Qualifications and Curriculum Authority

Evaluating assessment systems

Paul E. Newton
Head of Assessment Research
Regulation and Standards Division, QCA

Paper 1 – June 2007

Index

Contents **Page**

Executive summary		1
1	Introduction	2
2	What does evaluation entail?	2
3	What does ‘fit-for-purpose’ mean?	3
4	How many purposes are there?	4
5	Why does the large number of purposes matter?	6
6	Can we construct a framework for system-level evaluation?	7
7	How should we conduct system-level evaluation?	14
8	Is system-level evaluation feasible?	15
9	Using the framework to identify common limitations	16
10	Endnotes	19

Executive summary

The following report highlights five common confusions related to the evaluation of educational assessment systems:[i]

1. the difference between validation and evaluation, where validation (concerning the accuracy of inferences from results) is merely one component of evaluation
2. the meaning of 'purpose' in 'fitness-for-purpose', which can be interpreted in a variety of different ways, all of which are (differently) relevant to evaluation
3. the number of purposes which can be identified, which is much higher than tends to be appreciated (for example, national curriculum test results are probably used for at least 14 different purposes)
4. why it matters when results are used for many different purposes, which is because different uses require that different kinds of inference be drawn from results, so results that warrant accurate inferences for one purpose may not warrant accurate inferences for another
5. the many components of a rigorous evaluation, which include analysis from the perspectives of technical accuracy, moral defensibility, social defensibility, legal acceptability, economic manageability and political viability.

The report emphasises the importance of: distinguishing the different meanings of similar terms; distinguishing logically separable evaluation questions; and distinguishing the many alternative perspectives on evaluation.

Evaluating assessment systems

1 Introduction

1.1 This report explores the concept of evaluation, as it applies to educational assessment systems; and then presents a framework for evaluating them.[ii] It is intended as a tool for helping the Education and Skills Select Committee to grapple with the many questions which comprise its *New Inquiry into Testing and Assessment*, particularly the very broad ones, like:

- is the testing and assessment in 'summative' tests (for example, GCSE, AS, A2) fit for purpose?
- should the system of national tests be changed? If so, should the tests be modified or abolished?

1.2 This report does not offer a view on the legitimacy of our present national assessment systems. Instead, it offers a way of organising evidence and argument in order to reach such a view. It helps to identify what makes for a good evaluation question and what makes for a good evaluation conclusion: issues which can seem deceptively straightforward at first glance. In particular, it aims to expose a number of common confusions which can mislead the unwary inquirer.

1.3 This report offers generic insights, which apply in the same way across the spectrum of educational assessment systems (occupational, vocational, general; tests, examinations, teacher assessments; on paper, on-screen, online; etc.).

2 What does evaluation entail?

2.1 The first confusion to confront is the nature of evaluation itself. An easy mistake to make is to reduce the big programme of **evaluation** to the smaller programme of **validation**.[iii] The central question at the heart of validation is this: are the inferences that we draw from our assessment results sufficiently accurate (for the uses to which they will be put)? Or, less formally: are our results accurate or not? Although this is a necessary and fundamental component of evaluation, it is still only one component. Evaluation requires the inquirer to consider any question that might bear upon the legitimacy of the assessment system, such as:

- might the way in which test results are reported have positive or negative impacts (e.g., is it better simply to rank students or to tell them how much of a programme of study they have 'mastered')?
- might the fact of testing itself have positive or negative impacts (e.g., does the inevitable 'washback' support or detract from good teaching and learning)?

2.2 Evaluation entails marshalling as much relevant evidence and argument as possible, to judge whether systems work as they are intended to and in the best interests of participants, stakeholders and society. The central question at the heart of evaluation is this: are our assessment systems fit-for-purpose?

3 What does 'fit-for-purpose' mean?

3.1 A second confusion to confront is the meaning of fitness-for-purpose. Before exploring the concept of 'fitness' we need to work out what we mean by 'purpose'. This is not as straightforward as it might sound. Consider the following three interpretations.

1. The purpose of assessment is to generate a particular kind of **result**. For example, students sit an exam in GCSE science to rank them in terms of their end-of-course level of attainment.
2. The purpose of assessment is to enable a particular kind of **decision**. For example, students sit an exam in GCSE science so that we can decide whether they have learned enough of the basic material to allow them to enrol on an A level science course.
3. The purpose of assessment is to bring about a particular kind of educational or social **impact**. For example, students sit an exam in GCSE science to force them to learn the subject properly, and to force their teachers to align their teaching of science with the national curriculum.

