What is the price of consistency in educational assessment?

Dr. Andrew Davis
Research Fellow
School of Education
Durham University
Leazes Road Durham
DH1 1TA
a.j.davis@durham.ac.uk

Introduction

How far can *consistent* assessment capture all the worthwhile features of educational achievement? Are some important components of learning *necessarily* open to a range of potentially inconsistent judgments by different assessors? In this paper I develop a cautiously affirmative answer to the second question, hedging it about with a number of qualifications.

These issues are related to the familiar tension between reliability and validity. It is widely discussed (Gipps 1999, Isaacston 1999, Nystrom 2004), with a long history in the literature, going back at least to Cronbach 1960). Moreover, many commentators note that whether it is a problem depends on assessment purposes. 'High stakes' assessment prevails in many developed countries, with major effects on the lives of students, teachers and educational institutions. For such assessment, excellent levels of reliability must be expected, though putting a figure on this is problematic and contestable. Perfect reliability is impossible. So what level *is* necessary? Arguably there is no definitive answer. We are always challenged to weigh the consequences of assessment for a given purpose against 'costs' to its validity.

There are a number of ways of viewing consistency or reliability in assessment. My broad focus here is on reliability construed as levels of agreement between different assessors. For example, the reliability of a written test is the extent to which several examiners marking 'blind' would come up with similar scores for given pupils.

Empirical research (Harlen 2004) indicates that closely specified achievements are more likely to be measurable with high degrees of reliability than those characterised more loosely. For instance, a test of whether someone can run 100 metres in under a specified time can have a very high level of reliability; the timing device, what counts as finishing the 100 metres and so forth can be laid down in great detail, ensuring that different assessors will arrive at the same verdict.

Suppose it turns out that at least some aspects of 'important' educational achievement cannot be appraised consistently. A high stakes system with at least half an eye on reliability is unlikely even to attempt to deal with them. This may be why Speaking and Listening is not assessed in the UK National Curriculum English tests though I cannot offer evidence for such a speculation. Moreover, a familiar theme from empirical research is that teachers devote less time and efforts to untested learning outcomes. It is supported by Harlen 2004 in a research review of a large number of studies. (For discussion of this issue see Wiliam 2003).

Some commentators may concede that a proportion of significant achievements *are* ignored in very reliable systems but that the benefits of high stakes assessment make this worthwhile. The plausibility of their verdicts will depend in part on their justification for prioritising certain aspects of educational achievement over others. I return to this issue at the end of the paper.

Inconsistency as a defect

The main part of the paper examines whether inconsistency *must* be regarded as a defect in all cases in the context of educational assessment. However, I first briefly examine two contexts where inconsistency obviously *is* a problem. Often enough, the very fact that verdicts are not consistent with each other implies that they do not deserve to be taken seriously. An appreciation of why this is the case will help us understand later whether some cases of inconsistency within educational assessment might prove to be exceptions to this rule.

(1) Natural Science

In natural sciences such as physics and chemistry, results are expected to be replicable by any researcher. If not, they are discredited, or an explanation is required within the

terms of the relevant scientific theory to account for the discrepancies. If a scientist wants to say something true and objective about the size of a planetary body, an object's velocity, the voltage in a wire or what happens when iron filings are added to sulphuric acid, then it should be clear how the claims are supported. Moreover, anyone with appropriate skills, resources and instruments should be able to produce the same results. Without this there is nothing of significance, let alone anything approaching a truth value.

(b) *Normative judgements understood to be mere expressions of feelings.*

It is a commonplace that people frequently disagree when making value judgments. Many philosophers have urged that value judgements are 'mere' expressions of feeling, David Hume being just one representative of this long tradition. If he is right, then value disagreement is a symptom of the fact that we project our sentiments on a world whose fabric lacks value in principle. Of course, the sharpness of value disagreement in a Humean universe might sometimes be blunted by socialisation. Be that as it may, Shaper-Landau comments:-

'If moral facts were reports of 'objective' states of affairs, then we should expect in morality the breadth of convergence that emerges in some of the more rigorous empirical and theoretical disciplines.' (Russ Shaper-Landau 1994 p.331)

Inconsistency not linked to the absence of significance

In the light of these two examples I now critically examine in turn each of three arguments that significance is not inextricably linked to the consistency of educational assessment.

