (hence unjustified). Problems of comparability arise if grades from qualifications taken at different times or in different subjects are treated interchangeably.</p>
<p>5. Select which individuals should be offered places, from a larger pool of qualified applicants.</p>	<p>a) Offer of university place made to candidates with highest average GCSE score (or AS UMS score). b) Offer of university place made to candidates with highest predicted A-level grades.</p>	<p>The alignment between what is assessed by the prior qualification and subsequent likelihood of success probably varies according to context. In many cases, the relationship may be quite weak or unknown. This may make it less than ideal, but not necessarily unfair: using the best available predictor is fair, even if the prediction is not very good. On the other hand, if the relationship between grades at different levels is subject to bias from other factors, it will be unfair. Problems of comparability arise if grades from qualifications taken at different times or in different subjects are treated interchangeably.</p>

<p>6. Indicate the effectiveness of teachers or schools.</p>	<p>a) Use of pupils' examination performance to inform teacher-performance management.</p> <p>b) School-level floor targets for exam performance trigger inspection visits.</p> <p>c) Examination grades analysed and interpreted by inspectors as evidence of school quality.</p> <p>d) Examination performance used to inform parents' and children's school choices.</p>	<p>Attributing differences in student achievement to the effects of teaching, even with good adjustment for prior characteristics, is the subject of some controversy.³</p> <p>The ability of inspectors to interpret this kind of information appropriately may be questionable (Waldegrave and Simons 2014).</p> <p>Using value added for school choice decisions is also problematic (Leckie and Goldstein 2009).</p> <p>Problems of comparability arise if grades from qualifications that differ in difficulty are treated interchangeably.</p> <p>Aspects of a qualification that are otherwise valid and educationally sound, such as teacher-assessed elements, may become invalid when they form part of high-stakes assessment.</p>
<p>7. Evaluate the performance of the system or subgroups.</p>	<p>a) Changes in pass rates over time interpreted as evidence of system change.</p> <p>b) Differences between pupil subgroups (e.g. pupils on free school means versus those who are not) interpreted as evidence of the level of equity.</p> <p>c) Performance of subgroups which have experienced an intervention used for evaluation.</p>	<p>Problems of comparability arise if grades from qualifications taken at different times or in different subjects/qualifications are treated interchangeably.</p> <p>Comparisons of the size of a performance gap at different times require assumptions about the comparability and interval nature of the reporting scales, which are likely to be problematic for existing qualifications.</p>

3 Concerns about the interpretation and use of value-added data for teacher evaluation have come from both educationalists and economists (e.g. Haertel 2013; Raudenbush 2004; Raudenbush and Jean 2012; Sass, Semykina, and Harris 2014), while some economists are more positive (Chetty et al. 2014; Deming 2014; Deutch 2012).

Table 2: Impact-level purposes of national assessments

What impact should the assessment system have?	Examples	How well do current assessments do this?
<p>1. Motivate pupils to enjoy the course or work harder, and to develop a lifetime love of learning the subject.</p>	<p>a) Inclusion of assessment of coursework, practical work, or fieldwork in the qualification because it motivates pupils. b) Dividing the qualification’s teaching and assessment into a modular structure because it motivates pupils. c) Selection of curriculum content to be interesting or accessible to pupils.</p>	<p>A lack of systematic and robust evidence about what actually motivates pupils makes this difficult to judge, but anecdotal perceptions abound. We should distinguish between pupils’ enthusiasm for structures that lead to higher grades without more effort, and structures that actually motivate them to work harder or engage more authentically. There may be tensions between what is interesting or accessible, and what is important or valuable educationally.</p>
<p>2. Influence the time allocated, content focus, or curriculum approach of what is studied.</p>	<p>a) Teachers focus instruction on what is most likely to gain credit in the assessments. b) Schools and teachers are motivated to focus effort on getting all pupils to achieve proficiency in basic skills. c) Inclusion of the requirement for a language in the English Baccalaureate increases take-up of languages at GCSE.</p>	<p>Attaching high-stakes consequences to assessment outcomes tend to focus teachers’ attention on them very effectively. However, there is a danger that instruction can become narrowly focused on how to gain marks on a particular style of question and mark scheme. Also, being assessed confers value that in practice may override any wider educational values, such as when teachers defend asking pupils to only read ‘set books’. Large amounts of time may be devoted to practising past papers (e.g. in Year 6). Again there is a perception that this is an educationally barren experience, although testing can be one of the most effective ways to learn (Roediger and Karpicke 2006). If assessments are predictable in content and style, or give credit for regurgitation and compliance rather than requiring original, individual, high-order thinking, focusing instruction on them is likely to be educationally dysfunctional. Many of our existing national assessments are probably too much in the former category.</p>

3. Drive improvement in the system.	a) Making assessments harder in order to require greater effort and higher expectations from teachers and pupils.	Although the logic of this argument has superficial appeal, and seems attractive as a policy lever, the evidence does not really support the idea that we can achieve large-scale improvements by raising demand. ⁴
-------------------------------------	---	--

The judgements of current assessments in the third column of Table 1 and Table 2 present a rather mixed picture. In relation to some uses, our assessments are fit for purpose, while for others they leave a lot to be desired. Part of the problem is that many of the desired uses were not considered in the process of designing the assessments. The format and conventions of national assessments draw on a long tradition, and earlier templates continue to shape them even when they are revised. In addition, there is no expectation that exam boards consider or explicitly address a requirement to ensure that their assessments meet these criteria; the boards neither have to state what purposes their assessments are intended to support, nor do they have to produce any evidence showing how well they meet any such intentions.

