Asset Languages – a multilingual proficiency framework that supports learning

Neil Jones

ESOL, Cambridge Assessment, 1 Hills Road, Cambridge CB1 2EU, United Kingdom

Asset Languages is the system being developed by Cambridge Assessment to implement the Languages Ladder, "a new voluntary recognition system to complement existing national qualifications frameworks and the Common European Framework" which is a major element of the UK's National Languages Strategy. Asset sets out to accredit functional language proficiency within a can do framework. It is comprehensive, including at least 26 languages: those most commonly learned as "modern foreign languages", and those spoken by particular communities in the UK. It targets three contexts (Primary, Secondary, Adult), with skills assessed separately. It offers two assessment strands: external assessment at six major stages, and more informally accredited teacher assessment at 17 finer grades.

The challenges of developing this complex framework are not merely technical or logistical: they concern how to design tests and interpretations which enable valid and useful comparison across such widely differing languages and learner groups, and above all how to do this in a way which impacts positively on learning. This leads us to look critically at the framework metaphor in general, and at the Common European Framework in particular; and to propose some conceptual clarification and practical methods for framework construction.

The origins of Asset Languages

It is agreed that language learning and teaching in the UK faces serious problems. The National Languages Strategy (NLS) was launched in 2002 to tackle "a cycle of national underperformance in languages, a shortage of teachers, low take up of languages beyond schooling and a workforce unable to meet the demands of a globalised economy" (DfES 2002:10).

A key element of the strategy is the Languages Ladder: "a new voluntary recognition system to complement existing national qualifications frameworks and the Common European Framework" (DfES 2002). Asset Languages is the assessment system currently being developed by Cambridge Assessment[1] to implement the Languages Ladder.

The NLS itself originates in proposals made by the Nuffield Languages Programme (Nuffield Languages Inquiry 2000), and the recognition system, under the name "A Learning Ladder for Languages", was the subject of a subsequent feasibility study (Nuffield Languages Programme 2002)[2]. This study found that existing qualification frameworks were inadequate in two respects. Firstly, they found that many qualifications in languages were "confusing and uninformative about the levels of competence they represented" (Nuffield Languages Programme 2002:8), and that "beyond 14, student attainment in languages is mainly related to examination targets, and not to performance criteria in 'can do' terms, except in vocational courses" (idem:9). In consequence, they recommended that the new qualification framework should stress meaningful proficiency levels.

Secondly, they found that current qualifications did not support learning well, given their summative role at the end of an extended period of study. Consequently they recommended that the new framework should provide a “learning ladder” of bite-sized, accessible learning targets.

For both these purposes the Nuffield Inquiry identified the Common European Framework of Reference (CEFR) as a model to be followed (Council of Europe 2001).

Thus Asset Languages sets out to support the NLS in two important ways:

- by accrediting language ability within a functional, can do framework, so that levels are comparable across languages;
- by supporting language learning, providing a motivating “ladder” of learning targets which enables recognition of each step achieved.

These two purposes are distinct, and the inclusive Asset Languages framework reveals potential tensions between them. While this tension is perhaps not evident in a European context, it becomes so when one steps beyond the European family of languages, or seeks to bring very different kinds of learners into a single framework.

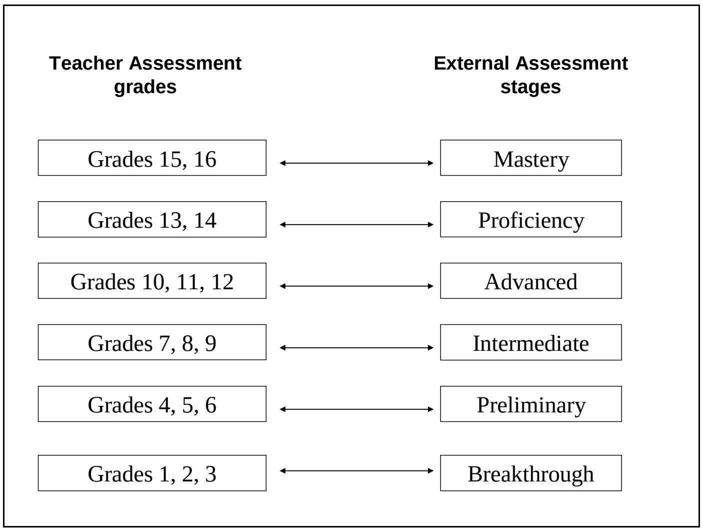
The system in outline

Asset Languages is produced jointly by two Cambridge Assessment business streams: OCR and Cambridge ESOL. Cambridge Assessment was awarded the tender for the Languages Ladder in October 2003 by the Department for Education and Skills (DfES).

