Graduate School of Education
Public Lectures
2005/2006

Comparative, International and Global Perspectives on Education

The Weaponry of Language Assessment: what are the threats?
Professor Pauline Rea-Dickins, 14 November 2005

A World of Difference? The social and cultural construction of the European teacher and Learner
Professor Marilyn Osborn, 21 November 2005

Constructing the Global Education Industry: markets or mayhem?
Professor Susan Robertson, 28 November 2005

Bridging Cultures and Traditions: perspectives from comparative and international research in education
Professor Michael Crossley, 9 February 2006
McNamara, T. 2005. 21st Century Shibboleth: language Tests, Identity and Intergroup Contact. (Mimeo, provided by author)


REFERENCES
See also http://www.bris.ac.uk/education/research/centres/creole/ethics
http://www.chna.org
See also http://www.bris.ac.uk/education/research/centres/creole/ethics
mass destruction in Iraq, there are – I suggest - in the area of language assessment too many examples where assessment results are both misused and abused in a thoroughly destructive manner: to promote political careers, to exclude, to repatriate and to deny significant numbers of school learners with educational opportunities that would allow them to compete on equal terms with their peers, both nationally and globally. Assessment can be destructive in other ways, too. Assessment cultures – and England is a prime but not exclusive example – have become obsessed with summative assessment of achievement or certification, encouraging continuous measurement of language knowledge. Jumping on and off the bathroom scales every minute does not lead to weight loss; by the same token the continuous measurement of student achievement does not necessarily lead to learning.

In 1974, Jacobovits wrote of language testing as the ‘reactionary wing of Applied Linguistics’. This has alarming currency some 31 years later and I doubt whether he could have conjectured at the time of his comment just how reactionary the broader domain of language assessment might become.

NOTES
1. This is not a ‘new’ phenomenon. For example, it is reported that the Chinese imperial examination lasted between 24 and 72 hours and was conducted in cell like rooms. “The personal suffering that individuals underwent both in the preparation and in the taking of these exams has become part of Chinese lore. …. Some committed suicide because of the disgrace that these failures brought to their families. Others continued taking exams even as very old, grey-haired men.” (http://www.china.org)

The Weaponry of Language Assessment: what are the threats?

Introduction
In this presentation, there will be a short introduction in which I explain what I mean by assessment and I provide my first example. Secondly, I will briefly consider ways in which the language assessment profession has tried to achieve quality in its assessment practices. This will be followed by a range of examples on the basis of which – in my concluding remarks - we can determine whether language assessment practices warrant being equated with weapons, and whether in fact they pose any kind of threat.

Historically, we ‘go back a long way’ in language assessment, to the Bible and the Book of Judges in fact. Spolsky writes of the ‘shibboleth test’ that used dialectical differences to recognise and dispose of enemies (Judges, 12, 4-6 cited in Spolsky (1995:242). The Chinese imperial examination system – dating from the Han Dynasty, (206BC-220) - is an interesting example on account of its meritocratic nature: whereby appointment to the civil service was gained through ability and not “through special or inherited privilege” (http://www.china.org). This was a examination system that lasted for 2000 years and must have been the first attempt to introduce ‘objectivity’ (for example, by “sealing examinees’ names on paper, blocking the examination site during the exam, and separate evaluation of each answer sheet by three different examiners.” (http://www.china.org). There were the Jesuits in the seventeenth century who sought to control school curricula through an examination system (Madaus, 1990, cited in Spolsky 1995). More recently, indeed currently in England, we are able to recognise government control of our education system through the use of National
Curriculum assessments and the way it uses these results to inform on literacy and other targets and policy delivery through the League Tables in the national press and, in the case of language assessment, the rather poor performances nationally in French. Before I move on to say something about how we may judge the quality of language assessment, I shall introduce my first example, from Australia in 1910, and to what McNamara calls “the notorious Australian Dictation Test” (McNamara, 2005:5).

Vignette 1: Assessment as a Tool for Exclusion
This assessment was used from 1901 until 1957 and nobody passed it from 1910 onwards. The dictation was not to be less than 50 words and could be in any language. The following is an English example of the dictated test (from McNamara, 2005:5):

The tiger is sleeker, and so lithe and graceful that he does not show to the same appalling advantage as his cousin, the lion, with the roar that shakes the earth. Both are cats, cousins of our amiable purring friend of the hearthrug, but the tiger is king of the family.

McNamara also provides some other interesting details about this procedure, in particular a letter dated 4th March 1927 from the Commonwealth Home and Territories Department to the Collector of Customs at Freemantle:

… The test when applied to an immigrant is intended to serve as an absolute bar to such a person’s entry into Australia, or as a means of depriving him of the right to remain in the Commonwealth if he has landed. The test should therefore be applied in a language with which the immigrant is not sufficiently acquainted to be able to write out at dictation.

