Outcomes-based assessment and reporting in language learning programmes: a review of the issues

Geoff Brindley

Language Testing 1998; 15; 45
DOI: 10.1177/026553229801500103

Outcomes-based assessment and reporting in language learning programmes: a review of the issues

Geoff Brindley *Macquarie University*

In recent years educational authorities in many countries have introduced outcomes-based assessment and reporting systems in the form of national standards, frameworks and benchmarks of various kinds which are used both for purposes of system accountability and for assessing individual progress and achievement in language learning. However, in some cases the introduction of these systems has proved problematic, owing to a number of political, technical and practical factors. These include the difficulty of combining formative assessment with summative reporting, the differing information requirements of different audiences, concerns about the validity and reliability of outcome statements and the lack of appropriate resources to support implementation. Such problems may be able to be alleviated by closer consultation between policy-makers, administrators and practitioners, by undertaking further research into the validity and consistency of outcome statements and by strengthening the links between assessment and reporting. A major investment in teacher professional development is necessary if teachers are to be responsible for carrying out their own assessments. Ongoing research needs to be conducted into the effects of outcomes-based assessment and reporting on student learning.

I Introduction

Over the last decade, increasing pressure has been placed on educational authorities by governments in many industrialized countries to meet national economic imperatives. Economic growth and international competitiveness are now conventionally seen as contingent on the capacity of education and training systems to produce a highly educated, flexible and literate workforce which is able to cope with the challenges of rapid social and technological change. In this climate, the widespread introduction of corporate management principles such as competition, productivity and cost-effectiveness into education has meant that educational policy and planning have

Address for correspondence: Geoff Brindley, National Centre for English Language Teaching and Research, School of English, Linguistics and Media, Macquarie University, NSW 2109, Australia; e-mail: gbrindle@laurel.ocs.mq.edu.au

Language Testing 1998 15 (1) 45–85 0265-5322(98)LT143OA © 1998 Arnold

become increasingly driven by considerations of economic accountability. As Blackmore (1988: 35) comments, ‘the primary concern of economic accountability is not process, but outcome. It is driven by utilitarian goals regarding the maximising of benefits with limited resources. It is market-oriented.’

In line with this emphasis on outcomes, government policy-makers and bureaucrats in many countries have in recent years put major efforts into the development of system-side indicators that allow the performance of educational institutions and systems to be closely monitored and evaluated (Fitz-Gibbon, 1990; Singh, 1990; Riley and Nuttall, 1994).

Figuring prominently amongst the key indicators used for system-level monitoring and reporting in many education and training contexts are statements of programme outcomes which describe expectations of learner performance standards at different levels of achievement. These descriptions, known under an assortment of names, including ‘standards’, ‘benchmarks’, ‘bandscales’, ‘attainment targets’ and ‘competencies’, have increasingly become the cornerstone of assessment and reporting systems in Europe, North America, Asia, Australia and New Zealand. Language education has not been immune from this trend and the last few years have seen the development of various systems for assessing and reporting learning outcomes in school and adult language education as well as in a range of vocational contexts (see, for example, Languages Lead Body, 1993; Citizenship and Immigration Canada, 1993; National Languages and Literacy Institute of Australia, 1994; Curriculum Corporation, 1994; Clark *et al*., 1994; New South Wales Adult Migrant English Service [NSW AMES], 1995; Scharer and North, 1992; de Jong, 1995; North, 1995a; 1995b).

However, while assessment and reporting mechanisms at the system level have become more outcomes-oriented, centralized and bureaucratic to serve national economic goals, at the classroom and local level the focus has shifted back to the individual learner. There has been a major move away from a reliance on mass standardized testing, particularly in the United States, towards the use of ‘alternative’ or ‘performance’ assessment which directly reflects learning activities and which is carried out in the context in which learning takes place. Performance assessment involves the use of a wide range of tools such as structured observations, progress profiles, various kinds of performance-based tasks and self-assessment to build up a complex and multidimensional picture of learner performance, in contrast to the single grade or score conventionally yielded by large-scale standardized tests (Moss, 1992; Gipps, 1994).

Teachers, thus, are now finding themselves under pressure from two directions. On the one hand, they need to carry out detailed assessments at the individual level for purposes of diagnosis and feedback to learners, a role which they are prepared to embrace because of the obvious beneficial effects on instruction (Broadfoot, 1992). However, at the same time, they are increasingly being called on to report learners' progress against national standards in order to meet accountability requirements.

A number of commentators have recently highlighted the inherent dilemma in trying to reconcile demands for national comparability with the need to relate assessment directly to the learning process (Brindley, 1989; Troman, 1989; Gipps, 1994; Nuttall and Stobart, 1994). In the context of achievement assessment in English as a second language programmes for immigrants in Australia, Brindley (1989: 44) has suggested that one way for educational institutions to resolve this tension is by developing a range of criterion-referenced procedures and instruments which can be used to monitor and assess achievement of the communicative objectives of a given course or unit of instruction. Such procedures would include rating scales based on course objectives, teacher-constructed tests, performance profiles and self-assessment check-lists. Brindley argues that the results of these forms of assessments can be used not only to inform the learning process but also to report outcomes to external audiences such as funding bodies, provided they are prepared to accept descriptions of achievement profiles which may contain some qualitative information (1989: 45).

Recently a number of outcomes-based assessment and reporting schemes have attempted to combine formative assessment with summative reporting in this way by using teacher assessments as the basis for determining learners' achievement in relation to externally determined standards or benchmarks (e.g., Curriculum Corporation, 1994; McKay, 1995b; Pierce and Stewart, 1996). Because of the recency of such innovations, however, their effects are still far from clear. Moreover, the introduction of state-mandated systems for assessing and reporting outcomes in some countries has been accompanied by considerable controversy (Moore, 1996). It therefore seems timely to consider the ways in which such systems have been developed and implemented, to document the problems which have emerged, and to examine the relationship between summative reporting and classroom-based performance assessment.

Section II of this article will discuss the nature of outcome statements, focusing particularly on language learning contexts, provide an overview of various ways in which these statements have been developed and briefly outline some of their claimed advantages and potential problems. Sections III and IV will provide an account of several recent projects which have set out to develop assessment and reporting schemes in adult, school and vocational contexts in the UK and Australia and describe some of the political, technical and practical problems that have arisen in the course of their development and implementation. Section V will highlight lessons which can be drawn from these experiences and suggest action that might be taken on each of these fronts in order to produce high-quality assessment information which can meet the requirements of different stakeholders in language programmes.

II Outcome statements in language learning

1 Definitions

Variously known, *inter alia*, as *standards*, *benchmarks*, *attainment targets*, *bandscales*, *profiles* and *competencies*, outcome statements are, broadly speaking, standards of performance against which learners' progress and achievement can be compared. They are usually developed on a national or system-wide basis, as in, for example, the National Language Standards relating to foreign language use at work in the UK (Languages Lead Body, 1993); the proposed European common framework scale of language competence (North, 1995b); the language competencies described in the Certificates of Spoken and Written English (CSWE) (NSW AMES, 1995), developed to meet the needs of the national Adult Migrant English Program (AMEP) in Australia; and the Toronto Benchmarks for school learners produced for the Toronto Board of Education in Canada (Rutledge, 1993). Although they differ somewhat in their format, their generality and in the level of detail they present, these statements generally consist of verbal descriptions of the types of learner performances which can be expected at different levels of ability in the language skill or 'strand' in question. To this extent, the general level descriptors used in some outcome statements are similar to those found in language-proficiency rating scales. These descriptors may then be broken down further according to different features of performance. For example, the English as a Second Language (ESL) bandscales developed by the National Languages and Literacy Institute of Australia (NLLIA) for primary and secondary school learners cover four skill areas – *Listening*, *Speaking*, *Reading* and *Writing*. The descriptors within these areas are organized so as to take into account the components of the language user's ability and characteristics of the language use context according to the framework of communicative language ability proposed by Bachman and Palmer (1996). Appendix 1 (McKay, 1995b: 45) shows the way in which the descriptors in the NLLIA bandscales have been ordered. Appendix 2 shows a sample level description from the scales.

Some systems for describing outcomes are accompanied by ‘pointers’ or indicators of the types of specific behaviour or tasks which might typically be observed at the level of ability in question, thus providing both a context-free general statement of ability and a context-dependent operationalization of the ability in question. Others which use *language competencies* as the overall unit of organization break down these competencies into *elements* which are linked, in turn, to a set of *performance criteria* describing the behaviour which must be demonstrated to indicate that the outcome has been achieved. In competency-based systems, learners’ achievement is often assessed against these criteria on a yes/no basis (that is, they are deemed either to have demonstrated the specified behaviours or not). Some competency-based schemes also include *range statements* specifying the conditions under which the performance occurs. An example of an oral language competency from the CSWE (NSW AMES, 1995) is given in Appendix 3.

Outcome statements frequently include principles or guidelines for assessment and/or observation aimed at assisting teachers to record evidence of the described behaviours and containing descriptions of sample assessment activities and formats (e.g., National Languages and Literacy Institute of Australia, 1994). Some benchmark or profile schemes used in school education employ professionally designed standardized tasks to assist teachers in assessing students’ achievements against the level descriptors (see, for example Forster, 1994, who describes a set of assessment resources linked to the National English Profile in Australian schools). In adult and vocational education, however, there appears to be less reliance on standardized tasks: it is the individual teacher’s role to select the assessment activities that will be used to determine whether the learners in their class have met the standard or achieved the competency in question.

2 Purposes

Outcome statements present a map of student progress which, in principle, can be used by a variety of audiences for a wide range of purposes. These include:

1) to establish expected standards of achievement;
2) to provide system-wide reference points to assist teachers in assessing individual progress;
3) to provide a common framework for curriculum development;
4) to provide more comprehensive information for reporting to

interested parties outside the classroom, such as parents, employers and educational authorities;

5) to clarify the kinds of performance that lead to academic success;
6) to support teachers in their implementation of curriculum objectives;
7) to provide a basis for identifying needs and targeting resource allocation;
8) to provide a resource for teacher professional development (Rutledge, 1993; Curriculum Corporation, 1994; McKay, 1995b).

