Meanings and Consequences: a basis for distinguishing formative and summative functions of assessment?

DYLAN WILIAM & PAUL BLACK, *King's College London*

ABSTRACT *The assessment process is characterised as a cycle involving elicitation of evidence, which when interpreted appropriately may lead to action, which in turn, can yield further evidence and so on. 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 prioritise the consistency of meanings across contexts and individuals. Aspects of the interplay of meanings and consequences are explored for each of the three phases, and it is suggested that this interplay may be fruitful in distinguishing the two functions. Tensions between summative and formative functions of assessment are illustrated in the context of the National Curriculum, and although it is shown that such tensions will always exist, it is suggested that the separation of the elicitation of evidence from its interpretation can mitigate that tension.*

Introduction

The terms formative and summative assessment are not very common in the technical literature on assessment. For example, the third edition of the classic text in the field, *Educational Measurement* (Linn, 1989), indexes only a single mention of each, both in the chapter by Nitko (1989). Part of this is explained by differences in the terms used on the two sides of the Atlantic, but in our view a much more significant factor is that the day-to-day activities of teachers have historically been of little interest to academic researchers in this area.

The term 'formative evaluation' had first been used by Michael Scriven (1967) in connection with the improvement of curriculum, but Bloom *et al.* (1971) were the first to extend the usage to its generally accepted current meaning. They defined as *summative evaluation tests* those assessments given at the end of units, mid-term and at the end of a course, which are designed to judge the extent of students' learning of the material in a course, for the purpose of grading, certification, evaluation of progress or even for researching the effectiveness of a curriculum (Bloom *et al.*, p. 117). They contrasted

0141–1926/96/050537–12

these with 'another type of evaluation which all who are involved—student, teacher, curriculum maker—would welcome because they find it so useful in helping them improve what they wish to do' (p. 117), which they termed 'formative evaluation'.

From the earliest use of these terms, it was stressed that the terms applied not to the assessments themselves, but to the functions they served. On the one hand, the results of an assessment that had been designed originally to fulfil a summative function might be used formatively, as is the case when a teacher administers a paper from a previous year in order to help students to prepare for an examination. On the other hand, one does not have to go far to find examples of assessments intended to have some formative value whose results are used simply as summative judgements of the achievement of students.

Subsequently, Airaisian & Madaus (1972) augmented the classification by the addition of *diagnostic* and *placement* functions for assessment. However, this fourfold classification is neither particularly useful nor illuminating since the relationship between the four kinds of function is not clearly drawn out, and in any case, the terms are used in rather different senses from those currently accepted, at least in the UK. Nitko (1989) does not use the term 'formative' for types of assessments, preferring to classify assessment decisions as *placement* decisions, *diagnostic* decisions, *monitoring* decisions, and *attainment* decisions.

Characteristics of Formative Assessment

Whatever the labels that are used to describe it, formative assessment itself is, of course, nothing new. Almost all successful teaching (and certainly any teaching that is successful over a sustained period) relies heavily on adapting the teaching in the light of evidence about the success of previous episodes. These adaptations vary in terms of both their scope (e.g. the number of students involved) and time-scale.

At one extreme, a teacher explaining something to an individual student may amend his or her approach almost instantaneously in response to a frown of puzzlement on the student's face or some other aspect of body language. The adaptations may occur in teaching situations with just as narrow a focus, but over a much longer time-scale, as happens when a music teacher decides that a particular approach to teaching the violin, which has worked with many students, is not working with a particular individual (narrow focus, longer time-scale). Conversely, when teaching to a whole class, many teachers use a 'reference group' of students (Dahllöf, 1971) to judge the pacing of their lessons (broader focus, short time-scale). At the other extreme, an adaptation can have both a broad scope and take place over a long time-scale, as in the case of a 5-yearly review of an undergraduate degree programme.