One is sorely tempted to reflect on recent examples of citizenship tests Not all assessment is ‘bad’ (but that would be for another talk!). Many assessments are, by definition, meant to discriminate but we do need to show that this is done fairly. The boundaries of language assessment now extend well beyond educational and language proficiency certification contexts into the domains of ‘language for citizenship’ and immigration policy. However, the ways in which the language assessment profession attempts to assure quality within its traditional boundaries have not yet been extended to those new domains in which language data on individuals or groups are captured. What concerns me is what happens to language assessment procedures and/or results once these are out of the hands of the ethical and responsible language assessor? As I have highlighted, there are social and educational consequences for learners whose linguistic needs are inappropriately assessed and I have not even mentioned the considerable damage caused by some assessment practices, and the ensuing loss of self-esteem that can stay with an individual for a lifetime. I have shown that language assessment is not limited to the classroom, with social and also political consequences. In a discussion of where responsibilities lie (i.e. begin and end) within the assessment profession, Davies (1997) has questioned whether you would put the blame on an architect if the building he had designed was being used to house ugly racist meetings. The inference we draw from his comment is clear. However, as language assessment researchers, we are in a position to ‘out’ unethical and questionable practice, and to ask difficult questions. Indeed, it is our responsibility to do so. As I observed at the outset, language assessment is not an exact science yet, on the basis of the pronunciation of a few words on a tape, decisions are taken to deport an individual.

In conclusion, therefore, unlike Hans Blix who failed to find weapons of
groups and individuals, can change through language contact and spread, and

- the possibility of the applicant accommodating to the interpreter’s dialect was not considered: in the case they report, the interpreter’s dialect was different from the applicant and the applicant appeared to alter his own pronunciation and choice of words in order to be understood by the Farsi-speaking interpreter. (Eades et al., 2003:18-19)

(Consider within this context my own lexical shifts when I am in America: from lift to elevator, from pavement to sidewalk, from tomato to tomayto or from chips to French fries: this does not make me American).

Evidence of the kind reported above suggests an abuse of linguistic assessment practices where individuals are denied basic human rights on untenable grounds. I am not saying here that a linguistic assessment should not be done. What I am saying is that any such assessment should be theoretically grounded, and demonstrate its validity. Returning to the question I asked at the beginning of this presentation and whether we could find any examples analogous to the Australian dictation test used as a tool for exclusion, we can. Indeed, it is deeply concerning that language assessment can serve as an agent against the legitimacy of individuals to reside in countries in safety.

And now, finally, to some conclusions about the contemporary realities of language assessment practices.

Contemporary Realities of Assessment: What are the Threats?
The problems I have documented in my presentation reside not with the assessments themselves but, rather, in how they are used and I am guided in these conclusions by Messick’s (1989) construction of validity that has at its core notions of educational and social consequences.

in England and the use of some questions that have appeared too challenging for members of the general public surveyed!)

What follows is some advice for a prospective administrator of the assessment in respect of a specific candidate:

Sir
I have the honour to inform you that a Chinese named Kang Yu Wei contemplates visiting Australia shortly. He was formerly a man of considerable standing in China. For various reasons some of which have regard to his moral reputation the Government have refused permission for him to land in Australia. I shall be glad, therefore, if you will issue instructions that in the event of this man arriving at Freemantle he is subjected to the dictation test which should be applied in such a manner as to ensure its efficiency. It is probable that Kang Yu Wei understands English. Inquiries should be made on that point, and your officers should be in readiness to apply the test say in Spanish or Italian.

Here we have a blatant example of an assessment used to prevent immigration – not just a misuse but an abuse of a language assessment system. But this wouldn’t happen now, would it? Or, would it? I shall come back to this question towards the end of my lecture.

Background
What is Assessment?
I want to make two general points here. Firstly, the term ‘assessment’ in my title, as opposed to ‘testing’, has been a deliberate choice. These terms are not synonymous, although both language testing and assessment have the aim of obtaining information about an individual’s language abilities. Tests are usually formal procedures - of the pencil and paper variety that could include a dictation procedure - but may also be in the form of speaking tests, either a 1-1 interview or a paired
Assessment, however, is a broader concept, of which tests are a part. It includes the function of certification (e.g. GCSEs, A levels) and documenting achievement (e.g. at the National Curriculum KeyStages) within a school context. There are also examination boards, such as Cambridge Assessment (England) and Educational Testing Service (ETS, US), that provide certification for proficiency in English as a Foreign Language but this form of assessment is divorced entirely from specific learning and teaching contexts. By way of contrast, assessment is also central to effective teaching and learning – as an integral part of classroom instruction. Much of this ongoing assessment will be informal and take place through interaction between teachers and their learners. In this way, teachers develop keen insights into the strengths and weaknesses of the language abilities of their class and plan instruction accordingly; and the students themselves learn from the feedback provided by their teachers or their peers. In this last function of assessment, I am referring to ongoing formative classroom assessment, the focus of much of my own recent research, which aims to nurture rather than measure language development (e.g. Rea-Dickins 2003a, 2003b, 2001). This dimension of assessment links directly to research in second language acquisition (e.g. Ellis 2003, 1997, 1994) and in general educational assessment (for example, Assessment Reform Group 2001; Broadfoot & Black, 2004). This university is also in the process of developing its policy and practice in assessment, teaching and learning to ensure a balance between the development of learning on the one hand and the measurement of learning on the other. Thus, a teacher’s observation of a learner’s language use in the classroom - known as observation-driven assessment - is an important means to assess linguistic progress.