Whether or not outcomes-based assessment and reporting systems can fulfil all these purposes simultaneously is, however, a matter of contention which will be taken up later in Section IV of this article.

3 Development

The procedures used to develop outcome statements for language learning programs differ according to the context in which they are commissioned and the resources available to carry out the work. In general, however, the process of development will include some or all of the following steps (see Griffin and McKay, 1992, for a fuller description of a range of methods for bandscale development):

1) A need is identified, usually by a national, state or provincial educational authority for a set of standards/benchmarks/profiles/competencies, often in the context of pressures from funding bodies for greater explicitness in reporting of programme outcomes and/or demands for comparability of outcomes and portability of qualifications across systems.
2) A project is commissioned and a research team sets out to develop the statements.

There are a number of possible starting points for defining the content of the standards/benchmarks/profiles/competencies:

a) existing curriculum objectives or proficiency descriptors may form the basis for the statements (e.g., Rutledge, 1993; North, 1995a; 1995b);
b) analyses of communication needs may be undertaken to identify the domains of language use and the tasks involved (e.g., Mawer, 1991);
c) theories of language use or language ability may be used as an organising framework for defining the domains and tasks (e.g., NSW AMES, 1995; McKay, 1995b).

The development of the statements is usually undertaken by curriculum or assessment specialists, in varying degrees of consultation with practitioners. Some outcome statements are produced in a ‘top–down’ fashion (Griffin and McKay, 1992: 19), whereby specialists are mainly responsible for defining the tasks which learners can perform at different levels and specifying indicators of ability. Sometimes, however, the outcome statements may be distributed in draft form to practitioners and other experts, such as applied linguists, for comment and editing (Griffin and McKay, op. cit. refer to this as the ‘top–down consultative’ method, p. 20). Depending on the predilections of the developers, psychometric-empirical methods may also play a major role in determining the placement of the tasks which comprise the outcomes on the continua (‘top–down empirical’). In this context, Rasch analysis has been used in a number of recent projects as a means of scaling descriptors (see, for example, Griffin, 1991; 1994; North, 1995a; 1995b). At the other end of the spectrum, some projects may rely largely on the subjective judgements of practitioners to identify the tasks and their placement on the continua of levels (‘bottom–up’). These methods, however, are not mutually exclusive and a given set of statements may reflect a combination of several of the above approaches.

3) The statements are given to selected practitioners to use on a trial basis and modified in the light of feedback. Further input may be sought at this point from a range of interested stakeholders on the political acceptability of the outcome statements and on their specific wording.
4) The statements are then published and given to practitioners to implement. Sometimes they may be accompanied by assessment tools, exemplars and/or guidelines aimed at assisting teachers to assess learners’ achievement. Depending on the resources available, professional development support in the form of assessment materials and in-service training may also be provided.
5) The implementation of the statements is monitored and/or formally evaluated (this practice, however, seems to be relatively rare).

4 *Advantages*

There would appear to be a number of potential benefits in using explicit statements of expected outcomes as a basis for assessment, monitoring and reporting. In the first place, the fact that outcomes are described in performance terms means that learners are focused on language as a tool for communication rather than on language knowledge as an end in itself. They are also able to obtain diagnostic feedback on the success of their learning since explicit performance criteria are provided against which they can judge their progress. Second, since there is a direct link between attainment targets, course objectives and learning activities, assessment is closely integrated with instruction: what is taught is directly related to what is assessed and (in theory at least) what is assessed is, in turn, linked to the outcomes that are reported. Third, teachers, by comparing students' progress and achievement with the standards statements, are able to make better-informed judgements about what individual learners need. Fourth, the use of the statements – provided they are couched in terms accessible to the relevant audiences – facilitates communication between all the stakeholders in the programme by providing a common language for discussing outcomes. Finally, from a programme management perspective, the specification of targets against which programme outcomes can be evaluated provides a rational and objective basis for determining programme needs and allocating resources.

5 Potential problems

Despite these apparent advantages of outcomes-focused approaches, a constant preoccupation with targets and 'terminal behaviour' brings with it a number of potential pitfalls. Not the least of these is the tendency for assessment to dominate teaching and learning to the point where any and every learning activity may become the object of an assessment. This tendency is amusingly illustrated by Rutledge (1993: 5) in the following anecdote:

> one day I entered the hall outside the principal's office to see that two boys about seven years old were ahead of me, waiting to see Laura Schissler, the principal. They weren't in any trouble. They had come to show her some written work, as was often done there. Laura kept a terrarium just outside her office with a few frogs in it, and the boys were very interested in those frogs. As they talked about them and pointed to them one boy went to reach in and pick one up, when the other boy exclaimed: 'Don't touch that frog or you'll have to write something about it'.

Hargreaves (1989: 135) puts the view that when assessment takes the form of constant observation and monitoring in relation to standards, it can become a form of surveillance:

> The building up of a dossier, of an extended case record, to be retrieved and referred to at any point where the difficult work of institutional guidance and channelling, control and correction needs to be done, comes uncomfortably close to some aspects of the formative process of recording pupils' personal achievements.

A second objection that has been voiced is that state-sponsored outcome statements define what students 'must be', thus submerging individual and contextual differences in learning and leading to a narrowing of the curriculum (Moore, 1996). Related to this is the concern that in high-stakes contexts where teaching is heavily geared to producing optimal outcomes, it is easy for the standards/benchmarks/profiles/competencies to assume the status of a covert curriculum, thus forcing teachers to 'teach to the test' in order to achieve the best results for their learners, and in the process losing valuable teaching time (Elliott, 1994).

In addition to these possible negative effects of outcomes-focused approaches on curriculum and instruction, a range of unresolved technical and practical problems continue to surround the theoretical basis and content of the assessment procedures which are used to locate learners' performance levels in relation to the targeted standards. In this context, the concerns that have been widely raised about performance assessment in general education apply equally to language education. These include:

1) lack of evidence for the construct and content validity of performance assessments (Messick, 1994);
2) low levels of consistency of rater judgements (Shavelson *et al.*, 1992; Mehrens, 1992).
3) limited generalizability across performance tasks in the same domain and across domains (Dunbar *et al.*, 1991);
4) problems in ensuring the comparability of assessments carried out by different practitioners in different contexts using different assessment tasks (Gipps, 1994; Brindley, 1994);
5) the administrative complexity and high financial costs of implementing assessment and reporting schemes which rely on teachers to collect large amounts of individual assessment information (Madaus, 1993);
6) the difficulty of setting cut-off points for determining achievement (Cizek, 1993).

These issues will be discussed at greater length further on.

III Developing and implementing assessment and reporting systems: some recent experiences

While outcomes-based approaches appear to offer a logical solution to the problem of integrating classroom-based performance assessment with system-wide reporting, in practice their introduction has in some cases proved problematic. Conflicts have arisen between practitioners and educational authorities not only over the purposes of outcome reporting but also over the content of the outcome statements themselves and the tests that are used with them. Dissension has also accompanied the introduction of government-sponsored outcomes-based schemes into contexts where systems for formative assessment already existed.

As a first step in trying to understand why these conflicts arose, I want to trace the history of several major assessment initiatives which have taken place recently in the UK and Australia. Some of these have evolved in response to practitioners' needs for diagnostic information, while others have been driven more by state or national accountability concerns.

1 UK National Curriculum assessment

Although this first example is not specifically related to language assessment, it nevertheless provides an instructive illustration of the particular pitfalls of politically driven assessment reform.

The National Curriculum was established in the UK in 1988 via the Education Reform Act. It comprises 11 subjects, including three core subjects (English, mathematics and science), seven foundation subjects and religious education. National Curriculum assessment, in line with the recommendations of the Task Group on Assessment and Testing (TGAT) set up by the government to advise on assessment policy, includes both school-based assessments conducted by teachers and national testing which is based on externally devised Standard Assessment Tasks (SATs). The SATs are used to determine learners' levels of attainment in relation to the national attainment targets. Assessment and reporting are carried out at four Key Stages: Key Stage 1 (7-year-olds); Key Stage 2 (11-year-olds); Key Stage 3 (14-year-olds) and Key Stage 4 (16-year-olds).

A great deal of controversy has surrounded the implementation of the National Curriculum since its introduction (see, for example, Gipps, 1994; Elliott, 1994; Nuttall and Stobart, 1994; Edwards, 1995). In particular, there has been considerable conflict between the government and teachers on the way in which assessment of children's learning at the various Key Stages should be carried out. The first major clash arose over the government's plans to test 7-year-olds at Key Stage 1, which teachers argued was inappropriate for this group and represented an unmanageable workload (in this regard Nuttall and Stobart report that in 1990 teachers were required to make over 200 detailed judgements per child!; 1994: 26). The government's decision to publish comparative test results in the form of 'league tables' which ranked the performance of schools also caused considerable concern. Similar problems arose with the assessment of 14-year-olds at Key

Stage 3. Here the construct and content validity of levels of attainment also came under question when the results revealed that some students who were successful in tasks classified at 'higher' levels were failing supposedly 'lower'-level tasks (Nuttall and Stobart, 1994: 27).

In 1993 the situation came to a head when English teachers boycotted the testing of Key Stage 3 English, declaring both the content and format of the assessment to be educationally unsound. At this point, one major teacher union mounted a legal challenge to national testing on the basis that it represented an unfair workload on teachers and this claim was upheld by the courts. A stalemate then followed until the government appointed Sir Ron Dearing to review National Curriculum testing. His interim report in July 1993 recommended that the prescribed content of the National Curriculum should be reduced, that teachers' assessments should be accorded equal weight in determining learner attainment levels and that testing and marking loads should be cut down. As a result, the statutory content and the number of attainment targets for each subject were reduced and the requirements for assessment and testing were made less detailed. In 1994 teachers were allowed to make judgements on learners' achievements using general level descriptors rather than the detailed descriptions of attainment contained within each attainment target. This increase in assessment responsibilities, however, highlighted the importance of providing professional development support to teachers, a need which was identified by Dearing in a subsequent review of National Curriculum assessment in 1995. His follow-up report also contained a range of recommendations aimed at ensuring consistent assessment standards and maximizing the utility of the test results to external audiences.