The common element in all of these assessment processes is that of *feedback*, defined by Ramaprasad (1983) as 'information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way' (p. 4). As Sadler (1989) notes:

> An important feature of Ramaprasad's definition is that information about the gap between actual and reference levels is considered as feedback *only when it is used to alter the gap.* If the information is simply recorded, passed to a third party who lacks either the knowledge or the power to change the outcome, or is too deeply coded (for example, as a summary grade given by the teacher) to lead to appropriate action, the control loop cannot be closed,

> and 'dangling data' substituted for effective feedback. (p. 121, emphasis in original)

Formative functions of assessment are therefore validated in terms of their consequences as much as their meanings. Until recently, however, the consequences of summative assessments have been excluded from validity arguments, but reformulations by Messick (1980), Cronbach (1988) and Madaus (1988) have emphasised that assessments are validated not simply by their outcomes, but also by what happens *as a result* of those outcomes:

> Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. (Messick, 1989, p. 13, emphases in original)

A consideration of the relationship between formative and summative functions of assessment must therefore take place in the light of discussion of how the more general concepts of validity and reliability impinge upon the relationship.

Validity and Reliability

If a response to (say) a single multiple-choice question or to a question asked in a class discussion is held to be inadequate, it is by no means clear what we are entitled to conclude. This is because the answer is a measure not only of pupil's ability in relation to the standard of understanding intended by the questioner, but also of the response to a variety of other features such as the language, the context of the question, or the pupil's tiredness at the time. The response is therefore a measure of a variety of properties and its validity is contaminated, because it does not measure only the feature intended (i.e. it embodies *construct-irrelevant variance*), or, to follow a different view of validity, action based on the response may not be appropriate.

Such difficulties may be dealt with in a variety of ways. One may ask a variety of questions, which differ in language and in context and have in common only the particular understanding that is of interest; one might go further and set such questions on different occasions so that the pupils may be in different states of mind. In such manoeuvres, the aim is to average out all those effects which are not desired and so enhance the signal to noise ratio for the effect of interest. This can be seen as enhancement of validity by repetition or averaging. This is what is commonly called reliability, and this first response to the problem of the mixed or contaminated validity of an item is the classical one for summative testing.

A different response may be more appropriate in the case of the question asked in the class discussion. The teacher may respond with a correction, or with a follow-up question that has a diagnostic purpose. This is a quite different type of response. It may be justified for two reasons. First, if the pupil's response were due (say) to a misunderstanding or misinterpretation of the original question, this may quickly become apparent. That is to say, the consequences allow correction for invalidity of the evidence. Secondly, for the teacher to ask the pupil several other questions in different language and contexts, in order to be sure that the interpretation of misunderstanding was well founded, would be impracticable, except in extraordinary circumstances—for example, where the teacher might judge that follow-up responses might simply create confusion if the issue had not been properly identified.

Thus, in the fluid action of the classroom where rapid feedback is important, optimum

validity depends upon the self-correcting nature of the consequent action. In the formal test, which is on a different time-scale and for which correction by follow-up is out of the question, optimum validity requires a collection of items. Consequently, in the way that this discussion is developed, reliability is not a separate issue—it is subsumed into validity.

However, these two scenarios, the classroom discussion and the formal test, are best regarded as two ends of a spectrum. There are many intermediates. One would be an oral examination—where fewer questions are usually possible than in a formal test but validity can be enhanced through the use of responsive and interactive dialogue. Another would be a reflective appraisal of a pupil's last three essays in order to give an overall grade. On one interpretation, the average over the three is 'reliable' because it might iron out influences on performance which are not relevant to interpreting the evidence as a prediction for future writing (i.e. it would iron out the invalidities). On another interpretation, the best of the three would be more valid as it might indicate what could be achieved if the pupil were strongly interested or motivated in the task. Or again, it might be the last of the three because this would indicate where the pupil had arrived after having the full advantage of the teacher's feedback. In this example, the choice between these is not to do with any tension between formative and summative functions, either in terms of the data collected or in terms of the criteria for interpretation. Furthermore, only a teacher who knew the particular student and the context of the work could make that best choice—he or she would be giving meaning to the data in the light of the predictive consequences that might follow from the result. There might also be feedback consequences. Sadler (1989) points out that a mechanical approach in which students' course-work marks are accumulated to give a summative result can have the effect of making the student unwilling to repeat work in order to improve it, because the same investment of time might add more to the total through production of yet another mediocre essay.

The Assessment Cycle

Both formative and summative functions of assessment require that evidence of performance or attainment is elicited, interpreted and acted upon, in some way. These actions may then directly or indirectly generate more evidence so that the cycle is repeated. The key agents in this process are, of course, the assessed and the assessor (often called the teacher and the student below, for simplicity), although sometimes the assessed and the assessor will be the same individual. However, many aspects of the process cannot be understood without acknowledging that this relationship between the assessed and the assessor is itself influenced by the relationships that each of the agents has to a wider social context.