It is therefore hardly surprising that existing national assessments meet only some of the stipulated requirements. The ones they do meet tend to be those that have been traditionally salient, or easiest to achieve, which may not be the purposes that would be seen as most important by groups such as employers, higher education institutions, parents, teachers, pupils, or members of the general public. To address this mismatch, our first policy recommendation is:

4 The problem here is not so much research that opposes the expectation of benefit, but a lack of clear evidence either way. Good evidence does support the positive impact of setting challenging and specific goals (Locke and Latham 2006), and the correlation between teachers' expectations and pupil attainment (Teddlie and Reynolds 2000). However, we also know that teachers' expectations are very resistant to change (Jussim and Harber 2005; Raudenbush 1984), and that requiring higher performance on particular measures can lead to improvements in those measures that are not matched by improvements in independent yardsticks (e.g. Klein et al. 2000). In the absence of any direct evidence of the causal effects of a national policy change in demand requirements, it is clearly difficult to predict whether such a change will work as intended.

Assessment developers should be required by the regulator (Ofqual) to state explicitly what interpretations, uses, and decisions their assessment outcomes are intended to support, and which are not appropriate. Evidence should be provided to show how well the assessments support the intended purposes.⁵

The issue of accountability

It is clear from the issues raised above that some of the key pressures on the quality of assessments and qualifications arise when they are used as part of an accountability system. The incentives in accountability structures are potentially powerful drivers of behaviour, for better or worse. It is therefore important to understand the consequences of accountability.

Potential advantages and problems of accountability systems

Arguments in favour of school accountability often draw on the claim that historically, due to the regulatory framework of education systems, schools have lacked strong extrinsic incentives to improve pupil achievement.⁶ In such a context, it is unlikely that resources are used efficiently, and questionable whether they matter much at all (Hanushek 2006). School accountability is one way of changing the extrinsic incentive structure within schools, in attempts to target quality deficiencies directly. By introducing carrots and sticks, with rewards and punishments depending on performance, the idea is that schools should have strong incentives to up their game.

Within the academic literature, proponents of school accountability are often economists who perceive the extrinsic incentive structure to be inadequate. Yet other economists and psychologists disagree, instead emphasising the strong potential for unintended consequences of accountability systems. Similarly, educationalists have also often been critical, also pointing to unintended

5 It is worth noting that the idea of explicitly stating and justifying the intended purposes of an assessment is the clear recommendation of the authoritative Standards for Educational and Psychological Testing (AERA, APA, and NCME 1999), and is, unfortunately, more likely to be established practice in assessment development in the US than in the UK.

6 In contrast to intrinsic motivation, which stems from direct enjoyment of performing tasks, extrinsic incentives refer to various forms of external pressure to perform the tasks well.

dysfunctional effects of accountability systems in qualitative research. Indeed, potential problems with accountability are widely documented, both in education and in fields such as health.⁷

The main perceived issues are:

1. Crowding out of intrinsic motivation

The introduction of extrinsic incentives may undermine intrinsic motivation to perform. This means that there might be no, or even negative, net effects of such incentives on the outcomes they target.

2. Narrowing

Examples of narrowing include focusing on borderline pupils at the expense of others, drilling pupils to pass a particular test without equipping them to sustain or transfer that performance to other tests, and focusing on short-term objectives at the expense of long-term success.

3. Gaming/cheating

Narrowing crosses a line into gaming when teachers help pupils too much with coursework, enter them for qualifications that have value only in accountability systems, or exclude pupils who are likely to be low attaining. Gaming, in turn, crosses a line into cheating when teachers or administrators engage in outright illegal manipulation of outcomes, such as changing pupils' answers after exams, or obtaining the official exam questions in advance and prepping pupils for these.

4. Unfairness

When doing the right thing is made more difficult or disadvantageous than something incentivised, this is fundamentally unfair. It may lead to feelings of helplessness (and hence reduced effort), or a tendency to do what leads to easy rewards rather than what is right. An example would be teachers or headteachers who are reluctant to take a job in a challenging school because they perceive that the accountability system unfairly penalises such schools.