Secondly, I wish to signal that language assessment is not an exact about life-long language learning and celebrating linguistic diversity then the stranglehold of the prestige languages needs to be loosened. Assessment mechanisms such as the CEF have the potential to do this but I am very sceptical about whether they will succeed in so doing.

For my final example, I return to the immigration context.

Vignette 9: Language Assessment for Deportation Purposes

Language assessments are now used around the world as a forensic tool to detect illegal immigrants. In other words, assessment is being used to determine nationality and ethnicity rather than proficiency in a language. Eades and colleagues (2003) have identified significant problems in current practices and have shown that:

- A person’s nationality cannot always be determined by the language he or she speaks
- A few key words (for example on a tape and not even in a face to face interview) and their pronunciation cannot normally reveal a person’s nationality or ethnicity – this is reminiscent of the shibboleth test in the Book of Judges I mentioned at the outset
- Everyday perceptions about pronunciation differences among groups of people cannot be relied upon
- Any analysis of pronunciation must be based on thorough knowledge of the language and region in question and must involve detailed phonetic analysis.

(Eades et al., 2003:7-8)

In an analysis of 58 cases where linguistic identification led to a decision not to grant the right to remain in a country, Eades et al. found evidence that:

- the linguistic identification was based on insufficient data
- the qualifications of the analysts were not provided
- the linguistic identification ignored the fact that languages, of both
unwanted washback on classroom language assessment practices, with undue teacher emphasis on the monitoring of language levels to the detriment of providing their students with opportunities to develop their language skills. This is something that is very obvious in my own classroom research in England. In my view, a restricted view of assessment that places systems to the fore has a tendency to obscure and stifle diversity in assessment contexts. Assessment procedures associated with identifying levels of achievement and standards have a certain homogeneity about them: not only do they tend to conform in terms of which languages are assessed, they have a tendency to conform in terms of approach, format, the skills assessed and so forth. Where, then, is the diversity we are requested to celebrate?

I therefore invite you to consider the impact of an assessment policy – which affirms the 'spread of English' and the use of other 'highly valued' foreign languages on:

1) the marginalisation of other languages
2) the devaluing of other languages which may, in turn, have a consequence for
3) the suppression of those languages which are perceived as unimportant (by an elite) relative to others which have more recognition, in that they are formally assessed. Is it a case of a language gaining its 'cachet of approval' through assessment.

We need, I suggest, to reflect on the appropriateness of our assessment mechanisms provided for our citizens and the diverse languages that they speak and how these might on the one hand inhibit, and on the other promote linguistic diversity. If the Council of Europe/EU is really science. We cannot evaluate an individual’s performance in a language with absolute precision, in the way we can measure the temperature of a liquid. One consequence of this is that in order to capture a ‘true’ reflection of language abilities, it is important to have different ways of capturing an individual’s language abilities through a range of different assessment procedures. In everyday life, if you wanted to know what someone looked like, you wouldn’t rely on just one photograph to show the different facets of a person, especially if it was taken first thing in the morning or late at night (or even at a lectern as I am now!): you would need a selection of photographs. In this respect language assessment is no different. We should be basing our decisions about an individual’s language abilities on several different types of evidence, preferably gathered over time.

I now turn briefly to how language assessors have approached achieving quality in their assessments: how do you know what constitutes quality in assessment?

How Good are the Assessment Procedures? An Early Perspective

I am not going to start – for obvious reasons – as far back as the Han Dynasty for my analysis but will highlight one significant influence on language assessment practices as a consequence of the pursuit of so-called ‘objective measurement’. In the early part of the 20th century, there was a move in the US towards standardised and ‘objective’ language testing. This was made possible by advances in the field of psychometrics more generally and educational measurement specifically. These had an enormous impact – that you could call both invasive and pervasive - in terms of the effects on assessment practices generally as well as, indirectly, on the field of language assessment.
In brief, any procedure to assess a set of skills (language or otherwise) had to show that it was both reliable and valid and the way that this was done was by analysing test results statistically, with the statistics informing on whether the test was ‘good’ or ‘bad’. Reliability referred to the consistency of the assessment procedure in measuring what it was supposed to measure. You would get a very different result if you used a metre stick on one occasion to measure someone’s height and a piece of elastic on another. So clearly we would chose the metre stick to get a more reliable result. The other concept, validity, was concerned with whether what was supposed to be assessed was actually being assessed. You would therefore want to show that an assessment was indeed measuring height – as in my example - and not hair length or boot size!

The most important point to retain here is that a good assessment tool was one that showed itself to be both reliable and valid through rigorous statistical procedures. Thus, statistical goodness of fit was the primary quality indicator: an assessment procedure was good to the extent that the statistics were right (e.g. good correlations, appropriate factor loadings and so forth). The assessment of language was, thus, sometimes treated as something that could be done with some certainty, accompanied by the belief that it is possible to achieve a ‘true’ score, i.e. a ‘true’ account of an individual’s language abilities. As my colleague Richard Kiely suggested when he read an early draft of this talk: “A factor in constructing this notion of certainty is the demand for it by users for making decisions – a test is a technology that has been constructed as certain and safe, especially in the context where decisions that affect people in negative ways are challenged and have to be defended” (personal communication).