Although there is no natural beginning or ending to the process of assessment, for the sake of discussion, it is convenient to start with the elicitation of evidence.

Eliciting Evidence

Before any inferences can be made, or actions taken, some evidence about the level of performance must be generated and observed. We can immediately distinguish between purposive and incidental evidence. Purposive evidence is that which is elicited as a result of a deliberate act by someone (usually the teacher) that is designed to provide evidence about a student's knowledge or capabilities in a particular area. This most commonly

takes the form of direct questioning (whether orally or in writing). Of course, this will not guarantee that if the student has any knowledge or understanding in the area being assessed, then evidence of this attainment will be elicited. One way of asking a question might produce no answer from the student, while a slightly different approach may elicit evidence of achievement. We can never be absolutely sure that we have exhausted all the possibilities, so that we can never be sure that the student does *not* know something, but some assessments will be better than others in this respect. The extent to which an assessment can be relied upon to yield evidence of attainment where it exists has been called the *disclosure of* the assessment (Wiliam, 1992a).

Disclosure can be regarded as a technical issue—essentially an aspect of test–retest reliability—but this would ignore the crucial fact that all assessments take place in essentially *social* settings. For a variety of reasons; students may choose to fabricate or withhold evidence of attainment (MacNamara & Roper, 1992), and this has led some to conclude that true formative assessment can occur only when no external agency is involved:

> The indispensable conditions for improvement are that the *student* comes to hold a concept of quality roughly similar to that held by the teacher, is able to monitor continuously the quality of what is being produced *during the act of production itself*, and has a repertoire of alternative moves or strategies from which to draw at any given point. (Sadler, 1989, p. 121, emphases in original)

Another example of difficulty with disclosure arises in the assessment of teaching practice for students engaged in courses of initial teacher education, where students often state to their tutors that everything is going smoothly even when it is not, because the potential gain (in terms of advice about possible solutions) is outweighed by the loss entailed in admitting that they are having difficulties.

Certainly locating the responsibility for elicitation, interpretation and action within the individual obviates some of the difficulties of disclosure (particularly that of the conscious withholding of evidence), although we should not assume that incomplete disclosure ceases to be a problem when no external agency is involved. We all have considerable capacity for deluding ourselves about the actual level of performance, particularly in emotionally-charged settings.

In the absence of any direct probing (whether by the individual interrogating his or her own performance or by an external agency), evidence of achievement is also spontaneously and continuously generated. This 'windfall' evidence, because it has not been generated as a result of particular probing, can often be more robust than evidence gained purposively: the fact that a student chose to use a particular skill, and did so successfully, is often evidence of deeper understanding than being able to apply that particular skill when told to do so. However, as the name 'windfall' implies, we may have to wait a very long time (or forever!) for the appropriate evidence to be generated. In this sense, *robustness* and *disclosure* are in tension: the more we make it clear what we want, the more likely we are to get it, but the less likely it is to mean anything.

As well as the means by which it is generated, evidence also differs in the *form* in which it is generated. Traditionally, only evidence that exists in some permanent form (as writing, artefacts, or on audio- or videotape) has been relied upon in formal assessment settings, with its concern to establish consistency across raters, while *ephemeral evidence* has been largely discounted. However, as far as formative assessment is concerned, inter-rater consistency is of secondary importance, and ephemeral evidence can be an entirely appropriate form of evidence.

Unfortunately, the evanescent nature of ephemeral evidence means that it must be captured immediately or lost. Where students are working in small groups within a classroom, they may well demonstrate very high-quality speaking and listening skills, but this may not be observed by the teacher, because she was in another part of the room at the time. For this reason, many teachers encourage students to write down, or record in some permanent form, the otherwise ephemeral evidence of their attainment. It is tempting to regard this as a process of conversion, from one form of evidence to another, but many students have difficulties in, for example, expressing in writing what they have articulated quite fluently in oral form, while others are 'tongue-tied' in classroom discussions, and can only really express themselves in writing. It is, therefore, perhaps more appropriate to regard the two forms of assessment as existing in parallel, with each being an imperfect representation of the quality of thought that gave rise to it (these can be regarded as issues of *fidelity*; see Wiliam, 1992a).