⁷ See, for example, Amrein-Beardsley et al. (2010); Baker and Linn (2002); Berliner (2011); Bevan and Hood (2006); Bird et al. (2005); Croft and Howes (2012); de Wolf and Janssens (2007); Fitz-Gibbon (1997); Frey and Jegen (2001); Jacob and Levitt (2003); Mansell (2007); O'Neill (2013); Smith (1995); and Wiggins and Tymms (2002).

5. Pressure

Accountability might cause undue pressure on individuals that undermines their ability to perform. This would be the case if, for example, good teachers take time off work because of stress caused by Ofsted inspections.

6. Legitimation

The importance of hitting targets and performance indicators might be seen as justification for dysfunctional or immoral behaviour, leading to an abdication of professional morality. For example, teachers might justify cheating on coursework on the grounds that it will benefit pupils if their school is judged outstanding. In this sense, bad behaviour drives out good: the perception that others are cheating makes it seem both more necessary and more acceptable.

7. Competition

Accountability systems may encourage schools or teachers to compete against each other, and discourage collaboration and mutual support. Some argue, therefore, that the overall impact on the system may be sub-optimal. On the other hand, others would argue that overly strict accountability systems constrain innovation and hamper genuine, potentially beneficial competition.

It is far from clear, therefore, whether accountability is a positive or negative development compared with the status quo. It certainly changes the extrinsic incentive structure in schools, but it is highly disputed whether or not this is a step in the right direction.

Evidence on the impact of accountability

Whether or not school accountability systems generate improvements in educational outcomes has been subject to increasing empirical research in the past decade. A meta-analysis by Lee (2008) finds a modest average positive impact of 0.08 standard deviations. If we were to translate this into international test scores in TIMSS and PISA, this is equivalent to 8 points, which is hardly transformative. However, the effect varies considerably across studies, and most studies reviewed suffer from significant limitations, particularly in their ability to attribute observed changes unequivocally to the introduction of accountability. Furthermore, all studies included were conducted in the US, and none looked

at any unintended side effects. Despite these limitations, Lee's review may be interpreted as giving slight support to the claim that high-stakes accountability raises performance, although a number of other interpretations are possible.

Broadly supporting Lee's (2008) conclusions, Figlio and Loeb (2011) reviews the American economic literature and finds that it indicates some positive effects on achievement, especially in mathematics, but that there are also studies that fail to detect any effects. In addition, there is also evidence of strategic behaviour among actors to artificially boost test scores. However, it is unclear how important and prevalent such strategic behaviour actually is – William (2010) finds the existing evidence for dysfunctional side effects 'inconclusive'. However, because of these uncertainties, as Lee (2008, p. 639) concludes, '[E]ducational policy makers and practitioners should be cautioned against relying exclusively on research that is consistent with their ideological positions to support or criticize the current high-stakes testing policy movement'.

Since long-term outcomes, such as earnings, are more difficult to manipulate, it is also worth mentioning Deming et al.'s (2013) recent research from Texas. The authors find that the long-term effects of accountability are mixed – upper-secondary schools on the verge of being judged 'low-performing' respond by raising their pupils' achievement, which later increases the likelihood that they attend university, and also raises their earnings by 1 per cent at the age of 25. This effect is equivalent to having a one standard deviation more effective teacher. But among schools that are not on the verge of being judged 'low-performing', accountability ratings do not have any effects overall – and actually lead to lower likelihood of university attendance and lower earnings among low-performing pupils. Clearly, therefore, we need more research on how different pupil types are affected by accountability.

What about England? One influential study is Burgess et al.'s (2013) analysis of the relative decline in GCSE attainment in Wales vis-à-vis England following Wales's decision to stop publishing league tables in 2001. The authors find positive effects of publication on GCSE results, equivalent to a modest but most-likely cost-effective effect size of 0.09 standard deviations, with no impact on school segregation. The effect was concentrated among schools in the lower 75 per cent in the ability and poverty distribution; schools in the top quartile of performance did not react at all, indicating that

the decision to stop publishing league tables also exacerbated inequality of achievement. The authors consider a range of possible alternative explanations for the observed difference, analyse them explicitly, but dismiss them all as unconvincing, although it is difficult to rule out such explanations entirely.⁸ Nevertheless, this study provides the best direct evidence we currently have of the impact of league tables in England.

Two other studies focus on the English inspection system, finding positive effects of failing an inspection (relative to schools that just passed) on subsequent GCSE outcomes with an effect size in the range of a modest 0.1 standard deviation (Allen and Burgess 2012; Hussain 2012). Both studies find that the positive impact occurs in core subjects, indicating that it is not the result of schools simply enrolling children in easier subjects. In addition, Hussain (2012) finds no evidence of narrowing, specifically that teachers exclude low-ability pupils from the tests or that they target borderline pupils only, and the positive effects also appear to persist in the medium term when the pupils are no longer in the failing school. At the same time, Allen and Burgess (2012) do find evidence of narrowing, indicating that the results in this respect are mixed. And, of course, there are other forms of manipulation the authors do not investigate. Furthermore, they only analyse the impact of accountability among pupils attending borderline failing schools, and, as Deming et al.'s (2013) research indicates, it is not possible to extrapolate the positive effects to other pupils.

