Assessment in Education
Vol. 14, No. 2, July 2007, pp. 149–170

Clarifying the purposes of educational assessment

Paul E. Newton*
Qualifications and Curriculum Authority, UK

This article concerns the importance of clarity in thinking and talking about certain core concepts of educational assessment. It begins by identifying three quite distinct interpretations of the term 'assessment purpose'. It continues by challenging the supposed distinction between 'formative' and 'summative'—arguing that the latter only applies to a kind of assessment result while the former only applies to a kind of use of assessment results. It ends by illustrating the wide range of uses to which assessment results might be put and stresses the importance of not concealing important distinctions by locating multiple discrete purposes within a small number of misleading categories.

Introduction

When considering optimal design characteristics for future assessment systems it is necessary to bear in mind the underlying purposes of those systems. The fact that a system which is fit for one purpose will not necessarily be fit for all purposes is a fundamental consideration when evaluating the legitimacy of proposals. It is one of the most important messages for policy-makers to understand. Unfortunately, it can be quite hard to convey the significance of this message; and part of the reason for this is a lack of clarity within the professional discourse of educational assessment. During recent discussions over the future of educational assessment in England, two major obstacles to effective communication became apparent to me:

(1) the term 'assessment purpose' can be interpreted in a variety of different ways
(2) the uses to which assessment results are put are often categorized misleadingly.

The fact that the term 'assessment purpose' can be interpreted in a variety of different ways is, perhaps, the most basic point for an assessment professional to appreciate, to ensure that her advice is not misunderstood. At least three can be identified.

*Head of Assessment Research, Regulation and Standards Division, Qualifications and Curriculum Authority, 83 Piccadilly, London, W1J 8QA, United Kingdom. Email: NewtonP@qca.org.uk

ISSN 0969-594X (print)/ISSN 1465-329X (online)/07/020149–22

DOI: 10.1080/09695940701478321

First, the *judgement* level—which concerns the technical aim of an assessment event (e.g., the purpose is to derive a standards-referenced judgement, expressed as a grade on a range from A to E). In England, this kind of usage is common in official documents, for example:

> The purpose of the [teacher] assessment shall be to determine the level of attainment achieved by the pupil in each [Attainment Target] specified for each core subject which applies to him… (HMSO, 1995, 4.2)

> The purposes of the tests and tasks are to: […] produce, for each child, a single test level for each of English, mathematics and science. (SCAA/DfE, 1994, p. 5)

Second, the *decision* level—which concerns the use of an assessment judgement, the decision, action or process which it enables (e.g., the purpose is to support a selection decision for entry to higher education). This is taken to be the most significant usage of the term 'assessment purpose' since it seems to be the level that is most frequently associated with it in the technical literature. For this reason, and for the sake of clarity, if the term 'purpose' is to be retained as a feature of assessment discourse—as it will be in the remainder of this paper—it ought to be restricted to the decision level. Subsequent sections will illustrate a range of purposes that occupy discourse at this level. Note that the purposes for which assessment judgements are used—by whomsoever happens to use them—need not be as intended, or as anticipated, by the system designer (a point that will be returned to at the end of this paper).

Third, the *impact* level—which concerns the intended impacts of running an assessment system (e.g., the purposes are to ensure that students remain motivated, and that all students learn a common core for each subject). These are impacts specifically attributable to the design of the assessment system, *per se*, rather than to features of the broader educational programme or system within which it operates. Again, note that the impacts of system operation need not be as intended, or as anticipated, by the system designer (although, strictly speaking, the use of 'purpose' at the impact level would apply only to intended impacts).

The important point to note is that each of the three discrete meanings hold distinct implications for the design of an assessment system. This means that each of them needs to be addressed separately, for example:

- to derive standards-referenced judgements, performance descriptions and exemplar materials need to be developed and shared
- to support selection decisions, assessment results need to have high reliability across the range of performance levels
- to ensure that students remain motivated, the assessment might be administered on a unit-by-unit basis with opportunity for re-taking; to ensure that all students learn a common core for each subject, the assessment might be aligned to a national curriculum

Where the three discrete meanings are not distinguished clearly, their distinct implications for assessment design may become obscured. In this situation, policy debate is likely to be unfocused and system design is likely to proceed ineffectively.

The remainder of this paper will address, in turn, the two major obstacles highlighted above, in an attempt to offer advice on how to sharpen the professional discourse of educational assessment. It will illustrate further the distinction between assessment judgements and the uses to which they are put, beginning with an analysis of the supposed distinction between formative and summative purposes.

The importance of clarifying levels of discourse

From one perspective, the distinction that has been drawn between formative and summative has been extraordinarily effective in carving an identity for assessment undertaken by students and teachers in the service of learning (in countries where assessment has traditionally been viewed as largely external to the teaching and learning process).

From another perspective, though, the distinction has been extraordinarily ineffective in carving the nature of assessment at its joints, i.e., the distinction has failed to advance the theory of assessment and might even have set it back somewhat. Why? Because, upon closer inspection, the distinction can be shown to be spurious. Moreover, the confusion that it has engendered has not been benign and, to some extent, has actually hindered the development of sound assessment practice. The following subsections will explain why, in an attempt to help begin to repair some of the damage.

The first thing that we need to consider is how the distinction entered the discourse of educational assessment and how its use has evolved. This will be illustrated through reference to the work of some of the leading theorists in the field.

How has the distinction been characterized?

The contrast between formative and summative assessment was first communicated to a wide audience during the early 1970s by Bloom *et al.* (1971) in their *Handbook of formative and summative evaluation of student learning*. They attributed the origin of the distinction to Scriven (1967). In fact, Scriven had originally drawn the distinction to highlight different approaches to programme evaluation. As he explained the terms (more recently):

> Formative evaluation… is typically conducted during the development or improvement of a program… and it is conducted, often more than once, for the in-house staff of the program with the intent to improve. (Scriven, 1991, pp. 168–169)

> Summative evaluation of a program… is conducted after completion of the program… and for the benefit of some external audience or decision-maker… (Scriven, 1991, p. 340)

Although Bloom *et al.* (1971) used the term 'evaluation' in the title of their *Handbook*, they were actually focusing primarily upon the process of student assessment, rather than upon the process of programme evaluation (which, incidentally, often involves student assessment).