Some degree of stability now appears to have been restored within the National Curriculum assessment arrangements, and educational authorities have accepted that national tests and teacher assessments 'have equal status and . . . provide complementary information about pupils' attainment' (Dearing, 1995). The implementation of National Curriculum assessment nevertheless graphically illustrates the difficulty of 'making assessment useful for teachers, but at the same time of fulfilling the government's objectives' (Nuttall and Stobart, 1994: 24).

2 The Primary Language Record

In contrast to National Curriculum assessment, which was heavily influenced by considerations of system accountability, the Primary Language Record (PLR) was introduced principally for purposes of assisting teachers to observe and monitor learners' progress.

Developed by the Inner London Education Authority (ILEA) in the UK beginning in the mid-1980s, and piloted for nearly a year in more than 50 schools, the PLR was intended as a record-keeping procedure for language and literacy teachers in primary education. It consists of a formal record called the Main Record, on which children's progress is recorded at a particular point during the year but which remains confidential within the school; and an informal record, entitled Observations and Samples, in which teachers keep their ongoing assessments based on classroom observation. These observations form the basis for the Main Record. There is also provision for recording the outcomes of teacher–parent and teacher–student conferences about students' progress (Barrs *et al.*, 1988).

A good deal of interest was shown in the PLR, although it made heavy demands on teachers in terms of the time it took to fill in the records. Myra Barrs, one of the creators of the record, comments that 'it does seem to be the case that it takes a full school year to "learn the forms", to internalise the ways of observing that they encapsulate and to see full value of this kind of recording' (Barrs, 1992: 56).

The PLR was gathering considerable momentum, when in 1992 the ILEA was abolished and National Curriculum assessment was introduced. Although teachers continued to use the PLR, the day-to-day assessment tasks became subordinated to the Standard Assessment Tasks (SATs) which accompanied the National Curriculum. According to Hilary Hester, another member of the PLR team, this shifted what was originally a system for ongoing record-keeping in the direction of more formal summative testing. Asked whether the PLR was compatible with the National Curriculum assessments, Hester (1992: 7) observed:

> It's not incompatible, no, but the problem is that the standards assessment is having a huge influence on the way teachers are seeing assessment, as it then becomes something that's set up – it's special, it's not part of a normal context, so it becomes assessment tasks, within which you have ticks for what students can do and crosses for what they can't do. You can see very clearly the kind of difference that's involved by looking at our reading scales, because we included a lot of -ing verbs in the descriptions, such as 'tackling known and predictable texts with growing confidence' or 'needing support', 'growing ability to predict meanings', 'developing strategies'. Well, within the National Curriculum they might not use these words, but it would be 'can predict meanings', 'are using strategies'. That's all having an effect on the way people are looking at assessment.

Hester (1992: 8) also noted that, as a result of National Curriculum assessment, teachers began to use the PLR to assess the effectiveness of programmes against the stages of English learning set out in the national attainment targets, even though its original purpose was to help teachers to diagnose needs and plan their teaching better.

3 The NLLIA bandscales

In late 1991, the National Languages and Literacy Institute of Australia (NLLIA) was commissioned by the Commonwealth Department of Employment, Education and Training (DEET) to develop an assessment and reporting package which could be used to describe the progress of learners of English as a second language in schools from K-12. This project arose in response to a long-felt need for a mechanism that could be used to assist teachers in making decisions on learner progress and achievement and at the same time serve as a reporting tool (McKay, 1994: 1). Drawing on a survey of a wide range of relevant literature and extensive consultation with ESL specialists, applied linguists, practitioners and researchers, the project developed a set of bandscales accompanied by exemplar assessment activities, observation guidelines and reporting formats to enable teachers to describe and report progress relative to the bandscales (see Appendix 2 for an extract from one of the bandscales). The scales are 'context-embedded' (McKay, 1995a) in so far as they seek to describe learner development within the mainstream learning context from junior primary to upper secondary. In order to reflect differences between these contexts, three separate bandscales were developed for junior primary, middle/upper primary and secondary learners.

4 The CURASS ESL scales

In the meantime, the Australian Education Council (AEC) – consisting of education ministers from the Federal and State governments – was formed in order to work towards the formulation of a national education programme and to conduct curriculum mapping exercises across states. National Curriculum profiles were developed for eight key learning areas under the aegis of the Curriculum and Assessment Committee (CURASS) of the AEC. In the wake of these developments, national ESL scales were commissioned, which had to conform to the format of scales used in other learning areas. The CURASS ESL scales which were subsequently developed differed from the NLLIA bandscales in a number of significant ways (McKay, 1994; 1995; Moore, 1996). First, they described outcomes of an assumed separate ESL curriculum, using a single set of scales for primary and secondary learners, rather than being anchored in the mainstream context. Second, they collapsed Listening and Speaking into a single category of Oral Interaction. Third, the levels in the CURASS scales described successful task performance rather than developing skills and knowledge (as in the case of the PLR referred to above). Fourth, they did not distinguish between the use of English in social and academic contexts. Fifth, they assumed uniform progress along a unidimensional scale: ‘The scales assume that the particular nature and order of outcomes will apply to all ESL learners regardless of age, grade and literacy background. Differences in learners will be reflected in their rate of attainment of outcomes and the outcome levels they ultimately attain’ (Curriculum Corporation, 1994: 8–9).

The CURASS scales are intended to be used for summative reporting of learner achievement – that is, for checking on progress at particular times by assigning a level on the basis of teacher-conducted observations and assessments (1994: 11). At the time they were published, they were not accompanied by specific assessment guidelines or tasks, although DEET subsequently allocated funding for professional development aimed at assisting teachers to use the scales and to develop accompanying assessment activities.

According to McKay (1994: 19), one of the authors of the NLLIA scales, the existence of the two scales aimed at the same target population brought about a ‘strong sense of confusion’ in teachers. Although both continue to be used, the fact that the CURASS scales are the property of the Commonwealth and State educational authorities and supported with resources for in-servicing means that they have received more public exposure and are thus likely to be more widely adopted than the NLLIA scales (Moore, 1996).

5 The Adult Literacy and Numeracy (ALAN) scales

The Adult Literacy and Numeracy (ALAN) scales were also the result of an initiative by DEET in Australia. They were intended to provide a reporting framework for practitioners to describe literacy and numeracy behaviours in government-funded literacy and numeracy programmes for adults. Five scales describing different levels of literacy and numeracy ability were developed – Reading, Writing, Quantitative Information Processing, Measurement and Basic Operations. The descriptors of literacy and numeracy competencies were derived from statements made by teachers about tasks and achievements that were considered to be relevant for adult literacy and numeracy students at a range of ability levels. Using the Rasch rating scale model, the descriptors were placed along a hypothetical continuum which aimed to describe a unidimensional model of literacy and numeracy. After the scales had been distributed to practitioners and experts for editing and comment and revised accordingly, they were subsequently field-tested in a range of different contexts by asking practitioners to rate the behaviour of learners. This exercise revealed inconsistencies in the way the scales were being applied, which led the scale developers to recommend that ‘extensive work [needed] to be carried out on the validity and reliability of judgements associated with the use of the scales' (Griffin and Forwood, 1991: 42).

Although the project team reported that the feedback they received during the development process on the theoretical basis and content of the scales was largely favourable (Griffin, 1991: 8), the scales were not universally embraced by the language teaching community. In particular they met with a scathing response from the Australian Council on Adult Literacy (ACAL), the professional association of Australian adult literacy teachers, which condemned the scales on the grounds that:

1) The methodology used to develop the scales was flawed, with consultation being used as 'little more than a data collection technique' and major theoretical approaches to the teaching of reading and writing being ignored (Australian Council on Adult Literacy, 1992: 3).
2) The scales represented a view of literacy which was 'both outmoded and restricted and translates into descriptors and notions of development that account for only one aspect of the performance, a form of individual mental processing' (p. 5).
3) The scales were divorced from any curriculum context and failed to recognize the diversity of needs in adult literacy and numeracy teaching (p. 5).
4) The scales were unusable since they were not accompanied by assessment tasks which would enable teachers to locate learners' level on the scales; no training and moderation procedures had been developed for raters and no procedures for monitoring the use of the scales had been put in place (p. 6).

In the face of concerns such as these, the government commissioned a team of academics to undertake a review of the scales. Following an extensive survey of users, large-scale trialling in a range of different sites and an examination of their content and theoretical basis, the review team identified a number of theoretical, empirical and practical problems with the ALAN scales. They concluded that a negative reaction by almost half the sample of participants in the review consulted 'constituted sufficient grounds for grave concerns over the implementability of the scales' (Freebody *et al*., 1993: 171) even if they were to be adapted. In addition, they noted 'serious and warranted concerns over the consistency of judgements associated with the use of the scales' (p. 171) and expressed doubts about their construct validity, questioning the appropriacy of the view of literacy learning which underlay the level descriptors: 'One question that then presents itself immediately is the extent to which judgements based on the scales depend for their reliability, validity and institutional intelligibility on the continued currency, or even dominance, of the Psycholinguistic approach to literacy education' (p. 58).

Finally, the review recommended that the use of the ALAN scales be discontinued and that DEET pursue the development of reporting frameworks 'by channelling and co-ordinating the efforts of currently uncoordinated groups working on competencies and on adult literacy and numeracy provisions' (p. 174). The scales were subsequently withdrawn and two major projects were commissioned: one to develop a national framework for describing adult English language, literacy and numeracy competence; and a second to design a system for reporting the outcomes of adult language, literacy and numeracy programmes, thus duplicating, at great public expense, the goal that the ALAN scale project had originally set out to achieve.

IV When State standards meet formative assessment: issues and problems

Although the various assessment and reporting schemes described above were developed in different contexts for different audiences, a number of common issues and problems emerge from the rather troubled history of their development and implementation. Specifically these issues relate to:

- the purposes and intended use of outcome statements;
- the properties of the statements themselves and the way they are interpreted and used;
- the means by which the statements are put into practice 'on the ground'.