The PISA results are another oft-cited piece of evidence about the benefits of accountability. Analysis of international country-level PISA data has been widely cited as showing a correlation between accountability and autonomy with high performance. For example, the DfE's (2013) announcement of its secondary school accountability reforms stated that 'OECD evidence shows that a robust accountability framework is essential to improving pupils' achievement'. In fact, the PISA report actually says almost the exact opposite, stating that 'there is no measurable relationship between...various uses of

8 For example, the substantial increases in school funding in England compared to Wales (BBC 2011) are directly controlled for, and the authors do not find evidence that abolishing league tables affects KS2 outcomes, which can be considered a 'placebo test' – Wales has never published KS2 results and these should therefore not be affected by the policy change.

assessment data for accountability purposes and the performance of school systems' (OECD 2010, p. 46). The confusion seems to have arisen because commentators and politicians have failed to grasp that the impact found in the PISA report is an interaction effect, which is very different. The OECD (2010, p. 105) finds positive effects of autonomy in countries that publish achievement data publicly, while there are negative effects of autonomy in countries that do not publish data. Accountability by itself, on the other hand, has no detectable relationship with achievement at the system level. Even the interaction effect evaporates, however, if state and independently-operated school pupils are analysed separately (Benton 2014).⁹

Overall, while there is some evidence to support positive effects of accountability on attainment, they are generally modest and seem to differ depending on school and/or pupil type. The evidence about possible unintended consequences is currently probably too limited to draw any clear conclusions. In general, therefore, the jury is still out on the overall effects of school accountability. Most likely, the relationship between different features of the system and other contextual factors will moderate any effects on performance and other outcomes. In short, accountability may be either good or bad – outcomes probably depend on system design. For example, a system that holds schools accountable for pupil progress may create different incentives from a system holding schools accountable for absolute achievement measures. For this reason, it is difficult to make strong arguments in favour or opposition of accountability without specifying what type of accountability one is talking about. And as argued below, we currently do not know enough about how these features might interact to make any safe predictions in any specific case.

Features of accountability systems

Since system design is likely to be key for the impact of accountability, it is important to discuss how different features impact on outcomes. All accountability

9 The interaction model including both state and independently-operated school pupils assumes that the control variables included, such as pupil background, have the same effect across the two sectors, which is far from clear. Further displaying problems with the OECD evidence, a sophisticated analysis of PISA data by Hanushek, Link, and Woessmann (2013) presents more nuanced conclusions on the impact of autonomy, finding its effects to depend on the level of countries' economic development.

systems create incentives around measures of achievement, which determines how actors within the education system react. When characterising these, it is useful to first separate the achievement measures from the accountability structure. Both impinge on the incentive structure. The achievement measures partly determine the type of incentives within the system – that is, to what goals schools are held accountable – whereas the accountability structure determines the strength of these incentives.

Measures may be used as targets or performance indicators, and typically consist of straightforward assessment outcomes, although some – such as value added or progress scores – first have to be constructed from those outcomes. Other measures may be composites, calculated by aggregating individual assessments in some way. For example, the ‘5 A*-C’ measure is based on five assessments in separate subjects. Another, more subjective measure used for accountability in England is the judgement of Ofsted inspectors. The box catalogues some of the questions that arise in relation to the suitability of measures for accountability purposes.

Key quality criteria for accountability measures

1. Do the measures represent valued outcomes?
2. Are there important outcomes not captured by the measures?
3. Is what is measured sensitive to changes in the desired behaviours (e.g. improvements in instruction or greater effort)?
4. Could performance on the measures reflect irrelevant or misleading confounds?
5. What are the limits of precision, misclassification, or consistency (reliability) of the measures?
6. Are the measures fair to all subgroups, including individuals with disabilities, different language, cultural, or social backgrounds, or to schools that serve different kinds of communities?
7. Could it be possible to improve performance on the measures without any real improvements in valued outcomes?

These quality criteria all concern aspects of validity and the fitness of the measures for accountability purposes, and they should be addressed by the evidence provided by assessment developers in following our first policy recommendation. There are plenty of accountability measures that fail to satisfactorily address these issues, and there might be limits to what is possible to achieve (Bevan and Hood 2006; Linn 2000; O'Neill 2013). As O'Neill (2013, p. 14) puts it,

Every time one performance indicator is shown to be inaccurate, or misleading, or likely to produce perverse results, some people claim that they can devise a better one that has no perverse effects. Experience suggests that they may well be as wrong as those who invented the last lot of indicators.

Nevertheless, if assessments are designed explicitly to be suitable for accountability purposes, it should be possible to improve current assessments to the extent they meet the criteria stipulated above. Exactly by how much this would improve the arrangements is less clear, as is the question of whether it can be achieved with the same assessments that meet the requirements for the purposes listed in Table 1. However, if assessment developers follow our first recommendation, this limitation should at least be explicit and evidence based. This leads to our second recommendation:

If assessments are used as part of accountability systems, they should be designed to meet quality criteria, such as those listed above. Part of the development process should include the collection of evidence about the extent to which an assessment does in fact meet these criteria.