Some verdicts in educational assessment contexts are purely factual. Here are three instances. The student's answer to question 10 was incorrect. The student only wrote 1200 words when the expected length of the essay was 5000. The student spelled 20 words incorrectly on his first page. I will assume without argument that inconsistency about *these* kinds of verdicts is obviously a problem, and that any satisfactory educational assessment process will seek to avoid it.

However, many components of educational assessment are normative. For example, an undergraduate essay might be said to show imagination, high levels of

critical reflection or cogently developed argument. It is with the normative elements that the following discussion is largely concerned.

Aesthetic judgments

The first argument for the *possibility* of legitimately combining significance with inconsistency may be summarised as follows: Some assessments resemble aesthetic judgments and significance in aesthetic judgments is not necessarily and comprehensively linked to consistency. So just how is this possible?

Clearly there are widespread disagreements between those appraising art, musical performances, dance, and literature. Of course, subjectivists insist that judgments about beauty, elegance, grace and the rest 'merely' reflect feelings and that this accounts for the range of verdicts. On this view there is no significance here *beyond* strong feelings. Given such a perspective, many would oppose relating *any* educational assessments to aesthetic judgments. Educational assessment surely must transcend mere expression of feelings, or so they would argue.

Subjectivism about aesthetic judgments denies the very legitimacy of Kant's 'judgements of taste'. These aspire to universal validity.

> ….he says that the thing is beautiful; and it is not as if he counts on others agreeing with him in his judgment of liking owing to his having found them in such agreement on a number of occasions, but he demands this agreement of them. He blames them if they judge differently, and denies them taste, which he still requires of them as something they ought to have; and to this extent it is not open to men to say: Every one has his own taste. This would be equivalent to saying that there is no such thing as taste, i.e. no aesthetic judgment capable of making a rightful claim upon the assent of all men.
>
> Kant 1928 p 52

One relatively weak counter to those dismissing subjectivism would be to *accept* the subjectivist story both for aesthetics and certain instances of educational

assessment, but to contend that the very existence of varied feelings is important. Hence an assessment regime which sidelines them is open to question. This move needs a little more explanation.

Suppose assessors drawn from a particular culture assess student achievement according to criteria such as 'level of insight', 'interpretive skills' and 'quality of critical reflection'. For the sake of argument, let us accord them appropriate levels of subject knowledge, professional experience and acquaintance with assessment processes. If the group work together over a period of time a reasonable consensus on verdicts should be perfectly possible, and indeed is a matter of common experience. Yet, given the criteria concerned, should we not *expect* a range of responses to at least *some* aspects of the student achievements concerned? Of course, a proportion of this disagreement will stem from everyday human failings. Assessors may be tired, irrationally swayed by the handwriting or the font size of the scripts, and so on. But in addition to this, some assessors will just *feel differently from others* about certain aspects of student performance. These feelings should not be ignored. This is how a subjectivist construal might run.

However, as I have already conceded, subjectivism about aesthetic judgments is a weak response. Moreover, from such a viewpoint, a claimed analogy with educational assessment is less than easy to swallow. Why should anyone *care* about examiners' *feelings*, if that is all they are? Consider just three possible uses to which assessment may be put: to inform teachers' decisions about what next to offer their students, to help select students for educational courses and to support employers who wish to make suitable appointments. The mere *feelings* of examiners should not play a substantial role in the fulfilling of these kinds of assessment purposes. To conclude, if a plausible case is to be made for a parallel between some aesthetic judgments and certain educational assessment verdicts, a crude subjectivism must be rejected.