Even when disclosure and fidelity are not problematic, the nature or timing of the data can limit the kinds of function that it can serve. For example, the information in the data might be too coarse-grained to be useful, or might just come at the wrong time. As Sadler (1989) points out, evidence that is elicited at the end of a course cannot serve a formative function for the students involved (although it could be formative in terms of the course for future students).

Interpretation

Of course evidence by itself is not information until it is interpreted, and the same evidence can be interpreted in different ways. In most classrooms, the interpreter is the teacher. She has a notion of what she would like the students to be able to do, and by examining the evidence, determines whether there is, in fact, a gap. For example, in a Key Stage 1 science lesson on 'floating and sinking', a student may correctly predict that a 'boat' made of metal foil floats, while predicting that a pellet of the same foil sinks. The teacher may conclude that the student has 'understood' floating and sinking. However, it could be that the student has a rather different set of conceptions: the student might believe that compressing the foil into a ball makes it heavier, which is why it sinks. The same pattern of responses is consistent with many different sets of student conceptions.

Wiliam (1992b) describes a situation where a pupil had generated some data about a mathematical relationship in which the number of free edges in an arrangement of octagonal tiles can be found by multiplying the number of tiles by three and adding eight.

> The student had stated:
>
> If you want to get the number of free edges, then you take the number of tiles, like 6, and times it by three, so you get 18 and then add eight so you get 26.
>
> One teacher inferred from this that the student had derived a general rule, but was presenting the example by virtue of a 'generic example', and felt that the activity demonstrated attainment of a particular statement in the mathematics national curriculum (make generalisations). The other teacher concluded that the student had only presented a *specific* example, and thus one could not infer that the student had made a general statement. (p. 11)

Both teachers observed the same evidence of attainment, but they disagreed about its

interpretation. For the first teacher there was no gap between the actual level and the reference level (in this case, whether the student could, in fact, make generalisations), but for the second teacher, there *was* a gap.

Questioning as a Turing test. The mathematician Alan Turing once proposed a simple test for deciding whether a machine was intelligent (Turing, 1950). He proposed that a person (the 'judge') should sit in a room with two keyboard terminals, one connected to a computer, and the other connected to a human operator. The judge would then ask a series of questions of the computer and the human operator, trying to discover which was which. If the judge were unable to determine which was which, then the computer could be said to be intelligent. In many ways, classroom questioning has this character. The teacher elicits and examines evidence of attainment based on his or her model of what it is to 'understand' the ideas in question, trying to establish whether the student(s) share this model. Provided the students' answers are consistent with the teacher's model, they will be regarded as having 'understood' the topic. However, as von Glasersfeld (1987, p. 13) has pointed out, all such a process establishes is that the teacher's schemas and the students' schemas both 'fit' the frame established by the questions, not that they 'match'. No amount of probing can establish conclusively that the schemas match, but the more demanding the 'Turing test', the more likely there is to be a match.

Action

Assessments yield evidence that can be interpreted in different ways for different purposes, but these interpretations are means to an end rather than ends in themselves. With very few exceptions, assessments are conducted for a purpose, and certain actions are contingent on the outcomes. Indeed, if nothing different can happen as the result of an assessment, there can be little point in conducting the assessment in the first place.

For *placement* decisions, the consequences can be acceptance for, or rejection from, employment or a course of study. Separate from, but related to these direct consequences are the social consequences of the decisions and the way they are made. However, such a function of an assessment is not formative according to the view presented here because Ramaprasad's definition of feedback requires that the information generated is actually used to close the gap between actual and desired levels of performance. If we discover that there is a gap, but have no idea about the nature of the discrepancy between actual and desired performance, then this (almost inevitably norm-referenced) information does not help us close the gap, and therefore fails to qualify as feedback. Such a process would be better described as simply *monitoring*.

To qualify as feedback, as well as alerting us to the existence of a gap, the information must actually be useful in *closing* the gap between actual and desired levels of performance. The information must therefore have embedded within it some degree of prescription about what must be done. The information must be related to a developmental model of growth in the domain being addressed—in short, it must be *construct*-referenced (Messick, 1975).