For the sake of the following discussion, I will categorize these as *political*, *technical* and *practical* issues. In this section I want to examine some of the difficulties which arose on each of these fronts and to try to draw out some general principles which might help to explain their origins.

1 Political issues

1) Formative assessment and summative reporting are different in purpose

The fact that both the PLR and the NLLIA bandscales were supplanted by nationally mandated outcomes-focused schemes developed primarily for accountability purposes suggests that the information provided by ongoing teacher-conducted observation and assessment cannot easily be used by educational authorities for summative reporting and evaluation. Moreover, the political reality seems to be that when there are two competing assessment schemes, system information needs will override those of formative assessment. In this regard, Barrs (1990: 45–46), reflecting on the TGAT's attempt to create a system of assessment for the UK National Curriculum 'that would meet the needs both of formative and of summative and evaluative assessment', concludes that:

> It seems clear that this compromise is unlikely to work: formative and summative assessment cannot possibly draw on the same information because their needs are very different and, where there is a conflict, it will always be the needs of the summative and evaluative assessment that prevail.

Although outcomes-based assessment schemes with a reporting focus look similar in many ways to classroom-focused systems, there are a number of key differences in the way they are structured which reflect their different purposes. As both the PLR and the NLLIA bandscales demonstrate, one of the principal features of assessment schemes that have been developed to support teacher observation and record-keeping is that is that they are 'context-embedded' (McKay, 1995a: 12) in the everyday world of the classroom and 'developmental' in the sense that they encourage teachers to observe developing, rather than 'terminal' behaviours. On the other hand, the schemes which replaced them, because they are aimed at providing system-level information on outcomes, are decontextualized, emphasizing successful task completion in relation to externally defined standards or stages of attainment and reporting achievement numerically on a unidimensional scale, thus reducing their utility as tools for classroom-based observation and assessment. As a result, formative recording which might formerly have been done for diagnostic purposes tends to become a series of summative observations of learners' attainment of outcomes (Hester, 1992; McKay, 1995a; Moore, 1996), provoking negative reactions from teachers if they feel that assessment is taking away teaching time.

2) Different audiences have different information needs

The difference in purpose between formative assessment and summative reporting is also reflected in the different forms in which information is required by users. Some bandscales or profiles used for formative assessment tend to describe individual learner achievement in quite complex ways so as to capture the nature of classroom learning. However, key consumers of assessment information for accountability purposes, such as policy-makers or educational bureaucrats, usually require information on outcomes which needs to be in simplified and aggregated form in order to assist them in making decisions on resourcing and to monitor programme performance. In the words of Ruby (1994: 9), ‘The characteristics of “good” information in the eyes of policy makers is information that is simple, comparable and timely.’ The level of detail that is provided by diagnostically oriented ongoing assessment schemes may be neither necessary nor interpretable for those audiences who require information on outcomes in summary form. In this regard, the history of testing unfortunately brings with it the strong expectation that achievement can be reported as a single score or grade and some audiences – for better or for worse – may continue to value assessments which can be presented in this form over more complex and qualitative descriptions of progress and achievement.

The challenge of using performance assessment as the basis for summative reporting for accountability purposes is to try to ensure that the information which is provided is not on the one hand so reduced through aggregation that its diagnostic value is lost or, on the other hand, so complex and technical that some users are unable to interpret it. It goes without saying that this constitutes a significant challenge. This point will be taken up in Section V.

3) Test appearance is a key factor affecting implementation

The controversies surrounding the tests used in National Curriculum English assessment and the debates over the content of the ALAN scale descriptors suggest that test appearance is a crucial factor in determining the extent to which outcome statements can be successfully implemented. If the theoretical underpinnings of the statements or the testing formats used are seen to be at variance with the strongly held views of powerful interest groups representing particular theoretical or pedagogical orientations, then their validity may well be publicly challenged, thus greatly reducing the likelihood of their adoption by practitioners and significantly increasing the risk of industrial action, especially if the work demands of the new system are thought to be unreasonable. The experiences described above thus suggest that it would be unwise to downplay test appearance.

2 *Technical issues*

1) The validity of level descriptors is an ongoing source of concern

The various assessment and reporting systems described in the previous section brought to light a range of concerns to do with the validity of outcome statements. In addition to the problems of test appearance alluded to above, issues of content and construct validity were also raised in relation to level or band descriptors. In this context, critics have pointed out the inadequacies of some of the ‘consultative’ methods for developing scales outlined earlier in this article.

Thus Barrs (1990: 44), commenting on the 'top–down consultative' method used to develop the statements of attainment in the UK National Curriculum, writes:

> The criteria represented by the statements of attainment of the national curriculum are not based on sound evidence of what children can do at different ages and stages; they represent a mixture of what the curriculum working parties thought was likely to represent the typical achievements of a particular age group, and what we might term 'pious hopes' – notions of what children ought (according to the working parties) to be able to do at these levels.

The clarity and internal coherence of level descriptors have also been the subject of criticism. Wolf (1991: 198), for example, notes that although the attainment targets in the National Curriculum were intended to convey clear standards, 'experience suggests the reverse. (Outcome) statements are decontextualized (and so would apply to many levels of difficulty). They are nonetheless assigned to one particular level.' In a similar vein, it has been argued that the standards described in some general language proficiency rating scales lack an empirical foundation since they are derived from experience and intuition rather than research (Brindley, 1991; Fulcher, 1996). The results of some research studies which have set out to investigate the relationship between level descriptors and actual language use in fact suggest that the constructs described are at odds with data derived from actual learner performance (Lee and Musumeci, 1988; Fulcher, 1987; Lantolf and Frawley, 1988).

Problems with level descriptors have also been attributed to a misplaced degree of confidence in the power of statistics. Here there is some evidence to suggest that teachers may be unwilling to accept scale descriptors which group together what appear to be heterogeneous phenomena in the same level, even though there may be statistical evidence of fit to a unidimensional scale. In this regard, McKay (1995b: 38) argues that the outcomes of the initial process used to develop the NLLIA bandscales (similar to that used in the ALAN scale project) lacked validity in the eyes of many teachers 'in part because learner characteristics (age, maturity, background knowledge etc.) and context (the nature of the task, degree of interlocutor support etc.) were not being adequately addressed when individual indicators were calibrated and grouped through a statistical process'. As a result, the descriptors which were subsequently developed included 'more integrated and context-related descriptions'.

The ALAN scale descriptors were questioned on similar grounds. In a critique of the psychometric characteristics of the scales, Cooksey (1993: 32) argued that many of the literacy and numeracy competencies which were grouped together on a unidimensional scale using the Rasch model were in fact tapping different abilities, thus reducing the utility of the model for scaling the competencies on the dimension of interest and diminishing the interpretability of scale levels.

2) Scores or ratings derived from a variety of different teacher-generated assessments of unknown validity and reliability are potentially invalid.

Cooksey (1993) also draws attention to a fundamental problem concerning the relationship between assessment and reporting in the ALAN scales. This point is of particular importance since it applies equally to all outcomes-based systems which attempt to map teacher assessments on to a system-wide reporting framework:

> The ALAN scales purport to yield a common reporting scheme for test results independent of the nature of the assessment procedure (henceforth referred to as indicator devices) used to obtain the scores. *Yet there is no objective and verifiable way of knowing how a particular score (be it a test score or ticked observation category) on an indicator device should be translated into a position in one of the bands or levels on a particular ALAN scale*. The process necessarily relies on human judgement to make the translation and there are no rules available to guide such judgements'. (Cooksey, 1993: 30, emphasis in original).

Without explicit procedures for relating teacher observations, assessments and tests directly to the outcome statements (or vice versa for that matter), the risk of misclassifying learners, according to Cooksey (1993: 35), becomes unacceptably high.

3) Subjective judgements of language performance are likely to show a good deal of variability

One of the major causes of variability in judgements of performance is unclear standards or level descriptors, as Gipps (1994: 141) points out in relation to the UK National Curriculum attainment targets. However, even in cases where assessment criteria are clearly specified and detailed rating schemes devised, rater consistency still seems to be a problem. In this respect, Griffin and Forwood (1991: 42) found considerable variation amongst raters in their application of the ALAN scale, thus undermining the validity of the assessment information provided. Not surprisingly, the same issue has arisen with the introduction of the CURASS scales, exemplified by the following observation by an ESL teacher on her attempts to use the scales in the classroom:

> How do I know that my interpretation of my students' presentations is what I assign rather than a reflection of my understanding of terminology and task, which may be subtly or clearly different from my similarly busy and therefore only occasionally consulted colleagues down the road? If there's no sense of

> consistent standardisation, how validly are we really applying the scales? (Woods, 1995: 30).

This teacher's concerns are borne out by a considerable body of research which suggests that assessment criteria will be interpreted differently by different audiences according to their previous experience, unconscious expectations and subjective preferences regarding the relative importance of different communicative criteria (see, for example, North, 1993; Gipps, 1994; Lumley and McNamara, 1995; Brown, 1995; Chalhoub-Deville, 1995).

3 *Practical factors*

1) Politically driven assessment and reporting systems which are perceived to make unreasonable demands on teachers' time are likely to be resisted.

In the final analysis it was logistical factors which brought about the downfall of the National Curriculum assessments. The sheer complexity and number of assessment tasks proved unmanageable and had already prompted a shift towards the adoption of more SATs even before the crisis of 1993. In Australia, the introduction of the CURASS scales in Australian schools has already provoked similar concerns, illustrated by the following comment from the teacher quoted above:

> As a learner to the process I have spent hours appraising one student in one strand. The passage of time and experience will inevitably streamline this process yet if the intentions of the scales and their consequent reporting are to be validly met, I wonder how we will find enough time to make it so without eroding the quality and creativity of face to face teaching with our students (Woods, 1995: 31).

However, it is not necessarily the case that teachers will reject an assessment system on the grounds that it is time-consuming. In this context, it is interesting to note that, although the PLR substantially increased teachers' workload, they were reportedly prepared to devote extra time to keeping detailed records because of the perceived value of the information to learners, teachers and parents and the contribution of the PLR to their own professional development (Barrs, 1992; Hester, 1992).