What about the accountability structure? A simple way to categorise the different types of structure would be to envisage a continuum from 'hard' to 'soft', a distinction that in turn has a number of dimensions (de Wolf and Janssens, 2007). Table 3 catalogues these dimensions.

Table 3: Characteristics of ‘hard’ and ‘soft’ accountability structures

	Hard accountability	Soft accountability
Relationship between measures and different types of incentives	Explicit Incentives are explicit/extrinsic and directly linked to measures. For example, this is the case when measures are used to determine performance-related pay and appraisal, or used for promotion, appointments or competence procedures.	Implicit Incentives are implicit/intrinsic and no direct consequences are attached to measures. The assumption is made that teachers and school leaders are already motivated to do their best, so additional incentives will not increase their performance, and/or that no measures can capture quality well enough to be directly incentivised.
Public openness of measures	Published Performance indicators based on measures are made public to increase their motivating force (e.g. ‘naming and shaming’), and to influence indirect consequences (e.g. induce fewer parents to choose lower-performing schools).	Confidential Performance indicators are kept confidential in order to prevent them from being distorted by strategic behaviour among school actors.
Location of evaluation	Objective data Performance indicators can be interpreted as measuring quality directly, without the need for interpretation. This avoids the risk of unpalatable messages being softened.	Professional judgement Performance indicators only indicate and must consequently be interpreted. They support judgement but do not replace it; they help us ask better questions rather than directly answering them.
Improvement mechanism	Consequences Direct, contingent rewards and sanctions that shape behaviour.	Feedback Feedback on performance indicators is used to inform improvement efforts, providing guidance, diagnosis, and prescriptions.
Prioritised actors	Consumers Parents and taxpayers are entitled to full information on the performance of services they use or pay for.	Professionals Supporting and trusting teachers to do their job will bring out the best in them.

Clearly, there can be a range of intermediate positions, and one could envisage a ‘pick and mix’ approach. It certainly seems likely that the impact of accountability on performance depends on the particular combination of these features and on the context in which they operate. At this stage, however, we do not know enough about how they interact to be able to make good predictions.

Given such ignorance, a policy of dictating a single accountability structure for all schools in England can hardly be described as evidence based. A more scientific approach would be to allow a range of variation in the factors identified in Table 3, within what is politically acceptable, and then randomly allocate different groups of schools to experience accountability systems that differ on these factors. We would then very quickly start building up robust knowledge of the conditions that would maximise the chances of accountability actually contributing to system-wide improvement. This leads to our third policy recommendation:

A programme of research should be undertaken with the aim of investigating what features of accountability structures lead to the best overall outcomes.

This experimental approach merely acknowledges that we cannot assume to be able to foresee the unintended consequences of accountability reforms. By trialling different structures, we effectively enter them in a competition with each other to find out which one works best.

Reconciling educational goals with demands of accountability

Similarly, it is also clear that a wide variety of approaches must be trialled to find out how we can reconcile the educational purposes of assessment to display pupil attainment of valued skills – and ensuring breadth in the curriculum studied – with the requirements of high-stakes accountability. This leads to our fourth policy recommendation:

Pilot projects featuring a range of strategies to square educationally desirable practices with high-stakes accountability should be introduced in order to determine what works.

Here, we consider one such approach that should be put forward for trialling: teacher assessment in the context of high-stakes accountability. This issue embodies the potential conflict between educational desirability and accountability perfectly: teacher assessment may very well be desirable from an educational view, but undesirable from an accountability standpoint.

Teacher assessment – based on coursework, practical work, and fieldwork – has been part of many GCSE courses since their inception. In recent years, however, fears about malpractice in setting, administering, marking, and moderating the teacher-assessed components have led to greater restrictions on how they are conducted, and ultimately to their being abolished in most subjects (Ofqual 2014). This decision has been controversial and opposed by some on the grounds that important aspects of learning in some subjects, such as speaking and listening in English or practical work in science, cannot be assessed appropriately in external exams (Adams 2014; Walker 2013). And if the teacher-assessed components are not included in the high-stakes assessments, it is likely that these are seen as less important – and consequently given less time and resources.

While it could therefore be desirable to allow teacher-assessed components from an educational perspective, it is likely that it encourages perverse incentives to engage in undesirable practices, for example grade inflation.¹⁰ This begs the question: is there any way teacher assessment can be made safe for accountability? We believe so, and the following suggestions are offered for consideration to be trialled in the pilots noted above:

1. Remove perverse incentives among teachers

Fix the distribution of total marks/grades at the centre level, according to ‘non-cheatable’ elements (e.g. external exams).¹¹ Effectively, this means that there would be a fixed-sum of teacher-assessed grades, so that teacher assessment only redistributes marks/grades among pupils within the centre. That means the teacher-assessed components are high-stakes for candidates, but low-stakes for teachers: increasing the grade for one pupil can only be done at the expense of another. This means that teachers have no incentive (or indeed ability) to inflate grades.