The Asset Languages framework is an extremely complex one. It includes 26 languages: those most commonly learned in school as modern foreign languages (MFL), and those spoken by particular communities in the U. K. It targets three contexts (Primary, Secondary, Adult), with differentiation of test content for each. It assesses and accredits reading, writing, listening and speaking separately.

Also importantly, given its intended positive impact on learning, Asset Languages offers two assessment strands: external assessment at six major stages, and more informally accredited teacher assessment at 17 finer grades.

Figure 1 Asset Languages Teacher and External assessment framework



Teacher Assessment grades	External Assessment stages
Grades 15, 16	Mastery
Grades 13, 14	Proficiency
Grades 10, 11, 12	Advanced
Grades 7, 8, 9	Intermediate
Grades 4, 5, 6	Preliminary
Grades 1, 2, 3	Breakthrough

Asset Languages is offered in computer-based and paper-based modes of delivery, and, given its “effectively on demand” availability, will need constantly renewed, multiple versions of test papers. Thus it can be seen that the system comprises potentially many thousands of tests. These must all be related to each other within the assessment framework using data-based, empirical methods, if the framework is to fulfil its function of accrediting equivalent levels of functional language ability across languages, levels, contexts of learning, contexts of use and modes of administration, relating them all to the can do levels described by the Languages Ladder.

Can this be done? Evidently the logistical problem is considerable, but the methodology exists, and is outlined below. But the question “can this be done?” is not simply one of logistics or methodology, but concerns the validity and meaningfulness of the whole enterprise.

The development of the “learning ladder” concept and the CEFR

It is useful at this point to review some history. The publication of the CEFR in 2001 is the culmination of decades of work by the Languages Policy Division of the Council of Europe, aimed at promoting the learning of languages by providing a sequence of accessible learning objectives – a “learning ladder”. This has had a profound effect on our conception of the role of assessment, which can be illustrated by reviewing the history of the Cambridge ESOL main suite of exams.

What can now be presented as a system in fact developed piecemeal over the best part of a century, beginning in 1913 with the highest level (Cambridge Certificate of Proficiency, now associated with CEFR C2). It was not until the 1930’s that the Local Exams Syndicate accepted the need for a certificate at a lower level. The First Certificate (CEFR B2), as it became, remained for many years the lowest level of foreign language proficiency considered to have any “social value” – i.e. to be worth certificating as a serious qualification.

Then in 1977 the influential learning objective Threshold Level was published by Van Ek and Trim, followed in 1979 by Waystage. These have since been brought into the CEFR as B1 and A2. Cambridge ESOL responded by adding the PET and KET exams at these levels. Also the CAE exam was added at CEFR C1, as a necessary objective on the long road from First Certificate to Proficiency. A Young Learner suite of tests also operate at around Breakthrough level (CEFR A1-A2). Thus the Cambridge ESOL learning ladder has come about.