2) Politically imposed deadlines for the implementation of new assessment systems may militate against the development of necessary support resources

One of the most commonly voiced criticisms of the UK National

Curriculum concerns the haste with which changes in curriculum and assessment were introduced and the lack of time for consultation with the various stakeholders (Elliott, 1994; Edwards, 1995). The same point was made in relation to the ALAN scale project (Australian Council for Adult Literacy, 1992: 3). Unrealistic timelines may also preclude careful piloting of assessment tools and accompanying professional development resources to assist teachers in interpreting and applying outcome statements. As a consequence, official outcome statements which are issued in the form of glossy finished products but which are unaccompanied by such resources may be deemed to be unusable. This was the case with the ALAN scales (Australian Council for Adult Literacy, 1992: 6), even though the developers were aware of the need for further work and made a number of recommendations to this effect on completion of the project.

V Learning from history

The history of the assessment and reporting systems described in this article represents an unfortunate saga of duplication and wastage of resources, accompanied in some cases by considerable professional and industrial conflict. Nevertheless educational authorities around the world continue to embark on the development of similar systems and in the current economic climate there is every reason to suppose that such schemes will proliferate as governments increase pressures on educational institutions to demonstrate accountability by adopting outcomes-based reporting.

In this section, I will outline some strategies based on the previous analysis which might be adopted on the political, technical and practical fronts in order to avoid repetition of some of the mistakes of the past. At the same time I will offer some suggestions which might assist in alleviating the tensions between the potentially conflicting purposes of formative assessment and summative reporting.

1 The political front

1) Adopting a broader view of evaluation and accountability

The preoccupation of governments with the establishment of national assessment and reporting systems can be traced back to the heavy emphasis which is traditionally placed on aggregated student outcomes as a measure of programme performance. As Taylor (1994: 255) observes, ‘the daily language of educators, policymakers, parents and students is comparative’, a fact which continues to be reflected in the publication in some countries of ‘league tables’ comparing school performances on the basis of public examination results. However, numerous educators over the years have argued strongly that student outcomes are only a partial and potentially misleading indicator of the overall merit of the programme (e.g., Kemmis and Stake, 1988; Broadfoot, 1992; National Forum on Assessment, 1992). Serious evaluation of programme effectiveness demands the collection of a far more comprehensive range of information about the operation of the programme. In the words of Nevo (1995: 270):

> In evaluating an educational object, outcomes or impacts should not be the only thing looked at. To get a good understanding of an educational evaluation object – be it a student, a teacher, a project or a school – we must also examine needs, goals, strategies, actual processes of implementation as well as the socio-political milieu of the object.

If the aim of monitoring and evaluating programme performance is not only to account for the use of resources but also to improve the quality of programme delivery, then educational authorities need to be persuaded that efforts must be directed to collecting these other types of information.

2) Clarifying information needs through genuine consultation between stakeholders

If an assessment and reporting system is to meet the varied information needs of multiple audiences, it is imperative that each of these audiences has a clear understanding of exactly what the purposes of the system are, how the results are to be used and what its effects on student learning are intended to be. At the same time, the potential practical consequences of assessment reform for the daily workload of both educational administrators and practitioners need to be carefully considered and worked through. Serious misunderstandings can arise when the purpose of assessment is unclear and there is no shared understanding of the benefits of the proposed changes to either teachers or learners. In order to clarify the information needs of all the stakeholders involved in language programmes, therefore, there needs to be dialogue and consultation between politicians, policy-makers, assessment specialists, administrators, practitioners and relevant others *before* new assessment systems are put in place, not after the political decision has been taken to implement the system. It would be very unusual for a company these days to put a new product on the market without doing any market research – yet in many cases major assessment reforms are put in place by government fiat without consideration of their likely effects. In this context, it is worth noting the potentially important role that professional associations can play in facilitating consultation by providing opportunities (for example, through public forums or targeted briefings) for policy-makers to receive input from the profession at large, rather than relying on advisers who may not be fully aware of the range of views and practices within the field.

Genuine consultation also means listening to those administrators and practitioners who will be responsible for putting the system into practice 'on the ground'. The considerable body of literature on educational change management suggests that if these people have no sense of ownership of the proposed reforms, they cannot be expected to be willing participants in the implementation process (Fullan, 1982; Fullan and Stiegelbauer, 1994). On this point, it is worth noting the following observation by Sir Malcolm Thornton, Conservative MP and Chairman of the House of Commons Select Committee on Education in the British Parliament. Reflecting on the set of circumstances leading to the teacher boycott of the National Curriculum assessment tasks, he comments:

> it seems to me self-evident that secretaries of state do not deliver educational reform, they pass laws. The education reform is delivered in a classroom, by a teacher and therefore the workability factor is something that has got to be looked at very, very carefully. I think that element of consultation, this sublime belief which people at top level have, that what they are doing is right, tends to be somewhat simplistic. By the time the new policy is handed down through various committees, it really does start to create enormous problems as far as practitioners are concerned. And if they have not been consulted then this question about the lack of that feeling of ownership is very much there. That, really, I think underpins the whole reason why Dearing had to be brought in to fill that vacuum' (Gilroy *et al.*, 1994: 274).

The problems experienced in implementing the National Curriculum in the UK seem to have brought about a realization that open discussion and sharing of assumptions, philosophies and goals are prerequisites to the introduction of major educational change. As a result, politicians have been more willing of late to participate in open forums and debates on educational issues (cf. Gilroy *et al.*, 1994). At the same time, in the UK and in other countries as well, educational policy-makers and bureaucrats are starting to make useful contributions to professional literature from a policy perspective (see, for example, Ruby, 1994). These are encouraging trends which can only improve communication between educational decision-makers and professionals and in the process assist in breaking down the traditional barriers which have separated the world of policy from the world of practice.

While this type of consultation and exchange can assist in clarifying the purposes of assessment and identifying information needs more precisely, the important question remains of how (or if) outcomes can be reported in aggregated form without losing or distorting key assessment information. Since this question is a technical issue as much as a political one, it will be addressed in the discussion which follows.

2 The technical front

1) Addressing validity issues

The debates surrounding the validity of outcome statements are to a certain extent inevitable since the theories (either implicit or explicit) of those responsible for the development of outcome statements will largely determine the skills or competencies which are deemed to constitute the domain of interest, the terms in which they are described and, by extension, the nature of the criteria which are used to assess performance. Given the wide range of competing views of language ability and language use which are currently held, it is not surprising that outcome statements containing quite different definitions of the same construct sometimes end up in circulation at the same time, causing considerable confusion amongst users and sparking ongoing professional debate (Brindley, 1994; McKay, 1994).

The only way to break this circle is to try to build up a body of independent evidence on language ability and use through investigation of samples of actual task performance. As Messick (1994: 19) notes: ‘We need to move beyond traditional professional judgement of content to accrue construct-related evidence that the ostensibly sampled processes are actually engaged by respondents in task performance.’

This suggests the need for a programme of research which would contain a number of components. First, in order to provide a more secure empirical basis for defining bands of increasing ability, further research is required into the validity of the task and text hierarchies which are typically described in existing outcome statements, and in particular into the correspondence between scale descriptors and task performance (Alderson, 1991; Brindley, forthcoming). Such research would include content analyses of tests which are linked to scale descriptors (Hill, 1995), investigations of the processes employed by learners in undertaking assessment tasks (Alderson, 1990; Cohen, forthcoming) and studies of factors affecting task and text difficulty in language tests (Shohamy and Inbar, 1991; Hamp-Lyons and Mathias, 1994). To complement this kind of test-based research, qualitative analyses of authentic language use are needed to shed more light on the nature of the language skills involved in performing the communicative functions described in outcome statements (Fulcher, 1996; Lazaraton, 1996). The results of such studies can then be used as a basis for developing sets of indicators or performance criteria against which learning outcomes can be assessed (Rutledge, 1993; Turner and Upshur, 1995).

A range of statistical tools such as factor analysis, multi-trait multimethod analysis and structural equation modelling can also be employed to investigate empirically the extent to which rating scales and other forms of outcome statements can be shown to demonstrate trait convergence with other measures of the ability in question (Bachman and Palmer, 1981; Brindley and Ross, 1996).

2) Linking assessment tasks to levels of achievement

As I have noted previously, a major problem with systems which require teachers to map their assessments on to reporting frameworks is the inherent variation in the kinds of assessment tasks that are used to arrive at the achievement levels (Cooksey, 1993: 30). A reporting system based on a wide range of independently conducted teacher assessments of unknown validity and reliability cannot be expected to provide valid and dependable information on student outcomes. In high-stakes contexts, these inconsistencies could lead to unfair decisions which could adversely affect people's lives.

In order to address these potential problems, a number of strategies can be adopted. First, if assessment results are to be used for high-stakes purposes such as certification or selection, banks of assessment tasks need to be developed which are systematically linked to the achievement levels in whatever reporting framework is being used, along the lines of the SATs used with the National Curriculum. Using statistical techniques such as Rasch analysis, the tasks can be calibrated in terms of known levels of difficulty, though the caveats mentioned by Cooksey (1993) and Buck (1994) concerning the potential multdimensionality of language abilities need to be borne in mind (see also North, 1995b for an account of some of the difficulties which can be encountered when attempting to scale language performance data using the Rasch model). The standardized tasks can also, if necessary, be used as 'anchors' to statistically moderate teacher assessments (Linn, 1993).

So that teachers can become familiar with typical performances at different levels of ability, they need to have the opportunity to familiarize themselves thoroughly with achievement levels and assessment criteria. A library of exemplars of student performance needs to be built up and a system established whereby teachers are given the opportunity to participate in regular moderation sessions at which interpretations and applications of the assessment criteria are compared and discussed. Such sessions also serve a valuable professional development function as they encourage reflection on the nature of student learning (Gipps, 1994). At the same time, if outcome reporting is to be based on teacher-conducted assessments as well as standardized tasks, then teachers have to be able to attend appropriately targeted professional development workshops in order to acquire skills in designing and evaluating assessment tasks. All of this implies a considerable investment for educational authorities both in terms of time and financial resources.