3.2 Obviously, to judge whether a system is fit-for-purpose, an evaluator needs to begin by identifying the purpose, or purposes, for which the system is supposed to be fit. However, if the evaluator is confused by the different possible meanings of 'purpose', no satisfactory conclusion will be reached. This is why it is essential to distinguish these different interpretations; all of which are perfectly reasonable; and all of which need to be considered in their own right when mounting an evaluation.

3.3 As it happens, there is yet another interpretation to be wary of:

4. The purpose of the qualification is to bring about a particular kind of educational or social **impact**. For example, students study GCSE science to support progression to a higher level of study (for those who wish to), and to equip all students with sufficient scientific literacy to function adequately as 21st century citizens.

Again, this is a perfectly reasonable interpretation of 'purpose'. However, strictly speaking, it is not within the scope of an evaluation into the legitimacy of an assessment system. Instead, it implies a broader evaluation remit, into the legitimacy of an educational programme. It would be perfectly possible to have a legitimate educational programme with an illegitimate assessment system; and *vice versa*. The two evaluation foci need to be kept quite separate.

4 How many purposes are there?

4.1 A third confusion concerns the number of purposes which need to be considered when evaluating an assessment system. This is best illustrated by considering the **uses** to which assessment results are put (interpretation 2 above). In England, we have become familiar with classification schemes such as that presented in 1988 by the Task Group on Assessment and Testing:

- formative uses (assessment for learning)
- summative uses (assessment of learning)
- evaluative uses (assessment for accountability)
- diagnostic uses (assessment for special intervention).

4.2 Although this kind of scheme is useful, it fails to convey the full complexity of the situation. In fact, there are many more categories of use to which educational assessment results might be put. Figure 1 illustrates twenty-two. The categories presented in Figure 1 are quite loose – and occasionally shade into each other – but the point isn't to present a definitive taxonomy, merely to illustrate just how many possible kinds of use there are. In fact, distinctions can often be made *within* categories, between uses which might recommend quite differently designed assessment systems (e.g., long-, medium- and short-term system monitoring).

Figure 1. Some examples of the many kinds of use to which assessment results can be put

1. **social evaluation** (to judge the social or personal value of students' achievements)
2. **formative** (to identify students' proximal learning needs, guiding subsequent teaching)
3. **student monitoring** (to decide whether students are making sufficient progress in attainment in relation to expectations or targets; and, potentially, to allocate rewards or sanctions)
4. **diagnosis** (to clarify the type and extent of students' learning difficulties in light of well-established criteria, for intervention)
5. **provision eligibility** (to determine whether students meet eligibility criteria for special educational provision)
6. **screening** (to identify students who differ significantly from their peers, for further assessment)
7. **segregation** (to segregate students into homogeneous groups, on the basis of aptitudes or attainments, to make the instructional process more straightforward)
8. **guidance** (to identify the most suitable courses, or vocations for students to pursue, given their aptitudes)
9. **transfer** (to identify the general educational needs of students who transfer to new schools)
10. **placement** (to locate students with respect to their position in a specified learning sequence, to identify the level of course which most closely reflects it)
11. **qualification** (to decide whether students are sufficiently qualified for a job, course or role in life – that is, whether they are equipped to succeed in it – and whether to enrol them or to appoint them to it)
12. **selection** (to predict which students – all of whom might, in principle, be sufficiently qualified – will be the most successful in a job, course or role in life, and to select between them)
13. **licensing** (to provide legal evidence – the licence – of minimum competence to practice a specialist activity, to warrant stakeholder trust in the practitioner)
14. **certification** (to provide evidence – the certificate – of higher competence to practise a specialist activity, or subset thereof, to warrant stakeholder trust in the practitioner)
15. **school choice** (to identify the most desirable school for a child to attend)
16. **institution monitoring** (to decide whether institutional performance – relating to individual teachers, classes or schools – is rising or falling in relation to expectations or targets; and, potentially, to allocate rewards or sanctions)
17. **resource allocation** (to identify institutional needs and, consequently, to allocate resources)
18. **organisational intervention** (to identify institutional failure and, consequently, to justify intervention)
19. **programme evaluation** (to evaluate the success of educational programmes or initiatives, nationally or locally)
20. **system monitoring** (to decide whether system performance – relating to individual regions or the nation – is rising or falling in relation to expectations or targets; and, potentially, to allocate rewards or sanctions)
21. **comparability** (to guide decisions on comparability of examination standards for later assessments on the basis of cohort performance in earlier ones)
22. **national accounting** (to 'quality adjust' education output indicators)