Conceptual challenges: linking to a framework

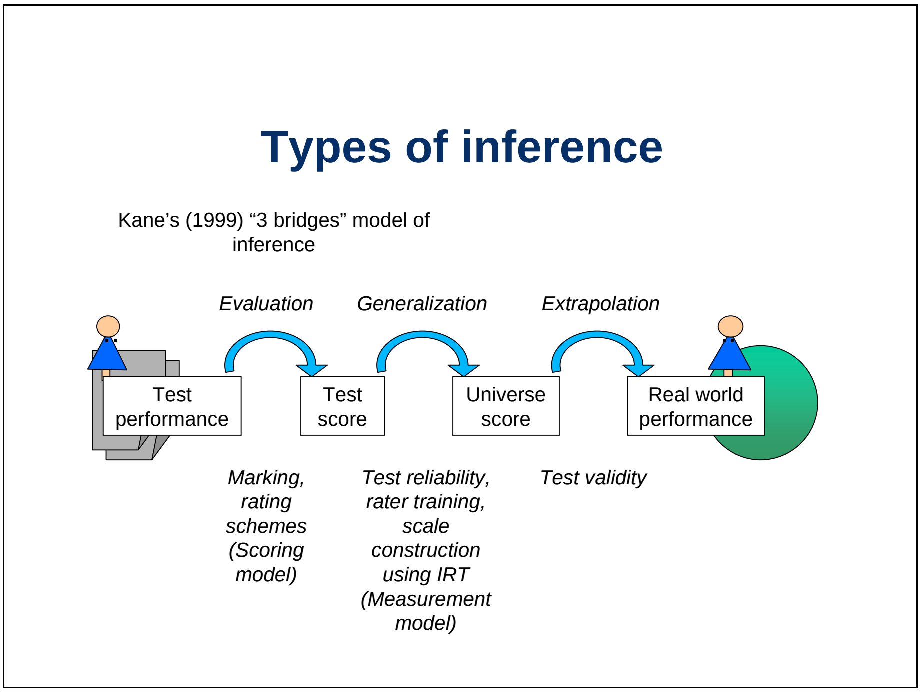
A test is valid, we may claim, if it supports the *inferences* which users of the results wish to make. Most tests are intended to support inference to a world beyond the test, in two respects:

- tests are necessarily limited samples of a much wider domain of skill or knowledge;
- test tasks are to an extent inauthentic, but relate to “real-world” tasks in some target situation of use.

The second point should generally be true of language tests because we learn languages (or should do) in order to use them and accomplish things through them.

Thus to demonstrate the validity of a test we need to develop an *inferential argument*. Kane (1999) has shown the structure of such an argument as a series of steps, or inferential bridges.

Figure 2 Kane's "three bridges" model of inference



Types of inference

Kane's (1999) "3 bridges" model of inference

Evaluation *Generalization* *Extrapolation*

Test performance → Test score → Universe score → Real world performance

Marking, rating schemes (Scoring model)

Test reliability, rater training, scale construction using IRT (Measurement model)

Test validity

What we can observe is test performance. From this we must derive a test score. This inference - called evaluation - depends on how we devise mark schemes and rating criteria - it constitutes our *scoring model*.

But the test score relates to a single event. Would a learner get the same score if they took a different version of the test? This inference to the universe score - generalization - depends on the reliability of the test, the training of the raters, etc. Such factors relate to our *measurement model*.

Finally we must relate the universe score to the real world. This inference - extrapolation - is at the heart of test validity.

This model has been expanded to find a place for theory, which explains how inferences are justified. The theory includes, in addition to the scoring and measurement models, a description of the *learners* we are testing: who are they, what are we testing them on, and why. For example, a very specific learner group might be: *children age 7-11 learning French in a formal setting with little exposure to authentic sources*.

Then we need a *test model*, which we use to select content and task types testing the relevant abilities and appropriate to the learner group.

If we then wish to relate this specific learner group and test situation to a general framework, such as the CEFR, it seems that a further inferential step is necessary – from the context-specific to the context-neutral. This should be seen as a kind of idealization, which adds value to the extent that the wider framework has greater currency.

What does this final inference depend on? Clearly, it is necessary to identify some basis of comparison which works for the specific learner group and the descriptors which define the framework. Not all may be relevant to a particular group – children, for example, do not have the instrumental needs or, indeed, the cognitive developmental stage of adults, and yet it may be possible to justify a comparison. A key issue then is how to go beyond identifying *similarity* (placing exams at the same level) to talk usefully about *differences* – in purpose, in target learner group, in construct of communicative language ability, and so on.

These issues are important. The prominence of the CEFR, in Europe but increasingly beyond Europe, has forced language testers worldwide to recognize the need to align their exams to the CEFR. It is not difficult to make a claim of alignment on the basis possibly of a cursory subjective judgment. So how can users of language exams choose between several exams, all claiming to be at the same level? And how can language testers competing to offer the highest-quality exams demonstrate that their claim to alignment is better – more meaningful and useful?

