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

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.

3. Allocate individuals to appropriate teaching groups.	a) Setting in Year 7 based on KS2 performance.	Notwithstanding the lack of evidence about the benefits of setting (Higgins et al. 2013), current assessments probably broadly meet this need.
4. Decide whether an individual is equipped to go on to a further course of study or employment.	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.	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.
5. Select which individuals should be offered places, from a larger pool of qualified applicants.	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.	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.

6. Indicate the effectiveness of teachers or schools.	a) Use of pupils' examination performance to inform teacher-performance management. b) School-level floor targets for exam performance trigger inspection visits. c) Examination grades analysed and interpreted by inspectors as evidence of school quality. d) Examination performance used to inform parents' and children's school choices.	Attributing differences in student achievement to the effects of teaching, even with good adjustment for prior characteristics, is the subject of some controversy.[3] The ability of inspectors to interpret this kind of information appropriately may be questionable (Waldegrave and Simons 2014). Using value added for school choice decisions is also problematic (Leckie and Goldstein 2009). Problems of comparability arise if grades from qualifications that differ in difficulty are treated interchangeably. 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.
7. Evaluate the performance of the system or subgroups.	a) Changes in pass rates over time interpreted as evidence of system change. 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. c) Performance of subgroups which have experienced an intervention used for evaluation.	Problems of comparability arise if grades from qualifications taken at different times or in different subjects/qualifications are treated interchangeably. 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.

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?
1. Motivate pupils to enjoy the course or work harder, and to develop a lifetime love of learning the subject.	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.	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.
2. Influence the time allocated, content focus, or curriculum approach of what is studied.	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.	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.

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 – Wiliam (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 polices. 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-alevel (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 Sonsultation 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., Katsipataki, 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 Prevalnce 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://carnegieknowledgenetwork.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 Tesearch 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.