2. Police bad behaviour

Conduct spot checks to ensure that pupils can replicate their performances in teacher-assessed components of their qualifications. It would be necessary to set

¹⁰ Some would argue that such incentives are themselves only present in systems with strong accountability, but it is clear that also in education systems with little accountability do teachers engage in test score manipulation (Angrist, Bettistin, and Vuri 2014).

¹¹ Exam-board centres are typically schools or colleges, but may be other institutions or groups of schools.

aside time for an external examiner to visit centres and supervise replication of coursework and practical tasks, such as music/drama practical work, English/history coursework tasks, and speaking and listening in languages. A proportion of spot checks could be ‘risk targeted’, based on anomalous data (e.g. teacher-assessed grades that are significantly above exam grades in previous years; surprisingly high average scores or low score variability; and implausible patterns of missing data).¹²

Whistle-blowing mechanisms should be created for teachers to enable them to report malpractice in both their own and other schools. Pupils, parents, and governors could also be given a way of reporting concerns.

Teachers, headteachers, and pupils should be asked to sign declarations that certain practices have not occurred. This helps to make clear where the line goes between acceptable and unacceptable behaviour and support. There must also be clearly outlined consequences for individuals who have signed such declarations, should malpractice later be revealed. For example, pupils who are caught lying would have their grades stripped for being complicit in cheating. On the other hand, those who had honestly reported any concerns would be awarded a grade based on any uncompromised elements of the qualification.

Introduce questionnaires for teachers and pupils, which probe a range of acceptable, grey-area, and unacceptable practices and perceptions. Statistical tests might later be able to signal ‘too good to be true’, ‘overly consistent’ or otherwise faked responses, and consequently trigger spot checks.

3. Build capacity through training and support

Teachers must be trained to enable them to assess pupils accurately. A range of evidence shows that valid teacher assessment is possible, but unlikely without substantial training for teachers (e.g. Stanley et al. 2009).

Introduce better moderation practices. More systematic use of cross-centre blind marking would increase confidence in the consistency and comparability of teacher-assessed marks from different teachers and centres.

12 See Jacob and Levitt (2003) and Angrist, Bettistin, and Vuri (2014) for examples of this kind of approaches in the American and Italian contexts respectively.

These suggestions are likely to go a long way in making teacher-assessed components consistent with the demands of high-stakes accountability. Again, however, it is crucial that our suggestions are not construed as policy recommendations for universal reforms at this point – they must first be trialled in a randomised pilot programme. Only if this is successful should we begin the discussion of scaling up the suggestions to national policy.

Conclusion

This chapter has discussed how we can improve the incentives in the English qualifications, assessment, and accountability system. In doing so, the framework for curricula would be greatly improved as well, since it is driven by what is demanded in high-stakes examinations as well as by the format of qualifications and assessment.

In order to evaluate whether a qualification or an assessment is fit for purpose, it is important to stipulate specific criteria for whether this is indeed the case. Without such criteria, it is difficult to assess empirically whether the qualification or assessment fulfils its intended function. For this reason, exam boards should be required to make explicit what purposes their assessments and qualifications are supposed to fulfil, and amass evidence to what extent they are successful in this respect.

However, it is by now clear that high-stakes accountability also puts additional demands on the qualifications and assessment system. In order to increase the likelihood that measures used for accountability purposes meet stipulated quality criteria, exam boards should explicitly design their assessments after such criteria while again amassing evidence to the extent they succeed in this endeavour.

Naturally, the structure of the accountability system is immensely important for the outcomes it produces. Since we know little about how to produce the optimal accountability structure, an experimental approach is favoured in which different schools are subject to different accountability features. Doing so would greatly increase our understanding of the conditions under which accountability may be a lever for school improvement – and the conditions under which it does not work as intended.

Similarly, in order to reconcile educationally desirable policies with demands of accountability, it is important to trial different approaches to find out what works and what does not. An example of such a policy is teacher-based assessment. To square this with high-stakes accountability, we have offered a couple of suggestions that should be tested.

Advocacy of evidence-based policy has become popular in the last couple of years. Yet, while politicians from left to right surely pay lip service to the idea, they rarely seem prepared to enforce it in practice. One important exception is the inception of the Education Endowment Foundation, which funds randomised trials on different policies. However, this organisation mainly focuses on trialling certain types of classroom-level practices. But if politicians are serious about evidence-based policy, there is no reason why this approach should not be used for trialling innovations in qualifications, assessment, and accountability at the system level too.