5 *Why does the large number of purposes matter?*

5.1 Confusion number four concerns why it matters that results can, and often are, used for multiple purposes. Surely, some would claim, as long as assessment results are accurate, then we ought to be able to use them for any purpose we like? Unfortunately, it's not quite as straightforward as that. The point is best illustrated by considering what it might mean to explore validity for different uses of results.

5.2 As mentioned earlier, the central question at the heart of validation is this: are the inferences that we draw from our assessment results sufficiently accurate (for the uses to which they will be put)? This has become the standard technical definition, and the word 'inference' is significant because different kinds of inference may be drawn – from the same assessment result – to support different kinds of use. This is not at all obvious, so it warrants a brief technical detour.

5.3 Assessment instruments are designed to support specific kinds of inference. So, an end-of-key-stage test will be designed primarily to support an inference concerning a student's 'level of attainment at the time of testing'. Let's call this the primary **design-inference**. And let's imagine, for the sake of illustration, that our assessment instrument – our key stage 2 science test – supports perfectly accurate design-inferences. That is, a student who really is a level X on the day of the test will definitely be awarded a level X as an outcome of testing.

5.4 In fact, when the test result is actually used, the user is likely to draw a slightly (or even radically) different kind of inference, tailored to the specific context of use. Let's call this a **use-inference**. Consider, by way of example, some possible use-inferences associated with the following result-based decisions/actions.

1. A placement/segregation use. The inference made by a key stage 3 head of science – when allocating a student to a particular set on the basis of a key stage 2 result – may concern 'level of attainment at the beginning of the autumn term'.
2. A student monitoring use. The inference made by a key stage 3 science teacher – when setting a personal achievement target for a student on the basis of a key stage 2 result – may concern 'level of attainment at the end of key stage 3'.
3. A guidance use. The inference made by a personal tutor – when encouraging a student to take three single sciences at GCSE on the basis of a key stage 2 result – may concern 'general aptitude for science'.

4. A school choice use. The inference made by parents – when deciding which primary school to send their child to on the basis of its profile of aggregated results in English, maths and science – may concern 'general quality of teaching'.
5. A system monitoring use. The inference made by a politician – when judging the success of educational policy over a period of time on the basis of national trends in aggregated results in English, maths and science – may concern 'overall quality of education'.

5.5 Each of these result-based decisions/actions is premised on the use of key stage 2 test results.[iv] Yet, in each case, a slightly different kind of inference is drawn from them. None of these use-inferences are precisely the same as the primary design-inference (the inference that the key stage 2 test result was primarily designed to support). Indeed, some of the use-inferences are radically different in nature from the design-inference.

5.6 So, when it comes to validation (establishing the accuracy of inferences from results for different purposes) the implication should be clear: accuracy needs to be established independently for each different use/inference. Results will inevitably be less accurate when used as indicators of future attainment than when used as indicators of attainment at the time of testing. And results may be less accurate still when used as indicators of general aptitude rather than as indicators of attainment. When it comes to using results as indicators of quality of teaching, or quality of education, we should expect less accuracy still, since the qualitative difference between the design-inference and the use-inference is so great.

5.7 This begins to ground the most important observation of the present report: an assessment system which is fit for one purpose may be less fit for another and could, conceivably, be entirely unfit for yet another.[v]

5.8 Recall that, for the sake of illustration, this section has focused purely upon the exploration of validity for different uses of results. The full story of evaluation needs to be far more embracing.

6 Can we construct a framework for system-level evaluation?

6.1 The fifth and final confusion concerns what an overall evaluation ought to look like. This is where we begin to explore the concept of 'fitness' in requisite detail. There are

at least six more-or-less discrete perspectives from which assessment systems need to be evaluated:

1. technical accuracy
2. moral defensibility
3. social defensibility
4. legal acceptability
5. economic manageability
6. political viability.

Each of these will be considered briefly below.