In 2001, for example, Extra and Gorter (2001, 12-13) wrote:

“the prognosis is that immigration populations will continue to grow as a consequence of the increasing number of political refugees, the opening of the internal European borders, and political and economic developments in Central and Eastern Europe and other regions of the world. It has been estimated that in the year 2000 about one third of the population under the age of 35 in urbanised Western Europe had an immigration background”.

Further, a survey of Roma in EU countries (reported in Bakker 2001) estimated that the official numbers in Greece to be between 150,000 and 300,000 and that approximately 90% of the Roma speak Romani. But, nowhere, can I find an example of the application of the CEF in the context of Romani language skills. This same comment would apply to a significant number of other languages spoken within the EU.

1. This assessment system is gathering considerable momentum, especially across higher income earning countries in Europe with the effect of affirming the ‘spread of English’. Although other languages - such as Spanish, Dutch, Greek, German, French, Italian and Portuguese - are assessed, these cannot be considered representative of linguistic diversity in the EU.

2. Whilst it is asserted that this framework may be used both to assess language achievement and to diagnose learner needs as the basis for nurturing language development, this latter function is far less well researched and articulated. There is thus a strong possibility of...
consequences may be that they fail their primary school leaving examinations and do not even have an opportunity to enter secondary education. Can this be called ‘fair’?

My penultimate example is a pan-European one.

**Context: An International Systems Approach to Language Assessment**

**Vignette 7: Uniformity, Conformity or Celebrating Diversity?**

The Council of Europe has always promoted life-long language learning amongst its citizens with the aim of facilitating communication for social or work purposes in different first languages. Three aims within European Policy are to achieve:

a) **quality**, in relation to improving the quality of language teaching and learning
b) **quantity**, in relation to an increasing number of citizens learning more languages more of the time
c) **diversity**, in relation to the promotion and maintenance of linguistic diversity.

In this example I want to consider the extent to which recent innovations in assessment in the European context in the form of the *Common European Framework for Assessment* (Council of Europe, 2001) promotes linguistic diversity (see Rea-Dickins, 2004). This assessment framework focuses on communicative language proficiency exemplified in terms of:

- 5 language skills: Reading, writing, listening, oral interaction and oral presentation, at
- 6 different levels of proficiency, from A1 = Breakthrough, A2; B1, B2; C1, C2 = Mastery

The intention is to be able to identify an individual’s language proficiency I now turn to some current perspectives on quality in assessment.

**Current Perspectives**

Reliability is still important. How could it not be? And of course we still need to know that what is being assessed is what we want to assess. But other considerations come into play that have to do with the use and *interpretation* of assessment results. I mention three here.

1. The notion of **consequences** from an assessment. The concept of *consequential validity* derive from the writings of Messick (1989) who was concerned with the social consequences of test use and how test data are interpreted. As Lynch observes:

   “Part of the goodness of a test, part of its validity, would be a consideration for the consequences of its use. If those consequences entail unethical or immoral outcomes, then the test and those associated with it are accountable, albeit to greater or lesser degrees.” (Lynch, 1997:317).

   This links with what Hamp-Lyons (1989:13, cited in Lynch, 1997) has identified as a core principle of ethical language testing that:

   “no test taker shall be harmed by the test”.

2. The notion of **fairness** and the need to give the individual a ‘fair chance’ to demonstrate his or her ability

3. The use of very different methodological approaches to establishing test validity, for example the analysis of qualitative data (i.e. this does not involve counting and tests of significance!) and, in particular, allowing the voices of different stakeholders in assessment processes to be heard. This could involve, for example, gathering information from both those who are assessed (i.e. individuals of all ages) and those who examine (i.e. examiners, raters, moderators) (see Rea-Dickins, 1997).

In summary, there is now a strong drive towards assuring, i.e. demonstrating, ethical assessment practice, and those involved in my profession adhere to a socially responsible code of ethics (e.g. The
Code of Ethics for the International Language Testing Association, 2002; American Psychological Association, 2005, USA). Thus, to conclude this first part of my presentation, we see a shift from a technicist approach, narrowly focused on the internal statistical properties of a test, to one that is concerned with how assessment procedures are used and the ethicality associated with such use in terms of consequences, i.e. consequential validity, for individuals or groups. As language assessors, we are not solely interested in the statistical properties of an assessment. We are also concerned about how assessments are used, by whom, for which purposes, in which contexts and with which consequences: this is the core of our ethical positioning.

So far, then, I have analysed some historical perspectives on the ways in which the quality of assessment procedures have been evidenced. I have also summarised some of the current thinking about how to gauge quality in the domain of language assessment. With this background in place, I wish, next, to share with you some examples of language assessment practices.