3) Dealing with demands for aggregated information

The reporting of outcomes in the aggregated form required by some audiences is potentially problematic. The danger here is that complex and detailed assessment information might be collapsed into a reduced form, thus stripping it of any educational value and rendering it virtually meaningless (Barrs, 1990: 45; Gipps, 1994: 86).

This problem is partly a political issue and partly a technical one and needs to be simultaneously addressed on both fronts. In the first instance, consultations between key stakeholders of the kind described above would allow various consumers of assessment information to state the form in which they required it. At this point, various strategies for presenting system-level information could be illustrated and examined and their consequences discussed, so that the full range of possibilities was evident. As Taylor (1994: 254) states:

> The first step in this process [of developing an assessment system] is to make explicit to all decision makers the choices to be made and the assumptions that underlie each choice. Unless the underlying assumptions of the models are put in language all educators and policy makers can understand, the tensions between assessment for standards and assessment for measurement can result in a failure to achieve the changes expected from the use of performance-based assessments.

On the question of aggregation, Gipps (1994: 85) comments:

> There would be no need to aggregate to a single level or grade were it not for the requirement to report in a simple quantifiable form. It is much more in keeping with the educational assessment function of criterion-referenced assessment that attainment be reported in terms of a descriptive profile, as a list of tasks completed, or a series of stages reached.

Many of the outcomes-based systems described in this article do allow for reporting in one or more of these forms. For this reason, it could be argued that outcomes statements are able to capture achievement without losing key diagnostic information. However, the quality of this information will be quite variable if the system contains no procedures for linking assessment indicators to outcome levels. In addition, profiles, stages, competencies, etc. may involve further layers of complexity which in some cases could prove quite difficult to translate into a manageable form for external users. For example, the

CURASS scales state that ‘students may well be at different levels across and within the strands and organisers of the scale’ (Curriculum Corporation, 1994: 10). This implies that programme outcomes would be presented as a multidimensional profile, a format which some users may have difficulty in interpreting. Once again, issues such as this would need to be negotiated between stakeholders as early as possible in the development process in order to avoid misunderstandings. At this point, practical implications for information storage and retrieval at both classroom and system level would also need to be explored.

Given the complexities of collecting and presenting multidimensional information, educational authorities that wish to use outcome information for system-wide reporting or programme evaluation purposes need to consider ways in which such information can be collated and summarized. Since it is clearly not the job of the individual teacher to do this, some institutions might need to develop specific rules and procedures to ‘translate’ assessment information into accessible aggregated form for those who require it in order to report on programme performance. Such reporting might involve, *inter alia*, sampling and summarizing the achievements of different groups, districts or schools; preparing qualitative reports on progress and achievement both at individual and group level; or gathering data for purposes of comparison. Examples of how some educational systems have addressed the issue of reporting performance assessments for accountability purposes are outlined by Moss *et al*. (1992) and Rutledge (1993).

4) Dealing with variability in judgements of performance

The question of how to improve the consistency of subjective judgements is a complex one which has been the subject of much recent debate in the assessment literature. While rater training and moderation sessions are the most commonly suggested ways of addressing this issue, there is evidence to suggest that training is unable to eliminate variations in rater judgements (North, 1993; Lumley and McNamara, 1995). Nor does training appear to prevent different groups from giving different weightings to assessment criteria (Moss *et al*., 1992; Brown, 1995). Given the complexity of the interaction between the language behaviour being rated, the personal characteristics of both the rater and the candidate and aspects of the setting in which the rating takes place, these findings are perhaps not surprising. However it leaves assessors in a dilemma: if variability in rater behaviour is the norm, can anything be done to ensure that learners are not unfairly disadvantaged by inconsistent judgements?

There are a number of ways of approaching the question of rater variation. One is by using sophisticated analytical techniques such as many-faceted Rasch measurement and its accompanying software packages, *FACETS* (Linacre, 1989), which adjusts candidate ability estimates to take account of raters' tendency to rate either harshly or leniently. In this way learners are compensated for having an excessively 'hard' or 'easy' marker. This technology is being increasingly used in the context of high-stakes language tests to analyse performance data and to monitor rater behaviour (Wigglesworth, 1993; Lumley and McNamara, 1995). However, it would be unrealistic to expect that many educational institutions would have the resources to invest in the training and expense associated with the use of these procedures.

Another way of addressing the problem of rater inconsistency is by using various forms of multiple rating or scoring which incorporate two or more perspectives on the same performance(s), combined with some kind of moderation. Thus North (1993: 45) describes a procedure for oral assessment using two assessors – one who knows the class in question (high sensitivity) and one who is familiar with the whole range of the level (low sensitivity). Ratings are carried out independently using both holistic and analytical marking 'with negotiation over grades between the two assessors as a final step to adjust for severity' (p. 45).

Moss (1994) takes a somewhat different perspective, arguing that the idea of conventional high-agreement reliability needs to be reconceptualized. Reliability, she claims, is only one of a number of ways of justifying subjective judgements:

> I am not advocating the abandonment of reliability. Rather, I am advocating that we consider it one alternative for serving important epistemological and ethical purposes – an alternative that should always be justified in critical dialogue and in confrontation with other possible means of warranting knowledge claims.

Although it is a little difficult to see how these views would translate into practice, Moss's comments are a timely reminder of the need to collect assessment information from a variety of sources and strengthen the case for using procedures such as portfolio assessment which present a broad range of learner performances.

Gipps (1994: 171) suggests that instead of reliability in the sense of 'accuracy', we should use the term *comparability*, which should be based on consistency. High-stakes assessments would need to demonstrate this quality:

> Consistency leading to comparability is achieved by assessment tasks being presented in the same way to all pupils assessed; assessment criteria being interpreted in the same way by all teachers; and pupil performance being evaluated according to the same rubric and standards by all markers.

Gipps argues that this approach to the question of reliability is more consonant with outcomes-based approaches than traditional procedures which focus on 'true scores', but warns that 'it is very time and labour intensive and relies on the involvement of teachers' (p. 164), an issue to which I now turn.

3 The practical front

1) Planning for change

The considerable literature on educational change demonstrates that the implementation of innovation at the system level is a long, complex and difficult process which needs to be carefully planned if the change is to last (Fullan, 1982; Nicholls, 1983). This is no less true of innovations in assessment than it is of changes to syllabuses, methods or materials.

One of the most important insights to emerge from research into the way in which change happens in education is the crucial role of the teacher in implementation. This is summed up succinctly by Fullan (1982: 107): 'Educational change depends on what teachers do and think – it's as simple and as complex as that.' As some of the previous exercises in assessment innovation demonstrate, if the proposed changes are imposed hastily without adequate consultation, are unaccompanied by professional support and at the same time appear to involve a heavy time commitment, teachers are unlikely to participate willingly in their implementation. The situation is exacerbated if they cannot see obvious benefits of the change at the level of the classroom.

These problems can partially be addressed through ensuring that adequate consultation takes place at the initial stages when needs and policies are discussed. If, following these consultations between stakeholders, it seems that an innovation – no matter how politically imperative or technically sophisticated – is neither justifiable in terms of its benefit to learning or translatable into practice with the resources available – then there would have to be strong grounds for rejecting it. However, as Fullan (p. 115) points out, even though an innovation and its benefits may be rationally explained at the beginning, 'it cannot be absorbed for teachers like anyone else do not learn new ideas all at once. Change is a process not an event.' For this reason, writers on change management emphasize the need to provide ongoing support to practitioners in the form of professional development and opportunities for interaction with colleagues throughout the life of a major innovation (Fullan and Stiegelbauer, 1991). In this way, they are able to work through the implications of the change for their daily practice. Careful piloting, over a long period if necessary, plays a critical role during the initiation phase in ensuring that the innovation can be adapted to meet classroom realities (Barrs, 1992: 54).

An example of the way in which changes to an assessment system can be managed at an institutional level is provided by Bottomley *et al*. (1994) in the context of the Australian Adult Migrant English Program. They describe the introduction of an outcomes-based curriculum framework (in this case the CSWE) into Adult Migrant Education Services (AMES) Victoria, an organization which had previously allowed teachers a good deal of latitude in determining curriculum objectives and in which learner achievement had been assessed using a 'curriculum-free' proficiency rating scale. Drawing on principles of change enunciated by Fullan and Stiegelbauer (1991), they show how carefully managed support to practitioners was used to facilitate the introduction of a competency-based assessment system which required teachers to assess and report against prespecified learning outcomes.

Yet it appears that not even the provision of high levels of institutional support will guarantee the universal adoption of new forms of assessment. In a follow-up study aimed at investigating the extent to which the CSWE had been adopted over a two-year period, Bamforth and Grieve (1996) found that tensions still existed between, on the one hand, teachers' desire to cater for individual needs and differences and, on the other, the institutional requirement to teach and assess specific competencies. They suggest, however, that it may be possible to reconcile these conflicting demands if the curriculum and the accompanying assessment procedures are open to constant critique and review.

VI Conclusion

In this article I have attempted to examine the nature of outcomes-based assessment and reporting systems with a particular focus on language education and to look at the problems which may arise in their development and implementation through the presentation of a number of 'case studies' of assessment innovations in the UK and Australia, all of which experienced problems at the political, technical or practical level.