Benjamin Bloom, Thomas Hastings and George Madaus. Bloom *et al.* (1971) began their *Handbook* by noting how, throughout the world and for many centuries, education had emphasized a selective function; that is, it had been concerned primarily with identifying the students to be dropped at each major stage. The authors wished, instead, to promote a view of education in which its primary function was to develop the individual. Reflecting this tension, they explained the distinction between formative and summative as follows:

> The main purpose of formative observations… is to determine the degree of mastery of a given learning task and to pinpoint the part of the task not mastered. … The purpose is not to grade or certify the learner; it is to help both the learner and the teacher focus upon the particular learning necessary for movement towards mastery. On the other hand, summative evaluation is directed toward a much more general assessment of the degree to which the larger outcomes have been attained over the entire course or some substantial part of it. (Bloom *et al.*, 1971, p. 61)

They identified three characteristics according to which formative and summative could be distinguished, having to do with:

(1) *purpose*—expected uses to which the outcomes will be put (formative assessment focuses on helping the learner learn while summative assessment focuses on grading or certification)
(2) portion of course covered—*timing* (formative assessment tends to be more frequent, focusing on smaller units of instruction and occurring during a course rather than at the end)
(3) level of *generalization* sought by items in the instrument used to collect data for the assessment (formative assessment focuses on testing for narrow components of proficiency while summative assessment focuses on testing for broad abilities).

And they noted that:

> We have chosen the term 'summative evaluation' to indicate the type of evaluation used at the end of a term, course, or program for purposes of grading, certification, evaluation of progress, or research on the effectiveness of a curriculum, course of study, or educational plan. … Perhaps the essential characteristic of summative evaluation is that a judgment is made about the student, teacher, or curriculum with regard to the effectiveness of learning or instruction, after the learning or instruction has taken place. …
>
> Formative evaluation is for us the use of systematic evaluation in the process of curriculum construction, teaching, and learning for the purpose of improving any of these three processes. (Bloom *et al.*, 1971, p. 117)

Although the *Handbook* was thoroughly revised a decade later (Bloom *et al.*, 1981) the above quotations and differentiating characteristics remained almost entirely unchanged. Perhaps the most useful clarification was to italicize 'after' in the quotation from page 117 (above) which helped to indicate that the essential characteristic of summative assessment was that it occurred *after* the learning had taken place, in contrast to formative assessment which took place *during* learning to improve it. This was in keeping with the defining characteristic of formative evaluation, as discussed by Scriven (1991).

Two points are interesting here: first, that Bloom *et al.* (1971) identified a range of distinguishing characteristics, including purpose, timing and generalization; second that they elevated timing as the essential characteristic of summative assessment. This is of interest because the early work is often read to elevate purpose as the essence of the distinction between formative and summative. In fact, in commenting upon the work of Bloom *et al.* (1971), Black and Wiliam (2003) noted that: 'From their earliest use it was clear that the terms "formative" and "summative" applied not to the assessments themselves, but to the functions they served' (p. 623).

Admittedly, it is hard to read Bloom *et al.* (1971) without getting the impression that their key dimension of interest *is* function/purpose/use rather than timing or generalization. On the other hand, they actually proposed that level of generalization was the key! 'Perhaps level of generalization is the factor which differentiates summative from formative evaluation most sharply' (Bloom *et al.*, 1971, p. 62).

So, according to Bloom *et al.* (1971), the essence of summative assessment was timing, while the essence of the distinction between summative and formative was level of generalization. Perhaps they felt that purpose was less central because they recognized a variety of *different* purposes for which results from summative assessments could be put? In their chapter on summative assessment, they cited purposes such as: to give grades; to certify competence; to provide feedback to students; to predict later success; (etc.).

Royce Sadler. Writing towards the end of the 1980s, Sadler (1989) was keen to develop the theory of formative assessment by exploring conditions for effective feedback. He distinguished formative and summative as follows:

> Formative assessment is concerned with how judgements about the quality of student responses (performances, pieces, or works) can be used to shape and improve the student's competence by short-circuiting the randomness and inefficiency of trial-and-error learning. Summative contrasts with formative assessment in that it is concerned with summing up or summarizing the achievement status of a student, and is geared towards reporting at the end of a course of study especially for purposes of certification. It is essentially passive and does not normally have immediate impact on learning, although it often influences decisions which may have profound educational and personal consequences for the student. The primary distinction between formative and summative assessment relates to purpose and effect, not to timing. (p. 120)

So, for Sadler, the distinguishing feature of formative assessment was the use to which results are put. For summative assessment, the case is slightly less clear, as he emphasized both the nature of the event itself (summing up) and an unspecified set of purposes (certification). Interestingly, he made a point of emphasizing that the distinction did not turn on timing but on purpose (and effect).

Paul Black. Paul Black has contributed a great deal to debates on formative and summative assessment. Comments from his book *Testing: friend or foe*? make a useful

point of reference. He distinguished three 'main types of purpose' (Black, 1998, p. 24), which he discussed under three headings:

(1) support of learning
(2) certification, progress and transfer
(3) accountability

Towards the end of his chapter on purposes, he defined the three main types more explicitly as:

(1) formative, to aid learning
(2) summative for review, transfer and certification, and
(3) summative for accountability to the public

Black concluded:

> Some have laid stress on the differences between the formative and summative purposes, and have argued that the assessment instruments and procedures needed for the one are so different from those for the other that neither can flourish without clear separation. On the other side, it can be argued [as Black does] that the two functions are two ends of the same spectrum and that there is no sharp difference, and that if the two functions are separated, then teachers' assessment work will be devalued. (Black, 1998, p. 34)

Here, Black talks straightforwardly as though the distinction between formative and summative is a matter of purpose/function. However, he also suggests that the purposes are at two ends of the same 'spectrum of practice in school-based assessment rather than two isolated and completely different functions' (p. 35). The implication seems to be that features of the assessment event—presumably including timing and generalization—are not characteristics according to which formative and summative can necessarily be distinguished, since the same classroom assessments can be used for both purposes. This proposal was considered further by Harlen and James (1997).