Uses, Misuses and Abuses of Language Assessment Data
Context: Assessment of Overseas Doctors’ Language Proficiency

Vignette 2: The ‘Language’ Test as ‘Proxy’

An early example is from language assessment in a medical context (see Rea-Dickins, 1987). I was working at the University of Lancaster funded by the General Medical Council to develop with a colleague, Dr Caroline Clapham, the first language assessment system in the UK for overseas doctors wishing to work within the National Health Service. We were also involved in the development of the interview, providing guidance for ‘medical-examiners’ on how to assess a doctor’s skill in

4. State policies generally provided a minimum of guidance to schools. (Rivera et al. 2000:xii)

Here we see that in spite of an Act to provide equitable educational opportunity for all students, accommodation policies and their implementation were shown to fall very short of the mark. Whilst the ELLs were now mostly included in the statutory assessments – a positive position - they may not be provided with adequate opportunities to show their potential and demonstrate fully their abilities and, thus, this is another example of a lack of fairness to the individual test taker that might have significant social and educational consequences, for reasons I have given above. I have shown in my own research with learners using English as an Additional language in mainstream classrooms in England (e.g. Rea-Dickins & Gardner, 2000) that if a child is classified as a grade higher than actual ability, one implication might be that additional language support may be either reduced or even withdrawn. On the other hand, where a child is assessed as being in need of additional language support, and in fact they are not, these learners may lose out in terms of being provided with more interaction with the curriculum and teaching that challenges rather than supports.

Further, I invite you at this point to reflect on how many school children throughout the world are learning their school subjects through a language that is not their first language. This will include the majority of children in Africa south of the Sahara where either English is the medium of instruction from middle primary school onwards, or an indigenous language – a lingua franca and not a child’s first language – is used. In some cases, children who have not yet fully developed their first language literacy start their learning careers in a language other than their first, with potentially stark consequences. Such educational
Accommodations are defined as changes that are made to an assessment that address student’s specific needs such that these students have access to the test to demonstrate fully their academic knowledge. Further:

*The use of accommodations is intended to level the playing field for students who receive them. In this case, the needs are linguistic because the students in question have still to develop full academic language proficiency.*

Examples of language-related accommodations include:

- The instructions for the assessments are read aloud
- The instructions are translated into the appropriate language
- The questions themselves are read aloud
- There are bilingual version(s) of the assessment
- Students dictate their answers.

An evaluation of state policies was conducted in 50 states, with some central findings as follows:

1. There was great variability in the degree to which state policies specifically addressed various aspects of ELLs participation in state assessment
2. States more frequently used accommodations that do not support the linguistic needs of the ELLs but were those carried over from Special Educational Needs (SEN) policies, such as additional time
3. The accommodations that can address the linguistic needs were least frequently allowed and most frequently prohibited, e.g. test translation, repetition of instructions

My first point concerns the use of the assessment data (by this I mean the results). Having been responsible for the analysis of the language assessment results, I was aware that in the first year of the assessment going live that only a handful – literally - of the overseas doctors failed the language component, with significant numbers underachieving on the tests of medical knowledge (i.e. the multiple choice written examination and/or the medical interview). Most curiously, however, when you read the reports in the press on these assessments, which were quite controversial at the time, the headlines in the press were along the lines of *Overseas Doctors Fail Language Tests*. I knew this to be erroneous: the failures in that first year should have been attributed to an lack in medical knowledge – but GMC press releases asserted that it was the doctor’s language that was not good enough. This was my first encounter and at the time a rather curious example of the use of a language test as a proxy for failure in another area but in fact analogous to my example of the Chinese candidate presenting for the Australian dictation test. I should add that Caroline and I were far too junior to exercise any refocusing of the press releases. I feel that with such a lapse of time I can speak openly: the GMC also refused permission for me to base my doctoral research on the development of these assessment procedures which were then really new in their approach to assessing communicative English language proficiency in the workplace. All the medical examiners at that time were male, of a certain age, including some specifically appointed to the Queen, with several most eminent and titled. How could a lowly applied linguistic researcher and female to boot hope to ‘put the record straight’!!
Vignette 3: ‘Undoing’ Authenticity

In this early example of an evidenced-based language test development project, we collected data from the Casualty departments at Lancaster Royal Infirmary and St Bartholomew’s in London. Our recordings revealed that both doctors and patients regularly encountered a wide range of different pronunciation and accents and to enhance the authenticity of the assessment (i.e. the validity of these language measures), the listening component of the test (developed by Caroline Clapham) included speakers with regional and international accents and intonation.

In keeping with the thinking at that time, the data from the tests were analysed statistically at Lancaster and only those parts that were found to be statistically reliable were retained but as this was done by the language specialists, we were able to ensure that the content coverage, or test syllabus, still represented those aspects of language use in a medical setting that were important to assess. However, within a relatively short space of time, there were vehement objections from the medical examiners about the regional accents and some parts of the listening test were re-recorded – with no involvement of the language specialists – and the ‘new’ listening test administered without any further analyses. Thus, whilst we were able – as applied linguists - to assure reliability and appropriate content coverage of the procedures, this no longer held once the items had been changed – to Received Pronunciation – when we were no longer involved with the assessments. In other words, this listening test was no longer valid in respect of the kinds of voices that doctors and patients would be hearing within a hospital casualty department and the reliability of these new items was ‘unknown’. This example would seem to be an instance of misuse of an assessment procedure based on – I would argue -

question and cut to the core of this Maths problem: multiply 40(p) by 60 (minutes). Some questions are extremely nuanced so that even the ‘more able’ English language learners may become confused and Scott (2005) in her doctoral research provides a wealth of interesting and sometimes worrying examples. For example, the language of the test question may be complex; it may contain superfluous information and distracting information; within one paper there may be examples of the same thing being said in different ways. It could be argued that being able to interpret such questions is embedded within what we define as mathematical competence. However, it has also been shown that far from providing a context for the learners to show what they do know, the reverse may also hold true in school assessment contexts (for example in Israel demonstrated by the work of Shohamy and colleagues). In cases where learners are denied the opportunity to show what they know, these cannot be considered fair assessments.