A more robust approach to aesthetic judgments would point up the complexity and richness of aesthetic disputes and argue that they cannot be located entirely within the realm of 'mere feeling'. A disagreement about whether a mathematical proof is 'elegant' might stem from differing views of proof itself – some favouring an algebraic approach and others taking especial delight in the use of visual representations of various kinds. A range of verdicts about whether the performance of a baroque violin sonata was 'expressive' might reflect a diversity of approaches to the 'meaning' of the piece. One assessor might pay much attention to what the

composer intended, as far as evidence could be gathered, another could attempt to interpret the piece in the context of the particular cultural setting in which it first emerged while yet another focused on the music as a conveyer of meaning to the listener in the 21st century. This scenario does not assume that aesthetic appraisals fall neatly into just one of these three categories, nor that the categories exhaust the possibilities. Many appraisers will draw on complex hybrids of these and other approaches.

Scruton 1997 explains one feature of the complexities pervading aesthetic judgment in terms of 'aspect perception'. He attributes this idea to Wollheim 1987, who calls it 'representational seeing'. When we see one thing in another, such as seeing a face in a picture (which itself is not a picture *of* a face) we have an instance of such aspect perception. Sometimes we entertain the idea that there is, for instance, a man in a picture. We need not be thinking that there is a man there; we may be restricted to an imaginative involvement. In his discussion, Scruton refers to ambiguous figures; many readers will associate these with Wittgenstein's treatment of 'seeing as' in *Philosophical Investigations* and his examples of the schematic cube and the duck-rabbit (Wittgenstein 1958). Such complexities open up the possibility of legitimately inconsistent judgments about the situations concerned.

Even if these examples from aesthetics are understood and accepted, it is, of course, a substantial further step to claim that such approaches are ever applicable in educational assessment contexts. However, in Higher Education and in many school subjects, assessors will be judging features such as critical thinking, imagination and analysis. Criteria for writing tasks included in English tests for 11 year olds included 'Length and focus of sentences varied to express subtleties in meaning and to focus on key ideas,' 'All aspects of the story are consistent and contribute to overall impact' and 'Viewpoint well-controlled, eg selection of detail to encourage reader to sympathize with the explorer; action portrayed from different viewpoints.' (QCA 2006) The descriptors here were littered with normative and even aesthetic components.

'Aspect perception' seems a rich and powerful idea to play with in this connection. Yet any attempt to liken verdicts about these achievements to aesthetic appraisals may provoke a negative reaction. The relevant assessments are supposed to be far removed from judgments about paintings, for instance, where, despite the importance of cognitive expertise, the role of the personal response is central and

entirely appropriate. If, for instance, university examining resembles the proceedings of art critics, then it is time for the process to be shaken up, to be made more 'transparent', open to external scrutiny and generally to be 'professionalised'. Or so it might be argued.

If the hostility to the comparison with aesthetic verdicts stems mainly from indignation about the very possibility of inconsistency, it might well be possible to exclude such inconsistency. For, within a particular society, a level of consensus *can* develop about the quality of novels, paintings, compositions and other expressions of the creative arts and, as we have already noted, assessors can develop a shared culture within which convergence of verdicts might be achieved.

Although this is perfectly possible, any attempt to *ensure* high levels of consensus in the arts arguably damages something at their very heart. It is no accident, no peripheral inconvenience, that the history of Western Art music includes the rejection of tonality by Schoenberg and the Second Viennese School and that Messiaen[1] fails to follow the canons of sonata development and the sense of progression over time which are central to the classical compositions of Haydn and Mozart. When these events occurred they were highly controversial and perhaps still are. Those making aesthetic appraisals of such works of Western Art Music will not achieve convergence in judgment. A sophisticated treatment of issues in the philosophy of music is beyond the scope of this paper. All that can be said here is that some judgments about whether, for instance a piece of music 'has a proper sense of direction' or 'develops appropriate tensions and satisfying resolutions of these' will involve making assumptions about the extent to which music *must* present, in any sense, a narrative over time. These assumptions will not and arguably *should* not be required to be shared universally by those reaching verdicts about quality in this area.