Wynne Harlen and Mary James. Harlen and James (1997) expressed concern that the important distinction between formative and summative was becoming clouded, so they set about some cloud-busting. They argued that the 'conflation of summative and formative purposes' in 'official documents' has smothered the essential differences between them and hindered the development of 'genuine formative assessment' (quotations from p. 365).

One particular official document which they singled out for attention was the report for UK Government of the Task Group on Assessment and Testing—a group headed by Paul Black—which set out a framework for assessing the new national curriculum (DES/WO, 1988). The report brought the terms formative, summative, diagnostic and evaluative into common parlance in the UK. According to Harlen and James, though, it went wrong by failing to distinguish *in kind* between formative and summative purposes (since it proposed that summative judgements could be derived from an aggregation of judgements made for formative purposes).

Harlen and James (1997) 'attempted to distinguish formative from summative assessment' (p. 372) by listing contrasting characteristics, for example, summative assessment needs to prioritize reliability, while formative assessment needs to prioritize validity and usefulness; formative assessment treats inconsistent evidence as informative, while summative assessment treats inconsistent evidence as error; etc.. They attempted to clarify the relationship between formative and summative as follows: it is not possible to aggregate assessment judgements made for formative purposes to derive summative judgements; however, it is possible to use the same corpus of assessment evidence to derive both formative and summative judgements, as long as that evidence is interpreted differently for each purpose.

Harlen (2005) subsequently developed this argument, and further clarified the distinction between formative and summative as follows:

> The two main purposes of assessment discussed in this article are for helping learning and for summarizing learning. It is sometimes difficult to avoid referring to these as if they were different forms or types of assessment. They are not. They are discussed separately only because they have different purposes; indeed the same information, gathered in the same way, would be called formative if it were used to help learning and teaching, or summative if it were not so utilized but only employed for recording and reporting. While there is a single clear use if assessment is to serve a formative purpose, in the case of summative assessment there are various ways in which the information about student achievement at a certain time is used. (Harlen, 2005, p. 208)

This definition makes some useful points for a summary of the story so far:

(1) the distinction between formative and summative is frequently prioritized in theoretical discussions (rather than, say, diagnostic versus summative, or formative versus evaluative); that is, the formative versus summative distinction is assumed somehow to be fundamental
(2) people often seem to think that the distinction turns on the nature of the assessment event itself
(3) it now seems to be generally accepted—at least within academic circles—that the distinction turns on the nature of the assessment purpose, i.e., the use to which assessment judgements will be put
(4) summative assessment (but not formative assessment) is associated with a variety of different purposes.

Over the past three decades or so, the distinction between formative and summative has proven extremely problematic to define with any sense of precision. Perhaps we ought to question why (as has Taras, 2005, although she developed a somewhat different explanation from the one presented below)? It begs the question of whether there is actually a meaningful distinction to be drawn at all.

Is the distinction actually meaningful?

Since the earliest days, most commentators have assumed that there *is* a meaningful distinction to be drawn between summative and formative. At least to my mind,

though, no one has yet managed to nail a definition. I believe that there is a simple reason for this: the term 'summative' can only meaningfully characterize a type of assessment judgement (i.e. it operates at the judgement level of discourse), while the term 'formative' can only meaningfully characterize a type of use to which assessment judgements are put (i.e. it operates at the decision level of discourse). The terms belong to qualitatively different categories; to attempt to identify characteristics that distinguish them—within a single category—is to make a category error.

There is no summative purpose. In drawing a distinction between summative and formative, it has often seemed that scholars are comparing assessment purposes at the decision level, since the distinction has typically been drawn to foreground the formative purpose. Yet, when referring to the alleged summative purpose, the same researchers tend simply to use the term as a catch-all expression for categorizing any of a variety of different purposes which are predicated on the use of individual summative assessment judgements. The term 'summative' basically evokes the nature of the assessment judgement typically used to support these purposes: a summing up. What it does not do is to highlight anything more significant about them.

This is remarkably obvious in even the most prestigious of texts. For example, the influential North American report of the Committee on the Foundations of Assessment (see Pellegrino *et al.*, 2001, pp. 37–42) described its three broad purposes of assessment as:

(1) assessment to assist learning—formative assessment
(2) assessment of individual student achievement—summative assessment
(3) assessment to evaluate programmes

Note how the second category employs 'of' rather than 'to', i.e., even the name makes no reference to how the assessment result will be used. Furthermore, although the distinction between summative and evaluative seems to tap into a conceptually important difference between purposes, it ultimately appears to turn on whether we are talking about *discrete* individual summative judgements or *aggregated* individual summative judgements.