There are, however, ways in which learners working through English as an additional language may be supported within the context of ‘high stakes’ assessment situations, as evidenced by my next example.

Vignette 6: Assessment/Best Practice Policy Interface

The Improving America’s Schools Act of 1994 (cited in Rivera et al. 2000) required states to:

Implement assessment systems ... that allow all [my emphasis] students the opportunity to demonstrate their skills and knowledge” Rivera et al. 2000:ix )

One of the policies that states were required to have in place has to do with accommodations for learners who do not have English as their first language (known in the states as English Language Learners, ELL).
shall say something about the nature of these tests in a moment), are placed in an even more vulnerable position running a high risk of educational failure because their specific linguistic and academic needs are not being addressed (cf. TESOL 1997; McKay et al. 1994; SACSA, 2002) as they would be if placed in programmes in which academic subjects are initially taught in Spanish (home language) while the children are also developing their English language proficiency. In summary, with reference to the assessment-policy interface, we evidence the abuse of assessment results to confirm a theoretically untenable position about learning entangled with political aspiration.

My next example is a related one. A central problem with school age children who are learning through a language that they are still acquiring is that they may not know the ‘language involved in the assessment’ but may have considerably more knowledge of the academic content than can be detected by the test, but you cannot know this by simply looking at assessment results.

**Vignette 5: A Question of Language or Maths?**

Consider this example from materials linked to the National Numeracy Strategy in England and published by the DfES:

Roy’s mobile phone costs 15p, 25p or 40p per minute depending on the time of day. His monthly bill comes to £146. He also pays £27.50 per month in rental charges. How much would a call of 1 hour cost at the most expensive rate? (DfES 2001c, cited in Scott, 2005:194)

Learners who are working through the medium of English as an additional language may well know the answer to this question but this knowledge or competence may not be performed, i.e. demonstrated, if their English is so limited that they are unable to understand the inappropriately held views about language. I can confirm that such attitudes prevail to this day. The chair of the Australian Medical Council fairly recently (1996) interrupted the presentation of a colleague of mine who developed the Occupational English Test in Australia (McNamara, 2005:10) with:

“The quality of mercy is not strain’d,  
It droppeth as the gentle rain from heaven …

Surely any educated doctor should be able to recognise that quote?”

At which point, I would ask: Wherein lies the validity of that particular claim for what constitutes ‘knowing a language’?

Given that I work in an Education department, it would seem fitting that some of my vignettes are taken from school-based settings. I shall present three: two from the US and one from England.

**Context: School-based Assessment**

**Vignette 4: The Assessment-Policy Interface**

As we all know from the practice in England, the public reporting of assessment outcomes – for the different Key Stages - is a central part in our government’s armoury that demonstrates ‘policy delivery’. As Leung & Rea-Dickins point out, the increasingly prominent role assigned by policy makers to assessment has drawn attention to the complex and increasingly “problematic interface between assessment as a specialist practice and assessment as policy instrument” (forthcoming 2006:1; see also Broadfoot & Pollard, 2000), as my next example from the US shows.

In Arizona, there is provision for some form of bilingual schooling
whereby learners from a largely Hispanic background have opportunities to learn through the medium of both English and Spanish. This, from a second language acquisition perspective, affords better opportunities for both school subject and English language learning than the ‘English only’ programmes also offered in the state. In 2004, the Arizona Education Department sent out a press release with a set of school performance data (in a fashion similar to our league tables) to show that their Structured English Immersion programmes (SIP) – i.e. the English only instruction with very little first language use - were producing higher levels of achievement in English Language, Reading and Mathematics than the bilingual programmes within that state.

Students in structured English immersion programs outperformed students in bilingual programs in that they were anywhere from one to four months ahead between second and fourth grade, as much as six months ahead in fifth grade, and over a year ahead from sixth grade on. This means that for students in sixth grade and above, students in structured English immersion programs were over a year ahead of students in bilingual programs. (Arizona Department of Education, 2004)

These data (based on results from standardised achievement tests in Reading, Maths, and English) were then used by the officials of the Education Department to justify their preference for the Structured English Immersion programmes, which were held to be capable of producing better results.

Coincidentally, however, it is reported (MacSwan 2004; downloaded 28/10/05, http://www.public.asu.edu/~macswan/ade/AR08-13-03.htm) that the schools’ Superintendent, Tom Horne, had run his campaign for office on the ‘anti-bilingual education’ ticket and subsequently claimed that his “department’s vigorous ban on Spanish was working wonders for students”.

“There now that we are enforcing the voter-approved requirement that students who are not proficient in English be in structured English immersion programs, hopefully there are no additional students subjected to these educationally inferior bilingual programs,” Horne stated.