Were musical developments to have been subject to a requirement that they evoked a consistent response from appropriately qualified judges, then many events crucial to the continued flourishing of musical culture could not have taken place. It would have been, in effect, to have erected a barrier between the appraisers and the essence of that which they were appraising.

Again, those reluctant to compare educational assessment to aesthetic judgment are likely to throw up their hands in disgust. They will say that any idea that the kinds of judgments involved in educational assessment could be subject to the

kind of revolution and radical subversion resembling paradigm shifts in the arts would be patently absurd.

This rejection of the analogy with aesthetic judgment arguably misses the point. It is not being suggested that continuous or even occasional revolution would be a good thing within the processes of educational assessment. At the same time, paradigm shifts in the arts are an extreme instance of something fundamental to their existence: the continuing possibility of value and significance that is independent of practices and sets of criteria developed by particular communities and cultures. The question under consideration is whether this idea is relevant to the status of at least some normative judgments in educational assessment.

Many involved in assessment have encountered cases which seem to relate to this point. Unusual and sometimes gifted pupils write essays or offer other products which cannot be properly graded according to the criteria laid down. Yet markers may claim to discern very real value. Not infrequently in these kinds of situations, markers disagree sharply, with one party identifying outstanding qualities, while others may feel that the assignment should fail. Sometimes the rules can be bent, and the student emerges with a high score, arguably appropriate to the quality of their work. On many other occasions, the system wins, and the student may receive a poor grade. Often enough the system agrees to 'split the difference', with the result that a potentially outstanding student gains a mediocre verdict. Although this situation will strike many as unsatisfactory, as we have already noted, we need to know the uses to which the assessment results will be put before deciding just how seriously we should take such problems.

I have not succeeded in establishing beyond doubt that some legitimate judgments in educational assessment resemble some aesthetic appraisals, and that because the latter are not comprehensively tied to consistency in judgments neither should the former be so tied. Moreover, we would still need to debate the relative importance of that potentially excluded by prioritising consistency. Nevertheless I suggest that enough has been said to justify taking the analogy with aesthetics seriously.

Holism and particularism in normative judgments

I now examine a second argument for the possibility of combining a degree of inconsistency with 'significance'. This argument stems from a type of *holism* about some assessment criteria.

One supposedly rigorous approach to assessment is sometimes known as using *weighted criteria.* A mark scheme specifies maximum marks under a range of headings, such as cogency of argument, clarity of expression, critical reflection, appropriate use of relevant literature, accuracy in the use of scholarly conventions and standard of written English. These exemplar headings might fit essays in the humanities; evidently assignments of other kinds in different subject areas would attract distinctive criteria.

The apparent virtues of this way of proceeding are obvious. Markers have a clear framework within which they can work. When discussing assessment with each other they have common guidelines; a developing consensus and consistency of verdicts seem very likely outcomes.

Yet it can be argued that there are serious theoretical problems here, at least within *some* subject areas, especially the humanities and social sciences. The meaning and significance of many of these criteria cannot be separated from that of their bedfellows. For instance, what counts as cogency of argument depends at least in part on what counts as clarity of expression, which, in turn, cannot easily be separated from standard of written English. Many of the criteria cannot be considered and assessed *on their own.* How they are applied depends on how related criteria are applied, and vice versa. There is a kind of hermeneutic circle here. We cannot understand any one criterion unless we understand others and how that one relates to the others. Yet, in turn, we cannot understand any one of the others either without understanding the rest and how they fit together.