The same confusion is more subtly disguised in the most recent update of Bloom's taxonomy of educational objectives:

> Formative assessment is concerned with gathering information about learning as learning is taking place, so that 'in-flight' instructional modifications may be made to improve the quality or amount of learning. In contrast, summative assessment is concerned with gathering information about learning after the learning should have occurred, usually for the purpose of assigning a grade. (Anderson & Krathwohl, 2001, pp. 101–102)

It is not uncommon for grading to be cited as a specific assessment purpose (at the decision level). This is unhelpful, though, since a grade is a type of assessment judgement and 'grading' is the process of assigning a grade. Grading is not a discrete *use* to which an assessment judgement can be put; it is a term that has application only

at the *judgement* level. In response, it might be argued that the term 'grading' often denotes both a process *and* a purpose. However, if so, then exactly what purpose or, perhaps, which purposes? Any purpose that is based on the use of a grade? Any purpose that is not formative?[1]

The confusion between summative and formative is even evident in Wiliam and Black's (1996) attempt to define the distinction formally:

> An assessment is defined as serving a formative function when it elicits evidence that yields construct-referenced interpretations that form the basis for successful action in improving performance, whereas summative functions prioritize the consistency of meanings across contexts and individuals. (p. 537)

Here, once again, the use of the assessment judgement appears to be central to the definition of the formative function but is not referred to at all in the definition of the alleged summative function.

The supposed distinction between formative and summative is not grounded in the use to which assessment judgements are put. Why? Simply because there is no meaningful distinction to be drawn. The rhetoric *appears* to distinguish between two conceptually distinct types of use to which results can be put; in fact, it simply foregrounds one particular type, the formative use.

There are no formative judgements. If you get to be a summative purpose simply by being based on summative judgements, then the definition is hollow: the idea of a summative purpose is conceptually redundant. Likewise, what if a 'summing up' judgement is used for a formative purpose? Would it simultaneously be both a summative judgement and a formative judgement? Again, if you get to be a formative judgement simply by being used for a formative purpose, then the definition is hollow: the idea of a formative judgement is conceptually redundant.

Obviously, formative uses do require assessment judgements. The point is simply that we risk confusing ourselves, and others, when we call them 'formative judgements' (since there is nothing about the judgement, *per se*, that is formative). It is an assessment judgement used for a formative purpose, not a formative judgement.

To avoid getting ourselves confused, and to avoid confusing others, we need to use the language of assessment with greater precision. We may talk of a formative purpose, to indicate the use to which a result is put, but we ought not to talk of a summative purpose. Likewise, we may talk of a summative judgement, but we ought not to talk of a formative one. To my mind, the following distinctions help to characterize the effective use of this terminology.

Characterizing judgements and uses

As noted earlier, it is important to distinguish between the aim of an assessment event—which concerns translating an observation of performance into a particular kind of assessment judgement—and the use to which that judgement is put.

Judgements. The term 'judgement' is used here, in a general sense, to refer to the overall *outcome* from an assessment event: the *representation* of an assessed person's knowledge, skill or understanding, i.e., the representation of their competence. It would therefore extend to:

- a realization by a student—based upon a discussion with her friend—that her friend had a misconception that the earth is flat and has an absolute frame of reference for up and down (the representation forms within the mind of the first student and might be externalized verbally in the form of an exclamation)
- an evaluation by a teacher—based upon a portfolio of work produced over an extended period—that a student had attained a certain level of attainment in relation to a set of performance descriptions (the representation forms within the mind of the teacher and might be externalized in the form of a grade, level, or such like)
- a pronouncement from a testing company—based upon responses to a multiple choice test—that a student had performed substantially better than her peers (the representation forms largely through the operation of automated procedures and might be externalized in the form of a standardized score, percentile rank, or such like).

For the sake of illustration, in subsequent sections, the term 'result' will tend to be used to refer to the outcome from an assessment event.

Judgements of educational attainment (based on evidence from performance) range along a continuum from summative to descriptive, where

- a *summative* judgement is characterized by *appraisal*—a decision concerning the value or quality of the (apparent) competence
- *descriptive* judgements are characterized by *analysis*—a reflection on the nature of the (apparent) competence.

At the summative end of the continuum are judgements that (purely) summarize the value of an educational attainment in essentially *quantitative* terms, for example:

- *self-referenced* judgements (e.g., attained better this time than before)
- *norm-referenced* judgements (e.g., attained at a higher level than n% of students).

At the descriptive end of the continuum are judgements that (purely) describe the nature of an educational attainment in essentially *qualitative* terms, for example:

- *concept-referenced* judgements (e.g., understands in this respect but not in that respect)
- *performance-referenced* judgements (e.g., succeeds in this respect but not in that respect)

In between these extremes we find judgements that combine elements of summary and description, for example:

- *criterion-referenced* judgements (e.g., can do x, cannot do y or z)
- *standards-referenced* judgements (e.g., likely to be able to do x, y and z).

In this last example, standards-referenced judgements would be closer to the summative end of the continuum, since the element of description is inevitably less faithful to the actual attainment (than for criterion-referenced judgements), and since the judgements tend to be more closely linked to quantitatively expressed levels of attainment.

Uses. The uses to which assessment judgements/outcomes/results might be put include: formative purposes; diagnostic purposes; placement purposes; system monitoring purposes; etc. (this list will be extended shortly).

The important point here is that there are *many* distinct assessment purposes. It is not the case that there are basically just two or three broad categories of purpose—formative, summative, evaluative—and a whole range of sub-purposes (again, this point will be developed shortly).

Distinctions. These distinctions are useful for helping to make clear why there is no such thing as, for instance, a formative judgement. Whatever the nature of a judgement there would be nothing formative happening unless the judgement was used in an attempt to improve learning. Moreover, both summative judgements and descriptive judgements can be used for formative purposes (although descriptive judgements are often more useful for formative purposes, when it is necessary to identify not simply that a person has failed, but why and consequently what to do about it). The term 'formative' resides at the decision level of discourse on assessment purposes (and, consequently, it is helpful to think of it as a genuine 'purpose').

Similarly, whatever the purpose to which (say) a norm-referenced result is put, it is still a summative judgement, since it involves a 'summing up' appraisal of performance. This is why there is no tension in using norm-referenced (summative) test results for diagnostic purposes, even though diagnostic purposes are often considered quite similar to formative ones. The term 'summative' resides at the judgement level of discourse on assessment purposes (and, consequently, it is not helpful to think of it as a genuine 'purpose').