There are three points I wish to make here. Firstly, the public and parents were presented with ‘overwhelming’ evidence in favour of English only opportunities in school and generally speaking this audience is not in a position to engage critically and determine whether a piece of research is sound or otherwise. However, and this is my second point, the research community is able to identify that the research was methodologically flawed: it ignored socio-economic differences, it did not take into account how much previous English language teaching the students had been exposed to, it relied exclusively on test results and failed to investigate causal effects between test results and the different kinds of bilingual provision available (for fuller details, see TESOL 2004). There is, thus, no way of knowing whether the reported gains were due to greater resourcing, to some initial advantage, to programme placement or to another variable. Most disturbing of all, as my third observation, is that the recommendation to abandon any form of bilingual instruction ignores the rights of children to benefit from a bilingual curriculum. In turn, this may limit their ability to participate on an equal footing alongside their monolingual peers. It is all the more alarming when you consider the wider statistics: the 1998 Population Survey showed that 44% of Hispanic 16-24 year olds born outside the USA dropped out of high school and of these 80% are reported as having low English language proficiency (MacSwann et al. 2002:1). In other words, children who are already ‘disadvantaged’ linguistically, on the basis of test scores (and I
whereby learners from a largely Hispanic background have opportunities to learn through the medium of both English and Spanish. This, from a second language acquisition perspective, affords better opportunities for both school subject and English language learning than the ‘English only’ programmes also offered in the state. In 2004, the Arizona Education Department sent out a press release with a set of school performance data (in a fashion similar to our league tables) to show that their Structured English Immersion programmes (SIP) – i.e. the English only instruction with very little first language use - were producing higher levels of achievement in English Language, Reading and Mathematics than the bilingual programmes within that state.

Students in structured English immersion programs outperformed students in bilingual programs in that they were anywhere from one to four months ahead between second and fourth grade, as much as six months ahead in fifth grade, and over a year ahead from sixth grade on. This means that for students in sixth grade and above, students in structured English immersion programs were over a year ahead of students in bilingual programs. (Arizona Department of Education, 2004)

These data (based on results from standardised achievement tests in Reading, Maths, and English) were then used by the officials of the Education Department to justify their preference for the Structured English Immersion programmes, which were held to be capable of producing better results.

Coincidentally, however, it is reported (MacSwan 2004; downloaded 28/10/05, http://www.public.asu.edu/~macswan/ade/AR08-13-03.htm) that the schools’ Superintendent, Tom Horne, had run his campaign for office on the ‘anti-bilingual education’ ticket and subsequently claimed that his “department’s vigorous ban on Spanish was working wonders for students”.

“Now that we are enforcing the voter-approved requirement that students who are not proficient in English be in structured English immersion programs, hopefully there are no additional students subjected to these educationally inferior bilingual programs,” Horne stated.


There are three points I wish to make here. Firstly, the public and parents were presented with ‘overwhelming’ evidence in favour of English only opportunities in school and generally speaking this audience is not in a position to engage critically and determine whether a piece of research is sound or otherwise. However, and this is my second point, the research community is able to identify that the research was methodologically flawed: it ignored socio-economic differences, it did not take into account how much previous English language teaching the students had been exposed to, it relied exclusively on test results and failed to investigate causal effects between test results and the different kinds of bilingual provision available (for fuller details, see TESOL 2004). There is, thus, no way of knowing whether the reported gains were due to greater resourcing, to some initial advantage, to programme placement or to another variable. Most disturbing of all, as my third observation, is that the recommendation to abandon any form of bilingual instruction ignores the rights of children to benefit from a bilingual curriculum. In turn, this may limit their ability to participate on an equal footing alongside their monolingual peers. It is all the more alarming when you consider the wider statistics: the 1998 Population Survey showed that 44% of Hispanic 16-24 year olds born outside the USA dropped out of high school and of these 80% are reported as having low English language proficiency (MacSwann et al. 2002:1). In other words, children who are already ‘disadvantaged’ linguistically, on the basis of test scores (and I
shall say something about the nature of these tests in a moment), are placed in an even more vulnerable position running a high risk of educational failure because their specific linguistic and academic needs are not being addressed (cf. TESOL 1997; McKay et al. 1994; SACSA, 2002) as they would be if placed in programmes in which academic subjects are initially taught in Spanish (home language) while the children are also developing their English language proficiency. In summary, with reference to the assessment-policy interface, we evidence the abuse of assessment results to confirm a theoretically untenable position about learning entangled with political aspiration.

My next example is a related one. A central problem with school age children who are learning through a language that they are still acquiring is that they may not know the ‘language involved in the assessment’ but may have considerably more knowledge of the academic content than can be detected by the test, but you cannot know this by simply looking at assessment results.

**Vignette 5: A Question of Language or Maths?**

Consider this example from materials linked to the National Numeracy Strategy in England and published by the DfES:

> Roy’s mobile phone costs 15p, 25p or 40p per minute depending on the time of day. His monthly bill comes to £146. He also pays £27.50 per month in rental charges. How much would a call of 1 hour cost at the most expensive rate? *(DfES 2001c, cited in Scott, 2005:194)*

Learners who are working through the medium of English as an additional language may well know the answer to this question but this knowledge or competence may not be performed, i.e. demonstrated, if their English is so limited that they are unable to understand the

Inappropriately held views about language. I can confirm that such attitudes prevail to this day. The chair of the Australian Medical Council fairly recently (1996) interrupted the presentation of a colleague of mine who developed the Occupational English Test in Australia (McNamara, 2005:10) with:

> “The quality of mercy is not strain’d, It droppeth as the gentle rain from heaven … Surely any educated doctor should be able to recognise that quote?”