A methodology for doing this is emerging through the development and piloting of a Council of Europe manual *Relating language examinations to the CEFR* (CoE 2003). Cambridge ESOL is among European language testers who have undertaken to take part in the pilot, providing case studies, including one for Asset Languages. What becomes clear in the context of Asset is the importance of focussing on specific groups of learners: one could say that finally it is learners, rather than tests, which are aligned to the framework.

A learner-centred view of proficiency

Breakthrough (CEFR A1) is difficult to define as a proficiency level, for the lower the level, the less different groups of learners have in common – differences in age, educational background or first language make for very different kinds of Breakthrough-ness. In practice, proficiency testing at lower levels involves increasingly specific guidance to candidates and teachers concerning the content of the test. Proficiency tests at low levels effectively become achievement tests.

A particular issue arises with languages which have difficult scripts, such as Chinese (one of the Asset languages). A Languages Ladder or CEFR description of reading or writing at Breakthrough implies a functional level of skill which will take much longer to achieve in Chinese than for another European language using the Roman alphabet – the level ceases to represent an attainable first learning target. To maintain an accessible progression we must be prepared to sacrifice functional equivalence at the lowest levels (which, as argued above, is not sacrificing very much) and provide for a staged acquisition of the script using controlled character sets. This is one way in which Asset Languages' inclusive framework points up tensions between the learning and proficiency framework which the CEFR is able to gloss over.

The lower levels show most clearly that conceptions of language proficiency must factor in considerations of the learning context. Pursuing this idea, we need a general model for test design which enables us to operationalise a notion of language proficiency for a given group of learners. Construct definitions of language proficiency used latterly in assessment have their basis in theories of communicative language competence (Bachman 1990, Canale & Swain 1980, Council of Europe 2001). For example, linguistic, sociolinguistic and pragmatic competences are posited as components of communicative language competence, with each comprising particular knowledge, skills and know-how. More recently assessment has recognized the need for more explicit theories of test construction in order to formalize the way that the various elements of a test design situation – the learner, the purpose, the measurement model etc. – are integrated into a sequence of design and implementation procedures (Mislevy et al 2004, Luecht 2004, Weir 2004). Cambridge ESOL, for example, is applying Weir's socio-cognitive framework to the constructs of reading, writing, speaking and listening across its range of exams (Weir and Shaw 2005).

It is this direction of enquiry which we need to develop further for the Asset Languages framework. We must be aware of the extent to which our constructs are declarative, and not merely descriptive. Thus underlying a construct is a rationale for defining it thus, which will reflect the needs of particular learner groups, the benefits of identifying achievable and useful outcomes at each level, pedagogic considerations of sequencing or selection of learning objectives, practical constraints, and so on.

The point is simply that our notions of language proficiency are socially constructed and relate to particular learners and goals. The challenge for the Asset Languages framework is thus to shape itself optimally to the needs of language learners from a range of contexts of learning, ages and backgrounds, but to avoid being driven too far by any one group. The task for Asset Languages' external assessment framework would be to find an acceptable match to different users' needs while preserving the wide interpretative framework which the Languages Ladder aims to provide.

Formative assessment in Asset Languages

It is a major step for an exam board to move from summative assessment at the end of a learning cycle to formative assessment which is embedded within that cycle – feeding forward into learning rather than simply looking back. But this is precisely the role which the original proponents of the Languages Ladder saw for this new accreditation framework. Within the Asset Languages system it is the teacher assessment strand, rather than the external assessment, which most closely resembles the proposals put forward in the Nuffield Inquiry and subsequent feasibility study: a finely-graded progression (three grades to each major stage) is certificated by teachers using light-touch tests. Teachers can administer the tests after completing a short training and accreditation process. Asset Languages teacher assessment is locally administered and certificated, and no central records are kept.