6.2 Technical accuracy

6.2.1 The first evaluation perspective is technical accuracy; essentially, the concept of validation. It poses the question: overall, how accurate can we expect inferences from results to be? And, as explained previously, this question needs to be explored independently, for each discrete use of results, i.e., for each discrete use-inference.

6.2.2 Unfortunately, it isn't always obvious which inference underlies (or ought to underlie) each use, which complicates the matter greatly. An example from system monitoring is helpful here. When considering trends in the percentage of students who attain at or above level 4 in science at key stage 2, are we (or ought we to be) drawing inferences concerning:

- the level of attainment of specific cohorts of students from one year to the next (where attainment is defined in terms of an explicit programme of study in science)?
- the level of proficiency of the national cohort over time (where proficiency is defined in terms of an implicit 'fuzzy set' of essential core competencies in science)?
- the level of performance of teachers of science over time?
- the overall effectiveness of policy and practice related to the teaching of science over time?

6.2.3 The first of the above use-inferences will be closest to the design-inference (being defined in terms of an explicit programme of study) and would, therefore, be likely to facilitate greatest accuracy. However, it's arguably of least interest as far as system

monitoring goes, because it's furthest away from the ultimate system monitoring ideal of identifying whether 'things are better now than before'. For instance, in the first years of a new curriculum for science: we would expect average attainment to increase gradually as teachers became better at delivering the new curriculum (with more practice and training in teaching the new elements, with an improved selection of curriculum-specific text books and resources, and so on); and we would expect average test performance to increase gradually as teachers became better at preparing students for the specific form of assessment associated with the new curriculum. Such gradual increases would seem to be inevitable.[vi] However, they would not necessarily imply that teachers were becoming better at teaching, *per se*; nor even that they were necessarily becoming better at teaching science; nor would it necessarily mean that students of the new curriculum were more accomplished than students of the old curriculum. As far as system monitoring is concerned, we probably ought to be validating in terms of more distant use-inferences (e.g., inferences concerning the performance of teachers, or the overall effectiveness of the system), since these have greater real-world significance. Unfortunately, these are correspondingly much harder to validate.

6.2.4 In theory, the analysis of accuracy is largely technical, using established methods for eliciting evidence of content validity, predictive validity, reliability, and so on. However, in practice, exactly how the various sources of evidence are synthesised into an overall judgement of accuracy is often not clear and, consequently, not that technical after all.

6.2.5 The logic of this perspective is essentially that: all other things being equal, more accuracy is better; and that accuracy must significantly exceed a threshold of chance.

6.3 Moral defensibility

6.3.1 The second evaluation perspective is moral defensibility. It poses the question: given the likelihood of inaccurate inferences from results, and the severity of consequences of error for those assessed inaccurately, is the specified use of results defensible?

6.3.2 This perspective starts by acknowledging that – within any assessment system – there will be a proportion of students who get assessed incorrectly and, consequently, for whom incorrect decisions will be made (be those selection decisions, provision eligibility decisions, placement decisions, and so on). It then proposes that – even if the system is just as far as most students are concerned – if it is sufficiently unjust for

a sufficiently high number of students, then the system may have to be judged morally indefensible. This is analogous to why many countries refrain from executing serial murderers. It's not that execution, *per se*, is necessarily judged to be morally indefensible; it's the risk of executing even a small number of innocent people. So the assessment parallel might be:

- when the stakes are low for students – as is often true of everyday formative assessment – it would not matter too much if it were fairly error-prone (such errors can often be identified quickly through ongoing dialogue)
- but when the stakes are high for students – as when examination results are used for selection – it would matter (such errors can often negatively affect life chances time and time again).

6.3.3 This results in a utilitarian analysis (emphasising the minimisation of 'horror' more than the maximisation of 'utility') for which two kinds of evidence need to be taken into account:

- technical judgements – concerning the amount of inaccuracy that might be expected (stemming from the analysis of technical accuracy)
- value judgements – concerning the severity of negative consequences for those students who *are* assessed incorrectly.

6.3.4 This final point raises a fundamental question for the evaluator: whose value judgements ought to be taken into account in this analysis, and how? The answer is far from clear, especially since different stakeholders (e.g., politicians, students, evaluators) are likely to have different values.

6.4 Social defensibility

6.4.1 The third evaluation perspective is social defensibility. It poses the question: is the trade-off between the positive and negative impacts from operating the assessment system sufficiently positive?