There is no need to exaggerate the position to make the basic point. We must, for instance, concede that some criteria are more 'atomic' then others. If we are awarding marks for spelling standards then these can be isolated from any marks we might award for critical reflection. (Though even this might be challenged in an extreme case. If the spelling is so appalling that the writing is almost impossible to follow then marks for critical reflection are hardly going to be readily available.)

Be that as it may, many of the key criteria used to judge an essay in English literature, history or philosophy are interrelated. A very similar story could be told for hosts of other assessment examples. What, then are the implications for the strength of the link between consistency and educational significance in assessment?

Perhaps it is possible for assessors to reach agreement with each other over a period of time about this kind of criteria jigsaw – about how the different elements should be seen as relating to each other when marking the essays in question. My concern about this is as follows. Might there be a *cost* to securing such an agreement? Might some features of the essay be ignored? For instance, when reaching a verdict about the quality of critical reflection, if we insist that this criterion is related in specific ways to others such as cogency of argument, are we confining ourselves to scrutinising the essay in one particular way and excluding other perfectly legitimate approaches? Are we, so to speak, seeing the assignment as illuminated by a particular kind of lighting when, if we were allowed different lighting, other aspects might become visible?

It is illuminating to relate this debate to issues that have arisen over the last two or three decades around the approach to meta-ethics known as moral particularism – a debate associated especially with Jonathan Dancy and John McDowell. When considering reasons favouring one action rather than another, for instance, Dancy claims that:

> '1. A feature or part may have one value in one context and a different or opposite value in another.
>
> 2. The value of a complex or whole is not necessarily identical with the sum of the values of its elements or parts.' (Dancy 2000, p 139)

Examples are contested by those opposed to particularism, but here are two simply to illustrate the type of thesis Dancy supports. The fact that an action results in pleasure can make it better in some circumstances *and worse* in others. Suppose a possible action of mine results in letting people watch hangings. If the people get pleasure from the spectacle then (it might be claimed) my action is morally worse than it otherwise would have been. A second case: 'That one of the candidates wants the job very much indeed is sometimes a reason for giving it to her and sometimes a reason for doing the opposite'. (Dancy 2000 p 132-3).

We need not be in a position to pronounce definitively on the philosophical strengths and weaknesses of moral particularism in order to draw some parallels with assessment contexts. Irony in an essay about certain postmodern postures might be a strength: that irony in a critical discussion of Frege on sense and reference could be wholly out of place and provide a reason for a lower mark than would otherwise have been earned. Rich imagery and metaphor might enhance a discussion of Keat's poetry, but could well detract from the quality of an analysis of the German economy between the two World Wars. One explanation for this 'particularism' lies in the holism and the 'hermeneutic circle' referred to above in the context of the discussion of weighted criteria. The weight and significance of irony, for instance simply cannot be appraised outside the context of the writing in which it occurs; its role and contribution depends crucially on other features of the assignment in question.

Arguably, there can be legitimate disagreement about the contribution of such features to the overall quality of the essay. If we insist on excluding such disagreement in order to achieve the requisite levels of reliability and to make the assessment process 'work', are we then able to examine *all* the significant features of the student products or achievements concerned? I suggest that this question deserves serious consideration.

Incomparability and incommensurability

The third argument for the possibility of combining inconsistency with significance draws on concepts of incommensurability. Isaiah Berlin defends a pluralism of values, as opposed to a monism according to which we can order, compare and contrast values within some kind of overall theory. '..human goals are many, not all of them commensurable..To assume that all values can be graded on one scale, so that it is a mere matter of inspection to determine the highest, seems to me to falsify our knowledge that men are free agents..' (Berlin 1969 p 171)