To sum up, in order to serve a formative function, an assessment must yield evidence that, with appropriate *construct*-referenced interpretations, indicates the existence of a gap between actual and desired levels of performance, and suggests actions that are in fact successful in closing the gap. Crucially, an assessment that is *intended* to be formative (i.e. has a formative *purpose*) but does not, ultimately, have the intended effect

(i.e. lacks a formative *function*), would not, with this definition, be regarded as formative.

The Relationship between Formative and Summative Functions

The definition of the formative function of assessment adopted here places conditions on both the interpretations made of the evidence and on the consequent actions. Any assessment must elicit evidence of performance, which is capable of being interpreted (however invalidly). Whether or not these interpretations and actions satisfy the conditions for formative functions, the fact that interpretable evidence has been generated means that the assessment can serve a summative function. Therefore *all* assessments can be summative (i.e. have the potential to serve a summative function), but only some have the *additional* capability of serving formative functions. The question is not, therefore, can an assessment serve both functions, but the extent to which serving one has an adverse effect on its ability to serve the other.

As noted above, summative and formative functions are, for the purpose of this discussion, characterised as the ends of a continuum along which assessment can be located. At one extreme (the formative) the problems of creating shared meanings beyond the immediate setting are ignored: assessments are evaluated by the extent to which they provide a basis for successful action. At the other extreme (summative) shared meanings are much more important, and the considerable distortions and undesirable consequences that arise are often justified by appeal to the need to create consistency of interpretation. Presenting this argument somewhat starkly, when formative functions are paramount, meanings are validated by their consequences, and when summative functions are paramount, consequences are validated by meanings.

Formative Assessment in the National Curriculum

In 1988, the British Government's National Curriculum Task Group on Assessment and Testing (TGAT) published its proposals for an assessment and reporting structure for the National Curriculum in England and Wales. The group took the view (NCTGAT, 1988) that a single assessment system could serve both summative and formative functions, provided the formative function was the foundation of the system:

> It is possible to build up a comprehensive picture of the overall achievements of a pupil by aggregating, in a structured way, the separate results of a set of assessments designed to serve a formative purpose. However, if assessments were designed only for summative purposes, then formative information could not be obtained, since the summative assessments occur at the end of a phase of learning and make no attempt at throwing light on the educational history of the pupil. It is realistic to envisage, for the purpose of evaluation, ways of aggregating the information on individual pupils into accounts of the success of a school, or LEA [local education authority] in facilitating the learning of those for whom they are responsible; again the reverse is an impossibility. (para. 25)

This view has been criticised by many who have asserted that formative and summative functions cannot co-exist in any meaningful way. It is certainly true that the involvement of external agencies creates difficulties in terms of disclosure as noted above, but many

other difficulties can be alleviated by separating the *elicitation* of the evidence from the *interpretation* of the evidence.

For example, in the first version of the National Curriculum, the attainment targets for mathematics and science were presented in terms of statements of attainment (296 for mathematics and 407 for science), each of which was allocated to one of the 10 levels of the National Curriculum. Many teachers devised elaborate record sheets that would allow them to indicate, for each statement of attainment, whether it had been achieved by a student. Originally, such a record sheet served a formative function: it gave detailed construct-referenced information on a student's current attainment, and, just as importantly, what had not yet been attained. While some teachers did question the notion of progression inherent in the allocation of the statements of attainment to levels, most seemed happy to accept that the student's next objectives were defined in terms of those statements just beyond the 'leading edge' of attained statements.

When a student produced evidence that indicated that she or he had partially achieved a statement (perhaps by demonstrating a skill in only a limited variety of contexts), then teachers would often not 'tick off' the statement, so that they would be reminded to re-evaluate the student's performance in this area at some later date. Since there are typically many opportunities to 'revisit' a student's understanding of a particular area, this seems a good strategy, especially since a false-negative attribution (assuming that a student does not know something they do, in fact, know) is, in an educational setting, likely to be far less damaging than a false-positive (assuming that they do know something they do not).

However, many schools subsequently chose to derive the summative levels required in National Curriculum assessment by the inflexible application of a formula—more often than not the 'n-1' rule suggested by the School Examinations and Assessment Council (SEAC) [2]. This immediately created a tension between formative and summative functions of the assessment. Where teachers had left statements 'unticked' in order to prompt them to return to those aspects at a later date, students who had relatively complete understandings were often regarded as not having met the criterion. In order to prevent this happening, teachers then stopped using the record sheets in this formative way, and started using them to record when the student had achieved a sufficient proportion of the domain addressed by the statement. The record sheets became entirely summative records of statements covered (a process that has been described as 'scalphunting'!)