But what was wrong with being a little confused?

Right from the outset, the point of distinguishing formative from summative was to foreground the former. It seems fair to conclude that this has been achieved on an international scale. In many countries, 'formative assessment' is now part of mainstream parlance within a range of educational communities. The problem that assessment professionals now tend to face is how to ensure that people who work in education attach correct meanings to the term, to ensure that when they think they

are doing 'formative assessment' they actually are using assessment to achieve a formative purpose.

This problem has been exacerbated by the lack of precision of our professional terminology. Given that we—as assessment professionals—are confused over how to define formative and summative, it is not surprising that those beyond the academy are even more confused and that important messages are being lost.

Wiliam (2004), for example, observed how 'formative' is often mistakenly viewed as relating to a kind of assessment event rather than a kind of assessment purpose:

> In the United States, the term 'formative assessment' is often used to describe assessments that are used to provide information on the likely performance of students on state-mandated tests—a usage that might better be described as 'early warning summative'. In other contexts it is used to describe any feedback given to students, no matter what use is made of it, such as telling students which items they got correct and incorrect (sometimes called 'knowledge of results'). (Wiliam, 2004, p. 4)

Exactly this concern has more recently been echoed by Arter and Stiggins (2005).

As it stands, we should not be surprised when people mistakenly do assume that 'formative' refers to a kind of assessment judegment, rather than to a kind of use to which assessment judgements are put. After all, we routinely characterize formative assessment by contrasting it with summative assessment and 'summative' clearly does refer to a type of assessment judgement, one which involves 'summing up' (the clue's in the name). Greater precision in the discourse of educational assessment should help practitioners and policy-makers to understand better the key principles at the heart of our proposals.

The importance of not categorizing assessment purposes misleadingly

As an assessment professional who works closely with policy-makers, my concerns go beyond general confusion over the precise meaning of certain terms. I worry that spurious distinctions between categories of purpose may negatively impact upon policy-making. By lumping together all purposes that are not formative and calling them 'summative' or 'evaluative', we risk misleading policy-makers into thinking that these purposes really do share something important in common. In particular, we risk giving the impression that a result which is fit for one 'summative purpose' is fit for 'summative purposes' more generally. It is not hard to see how statements of the following kind might mislead the unwary reader toward this kind of inference: '… can this rich but sometimes inconsistent information be used for summative assessment purposes as well as for formative assessment, for which it is so well suited?' (Harlen, 2005, p. 218)

As illustrated earlier, there has been ongoing controversy in the UK over whether assessment evidence elicited for formative purposes can also be used for (so-called) summative ones. Unfortunately, this seems to have diverted attention from the more general and far more fundamental debate over the extent to which evidence elicited for *any* particular purpose can legitimately be used for *any* other. As the number of purposes for which assessment judgements are being used increases, so too does the

importance of reinvigorating this debate. This will mean emphasizing a broader range of assessment purposes and the different assessment design principles that enable them to be achieved. The following discussion focuses specifically upon the decision level of discourse on purposes; although, of course, implications for assessment design will also follow from goals at the impact level too.

An unabridged story of assessment purposes

We give the wrong message when we try to simplify assessment purposes by allocating them to a small number of categories (such as formative, summative and evaluative): we imply that the sub-purposes within those categories are importantly alike. This risks the impression that results which are fit for one sub-purpose within a category will be fit for the other sub-purposes as well. This is a very dangerous impression to present, for it is contrary to the impression that ought to be given to policy-makers, to ensure that wise decisions are made.

It is not simply the case that results which are fit for one purpose (e.g., selection) may not be fit for another (e.g., placement), it is even the case that results which are fit for one instance of a particular purpose (e.g., short-term system monitoring) may not be fit for another (e.g., long-term system monitoring). We need to convey the complexities of assessment design and fitness-for-purpose; we should not allow those complexities to be over-simplified. The following discussion aims to draw clearer distinctions between assessment purposes and to emphasize that the different uses to which results are put often require substantially different assessment processes, even when those uses appear to be quite similar.

Categories of purpose. There are very many purposes for which educational assessment judgements might be used, of which the following is merely a selection. They are explained more fully in Table 1. Whereas, in the past, we have tended to want to classify them into a smaller number of categories, it is probably more constructive to consider each a category in its own right.[2]

(1) *Social evaluation* uses
(2) *Formative* uses
(3) *Student monitoring* uses
(4) *Transfer* uses
(5) *Placement* uses
(6) *Diagnosis* uses
(7) *Guidance* uses
(8) *Qualification* uses
(9) *Selection* uses
(10) *Licensing* uses
(11) *School choice* uses
(12) *Institution monitoring* uses
(13) *Resource allocation* uses

(14) *Organizational intervention* uses
(15) *Programme evaluation* uses
(16) *System monitoring* uses
(17) *Comparability* uses
(18) *National accounting* uses

Some might be surprised by the omission of 'certification' from this list of purposes. However, unlike the others on the list, its use often fails to implicate a specific decision, action or process. Sometimes 'certification' is used as loosely as 'grading' and fails to refer to a specific purpose for the reason described earlier. Other times it is used somewhat more precisely to indicate—and to testify to—the attainment of a general competence or profile of competencies. This is when it begins to hint at a specific use, although none is made explicit; even so, it is not clear that the implicated use(s) would be distinct from others on the list. In fact, in many ways, certification might be read as an implicit qualification purpose: certification indicates that a student has satisfied the demands of a course of instruction; but the course is likely to have been designed such that a student who satisfies those demands is, thereby, at least minimally qualified, if only for an aspect of everyday citizenship. Notably, in the USA *Standards* document (AERA/APA/NCME, 1999), the term 'certification' is used precisely, but in an occupational context and in relation to the assessment of higher professional competencies.

The fact that it is not uncommon to hear 'grading' and 'certification' discussed loosely (particularly in the school testing context)—as though they were discrete purposes in their own right—is worrying. It suggests either an ambivalence toward the uses of assessment results, or the false assumption that putting them to any use whatsoever will be unproblematic.