6.4.2 On the one hand, there will inevitably be a range of intended positive outcomes. In particular, the assessment results ought to empower users to make important educational and social decisions appropriately (such as selection decisions, placement decisions, school choices, and so on); to enable society to function more fairly and effectively than it otherwise would. Although, in theory, it may be judged

entirely possible to draw sufficiently accurate inferences to support a range of important decisions; in practice, that doesn't guarantee that users will actually do so. So this needs to be investigated empirically. In addition, features of the assessment system itself may well be designed to facilitate important educational and social impacts (such as the improved attainment of students when assessed through modular rather than linear schemes) and these impacts need to be investigated as well.

6.4.3 On the other hand, there will inevitably also be a range of unintended, and possibly unanticipated, negative outcomes. In particular, features of the assessment system which appear to be innocuous may turn out not to be so. Consider, for example, standards-referenced systems, which employ a single scale to report absolute level of attainment at key stages of an educational experience that may span many years (e.g., the national curriculum assessment system). The theory is that this should be motivating for even the lowest-attaining students; since it enables them to see that they are making progress as time goes by.[vii] However, such systems could conceivably turn out to be demotivating for precisely this group of students. Not only does the assessment reveal them to have attained lower than their peers at each key stage; they also see the gap between themselves and others widen on each assessment occasion.

6.4.4 The evaluator needs to be careful to distinguish those impacts which relate to the legitimacy of the assessment system, *per se*, and those which relate primarily to broader evaluation questions; for example, those concerning the legitimacy of educational or social policies or practices. School choice, for example, (even if based upon entirely accurate inferences concerning the general quality of teaching at a school) could conceivably result in a more socially divided society, which might be judged to be a bad thing. These are obviously important issues to be evaluated. However, they are issues for an evaluation of the policy of school choice, rather than for an evaluation of the assessment system which enables it. In practice, it is actually quite complicated to judge which impacts bear primarily upon the legitimacy of an assessment system and which relate primarily to broader evaluation questions; but it is useful to recognise the distinction and to try to work towards separation where possible. A rough rule-of-thumb might be: would we expect a different kind of impact if an alternative assessment system was in operation? If so, then the impact probably ought to be considered. If not, then the impact is probably attributable primarily to a broader policy or practice and, therefore, probably ought not to be considered. In the example above, relating to school choice, the 'divided society' impact might be

expected to occur regardless of the system used to generate results; so this impact might therefore not be relevant to scrutinise during an evaluation into the legitimacy of the underlying assessment system.

6.4.5 As with the moral defensibility perspective, the social defensibility perspective requires that two kinds of evidence be taken into account:

- empirical evidence – concerning the nature and prevalence of relevant intended and unintended impacts
- value judgements – concerning the costs of the negative impacts and the benefits of the positive impacts.

The synthesis of this evidence is based upon the utilitarian principle that: if, on balance, there appears to be too little benefit, for too few individuals, then the system may have to be judged socially indefensible.

6.5 Legal acceptability

6.5.1 The legal acceptability perspective asks: can the assessment system be operated without contravening the law?

6.5.2 This is becoming increasingly salient, both nationally and internationally. In England, the 1995 Disabilities Discrimination Act introduced legal rights for people with disabilities covering employment, access to services, education, transport and housing. The 2005 version of the Act included a new chapter which specifically covered qualification bodies; a provision which is intended to be extended to general qualifications from 1st September 2007.

6.5.3 The new legislation raises questions such as whether it is legally acceptable, within high-stakes general qualifications like GCSE English, to require specific forms of competence. For example, to be competent in English, is it absolutely necessary to be able to speak and listen fluently? Might there be a legal right for speaking- and hearing-impaired students to access this crucial ‘gatekeeper’ qualification? Nowadays, we routinely need to consider whether our systems can be designed to be more inclusive without unduly compromising them.

6.5.4 The analytical bases for evaluation, from this perspective, are the principles and precedents of law; the basic premise being that contravention of the law is

unacceptable. Significantly, judgements from the legal acceptability perspective can, and sometimes will, contradict judgements from the perspective of technical accuracy. Indeed, it may occasionally be necessary to make an assessment less valid in order for it to comply with the law. This is because technical analyses typically elevate the majority (sometimes at the expense of minorities) while legal analyses often elevate minorities (sometimes at the expense of the majority).[viii] Legal experts and assessment experts do not always share the same concept of fairness.