One of the messages from these innovations is that reconciling the purposes of formative assessment and summative reporting is going to be very difficult to achieve. It seems clear, as Gipps (1994: 14) notes, that 'assessment for formative purposes has quite different properties and qualities from that used summatively for accountability purposes'. For this reason, some have argued that the two should be kept separate and distinct (e.g., Barrs, 1990). However the danger of such a separation is that it could lead to a return to test-based 'curriculum-free measurement' and single-score reporting. Though this might be an attractive option to some audiences, it would be at variance with many of the principles which underlie the widespread adoption of performance assessment. If the goal of assessment is to provide information that can be used to improve the quality of instruction, then the outcomes which are reported should be able to be related to the learning which actually occurred. If some audiences are either unwilling or unable to deal with the type of detailed qualitative information that is yielded by performance assessments, procedures are required by which data on individual achievement can be translated into a reduced form that is nevertheless meaningful in terms of the learning that took place. Some systems – for example, those that use tasks or competencies as the unit for describing outcomes – are able to do this by providing summary reports listing the competencies achieved. These can then be related to fuller descriptions of each competency, thus enabling the users of information to build up a detailed picture of what was learnt in a record of achievement. Other examples of ways in which information from performance assessments has been used for accountability purposes can be found in the Toronto benchmarks project in Canada (Rutledge, 1993), the General Certificate of Secondary Education (GCSE) in the UK (Gipps, 1994), the Victorian Ministry of Education Literacy Profiles in Australia (Griffin, 1994) and the Vermont portfolio project in the USA (Moss *et al*., 1992).

However, the quality of the information provided on outcomes will only be as good as the assessments on which the reporting is based. For this reason it is important, particularly in high-stakes contexts, to ensure that quality control procedures are put in place for the development and administration of assessments. This means providing adequate professional development, materials development, rater training and systems for moderation to ensure comparability. At the same time, a good deal of research will need to be undertaken into the unresolved technical and practical questions which continue to surround performance assessment and outcome reporting systems. These include the validity of outcome statements, the consistency of judgements of performance and the feasibility of adding further assessment responsibilities to practitioners' day-to-day teaching loads.

Some of the political difficulties which have been experienced in implementing assessment and reporting systems can be mitigated through improving channels of communication between decision-makers and language professionals and by careful planning of the proposed changes. In cases where state-mandated assessment reform is governed by short-term electoral cycles and political imperatives, however, it is probably unrealistic to expect large-scale investment on the part of educational authorities in long-term planning and support systems. But if top–down change continues to be introduced in an unplanned way with inadequate consultation and resourcing, there is considerable evidence from studies of educational innovation to suggest that the kinds of controversies I have described in this article can be expected to recur (Fullan and Stiegelbauer, 1991).

Finally, although there has been a good deal of recent discussion and debate concerning the merits and shortcomings of different outcomes-based approaches, to date there have been relatively few attempts to explore their impact on day-to-day teaching and learning. Little is known about how information on learner outcomes is being collected, interpreted and used at either classroom or system level. Nor have the reactions of teachers, learners and external stakeholders to the introduction of national standards, benchmarks, competencies, etc. been documented. It is only through systematic investigation of such questions, along with ongoing empirical research into the validity and consistency of the assessment instruments themselves, that it will be possible to gauge the extent to which outcomes-based assessment and reporting can assist in improving the quality of learning.

Acknowledgement

This article is a revised version of a plenary presentation given at the International Conference on Testing and Evaluation in Second Language Education, Hong Kong University of Science and Technology, 21–24 June 1995. I would like to thank two anonymous reviewers for their helpful comments on an earlier version.

VII References

Alderson, J.C. 1990: Testing reading comprehension skills (part two). *Reading in a Foreign Language* 7 (1), 465–503.

—— 1991: Bands and scores. In Alderson, J.C. and North, B., editors, *Language testing in the 1990s*, London: Macmillan, 71–86.

Australian Council for Adult Literacy 1992: *The Adult Literacy and Numeracy scales: the ACAL view*. Northcote, Victoria: Australian Council for Adult Literacy Inc.

Bachman, L.F. and **Palmer, A.S.** 1981: A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading. In Palmer, A.S., Groot, P.J.M. and Trosper, G., editors, *The construct validation of tests of communicative competence*, Washington, DC: Teachers of English to Speakers of Other Languages, 149–65.

—— 1996: *Language Testing in Practice*. Oxford: Oxford University Press.

Bamforth, J. and **Grieve, E.** 1996: *From implementation to continuation: a case study of the institutionalisation of a curriculum innovation*. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Barrs, M. 1990: *Words not numbers: assessment in English*. Exeter: Short Run Press.

—— 1992: The Primary Language Record: what we are learning in the UK. In Bouffler, C., editor, *Literacy evaluation: issues and practicalities*, Sydney: Primary English Teaching Association, 53–62.

Barrs, M., Ellis, S., Hester, H. and **Thomas, A.** 1988: *The Primary Language Record*. London: Centre for Language in Primary Education.

Blackmore, J. 1988: *Assessment and accountability*. Geelong: Deakin University Press.

Bottomley, Y., Dalton, J. and **Corbel, C.** 1994: *From proficiency to competencies: a collaborative approach to curriculum innovation*. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Brindley, G. 1989: *Assessing achievement in the learner-centred curriculum*. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

—— 1991: Defining language ability: the criteria for criteria. In Anivan, S., editor, *Current developments in language testing*, Singapore: Regional Language Centre, 139–64.

—— 1994: Competency-based assessment in second language programs: some issues and questions. *Prospect* 9 (2), 41–55.

—— forthcoming: Describing language development? Rating scales and second language acquisition. To appear in Bachman, L.F. and Cohen, A.D., editors, *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press.

Brindley, G. and **Ross, S.** 1996: Trait-method comparisons across three language test batteries using exploratory, MTMM and structural equation modelling approaches. Unpublished manuscript, Macquarie University.

Broadfoot, P. 1992: *A question of quality: the changing role of assessment in education*. Workshop Report no. 4 Belconnen, ACT: Australian Curriculum Studies Association.

Brown, A. 1995: The effect of rater variables on the development of an occupation-specific language performance test. *Language Testing* 12 (1), 1–15.

Buck, G. 1994: The appropriacy of psychometric models for testing second language listening comprehension. *Language Testing* 11 (2), 145–70.

Chalhoub-Deville, M. 1995: Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12 (1), 16–33.

Citizenship and Immigration Canada 1993: *Language benchmarks: English as a Second Language for adults*. Ottawa: Citizenship and Integration Policy Division, Citizenship and Immigration Canada.

Cizek, G.J. 1993: Some thoughts on educational testing: measurement policy issues into the next millenium. *Educational Measurement: Issues and Practice* 12 (3), 10–22.

Clark, J.L., Scarino, A. and **Brownell, J.A.** 1994: *Improving the quality of learning: a framework for target-oriented curriculum renewal in Hong Kong*. Hong Kong: Hongkong Bank Language Development Fund/Institute of Language in Education.

Cohen, A. Forthcoming: Strategies and processes in test-taking and SLA. To appear in Bachman, L.F. and Cohen, A.D., editors, *Interfaces between second language acquisition and language testing research*, Cambridge: Cambridge University Press.

Cooksey, R. 1993: Psychometric characteristics of the Adult Literacy and Numeracy scales: observations and commentary. In Freebody, P., Cumming, J., Falk, I., Muspratt, S., Doyle, S., Flaherty, J. and Lee, N., *A trialing and evaluation of the Adult Literacy and Numeracy (ALAN) scales*. Volume 2, Appendices. Nathan, Queensland: Faculty of Education, Griffith University.

Curriculum Corporation 1994: *ESL scales*. Carlton, Victoria: Curriculum Corporation.

Dearing, R. 1995. *Review of assessment and testing: report from the School Curriculum and Assessment Authority to the Secretary of State for Education and Employment*. London: School Curriculum and Assessment Authority.

de Jong, J. 1995: The need for standards in language education. *System* 23 (4), 441–44.

Dunbar, S., Koretz, D. and **Hoover, H.** 1991: Quality control in the development and use of performance assessments. *Applied Measurement in Education* 4 (4), 289–304.

Edwards, J. 1995: British league tables and lessons for Australia: assessment, quality and social justice. *Curriculum Perspectives* 15 (1), 9–18.

Elliott, J. 1994: The teacher's role in curriculum development: an unresolved issue in English attempts at curriculum reform. *Curriculum Studies* 2 (1), 43–69.

Fitz-Gibbon, C. editor, 1990: *Performance indicators*. Clevedon, Avon: Multilingual Matters.

Forster, M. 1994: DART: assisting teachers to use the English profile. In Warhurst, J., editor, *Assessing and reporting in teaching and learning: implementing the national profiles*, Belconnen, ACT: Australian Curriculum Studies Association, 34–47.

Freebody, P., Cumming, J., Falk, I., Muspratt, S., Doyle, S., Flaherty, J. and **Lee, N.**, 1993: *A trialing and evaluation of the Adult Literacy and Numeracy (ALAN) scales*, Volume 1, Nathan, Queensland: Faculty of Education, Griffith University.

Fulcher, G. 1987: Tests of oral performance: the need for data-based criteria. *English Language Teaching Journal* 41, 287–91.

—— 1996: Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13 (2), 208–38.

Fullan, M. 1982: *The meaning of educational change*. Toronto: Ontario Institute for Studies in Education.

Fullan, M. with **Stiegelbauer, S.** 1991: *The new meaning of educational change*. London: Cassell.

Gilroy, P., Price, C., Stones, E. and **Thornton, M.** 1994: Teacher education in Britain: a *JET* symposium with politicians. *Journal of Education for Teaching* 20 (3), 261–300.

Gipps, C. 1994: *Beyond testing*. London: The Falmer Press.

Griffin, P. 1991: The Adult Literacy and Numeracy (ALAN) scales: validity issues in assessment and reporting. Melbourne: Assessment Research Centre, Phillip Institute of Technology.

—— 1994: Teachers working with profiles: findings from current research and development. In Warhurst, J., editor, *Assessing and reporting in teaching and learning: implementing the national profiles*, Belconnen, ACT: Australian Curriculum Studies Association, 19–33.

Griffin, P. and **Forwood, A.** 1991: *Adult Literacy and Numeracy Competency scales*. Melbourne: Phillip Institute of Technology.

Griffin, P. and **McKay, P.** 1992: Assessment and reporting in the ESL Language and Literacy in Schools Project. In National Languages and Literacy Institute of Australia, *ESL development: language and literacy in schools project*, Volume 2, Canberra: National Languages and Literacy Institute of Australia, 9–28.

Hamp-Lyons, L. and **Mathias, S.P.** 1994: Examining expert judgements of task difficulty on essay tasks. *Journal of Second Language Writing* 3 (1), 49–68.