But how can an exam board find a role inside the classroom, and what kind of assessment should we qualify as formative - as *assessment for learning*? For Leung (2004:21) the essence of formative assessment is that it adapts flexibly to local, immediate learning contexts. It should not consist in measuring achievement against an inventory of externally-defined attainments. Current conceptions of "formative" assessment in the US context have been criticized: Wiliam (2004:4) writing of the United States, says "the term 'formative assessment' is often used to describe assessments that are used to provide information on the likely performance of students on state-mandated tests – a usage that might better be described as 'early-warning summative'". Shepard (2005) speaks of the term being hijacked. Pellegrino (2003) rechristens the *No child left behind* program *No child left untested.*

The intention of Asset Languages teacher assessment is to fulfill a formative role by fitting flexibly into existing schemes of work. That is, it aims to enable teachers to adapt tests to suit their local, immediate contexts. Materials currently comprise a pack for each language and level. Skills are accredited separately. For each grade a small number of test tasks must be selected and administered. Teachers are encouraged to adapt some of the tasks where necessary, for example, to use already-taught vocabulary. In this way it is hoped to develop a system which teachers can use formatively.

Asset Languages' formative role will be achieved to the extent that it empowers teachers to use the Languages Ladder creatively, while preserving a sufficiently strong link to the external assessment and the important interpretative framework which it aims to provide.

Constructing the framework

This section gives the briefest outline of the methodology being used for Asset Languages. Constructing an assessment framework is conceptually a two-stage process. The first task is to construct a scale for each language and skill, linking all the levels.   Having constructed these scales one can proceed to set standards, seeking to ensure comparability across languages and learning contexts. Scale construction for Asset is done differently for subjectively-rated skills (writing and speaking) and objectively-marked skills (reading and listening). Both teacher and external assessments must relate to the same framework, but it is the external strand, given its requirement to ensure comparability across languages, which is the main focus in empirical scale construction and standard setting.

Subjectively rated tests (speaking and writing)

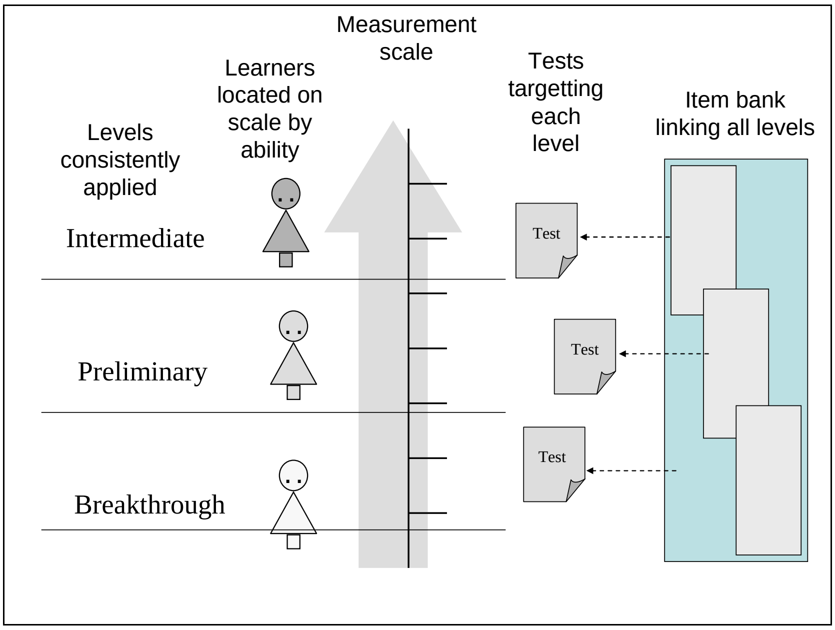
The case of subjectively-rated performance skills seems superficially straightforward. Progression, it seems, can be described in terms of *what* things people can do and *how well* they can do them. The first step in constructing a scale is thus to select tasks which are relevant to our construct of writing or speaking and which offer an appropriate degree of challenge for each target level, to elicit performance which can be rated with respect to that level. In practice the factors of task difficulty (the *what*) and performance quality (the *how well*) are very difficult to disentangle in performance assessment. This is why can do descriptions of level such as the Languages Ladder or the CEFR do not translate in any simple way into rating instruments. In this situation it is the use of exemplar scripts, and training and standardisation centred on these, which in practice guarantees the application of consistent standards across exams and across sessions.

Objectively-marked tests (reading and listening)

Objectively-marked skills offer different problems of interpretation, because a greater inferential leap is necessary from what learners have to do in tests to what they are able to do in the real world. But the capture of item-level response data enables us to construct scales based on a measurement model, which is potentially a huge step forward in terms of consistency of standards and richness of interpretation of test performance. Item response theory (IRT) provides the statistical approach; item banking is the term used to describe the test construction methodology based on it (Hambleton et al 1991, Wright and Stone 1979, Bond and Fox 2001).