Could a measure of disagreement between assessors stem from attempts to compare at least *some* features which are inherently resistant to comparison*?* If, for instance people are asked to compare the Taj Mahal with the Sydney Opera House there will be a range of reactions. These will include sheer bewilderment, verdicts favouring the Opera House, verdicts according the two buildings the same value and judgments on the side of the Taj Mahal. The diversity of responses in itself proves

nothing. However, it might well be symptomatic of the fact that people think that in such examples comparisons are odious. Similarly, they may feel that a Mozart Opera ought not to be compared with a Coldplay item, and Jane Austen's novel *Emma* ought not to be weighed against Tolkien's *Lord of the Rings.*

Note, however, that the opening formulation above speaks cautiously of *at least some features* that should not be compared. If we are determined to do so, we can find 'covering values' (Chang 1997) against which the items *could* be compared, at least in theory. We cannot compare *Emma* with *Lord of the Rings* in respect of their respective depictions of the *nouveaux riche* or the verbal painting of landscape, but perhaps we can make more progress if we consider them under the heading 'Worth of literature dealing with moral themes'. Would this move allow sensible comparisons to be made after all?

It may be objected that we cannot compare the value of the treatment of the *nouveaux riche* in *Emma* and in the *Lord of the Rings* because, obviously, only *Emma* attempts it. Similarly, Tolkien achieves (with consummate skill and artistry) the verbal painting of landscape, while Jane Austen is otherwise employed. What is needed for the argument that comparisons are genuinely odious, are features which are in a sense common to both works yet where comparability is clearly open to question. It is true that both works deal with *morality*. Yet, arguably, Tolkien's epic portrayal in a fantasy world of the struggle of good against evil should not be considered alongside Jane Austen's subtle and gently ironic treatment of Emma's moral failings. Superficially, we have found a covering value. However, the works are simply not doing the same kind of thing here – and insisting that we must reach a verdict on which work does this 'better' seems to involve a fundamental distortion of the distinctive qualities to be found in each work.

I suggest that in our reactions to these examples, we are grappling with incommensurability. Lukes 1997 explains this as existing where it would be 'inappropriate' to make comparisons although they would not necessarily be unintelligible or meaningless. So, although a comparison need not be incoherent, we nevertheless hold back from making one; 'we do sometimes refuse to commensurate or compare alternatives' … 'such a refusal can display our understanding of what is involved in certain relationships..' (Lukes *ibid)*

To set Lukes's views in context, we need to compare and contrast them with other recent treatments of incomparability and incommensurability. For instance, John

Broome (2001) thinks of incommensurability as involving alternatives which realize such different values that

> 'it is impossible to weigh them against each other precisely….When values are incommensurable, it may not be determinate which of the two alternatives is better. It may be that neither is better than the other, yet we also cannot say they are equally good.'
>
> Broome 2001 p 12

He goes on to illustrate incommensurability in this sense with the example of God telling Abraham to sacrifice his son, claiming that submitting to God's will *cannot* be weighed against saving Isaac.

Compare this treatment with Chang's (2002), who argues that the failure of the trichotomy of possible relations between A and B – that A is neither better nor worse than B and that A and B are not equally good does not prevent them being what she calls 'on a par' – they could still in some sense be comparable. She dubs those prepared to countenance a *fourth* relation between A and B 'tetrachotomists', observing that 'the tetrachotomist thinks that even if one item is neither better nor worse than another and that the items are not equally good, there may nevertheless be an evaluative difference between them..' (p. 664) Even if she is right about, this we are now contemplating comparability of a radically different kind from that in which we can compare A and B according to any kind of common measure.

I suggest that one of the reasons for our refusal to rank certain items on a common scale is our awareness of the *distinctive* quality of the features concerned. It is not a global rejection of ranking procedures, nor of the possibility of covering values in even the majority of cases. It is rather the appreciation that *some* aspects of morality and of works of art resist ordering on a scale in terms of a common value. This is not the claim that 'distinctiveness' is a sufficient condition for significant value. It is the much more modest contention that it is a *necessary* feature of *some* kinds of significant value.