Hargreaves, A. 1989: *Curriculum and assessment reform*. Milton Keynes: Open University Press.

Hester, H. (interviewed by H. Moore) 1992: Late night live. *TESOL in Context* 2 (2), 6–10.

Hill, K. 1995: Scales and tests – competition or co-operation? *Melbourne Papers in Language Testing* 4 (2), 43–59.

Kemmis, R. and **Stake, R.** 1988: *Evaluating curriculum*. Geelong, Victoria: Deakin University Press.

Languages Lead Body 1993: *Introduction to the national language standards*. London: Languages Lead Body.

Lantolf, J. and **Frawley, W.** 1988: Proficiency: understanding the construct. *Studies in Second Language Acquisition* 10 (2), 181–95.

Lazaraton, A. 1996: Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing* 13 (2), 151–72.

Lee, J. and **Musumeci, D.** 1988: On hierarchies of reading skills and text types. *Modern Language Journal* 72 (2), 173–87.

Linacre, J.M. 1989: *Many-faceted Rasch measurement*. Chicago: MESA Press.

Linn, R.L. 1993: Linking results of distinct assessments. *Applied Measurement in Education* 6 (1), 83–102.

Lumley, T.J.N. and **McNamara, T.F.** 1995: Rater characteristics and rater bias: implications for training. *Language Testing* 12 (1), 54–71.

McKay, P. 1994: Two ESL scales at the national level? What has happened and why? *Australian Language Matters* 2 (2), 1–19.

—— 1995a: A response to the ESL scales. *Australian Language Matters* 3 (1), 12–13.

—— 1995b: Developing ESL proficiency descriptions for the school context: the NLLIA ESL bandscales. In Brindley, G., editor, *Language assessment in action*, Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Madaus, G. 1993: A national testing system: manna from above? An historical/technological perspective. *Educational Assessment* 1 (1), 9–26.

Mawer, G. 1991: *Language audits and industry restructuring*. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Mehrens, W. 1992: Using performance assessments for accountability purposes. *Educational Measurement: Issues and Practice* 11 (1), 3–9.

Messick, S. 1994: The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23 (2), 13–23.

Moore, H. 1996: Telling what is real: competing views in assessing English as a Second Language development. *Linguistics and Education* 8 (2), 189–228.

Moss, P. 1992: Shifting conceptions of validity in educational measurement: implications for performance assessment. *Review of Educational Research* 62 (3), 229–58.

—— 1994: Can there be validity without reliability? *Educational Researcher* 23 (2), 5–12.

Moss, P., Beck, J., Ebbs, C., Matson, B., Muchmore, J., Steele, D., Taylor, C. and **Herter, R.** 1992: Portfolios, accountability and an interpretive approach to validity. *Educational Measurement: Issues and Practice* 11 (3), 12–21.

National Forum on Assessment 1992: Criteria for evaluation of student assessment systems. *Educational Measurement: Issues and Practice* 11 (1), 32.

National Languages and Literacy Institute of Australia 1994: *ESL development: language and literacy in schools project*, Volume 1, *Teachers' manual*. Canberra: National Languages and Literacy Institute of Australia.

Nevo, D. 1995: *School-based evaluation*. New York: Pergamon.

New South Wales Adult Migrant English Service 1995: *Certificates of Spoken and Written English*. Sydney: Author.

Nicholls, A. 1983: *Managing educational innovations*. London: Allen and Unwin.

North, B. 1993: *The development of descriptors on scales of language proficiency*. Washington, DC: The National Foreign Language Center.

—— 1995a: Scales of language proficiency. *Melbourne Papers in Language Testing* 4 (2), 60–111.

—— 1995b: The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System* 23 (4), 445–65.

Nuttall, D. and **Stobart, G.** 1994: National curriculum assessment in the U.K. *Educational Measurement: Issues and Practice* 13 (3), 24–39.

Pierce, B. and **Stewart, G.** 1996: The development of the Canadian Language Benchmarks Assessment (CLBA). Unpublished manuscript, University of British Columbia.

Riley, K.A. and **Nuttall, D.L.**, editors, 1994: *Measuring quality: education indicators – United Kingdom and international perspectives*. London: Falmer Press.

Ruby, A. 1994: Officials, ministers and the demand for information. In Riley, K.A. and Nuttall, D.L., editors, *Measuring quality: education indicators – United Kingdom and international perspectives,* London: Falmer Press, 6–16.

Rutledge, D. 1993: *Benchmarks: a standards framework for judging student performance*. ACSA Workshop Report no. 7. Canberra: Australian Curriculum Studies Association.

Scharer, R. and **North, B.** 1992: *Towards a common European framework for reporting language competency*. Washington; National Foreign Language Center.

Shavelson, R., Baxter, G. and **Pine, J.** 1992: Performance assessment: political rhetoric and measurement reality. *Educational Researcher* 21 (4), 22–27.

Shohamy, E. and **Inbar, O.** 1991: Validation of listening comprehension tests: the effect of text and question type. *Language Testing* 8 (1), 23–40.

Singh, M.G. 1990: *Performance indicators in education*. Geelong: Deakin University Press.

Taylor, C. 1994: Assessment for measurement or standards: the peril and promise of large-scale assessment reform. *American Educational Research Journal* 31 (2), 231–62.

Troman, G. 1989: Testing tensions: the politics of educational assessment. *British Educational Research Journal* 15 (3), 279–95.

Turner, C. and **Upshur, J.** 1995: Constructing rating scales for second language tests. *English Language Teaching Journal* 49 (1), 3–12.

Wigglesworth, G. 1993: Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing* 10 (3), 305–36.

Wolf, A. 1991: Assessing core skills: wisdom or wild goose chase? *Cambridge Journal of Education* 21 (2), 189–201.

Woods, A. 1995: Observing and reflecting on a year 12 learner's oral language development. *English as a Second Language Teachers' Association Journal* 11 (3), 25–31.

Appendix 1 Framework for ordering of descriptors in NLLIA ESL bandscales

GENERAL STATEMENT
(designed to provide an overview of the level and to describe global aspects of language use).
e.g., range of activities/texts, nature of the context, e.g., nature of interlocutor support, whether face-to-face/group/class, learner's familiarity with content.

SKILLS/STRATEGIES AND FEATURES OF SECOND LANGUAGE ACQUISITION ('THE PROCESS DIMENSION') REFERRED TO IN THE BACHMAN MODEL
e.g. specific skill development in Reading, Writing, Speaking, Listening; descriptions of second language acquisition strategies, including role of L1.

FEATURES OF LANGUAGE PERFORMANCE
e.g., salient details of discourse features, vocabulary, syntax, pronunciation.

HIGHER LEVEL FEATURES OF PERFORMANCE
e.g., ways in which learners are dealing with cultural aspects within the language such as idiom.
e.g., level of precision.

NOTES ON THE LEVEL
e.g., notes on specific learner groups, implications for placement, appropriateness to purpose and context.

Source: McKay (1995b)

Appendix 2 Sample from NLLIA bandscales level descriptors

Middle and Upper Primary ESL S3

Speaking: Level 3

Developing a range of social and classroom language and moving into learning through English: drawing on knowledge of the world in L1 and English, and on L1 and English language and literacy (to varying degrees).

Can participate in face-to-face interactions on familiar personal topics but with frequent breakdowns in fluency and meaning due to limited English resource. Are able to generate own language beyond formulae and two or three word utterances.

Are able to participate in class interactions on familiar topics where teacher directs questions towards them individually, using contextual support, repeating and rephrasing and elaborating; but will have difficulty participating in discussions between teacher and learners at native-speaker speed. Can participate in group learning activities with mainstream peers only with a clear role definition (e.g., as the time keeper) which does not require much speaking in English.

Will rely heavily on the context, and on the interlocutor for support (e.g., allowing time for the learner to process what they want to say, giving supportive gestures and facial expressions, rephrasing questions etc.).	SAMPLE *G* *Daymack? is that another country? Denmark?* *R* *yeah-s that's another country that was Arman Damack Arman is my cousin living there and then Damark she's coming my cousin from Arman and my anty she's living there she was coming here*
Are taking more risks with language. Can now take longer turns and reformulate utterances for better understanding, if confident with the interlocutor and scaffolding between the learner and the interlocutor occurs.	*G* *mm* *R* *they didn't let her they left* *G* *they wouldn't let her come in* *R* *stralia*
Can initiate interactions with teachers and peers. May negotiate meaning (e.g., express the need for repetition). May continue to use L1 with L1 peers and adults, for example to communicate ideas and concepts that are cognitively demanding, and to clarify issues of content and procedure.	*G* *they wouldn't let her come to Australia so they wouldn't let her go to Turkey and then they wouldn't let her come to Australia* *R* *yean they dunna let and now* *G* *why's that* *R* *because um I don't know I gotta tell my mum* (Ramona)

Source: National Languages and Literacy Institute of Australia (1994: C12)

Appendix 3 Sample of oral competency from Certificate of Spoken and Written English II
Competency 6 Can negotiate an oral transaction to obtain information

Elements	Performance Criteria	Range Statements	Evidence Guide
i. can use appropriate strategies to negotiate transaction	• uses appropriate strategies to negotiate transaction e.g. opening and closing, confirming, checking	• familiar and relevant	Sample Tasks
ii. can use appropriate vocabulary	• uses appropriate vocabulary	• dialogue with 2 speakers	• Learners ask about enrolling in a class.
iii. can request information	• requests information as required using questions or statements	• sympathetic interlocutor	• Learners ask about childcare or enrolling child in school.
iv. can provide information	• provides information as required	• face-to-face • telephone for Distance Learning students • may include a few grammatical errors but errors do not interfere with meaning • short exchange – approximately 1 minute • recourse to clarification/repetition	

Phonology
Specific performance criteria related to phonology have not been included. However it is assumed that:

- articulation of some phonemes and clusters as well as intonation, stress and rhythm in longer phrases and clauses may often be inaccurate or unconventional
- teaching programs will pay attention to:
  phonological features of longer utterances
  developing learner self-monitoring and repair/correction strategies

Source: New South Wales Adult Migrant English Service (1995)