In this approach a bank of *calibrated* items is assembled – that is, items whose difficulty is known, initially from pretesting. Response data is collected so as to ensure a link across items at all levels. Thus a single measurement scale can be constructed to cover all levels. This scale, unlike a scale that relates directly to raw test scores, has useful features akin to those which we take for granted in the direct measurement of physical properties such as weight or temperature: it is linear and can be extended as far as needed; intervals on it can be meaningfully compared. It relates different testing events within a single frame of reference, greatly facilitating the development and the consistent application of standards. Tests are generated from an item bank to specific target difficulties, and learners' scores on these tests locate them directly on the underlying measurement scale. Figure 3 illustrates.

Figure 3 Item banking approach to framework construction



Measurement scale
Learners located on scale by ability
Levels consistently applied
Tests targetting each level
Item bank linking all levels
Intermediate
Preliminary
Breakthrough
Test
Test
Test

Item banking has been at the core of Cambridge ESOL's methodology for test construction and interpretation since the early 1990s, and is now used for almost every Cambridge ESOL exam. Using item banking to develop the Asset Languages framework is however a significant innovation for British educational assessment, where it is still relatively little known.

Standard setting

While scale construction for objective tests is essentially an algorithmic process, standard setting always involves judgment – but judgment which must be as constrained as possible, if the resulting framework is to be coherent.

Standard setting approaches can generally be divided into the task-centred or learner-centred. The former requires experts to study the content of a test and make a judgment about which scores indicate what level of competence; the latter relate learners' scores on a test to evidence of their abilities from beyond the test. Despite their wide use, task-centred procedures have serious shortcomings for the test equating purpose implicit in constructing a multilingual proficiency framework (Jones 2005). The most valid target of standard setting judgment is the real-world language skills of learners (even if the real world may be limited to the classroom). It is these which are the object of interest, rather than features of tests or tasks, which relate to the real world only indirectly.

Learner-centred standard setting approaches are thus important for the Asset Languages framework As a starting point we have used teachers' estimates of learners' national curriculum level collected during pretesting, as the best available indicator of their functional language proficiency. We are also developing approaches based on the use of can do questionnaires, building on experience from the ALTE Can Do Project, a major study which supported the linking of Association of Language Testers in Europe (ALTE) members' exams to the CEFR (Jones 2000, 2001, 2002). One study uses a plurilingual design where learners taking two foreign languages at secondary school self-assess their relative abilities in the two languages. For speaking and writing direct comparisons of performance can be made by suitable plurilingual informants. One project within Asset is developing a bank of English speaking and writing exemplars (although English is not one of the

Asset languages), with the intention of using it as a point of reference particularly for less commonly spoken languages.

Conclusion

The Asset Languages framework is still very much under construction. In this paper I have discussed conceptual issues which require practical solutions.

Firstly, the relation of assessment to learning objectives: how explicitly should test content be specified? How should assessment criteria be communicated? What kind of test preparation should be considered appropriate? We want to test learners on things they have had a chance to learn; but at the same time we must not lose the connection with communicative language use.

Secondly, how can we validly compare different learner groups? We must identify and describe the different needs and characteristics of different learner groups, developing the rationale for assessments explicitly to meet these. Then both quantitative and qualitative evidence can be used to build the claim to comparability within the framework.

Finally the success of the Asset Languages scheme will depend on how it comes to be used. It is well-suited to testing when ready, which for learner motivation should be far better than stage-based testing. Will it be used in this way?

We recognize that language learning in the UK faces serious challenges. Our hope is that Asset Languages will prove to be part of the solution rather than part of the problem.

References

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press

Bond, T. G. and C. M. Fox (2001). *Applying the Rasch model*. NJ: Lawrence Erlbaum Associates.

Canale, M. and M. Swain (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1.1. 1-47.