We must be cautious about how we deal with incomparability here. Crowder 1998 argues that if values *could not* be compared then we could not make comparative judgments about them in particular cases. Yet it is part of our moral

experience to succeed in doing just this, or so Crowder contends. He seems to be right about this; even if we cannot explicitly codify how we weigh, for instance the importance of a heart operation against an expensive drug for the treatment of schizophrenia, we constantly do make such decisions, and would hope that many of them are not just arbitrary.

There are yet other conceptions of incommensurability in recent philosophical literature. I briefly draw attention to just one more here. Incommensurability may be seen as involving the possibility of disagreement even given reasonable reflection – 'Incommensurability marks a practical limit on the power of reason alone to decide value conflicts' (Plaw 2004 p 113) As stated, this is compatible with a Humean approach to normativity but Plaw likens political judgments involving incommensurability in this sense to aesthetic judgments and he clearly does not think of the latter as merely subjective.

Suppose, then, that our understanding of selected values shows itself in an appreciation that we ought not to weigh them against each other on any kind of common scale. (I certainly cannot claim to have established this here – but would contend that the examples cited at least hint at this possibility.) I suggested at the beginning of the section that a proportion of disagreements between appraisers could be symptomatic of an insight (whether conscious or otherwise) that they are being asked to weigh features against each other on a common scale, features which should not be so weighed. The 'training' of relevant assessors to improve consistency in judgments, to strengthen the reliability of the process will tend to exclude the consideration of qualities with 'incommensurability' aspects.

Conclusion

The approach in this paper has been deliberately cautious and tentative. I have considered three arguments for the claim that significant components of educational achievement cannot be captured by a highly reliable assessment system. I have concluded that these arguments should be taken seriously, even if they are not conclusive.

Those hostile to comparisons between aesthetic judgment and educational assessment would have a short way with all this. They could concede that we might offer a faint nod in the direction of incommensurability when dealing with arcane

aesthetic or value issues, but urge that in the everyday business of educational assessment we should ignore it.

However, surely we should worry if reliability seems to exclude some kinds of value. Admittedly, the potential losses seem less salient in some subject areas and at some levels. Yet, whenever there are evaluative elements in assessment, and arguably there usually are, concerns about potential losses are not easily dispelled. Ultimately, however, the seriousness of all this can only be judged in the light of our educational aims. For instance, those with thin instrumental objectives for education seem unlikely to be troubled by the arguments presented here. In the light of their aspirations for education, they will argue that the advantages of a suitably reliable high stakes assessment system outweigh the disadvantages. Nevertheless I have urged elsewhere (Davis 1998) that such advantages are greatly overrated, even given the assumption that education should support modern industrial economies. There is no scope in this paper to rehearse these considerations.

Those sufficiently exercised by my arguments may still wonder what practical steps could be taken as a result to modify testing and examination systems. I suggest that responses can only be given on a case by case basis, having regard to the purposes of the assessments concerned.

In Higher Education there is, arguably, a case for allowing inconsistent verdicts to stand on occasion, and for allowing particular weight to be given to very positive verdicts from particular assessors, even if others disagree. The strength of this case will vary from subject to subject and will also depend on the nature of the student assignment or examination. Examples in mathematics and science seem less likely, though I would not rule them out. The humanities and social sciences might well provide a rich mine of cases. Of course, universities implementing such a policy might encounter difficulties which would include a greater likelihood of student appeals. They would have to be very convinced of my concerns to risk this and I am not optimistic that I could persuade them.

Another way of scrutinising how we deal in practical terms with the implications of legitimate inconsistency is this: awareness of the very possibility may change some of the ways we decide to use assessment in the first place, depending as always on what we think education is actually for. The example noted early in the paper of English tests for 11 year olds is an important one to consider – and countless others could be cited. The 'backwash' of high stakes testing is an acknowledged

phenomenon. If some features of children's writing cannot be reliably assessed, then they will not be assessed. Teachers will know that they are not, and their approaches to teaching and the curriculum will be influenced accordingly. One obvious way out of this problem is to *stop*, at a stroke, the high stakes uses to which these tests are being put.