References

- Adams, R. (2014) 'Science Community Dismayed at Decision to Axe Lab Work from A-levels', *The Guardian*, 9th April 2014, <http://www.theguardian.com/education/2014/apr/09/science-community-dismay-axe-lab-work-a-level> (accessed 15th June 2014).
- Allen, R. and S. Burgess (2012), 'How Should We Treat Under-performing Schools? A Regression Discontinuity Analysis of School Inspections in England'. Working Paper No. 12/287, Centre for Market and Public Organisation, University of Bristol.
- American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999), *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Amrein-Beardsley, A., D.C. Berliner, and S. Rideau (2010), 'Breaking Professional Law: Degrees of Cheating on High Stakes Tests', *Education Policy Analysis Archives*, 18(14). Retrieved 24th June 2010 from: <http://epaa.asu.edu/ojs/article/view/714>.
- Angrist, J. D., E. Battistin, and D. Vuri (2014), 'In a Small Moment: Class Size and Moral Hazard in the Mezzogiorno'. NBER Working Paper No. 20173, National Bureau of Economic Research, Cambridge, MA.
- Baker E. L. and R. L. Linn (2002), 'Validity Issues for Accountability Systems'. CSE Technical Report 585, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), UCLA.

- BBC Wales News (2011), 'Wales-England School Funding Gap is £604 Per Pupil'. 26th January 2011, <http://www.bbc.co.uk/news/uk-wales-12280492> (accessed 11th June 2014).
- Benton, T. (2014), 'A Re-evaluation of the Link Between Autonomy, Accountability, and Achievement in PISA 2009', Discussion Paper, Research Division, Cambridge Assessment.
- 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.
- Bevan, G. and C. Hood (2006), 'What's Measured is What Matters: Targets and Gaming in the English Public Health Care System', *Public Administration* 84(3):517–38.
- Bird S. M., D. Cox, T. F. Vern, H. Goldstein, T. Holt, and P. C. Smith (2005), 'Performance Indicators: Good, Bad, and Ugly', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 168(1):1–27.
- Burgess, S., D. Wilson, and J. Worth (2013), 'A Natural Experiment in School Accountability: The Impact of School Performance Information on Pupil Progress' *Journal of Public Economics* 106:57–67.
- Chetty, R., J. N. Friedman, and J. E. Rockoff (2014), 'Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates'. NBER Working Paper No. 19423, National Bureau of Economic Research, Cambridge MA.
- Croft, J. and A. Howes (2012), 'When Qualifications Fail: Reforming 14–19 Assessment'. Discussion Paper No. 1, Centre for Market Reform of Education, London.
- Deming, D. J. (2014), 'Using School Choice Lotteries to Test Measures of School Effectiveness'. NBER Working Paper No. 19803, National Bureau of Economic Research, Cambridge MA.
- Deming, D. J., S. Cohodes, J. Jennings, and C. Jencks (2013), 'School Accountability, Postsecondary Attainment and Earnings'. NBER Working Paper No. 19444, National Bureau of Economic Research, Cambridge, MA.
- Department for Education (DfE) (2013), 'Reforming the Accountability System for Secondary Schools: Government Response to the February to May 2013 Consultation on Secondary School Accountability'. Report, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/249893/Consultation_response_Secondary_School_Accountability_Consultation_14-Oct-13_v3.pdf (accessed 15th June 2014).
- Deutch, J. (2012), 'Using School Lotteries to Evaluate the Value-Added Model'. Job Market Paper, University of Chicago, http://harrisschool.uchicago.edu/sites/default/files/Job%20market%20paper-%20Jonah%20Deutsch_0.pdf (accessed 15th June 2014).

- De Wolf, I. F. and J.G Janssens (2007), 'Effects and Side Effects of Inspections and Accountability in Education: An Overview of Empirical Studies', *Oxford Review of Education* 33(3):379–396.
- Figlio, D. and S. Loeb (2011), 'School Accountability' Pp. 383–421 in *Handbook of the Economics of Education, Volume 3*. Elsevier.
- Fitz-Gibbon, C.T. (1997) *The Value Added National Project: Feasibility Studies for a National System of Value Added Indicators (Final Report)*. London: SCAA.
- Frey, B. S. and R. Jegen (2001), 'Motivation Crowding Theory', *Journal of Economic Surveys* 15(5):589–611.
- Haertel, E. H. (2013), 'Reliability and Validity of Inferences about Teachers Based on Student Test Scores'. Report based on the 14th William H. Angoff Memorial Lecture at the National Press Club, Educational Testing Service, Princeton, NJ.
- Hanushek, E. A. (2006), 'School Resources', pp. 866–906 in *Handbook of the Economics of Education, Volume 2*. Elsevier.
- Hanushek, E., S. Link, and L. Woessmann (2013), 'Does School Autonomy Make Sense Everywhere? Panel Estimates from PISA', *Journal of Development Economics* 104:212–32.
- Higgins, S., Katsipatakis, M., Kokotsaki, D., Coleman, R., Major, L.E., and Coe, R. (2013), 'The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit'. London: Education Endowment Foundation. Available at <http://www.educationendowmentfoundation.org.uk/toolkit> (accessed 14th June 2013).
- Hussain, I. (2012), 'Subjective Performance Evaluation in the Public Sector: Evidence from School Inspections'. CEE Discussion Paper No. 135, Centre for the Economics of Education, London School of Economics.
- Jacob, B. A. and S. D. Levitt (2003), 'Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating', *Quarterly Journal of Economics* 118(3): 843–77.
- Jussim, L. and K. D. Harber (2005), 'Teacher Expectations and Self-fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies', *Personality and Social Psychology Review* 9(2):131–55.
- Klein, S.P., Hamilton, L.S., McCaffrey, D.F. and Stecher, B.M. (2000), 'What do Test Scores in Texas Tell us?' *Education Policy Analysis Archives*, 8(49), <http://epaa.asu.edu/epaa/v8n49/>
- Leckie, G. and H. Goldstein (2009). 'The Limitations of Using School League Tables to Inform School Choice', *Journal of the Royal Statistical Society, A* 172:835–51.
- Lee J. (2008), 'Is Test-Driven External Accountability Effective? Synthesizing the Evidence From Cross-State Causal-Comparative and Correlational Studies', *Review of Educational Research* 78(3):608–44.