Council of Europe (2001): *Common European Framework of Reference for Languages*, Cambridge: Cambridge University Press

Council of Europe (2002): *Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Case Studies.* Strasbourg: Council of Europe Publishing

Council of Europe (2003) Relating language examinations to the CEFR. Manual; Preliminary Pilot Version. Retrieved from: http://www.coe.int/T/E/Cultural Co-operation/education/Languages/Language Policy/Manual/default.asp

DfES (2002). *Languages for all: Languages for life*. Retrieved from http://www.dfes.gov.uk/languagesstrategy/

Hambleton, R. K., H. Swaminathan, H. J. Rogers (1991). *Fundamentals of item response theory, Volume 2.* Newbury Park, CA:Sage.

Jones, N (2000) Background to the validation of the ALTE Can Do Project and the revised Common European Framework. *Research Notes* Issue 2 pp 11-13. Cambridge: CambridgeESOL. Retrieved from: http://www.cambridgeesol.org/rs_notes/offprints/pdfs/RN2p11-13.pdf

Jones, N (2001) The ALTE Can Do Project and the role of measurement in constructing a proficiency framework. *Research Notes* Issue 5 pp 5-8. Cambridge: CambridgeESOL. Retrieved from: http://www.cambridgeesol.org/rs_notes/offprints/pdfs/RN5p5-8.pdf

Jones, N (2002) Relating the ALTE Framework to the Common European Framework of Reference. In Council of Europe 2002: 167-183

Jones, N (2005). Raising the Languages Ladder: constructing a new framework for accrediting foreign language skills. *Research Notes* No 19. Cambridge ESOL. Retrieved from: http://www.cambridgeesol.org/rs_notes/rs_nts19.pdf

Leung, C. (2004). Developing formative teacher assessment. Language *Assessment Quarterly 1.1*, New Jersey: Erlbaum

Luecht, R. (2004): Multistage complexity in language proficiency assessment: a framework for aligning theoretical perspectives, test development and psychometrics. *Foreign Language Annals* Vol. 36, No. 4

Mislevy, R. J. Steinberg, L. S. and Almond R. G. (2002): Design and analysis in task-based language assessment. *Language Testing* 19 (4) 477-496

Nuffield Languages Inquiry (2000) *Languages: the next generation*. London: The Nuffield Foundation. Retrieved from: http://languages.nuffieldfoundation.org/filelibrary/pdf/languages_finalreport.pdf

Nuffield Languages Programme (2002). A Learning Ladder for Languages: possibilities, risks and benefits. Retrieved from: http://languages.nuffieldfoundation.org/filelibrary/pdf/learning_ladder.pdf

Pellegrino, J. (2003).*The "Second Transformation" – issues in combining advances the learning sciences with IT capablilities*. Presentation to NRC ILIT Committee.

Shepard, L.A. (2005). Competing paradigms for classroom assessment: Echoes of the tests-and-measurement model. Lecture and Powerpoint slides presented at the annual meeting of the American Educational Research Association, Montreal, April 2005.

Van Ek, J. A. and Trim J. L. M. (1990) *Threshold 1990*. Cambridge: Cambridge University Press.

Van Ek, J. A. and Trim J. L. M. (1990) *Waystage 1990*. Cambridge: Cambridge University Press.

Weir C (2004): *Language Testing and Validation*, Basingstoke: Palgrave Macmillan Ltd.

Weir, C, and S. D. Shaw (2005). Establishing the Validity of Cambridge ESOL Writing Tests: towards the implementation of a socio-cognitive model for test validation. *Research Notes* Issue 21 pp 10-14. Cambridge: CambridgeESOL Retrieved from: http://www.cambridgeesol.org/rs_notes/offprints/pdfs/RN21p10-14.pdf

Wiliam, D. (2004). *Keeping learning on track: Integrating assessment with instruction.* Invited address to the 30th annual conference of the International Association for Educational Assessment held in June 2004 in Philadelphia.

Wright, B. D. and M. H. Stone (1979). *Best test design*. Chicago, IL: MESA Press.

[1] Cambridge Assessment is the brand name of University of Cambridge Local Examinations Syndicate, a department of the University of Cambridge. Cambridge Assessment is a not-for-profit organisation.

[2] The Nuffield Languages Inquiry is a project initiated by the Nuffield Foundation, a UK charitable trust established in 1943 by William Morris (Lord Nuffield) to 'advance social well being', particularly through research and practical experiment.