The tension between summative and formative functions arose in this situation because of the inflexible application of a mechanical rule for aggregation that had the effect of conflating the elicitation of evidence with its interpretation. The distorting effect of the summative assessment can be mitigated if, instead of using an algorithmic formula, aggregation is by a process of *reassessment* (Wiliam, 1995). In other words, instead of relying on the results that already incorporate interpretations in order to serve a formative function, the teacher could look at the original evidence. The same point can be made in the context of the 'three-essay' example discussed earlier. Using some predetermined algorithm applied to the three scores already allocated to the three essays is one method of deriving an aggregate score, which may serve some purposes, but others may be better served by going back to the original essays. Evidence that was interpreted one way to serve a formative function can be interpreted quite differently to serve a summative function.

As long as a distinction is maintained between the elicitation and the interpretation of evidence, formative functions need not be incompatible with National Curriculum

assessment. However, having said this, it is worth noting that there is a large and growing body of research evidence that shows how difficult it is to introduce effective formative assessment into classroom practice (see, for example, Andrews, 1987, 1988; Torrance, 1991; Pole, 1993; Fairbrother, 1995).

Conclusion

There can be no doubt that significant tensions are created when the same assessments are required to serve both formative and summative functions. One response to this would be to say that the two functions require completely different approaches to elicitation, so that there is no prospect that different interpretations of the same evidence, or different actions based on the same interpretation can ever serve formative and summative functions adequately. This would have one of two effects. One possible outcome is that teachers' activities would be restricted to formative functions, with all summative assessment being undertaken by external agencies—an option teachers do not want (Brown *et al.*, 1995). The other possible outcome is that teachers are required to administer two parallel but completely separate assessment systems. Quite apart from the workload, which is likely to be inordinate, it seems likely that there would be serious backwash into teaching and learning.

The other response, and the one we have explored here, is to regard the two functions as the extremes of a continuum. Clearly, the problems identified above with regard to elicitation of evidence mean that not all evidence generated to serve a formative function can be used to serve a summative one. It would be very difficult to argue that responses to an 'off-the-cuff' question to a class in the middle of an episode of teaching would have any significance beyond the immediate context of the classroom. Conversely, evidence elicited at the end of a sequence of teaching can have very little formative influence on the students assessed. However, between these clear cases, it seems to us that there may be some common ground between the formative and summative functions. Finding this common ground will be difficult, since the issues are subtle and complex, and we have made only a small contribution here.

We are more confident that we have established that there are clear benefits to be gained in separating the interpretation of evidence from its elicitation, and the consequent actions from the interpretations, although we are still aware that more needs to be done in clarifying these issues. We hope that others will join in this debate, for we believe the potential advantages are significant, while the alternatives—teachers completely disconnected from all summative assessment or required to administer two separate assessment systems—would be disastrous.

Correspondence: Dylan Wiliam, King's College London, School of Education, Cornwall House, Waterloo Road, London SE1 8WA, UK. Email:dylan.wiliam@kcl.ac.uk

NOTES

This paper was presented as part of a Symposium at the 1995 British Educational Research Association conference entitled 'Formative and summative assessment: resolving the tension'.

[1] In much of the technical literature a distinction is drawn between *selection* decisions, which are made when deciding whether a candidate has the necessary potential to benefit from a college degree course on the one hand, and *placement* decisions, which are taken subsequently to decide which course would be most suitable.

[2] This rule suggested that where there were one or two statements of attainment at a particular level in an attainment target, all had to be attained in order for the student to be awarded that level, and where there were three or more statements, all but one had to be attained. It was widely assumed by schools that the use of this rule was a legal requirement, although this was not, in fact, the case (Wiliam, 1992c).

REFERENCES

AIRASIAN, P.W. & MADAUS, G.F. (1972) Functional types of student evaluation, *Measurement and Evaluation in Guidance*, 4, pp. 221–233.

ANDREWS, S. (1987) *The Achievements of Robert Arthur Essex: sixty employers give their perceptions of the Essex Records of Achievement Summary Portfolio* (London, Industrial Society).

ANDREWS, S. (1988) *Records of Achievement: R. A. Essex leaves school (sixty school/college leavers speak about involvement in Records of Achievement* (London, Industrial Society).