6.6 Economic manageability

6.6.1 From the perspective of economic manageability, the evaluator asks: is the burden of the assessment system upon society manageable?

6.6.2 The idea of burden does not reduce simply to financial cost, but also extends to issues such as: human resources (e.g., the availability of skilled examiners); workload (e.g., the time spent by students and teachers in preparing coursework); processing infrastructure (e.g., the demands made of the postal system when delivering scripts); and even ecological impact (e.g., the 'rainforest cost' of the paper which flows through the system each year).

6.6.3 The analytic basis for answering this kind of evaluation question is economic, grounded in the principles that: all other things being equal, less expense and consumption is better; and that there will be a threshold of expense and consumption which cannot reasonably be exceeded.

6.7 Political viability

6.7.1 The final perspective is political viability, which poses the question: is society prepared to buy into the assessment system?

6.7.2 Clearly, if society is not prepared to buy into the system then – no matter how good it might seem to be from the other perspectives – it will remain unviable. Unfortunately, such failures are not uncommon in the world of educational assessment. In England, the following might be mooted as examples: S papers; Records of Achievement; the Certificate of Pre-Vocational Education; parity of esteem between academic and vocational qualifications.

6.7.3 Unlike the other perspectives, the underlying principle here is essentially arational. It is best illustrated by platitudes of folk psychology such as: the customer is always right; or, you can lead a horse to water but you can't make it drink.

7 How should we conduct system-level evaluation?

7.1 Turning the framework for system-level evaluation into a real-life evaluation is far from straightforward. As noted, each discrete use of results ought to be evaluated, independently, from each of the six perspectives. This clearly implies a very large amount of research; and the more uses to which results are put, the more research is required. Moreover, it ought not to be restricted to the intended or 'official' uses either, since the unintended uses and even the proscribed ones are important too.

7.2 The example of national curriculum testing is useful here. Certainly, test results are not used for licensing nor for the certification of higher professional skills. Diagnosis and provision eligibility probably require results from more specialist tests; while selection and qualification would typically be based upon results from exams taken later in an educational career. Whether test results have (or ought to have) a role in guidance and national accounting is less clear. What does seem likely, though, is that results from national curriculum tests are used for the remaining 14 purposes, whether legitimately or not. Again, the question of legitimacy would need to be explored independently for each use.

7.3 At some point, evidence and argument from the independent analysis of specific uses, and specific impacts, needs to be brought together into an overall evaluation argument. This will require judgements concerning the acceptability of compromises and trade-offs, with reasoning along the lines of: 'the system may not be particularly good for this use, but it's probably better than nothing; admittedly it does have a big negative impact for a small number of students, but perhaps not too many; and, ultimately, the system is quite good for that purpose, and that's the principal purpose, after all...' (obviously, this is simply a caricatured microcosm of an overall evaluation argument).

7.4 The previous paragraph hints at another important point: to reach overall evaluation conclusions, it is necessary somehow to weight the importance of alternative uses and impacts. There needs to be some indication of which are the most valued uses of results, and impacts of system operation, and which are more like fringe benefits. This might be a problem if there is neither any consensus among stakeholders nor

formal specification from policy makers. Again, whose value judgements ought to be taken into account in this analysis, and how?

7.5 Ultimately, the legitimacy of the assessment system cannot be judged in isolation, but only in relation to:

1. a new-improved version of the same assessment system; or
2. an entirely different assessment system; or
3. a suite of more tailored systems, operating in parallel; or
4. no assessment system at all (which, admittedly, would be unlikely ever to triumph as an evaluation conclusion, but which is still useful as an anchor point).

7.6 This raises yet another complication: that each alternative will need to be put through the evaluation mill in its own right. That is, an overall evaluation argument will need to be constructed for each of the alternatives, to pit them against the present state of affairs. Unfortunately, since these are likely to be largely hypothetical at this stage, the construction of evidence and argument will inevitably be patchy and indirect.