- Linn, R. L. (2000), 'Assessments and Accountability', *Educational Researcher* 29(2):4–16.
- Locke, E. A. and G. P. Latham (2006), 'New Directions in Goal-setting Theory', *Current directions in psychological science* 15(5):265–8.
- Mansell, W. (2007), *Education by Numbers: The Tyranny of Testing*. London, Politico's Publishing.
- Newton, P. E. (2007), 'Clarifying the Purposes of Educational Assessment', *Assessment in Education: Principles, Policy & Practice* 14(2):149–70.
- Newton P.E. and Shaw S.D (2014), *Validity in Educational and Psychological Assessment*. London: Sage.
- OECD (2010), 'PISA 2009 Results: What Makes a School Successful? Resources, policies and practices, Volume IV', <http://www.oecd.org/pisa/pisaproducts/48852721.pdf> (accessed 14th June 2014).
- Ofqual (2014), 'An Update on the Reforms Being Made to GCSEs'. Report (Ofqual/14/5404), London.
- O'Neill, O. (2013), 'Intelligent Accountability in Education', *Oxford Review of Education* 39(1):4–16.
- Pellegrino, J. W., N. Chudowsky, and R. Glaser (eds.) (2001), *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.
- Raudenbush, S.W. (1984), 'Magnitude of Teacher Expectancy Effects of Pupil IQ as a Function of Credibility of Expectation Induction: A Synthesis of Findings from 18 Experiments', *Journal of Educational Psychology* 76(1): 85–97.
- Raudenbush, S.W. (2004), 'What Are Value-added Models Estimating and What Does This Imply for Statistical Practice?', *Journal of Educational and Behavioral Statistics* 29(1):121–9.
- Raudenbush, S.W. and M. Jean (2012), 'How Should Educators Interpret Value-Added Scores?', Carnegie Knowledge Network, <http://carnegieknowledge.org/briefs/value-added/interpreting-value-added/> (accessed 14th June 2014).
- Roediger, H. L., and J. D. Karpicke (2006), 'The Power of Testing Memory: Basic Research and Implications for Educational Practice', *Perspectives on Psychological Science*, 1(3):181–210.
- Sass, T. R., A. Semykina, and D. N. Harris (2014), 'Value-added Models and the Measurement of Teacher Productivity', *Economics of Education Review* 38:9–23.
- Smith, P. (1995). 'On the Unintended Consequences of Publishing Performance Data in the Public Sector', *International Journal of Public Administration* 18(2/3): 277–310.

- Stanley, G., R. MacCann, J. Gardner, L. Reynolds, and I. Wild (2009), 'Review of Teacher Assessment: Evidence of What Works Best and Issues for Development'. Report for the Qualifications and Curriculum Authority.
- Teddlie, C. and D. Reynolds (2000), *The International Handbook of School Effectiveness Research*. London: Falmer Press.
- Waldegrave, H. and J. Simons (2014), 'Watching the Watchmen: The Future of School Inspections in England'. Report, Policy Exchange <http://www.policyexchange.org.uk/images/publications/watching%20the%20watchmen.pdf> (accessed 14th June 2014).
- Walker, P. (2013), 'GCSE English to Drop Speaking and Listening Components', *The Guardian*, 29th August 2013, <http://www.theguardian.com/education/2013/aug/29/gcse-english-speaking-listening-drop> (accessed 13th June 2014).
- Wiggins, A., and P. Tymms (2002), 'Dysfunctional Effects of League Tables: A Comparison between English and Scottish Primary Schools', *Public Money and Management* 22(1):43–8.
- Wiliam, D. (2010) 'Standardized Testing and School Accountability', *Educational Psychologist* 45(2):107–22.