The importance of distinguishing between purposes. Table 1 does not aim to present an exhaustive list of purposes (see Newton, forthcoming, for a slightly longer list.) Furthermore, even within each of the purposes identified, it is possible to identify subtly different sub-purposes, each with different assessment design implications.

Differences within categories. The system monitoring purpose, for example, could be further subdivided into monitoring short-term, medium-term or long-term trends. A system that was intended to monitor long-term trends in attainment might need to be designed quite differently from one intended to monitor short-term trends. In particular, the system for monitoring long-term trends would need to be designed to accommodate the fact that certain questions and topics will change in educational significance over time. This accommodation might operate at the level of inferences drawn (where exactly the same test is administered over time), or it might operate at the level of item selection and equating (where items are replaced over time).

Table 1. An illustration of the variety of categories of educational assessment purpose (decision level)

Category of purpose/use	Typical user group	Decision, action or process supported by assessment results	Illustrative comment on characteristics of assessment results (assessment design implications)
Social evaluation	Students, peers, parents	Individual results are used to evaluate the social value of personal educational achievements [self evaluation or evaluation by others]	Entirely norm-referenced reporting systems can be problematic, since they tend to devalue achievement for the lowest attaining students
Formative	Students, teachers, parents	Individual (or aggregated) results are used to identify student (or group) learning needs, to direct subsequent teaching and learning	Frequent, fine-grained analyses of attainment in small sub-domains will typically be required to guide interventions for individual students, and reporting will often occur through verbal feedback
Student monitoring	Students, teachers, parents	Individual results are used to decide whether students are making sufficient progress in attainment over time [sometimes linked to performance targets for students]	Results need to be particularly reliable, across the full range, given the threat of error at both the earlier and the later assessment points
Transfer	Teachers	Individual results are used to tailor educational provision to the general educational needs of students who transfer to new classes/schools	Where transfer to a new teacher occurs, results need to be expressed in a 'language' which can easily be shared, i.e., detailed, with common criteria and procedures for grading (see Black, 1998, p.28)
Placement	Teachers	Individual results are used to place students in teaching groups, or educational programmes, that will be most suitable for them [the decision over which tier to enter a student for, within a multi-tiered examination, might be viewed as a special case of placement]	Reasonable reliability is required, since the decisions have potentially high stakes for students, although—where placement decisions are revisited fairly frequently—reliability may be lower since misplacement can be rectified
Diagnosis	Educational psychologists	Individual results are used to diagnose learning difficulties	Diagnostic assessment can lead to radical intervention (and potential stigmatization for those labelled with learning difficulties) which means that high validity and reliability are very important

Table 1. *(Continued)*.

Category of purpose/use	Typical user group	Decision, action or process supported by assessment results	Illustrative comment on characteristics of assessment results (assessment design implications)
Guidance	Students, parents, career guidance counsellors	Individual results are used to guide future educational and employment decisions	Where, for example, decisions concern which of a range of subject areas to pursue, results need to represent relative standing effectively (so that students have a good indication of which subjects they are best at), i.e., the results need to be comparable across subjects
Qualification	Admissions tutors/officers, employers	Individual results are used to judge whether a person is sufficiently qualified for a job, course of instruction or role in life, i.e., whether or not they are equipped to succeed in it	Results may best be represented as pass/fail, with high reliability required around that cut-score (but not necessarily at other points on the mark scale)
Selection	Admissions tutors/officers, employers	Individual results are used to predict which applicants—all of whom might, in principle, be sufficiently qualified—will be most successful in a job or course of instruction (and, consequently, provide a basis for choosing between them)	Results need to discriminate reliably across the full range, particularly where different courses/jobs tend to select between students from different ability ranges
Licensing	Professionals, those who are practised upon	A positive individual result provides a licence to practice, and some guarantee for those who are practised upon	Compensatory aggregation is not to be recommended, where practitioners need to be minimally competent across all elements of a domain (there is little security in knowing that a pilot is excellent at taking off if she is also very poor at landing)
School choice	Parents, students	Aggregated results are used to identify the most desirable school for a child to attend	Aggregated results must be sufficiently reliable to detect real differences in effectiveness between all schools (and must be accompanied by reasonable estimates of 'chance' variability which are sufficiently straightforward for all users to understand)

Table 1. *(Continued)*.

Category of purpose/use	Typical user group	Decision, action or process supported by assessment results	Illustrative comment on characteristics of assessment results (assessment design implications)
Institution monitoring	Teachers, school managers, local authorities	Aggregated results are used to decide whether institutional standards are rising or falling over time, e.g., at the level of classes or schools [sometimes linked to performance targets for teachers and/or managers, and on the basis of which rewards or penalties might be allocated]	Where, for example, results are used to allocate rewards/ penalties to teachers in different departments of a school, there should be comparability of assessment standards across subjects (i.e., common evaluation criteria)
Resource allocation	School managers, local authorities	Aggregated results are used to identify needs and, consequently, provide a basis for allocating resources	Results should be sufficiently fine-grained to enable the identification of specific weaknesses
Organizational intervention	Inspectors, local authorities	Aggregated results are used to identify failure and, consequently, provide a basis for justifying intervention	Aggregated results must be sufficiently reliable to detect real differences in effectiveness between the lowest attaining schools (and must be contextualized by data on other factors that might affect the interpretation of results)
Programme evaluation	Researchers, policy-makers	Aggregated results are used to evaluate the success of educational programmes or initiatives, nationally or locally	Results should accurately reflect the construct at the heart of the evaluation
System monitoring	All stakeholders	Aggregated results are used to decide whether system standards are rising or falling over time, e.g., at the local or national level [sometimes linked to performance targets for civil servants and/or politicians, and on the basis of which rewards or penalties might be allocated]	Results must demonstrate long-, medium- or short-term comparability of assessment standards over time (depending on the kind of monitoring undertaken)
Comparability	Exam board statisticians	Aggregated results from earlier assessments (e.g., exams at age 16) are used to guide decisions on comparability of standards for later assessments (e.g., exams at age 18)	Where large numbers of students tend to be assessed by different exam boards—for both earlier and later assessments—there must be sufficient comparability of standards between boards for the earlier assessment
National accounting	National accountants	Aggregated results are used to 'quality adjust' education output indicators	Trends in examination performances over time must unambiguously reflect changes in the quality of education