7.7 Finally, it is worth emphasising that the aspiration of system evaluation is not perfection, but legitimacy; and this is also true for the flip side of evaluation, design. This legitimacy is a real-world, pragmatic aspiration, which might be characterised as: ‘overall, at least satisfactory, and preferably good, but inevitably not perfect’. So, for example, whereas the principle of maximising validity (from a technical accuracy perspective) might go so far as to recommend a separate system for each discrete use of results, the principle of minimising burden (from an economic manageability perspective) might recommend just one. The overall evaluation conclusion (bearing in mind all perspectives) might recommend something in-between; say, two or three separate systems, operating in parallel, each supporting a distinct set of three or four different uses of results, and each with its own particular impacts. Compromise and trade-off are fundamental to the design and evaluation of assessment systems.

8 Is system-level evaluation feasible?

8.1 Given all of the above, is it humanly possible to undertake a rational and rigorous system-level evaluation? Is it possible to reach a straightforward conclusion to a straightforward question like ‘should the system of national tests be changed?’ This is a challenging issue. It’s probably true to say that no educational assessment system has ever been evaluated quite as rigorously as recommended above. Indeed, it’s an

inevitability of real life that decisions are generally made in the absence of complete evidence and argument; and the world of educational assessment is no different in that respect. Having said that, the inevitability of falling short of the ideal evaluation does not detract from the importance of constructing as rigorous an evaluation as is possible.

8.2 Frameworks like the one presented above can help inquirers to scaffold useful answers to thorny evaluation questions. They can be particularly helpful for identifying holes in the overall evaluation argument: where research still needs to be undertaken; and where argument still needs to be constructed. And they can also help stakeholders to reflect upon, to clarify and to articulate their different priorities for national assessment; to distinguish between the crucial uses and impacts and those which are more like fringe benefits.

9 Using the framework to identify common limitations

9.1 Finally, frameworks like the one presented above can also help inquirers to identify limitations in evaluation arguments presented to them by others. In this last section, a few common limitations will be illustrated.

9.2 The conflation of different evaluation questions

9.2.1 In constructing a robust evaluation argument, it is important to put to one side issues which appear to be relevant, but which actually fall under a broader evaluation remit. For example, when evaluating the use of test results for school choice purposes, it is clearly relevant whether the system supports sufficiently accurate inferences concerning differences in the general quality of teaching between institutions. However, as suggested earlier, the positive and negative impacts arising from school choice, *per se*, are probably not directly relevant and, as such, should not enter into the evaluation argument.[ix] Of course, they are crucial to evaluating the policy of school choice, and this evaluation needs to happen independently.

9.2.2 A particularly common limitation of many formal and informal evaluation arguments is the failure to distinguish between the impacts attributable to testing, *per se*, and the impacts attributable to the high-stakes uses of results which the testing is designed to support. So, for example, to the extent that high stakes can trigger behaviour which corrupts the validity of test results and the effectiveness of teaching, high stakes can similarly trigger behaviour which corrupts the validity of teacher assessment results

and the effectiveness of teaching. In short, it may not be the operation of the assessment system, *per se*, which is problematic, but the policies or culture underlying those high-stakes uses. Having said that, there are important differences in this situation from the one described above. First, although the impacts might be primarily attributable to the high-stakes uses, they directly affect the accuracy of results from the system; which thereby renders those impacts directly relevant to the evaluation. Second, the impacts upon teaching and learning, even if primarily attributable to the high-stakes uses, are likely to be different across different assessment systems; which again recommends that they enter into the evaluation.

9.3.3 When there is a range of equally valid, although logically separable, evaluation questions to ask, then these ought somehow to be arranged within a meta-framework. For example, it makes sense to interrogate the purposes of curriculum and qualification, before interrogating the purposes of assessment. Or, to put it less formally: the assessment-tail should not wag the curriculum-dog. At least, not too much; where the rider 'too much' is essential. In fact, the process of meta-evaluation needs to be iterative and will necessitate inevitable trade-offs and compromises. By way of extreme example, it would not be legitimate to promulgate a radically new curriculum for school-leaving examinations, if the learning outcomes which were elevated could not be assessed with sufficient accuracy: we need to remember that the examination results have important functions in their own right, as the basis for making the kind of qualification and selection decisions that are necessary to support a fair society.

9.3 The lack of a specified alternative

9.3.1 A common limitation of evaluation arguments is the lack of a specified alternative system. It is not foreseeable that society would tolerate the rejection of educational assessment entirely. So it would seem to be incumbent upon any critic of present arrangements to explain, in some detail, how an alternative system would, on balance, be more legitimate. The key issue, here, is one of detail. For instance, the 'test versus teacher assessment' debate is literally meaningless unless the detail of the alternative system is spelled out.