I will give the last word to Bernard Williams. He notes that 'if there are many and competing genuine values, then the greater extent to which a society tends to be single-valued, the more genuine values it neglects or suppresses'. (Williams 1980)

Notes

1. 'Where a conventional Western composition will seem to unfold as a thread through time, Messiaen's discontinuous music rather provides an environment within which time itself can be observed, 'coloured', as he would say, by rhythm; time suspended, in his slow movements, or time racing forwards, in his scherzos and dances, or, most frequently, time changing its rhythmic colour from moment to moment. Instead of affirming the orderly flow of everyday existence, this is music which acknowledges only two essences: the instantaneous and the eternal.' (Griffiths 2006)

References

Berlin, I. (1969) Two concepts of liberty, in *Four Essays on Liberty* Oxford, Oxford University Press.

Broome, J. (2001) Are intentions reasons? And how should we cope with incommensurable values? In Christopher Morris and Arthur Ripstein (eds) *Practical Rationality and Preference: Essays for David Gauthier* (Cambridge University Press, Cambridge) pp 98- 120

Chang, R.(1997) Introduction, in Ruth Chang (ed) *Incommensurability, Incomparability and Practical Reason* Cambridge Mass, Harvard University Press

Chang, R. (2002) The Possibility of Parity *Ethics* 112 4 pp 659-688

Cronbach, L. (1960) *Essentials of Psychological Testing* New York, Harper and Row

Crowder, G. (1998) John Gray's Pluralist Critique of Liberalism *Journal of Applied Philosophy* 15 3 pp 287-298

Dancy, J. (2000) *The Particularist's Progress* in B. Hooker and M. Little (Eds) *Moral Particularism* Clarendon Press, Oxford.

Davis, A. (1998) *The Limits of Educational Assessment* Blackwell, Oxford

Griffiths, P. (2006) Messiaen, Grove Music Online ed. L. Macy (Accessed 7.1.06) http://www.grovemusic.com

Harlen, W. (2004) A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In *Research Evidence in Education Library (REEL).* London: EPPE-Centre, Social Science Research Unit, Institute of Education.

Isaacston, S. (1999) Instructionally relevant writing assessment *Reading and Writing Quarterly: Overcoming Learning Difficulties* 15 29-48

Kant, I. (1928) *Critique of Judgment* (Oxford, Oxford University Press)

Lukes, S. (1997) Comparing the Incomparable: Trade-offs and Sacrifices in Ruth Chang (ed) *Incommensurability, Incomparability and Practical Reason* Cambridge Mass, Harvard University Press

Nystrom 2004) Reliability of Educational Assessments: The case of classification accuracy *Scandinavian Journal of Educational Research* Vol. 48, No. 4, September 2004

Plaw, A. (2004) Why Monist Critiques feed Value Pluralism: Ronald Dworkin's critique of Isaiah Berlin *Social Theory and Practice* 30 1 pp 105-126

QCA (2006) at http://www.qca.org.uk/downloads/3815_ks2_en_newworld.pdf

Scruton, R. (1997) *The Aesthetics of Music* Oxford Oxford University Press

Shaper-Landau, R. (1994) Ethical Disagreement, Ethical Objectivism and Moral Indeterminacy *Philosophy and Phenomenological Research* 54, 2

Wiliam, D. (2003) National curriculum assessment: how to make it better *Research Papers in Education* 18 2 pp 129-136

Williams, B. (1980) Introduction to Isaiah Berlin, in Henry Hardy (ed) *Concepts and Categories* Oxford, Oxford University Press, p xvii

Wittgenstein, L. (1958) *Philosophical Investigations* Oxford, Blackwell

Wollheim, R. (1987) *Painting as an Art* (London, Thames and Hudson)