BLOOM, B.S., HASTINGS, J.T. & MADAUS, G.F. (Eds) (1971) *Handbook on the Formative and Summative Evaluation of Student Learning* (New York, McGraw-Hill).

BROWN, M.L., MCCALLUM, E., TAGGART, B., BRANSON, J. & GIPPS, C.V. (1995) Validity and impact of national tests in the primary school: the teacher's view, paper presented at the twenty first *Annual Conference of the British Educational Research Association* held at University of Bath, September (London, King's College London School of Education).

CRONBACH, L.J. (1988) Five perspectives on validity argument, in: H. WAINER & H. I. BRAUN (Eds) *Test Validity*, pp. 3–17 (Hillsdale, NJ, Lawrence Erlbaum Associates).

DAHLLÖF, U. (1971) *Ability Grouping, Content Validity and Curriculum Process Analysis* (New York, NY, Teachers College Press).

FAIRBROTHER, R.W. (1995) Pupils as learners, in: R. W. FAIRBROTHER, P. J. BLACK & P. N. G. GILL (Eds) *Teachers Assessing Pupils: lessons from science classrooms*, pp. 105–120 (Hatfield, Association for Science Education).

LINN, R.L. (Ed.) (1989) *Educational Measurement*, 3rd edn (Washington DC, American Council on Education/Macmillan).

MACNAMARA, A. & ROPER, R. (1992) Attainment target 1—is all the evidence there? *Mathematics Teaching*, 140, pp. 26–27.

MADAUS, G.F. (1988) The influence of testing on the curriculum, in: L. N. TANNER (Ed.) *Critical Issues in Curriculum: the 87th yearbook of the National Society for the Study of Education (part 1)*, pp. 83–121 (Chicago, IL, University of Chicago Press).

MESSICK, S. (1975) The standard problem: meaning and values in measurement and evaluation, *American Psychologist*, 30, pp. 955–966.

MESSICK, S. (1980) Test validity and the ethics of assessment, *American Psychologist*, 35, pp. 1012–1027.

MESSICK, S. (1989) Validity, in: R. L. LINN (Ed.) *Educational Measurement*, pp. 13–103 (Washington DC, American Council on Education/Macmillan).

NATIONAL CURRICULUM TASK GROUP ON ASSESSMENT AND TESTING (1988) *A Report* (London, Department of Education and Science).

NITKO, A.J. (1989) Designing tests that are integrated with instruction, in: R. L. LINN (Ed.) *Educational Measurement*, pp. 447–474 (Washington DC, American Council on Education/Macmillan).

POLE, C.J. (1993) *Assessing and Recording Achievement: implementing a new approach in School* (Buckingham, Open University Press).

RAMAPRASAD, A. (1983) On the definition of feedback, *Behavioural Science*, 28, pp. 4–13.

SADLER, D.R. (1989) Formative assessment and the design of instructional systems, *Instructional Science*, 18, pp. 145–165.

SCRIVEN, M. (1967) *The Methodology of Evaluation* (Washington DC, American Educational Research Association).

TORRANCE, H. (1991). Records of achievement and formative assessment: some complexities of practice, in: R. E. STAKE (Ed.) *Advances in Program Evaluation: using assessment to reform education* (Greenwich, CT, JAI Press).

TURING, A.M. (1950) Computing Machinery and Intelligence, *Mind*, 59(236), pp. 433–460.

VON GLASERSFELD, E. (1987) Learning as a constructive activity, in: C. JANVIER (Ed.) *Problems of Representation in the Teaching and Learning of Mathematics* (Hillsdale, NJ, Lawrence Erlbaum).

WILIAM, D. (1992a) Some technical issues in assessment: a user's guide, *British Journal for Curriculum and Assessment*, 2(3), pp. 11–20.

WILIAM, D. (1992b) Inset for national curriculum assessment: lessons from the key stage 3 SATs trials and pilots, *British Journal for Curriculum and Assessment*, 2(2), pp. 8–11.

WILIAM, D. (1992c) National curriculum assessment arrangements—the legal minimum, *Education and the Law*, 4, pp. 135–144.

WILIAM, D. (1995) Combination, aggregation and reconciliation: evidential and consequential bases, *Assessment in Education: principles policy and practice*, 2, pp. 53–73.