9.4 Too incomplete an analysis of uses and impacts

9.4.1 Even when two or more systems are specified in sufficient detail, and pitted against each other, it is often the case that the evaluation argument remains incomplete,

through omission of central components. This frequently occurs when an alternative system is proposed which is particularly effective in relation to certain uses and impacts – perhaps genuinely more so than the present system – but which leaves crucial other uses or impacts unaddressed. In England, numerous protagonists have argued recently for employing moderated teacher assessment (for certain uses and impacts) alongside a national monitoring unit (for others); instead of the present system of national curriculum testing. Few protagonists, though, have also grappled effectively with how best to support uses which require the comparison of teachers and schools in a high-stakes context. This particular debate is very important – because the arguments in favour of certain forms of teacher assessment alongside a national monitoring unit are persuasive. However, due attention also needs to be paid to satisfying the demand for trustworthy data on school effectiveness.

9.4.2 Another limitation of many evaluation arguments is the lack of available evidence, or a reliance upon evidence which is easy to challenge. A particular example of this at present is the impact of national curriculum testing upon teaching and learning, especially at key stage 2. Despite the system having been in operation for over a decade, and despite considerable anecdotal evidence of negative washback, there is remarkably little systematically documented evidence. This greatly hinders effective evaluation.

9.5 The gulf between real and hypothetical

9.5.1 Finally, while extant systems must inevitably be evaluated in the context of real-world operation – mired in the kind of intricate relationships which give rise to unforeseen problems – alternative systems will typically be evaluated as promising-hypothetical. In this context, it is easy to give the alternative system undue benefit of the doubt, without recognising that its implementation will inevitably necessitate certain compromises and will result in its own unforeseen problems. At the very least, the root cause of the problems which beset the extant system need to be extrapolated to the promising-hypothetical.

Endnotes

[i] The term ‘assessment’ is used generically, to refer to any instrument or process through which student competence or attainment is evaluated (e.g., test, teacher assessment, examination, etc.). The term ‘system’ is used to encapsulate, in a broader sense, the detail of the structure and mechanism through which students are assessed. In relation to national curriculum testing, for instance, this detail would include procedures for test development, distribution, administration, marking, reporting, evaluating (and so on), as well as the technical, professional, managerial and administrative employees required to develop and operate those procedures.

[ii] It is based upon insights from the international literature on validation and evaluation, although references to specific sources have not been included (further information can be provided on request).

[iii] Although there is a huge debate in the technical literature on the precise extension of the term ‘validation’, this does not significantly affect the tenor of the argument developed in this report.

[iv] In reality, it would be advisable to use more than one source of evidence to support important decisions (such as placement, monitoring, guidance and so on). Indeed, assessment professionals are increasingly preaching this dictum. However, that does not change the basic principle that, when results are used to support different purposes – whether alone or in combination with other sources of evidence – different kinds of inference are drawn from them.

[v] There are different ways of emphasising the point that results which are fit for one purpose may not be fit for another. The approach in the text is to focus upon the different inferences which need to be drawn from results. Another approach would be to stress that systems need to be designed differently for different purposes and different design compromises will be made. (Compromises are made so as not to over-engineer the system, because increased accuracy comes at a price; assessment design aspires to sufficient accuracy, for a specific purpose, rather than maximum accuracy.) Ultimately, design characteristics and compromises which are legitimate for one use may be illegitimate for another.

[vi] Note that this is not to implicate the phenomenon of ‘teaching-the-test’ (whereby, over time, teachers reduce the scope of their teaching, excluding those aspects of the curriculum that the tests tend not to cover). This practice is neither appropriate nor inevitable. Were it to occur, it would occur *in addition* to the impact of practice, training and improved resources (described in the text).

[vii] This contrast with norm- or cohort-referenced systems, in which the lowest-attaining students may be awarded the same very low rank at every stage of their educational career,

despite making real progress in learning and despite achieving respectably given their particular situations.

[viii] Any technical analysis which is based upon an average (which is frequently the case for large-scale educational assessments) thereby tends to elevate the majority.

[ix] Other than when considering the negative impacts which arise from inappropriate school choices, consequent upon inaccurate results data (the moral defensibility perspective).