Equally, though, the system monitoring purpose could be subdivided differently; for example, into systems that are oriented more towards monitoring overall educational standards (to determine whether or not education policy is succeeding) versus systems that are oriented more towards curriculum evaluation and planning. To monitor overall educational standards it is important that subject-level aggregate measures can be computed, which might well necessitate the use of complex statistical modelling; which, in turn, might limit the breadth of assessment formats that can be employed (to those that generate results which minimize violation of the assumptions upon which the modelling is based). By way of contrast, for curriculum evaluation and planning, results do not need to be aggregated to subject-level, since performance in the discrete components of a subject will be the primary concern. Radically different design characteristics, such as these, are respective features of the USA's National Assessment of Educational Progress (regarding overall educational standards) and New Zealand's National Education Monitoring Project (regarding curriculum evaluation and planning).

Differences between categories. Within Table 1, there are a number of uses which appear to be quite similar, for example: transfer and placement; selection and qualification; resource allocation and organizational intervention. However, once again, even for such similar purposes, it would be wrong to assume that a system designed to satisfy one would necessarily be optimal for satisfying the other. An example based upon the major 16+ and 18+ examinations in England is of significance here.

Eckstein and Noah (1993) described the (16+) General Certificate of Secondary Education (GCSE) as a 'selecting-in' examination (providing a credential for the majority of students completing secondary schooling) and the (18+) Advanced level (A level) as a 'selecting-out' examination (providing university selectors with a basis for choosing between the most able students). This suggests that the GCSE is focused more on qualification, while the A level is focused more on selection. Interestingly, though, the two assessments have traditionally operated quite similarly, which raises the question of whether either examination is optimally designed for the purpose which it appears primarily to serve.

In fact, raising exactly this kind of query, a high-profile working group has recently criticized the A level for failing as an instrument for selection purposes, since it tends not to discriminate well between the ablest candidates (Working Group on 14-19 Reform, 2004). And the GCSE has also recently been criticized for failing as an instrument for qualification purposes, since even good grades do not guarantee a minimum level of competence in skills deemed essential for future learning and employment (DfES, 2005). Proposals are presently being considered which will make the GCSE more suitable for qualification and the A level more suitable for selection, thus making the two examinations less similar than at present.

The point here is that even ostensibly similar purposes, such as qualification and selection, require different assessment design decisions. A system fit for one of the purposes will not necessarily be fit for the other.

The expanding range of assessment purposes. The sheer range of purposes which can be identified needs to be emphasized, since it seems to be growing all the time. The experimental use of examination results for adjusting national accounts is a good case in point, which began in England toward the end of the 1990s (e.g., Caplan, 1998; Pritchard, 2003). To explore the economic management of the country, government inputs and outputs can be compared to provide an indication of productivity for various functions (e.g., education, health, defence). Implicit in the approach that is being adopted by national accountants at the Office for National Statistics is the idea that increasing the quality of the output itself represents more output (Pritchard, 2003, p. 28). For education, the basic output measure is a unit of volume describing the number of full-time equivalent pupils in education each year. However, based upon an assumption that the quality of education has improved over time, this output measure is quality adjusted:

> There is a key issue with education in that there is significant evidence that educational standards have been rising over a number of years. This is demonstrated by, amongst other indicators, increases in the average point score of pupils taking GCSE exams at aged 16. A number of factors may have led to this increase in standards. However, there is evidence that the quality of teaching is rising. For this reason, it was felt necessary to adjust the educational output series for quality change. There is no single way to do this and certainly no method is infallible. Using information on GCSE point scores and by assuming that changes result from quality increases over the 11 years of compulsory education, it was decided to add a quality factor of 0.25 per cent for schools for each year. (Caplan, 1998, p. 48)

The examination system in England was not designed with the adjustment of national accounts in mind. It is at least questionable whether GCSE results possess the stability of currency that would be required to support this function, even approximately (see Cresswell, 2003).

Conflicting design characteristics

Since different purposes require different assessment design decisions it is important for the system designer, in collaboration with policy-makers, to define the *primary purpose* for which results are intended to be used. Where there is an aspiration for the assessment system to support a number of different purposes (assuming that they are not logically incompatible) an explicit *prioritization* of purposes should be defined. The primary purpose and (to the extent possible) other high priority purposes will determine optimal design characteristics for a system that will enable users to draw sufficiently valid inferences from results.

In making this recommendation, the emphasis here is slightly different from that implied by Pellegrino *et al.* (2001). These authors noted that the more purposes an assessment system is intended to serve, the more each purpose will be compromised by trade-offs during the process of assessment design. Their implication at least seems to be that each purpose would have a similar weight in the battle over assessment design features, such that an assessment system 'designed' to support a large

range of purposes might end up being fit for none. To prioritize purposes in advance—and to orient design features to the highest-weight purpose—represents an alternative proposition (and, arguably, a more profitable approach). Of course, where more than one high-weight purpose is prioritized, trade-offs would still need to be made.

Once a system has been designed with an explicit prioritization of purposes in mind, the operational problem will then become how to ensure that results are not used for inappropriate purposes. This may prove extraordinarily hard to prevent. However, at the very least, I suggest that the system designer has an obligation to identify, for all stakeholders, those purposes for which results are unfit (not simply those for which results are fit). Stakeholders should be deprived of ignorance as an excuse for misuse.

These are the kinds of message that policy-makers need to hear; messages that openly acknowledge, rather than conceal, the true complexity of assessment system design and operation. These messages need to be communicated in ways that steer policy-makers away from inferential errors rather than towards them. And these messages need to include warnings concerning inappropriate practices, as well as advice concerning appropriate ones. On the one hand, this is to argue against collapsing qualitatively different purposes into misleading categories such as formative, summative and evaluative. On the other hand, it is to argue for a greater degree of transparency and openness concerning both the legitimate and the illegitimate uses of assessment results.

Acknowledgements

I am very grateful to Andrew Boyle, Jeremy Tafler and Mike Kane for comments on an earlier version of this paper. None of the views expressed should be taken to represent those of my employer, the Qualifications and Curriculum Authority.

Notes

1. Perhaps more than most assessment terms, the term ‘grading’ will have different connotations in different countries. For example, in the USA, it is often associated with the generation of grade point averages, which are used for purposes such as selection. Even in this context, though, it tends not to denote one specific purpose (and exclude others).
2. A reviewer of the first draft of this paper posed a quite reasonable question: why had it not even mentioned the distinction – in Table 1 – between categories of purpose based upon individual judgements (e.g., selection) and categories based upon aggregated judgements (e.g., system monitoring). My response is straightforward. The motivation behind this paper was to discourage spurious categorization and to focus attention upon levels at which distinctions really matter, i.e., matter for the design of assessment systems and for the intelligent use of assessment results. This means looking beyond simplistic distinctions (e.g., individual versus aggregated) to finer-grained distinctions (e.g., system monitoring versus comparability) and to even finer-grained distinctions still (e.g., long-term versus short-term monitoring).

Notes on contributor

Paul Newton is Head of Assessment Research at the Qualifications and Curriculum Authority, London. His interests lie in the evaluation of large-scale educational assessment systems and in the public understanding of assessment.

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999) *Standards for educational and psychological testing* (Washington, DC, American Educational Research Association).

Anderson, L. W. & Krathwohl, D. R. (2001) *A taxonomy for learning, teaching and assessing: a revision of Bloom's taxonomy of educational objectives* (Complete edn.) (New York, Addison Wesley Longman).

Arter, J. & Stiggins, R. (2005) Formative assessment as assessment FOR learning, *National Council on Measurement in Education Newsletter,* 13(3), 4–5.

Black, P. J. (1998) *Testing: friend or foe? The theory and practice of assessment and testing* (London, Falmer Press).

Black, P. J. & Wiliam, D. (2003) 'In praise of educational research': formative assessment, *British Educational Research Journal,* 29(5), 623–637.

Bloom, B. S., Hastings, J. T. & Madaus, G. F. (1971) *Handbook on formative and summative evaluation of student learning* (New York, McGraw-Hill).

Bloom, B. S., Madaus, G. F. & Hastings, J. T. (1981) *Evaluation to improve learning* (New York, McGraw-Hill).

Caplan, D. (1998) Measuring the output of non-market services, *Economic Trends,* 539, 45–49.

Cresswell, M. J. (2003) *Heaps, prototypes and ethics: the consequences of using judgements of student performance to set examination standards in a time of change* (London, University of London Institute of Education).

Department for Education and Skills (2005) *14–19 education and skills.* Cm 6476 (London, Stationery Office).

Department of Education and Science and the Welsh Office (DES/WO) (1988) *National Curriculum: Task Group on Assessment and Testing: a report* (London, Department of Education and Science).

Eckstein, M. A. & Noah, H. J. (1993) *Secondary school examinations: international perspectives on policies and practice* (New Haven, Yale University Press).

Harlen, W. (2005) Teachers' summative practices and assessment for learning—tensions and synergies, *The Curriculum Journal,* 16(2), 207–223.

Harlen, W. & James, M. (1997) Assessment and learning: differences and relationships between formative and summative assessment, *Assessment in Education: principles, policy and practice,* 4(3), 365–379.

HMSO (1995) *Statutory Instrument 1995 No. 2073. The Education (National Curriculum) (Assessment arrangements for the core subjects) (Key stage 3) (England) Order 1995* (London, The Stationery Office Limited).

Newton, P. E. (forthcoming) The multiple purposes of assessment, in: B. McGaw, P. L. Peterson & E. Baker (Eds) *International encyclopedia of education* (3rd edn.) (Elsevier).

Pellegrino, J. W., Chudowsky, N. & Glaser, R. (Eds.) (2001) *Knowing what students know: the science and design of educational assessment* (Washington, DC, National Academy Press).

Pritchard, A. (2003) Understanding government output and productivity, *Economic Trends,* 596, 27–40.

Sadler, D. R. (1989) Formative assessment and the design of instructional systems, *Instructional Science,* 18, 119–144.

School Curriculum and Assessment Authority/Department for Education (SCAA/DFE) (1994) *Key stage 2 assessment arrangements 1995* (London, SCAA).

Scriven, M. (1967) *The methodology of evaluation* (Washington, DC, American Educational Research Association).

Scriven, M. (1991) *Evaluation thesaurus* (4th edn.) (Newbury Park, Sage).

Taras, M. (2005) Assessment—summative and formative—some theoretical reflections, *British Journal of Educational Studies*, 53(4), 466–478.

Wiliam, D. (2004) Keeping learning on track: integrating assessment with instruction. Invited address to the *30th Annual Conference of the International Association of Educational Assessment*, Philadelphia, PA, USA, 14–18 June.

Wiliam, D. & Black, P. J. (1996) Meanings and consequences: a basis for distinguishing between formative and summative functions of assessment? *British Educational Research Journal*, 22(5), 537–548.

Working Group on 14–19 Reform (2004) *14–19 curriculum and qualifications reform: final report of the Working Group on 14–19 Reform* (London, DfES Publications).