At which point, I would ask: Wherein lies the validity of that particular claim for what constitutes ‘knowing a language’?

Given that I work in an Education department, it would seem fitting that some of my vignettes are taken from school-based settings. I shall present three: two from the US and one from England.

**Context: School-based Assessment**

**Vignette 4: The Assessment-Policy Interface**

As we all know from the practice in England, the public reporting of assessment outcomes – for the different Key Stages - is a central part in our government’s armoury that demonstrates ‘policy delivery’. As Leung & Rea-Dickins point out, the increasingly prominent role assigned by policy makers to assessment has drawn attention to the complex and increasingly “problematic interface between assessment as a specialist practice and assessment as policy instrument” (forthcoming 2006:1; see also Broadfoot & Pollard, 2000), as my next example from the US shows.

In Arizona, there is provision for some form of bilingual schooling
Vignette 3: ‘Undoing’ Authenticity

In this early example of an evidenced-based language test development project, we collected data from the Casualty departments at Lancaster Royal Infirmary and St Bartholomew’s in London. Our recordings revealed that both doctors and patients regularly encountered a wide range of different pronunciation and accents and to enhance the authenticity of the assessment (i.e. the validity of these language measures), the listening component of the test (developed by Caroline Clapham) included speakers with regional and international accents and intonation.

In keeping with the thinking at that time, the data from the tests were analysed statistically at Lancaster and only those parts that were found to be statistically reliable were retained but as this was done by the language specialists, we were able to ensure that the content coverage, or test syllabus, still represented those aspects of language use in a medical setting that were important to assess. However, within a relatively short space of time, there were vehement objections from the medical examiners about the regional accents and some parts of the listening test were re-recorded – with no involvement of the language specialists – and the ‘new’ listening test administered without any further analyses. Thus, whilst we were able – as applied linguists - to assure reliability and appropriate content coverage of the procedures, this no longer held once the items had been changed – to Received Pronunciation – when we were no longer involved with the assessments. In other words, this listening test was no longer valid in respect of the kinds of voices that doctors and patients would be hearing within a hospital casualty department and the reliability of these new items was ‘unknown’. This example would seem to be an instance of misuse of an assessment procedure based on – I would argue –

question and cut to the core of this Maths problem: multiply 40(p) by 60 (minutes). Some questions are extremely nuanced so that even the ‘more able’ English language learners may become confused and Scott (2005) in her doctoral research provides a wealth of interesting and sometimes worrying examples. For example, the language of the test question may be complex; it may contain superfluous information and distracting information; within one paper there may be examples of the same thing being said in different ways. It could be argued that being able to interpret such questions is embedded within what we define as mathematical competence. However, it has also been shown that far from providing a context for the learners to show what they do know, the reverse may also hold true in school assessment contexts (for example in Israel demonstrated by the work of Shohamy and colleagues). In cases where learners are denied the opportunity to show what they know, these cannot be considered fair assessments.

There are, however, ways in which learners working through English as an additional language may be supported within the context of ‘high stakes’ assessment situations, as evidenced by my next example.

Vignette 6: Assessment/Best Practice Policy Interface

The Improving America’s Schools Act of 1994 (cited in Rivera et al. 2000) required states to:

Implement assessment systems … that allow all [my emphasis] students the opportunity to demonstrate their skills and knowledge” Rivera et al. 2000 ix)

One of the policies that states were required to have in place has to do with accommodations for learners who do not have English as their first language (known in the states as English Language Learners, ELL).
Accommodations are defined as changes that are made to an assessment that address student’s specific needs such that these students have access to the test to demonstrate fully their academic knowledge. Further:

_The use of accommodations is intended to level the playing field for students who receive them. In this case, the needs are linguistic because the students in question have still to develop full academic language proficiency._

Examples of language-related accommodations include:

- The instructions for the assessments are read aloud
- The instructions are translated into the appropriate language
- The questions themselves are read aloud
- There are bilingual version(s) of the assessment
- Students dictate their answers.

An evaluation of state policies was conducted in 50 states, with some central findings as follows:

1. There was great variability in the degree to which state policies specifically addressed various aspects of ELLs participation in state assessment
2. States more frequently used accommodations that do not support the linguistic needs of the ELLs but were those carried over from Special Educational Needs (SEN) policies, such as additional time
3. The accommodations that can address the linguistic needs were least frequently allowed and most frequently prohibited, e.g. test translation, repetition of instructions

communicating with patients. We were called in to act as patients and to moderate the language assessment part embedded within the ‘medical’ interview.

My first point concerns the use of the assessment data (by this I mean the results). Having been responsible for the analysis of the language assessment results, I was aware that in the first year of the assessment going live that only a handful – literally - of the overseas doctors failed the language component, with significant numbers underachieving on the tests of medical knowledge (i.e. the multiple choice written examination and/or the medical interview). Most curiously, however, when you read the reports in the press on these assessments, which were quite controversial at the time, the headlines in the press were along the lines of **Overseas Doctors Fail Language Tests**. I knew this to be erroneous: the failures in that first year should have been attributed to an lack in medical knowledge – but GMC press releases asserted that it was the doctor’s language that was not good enough. This was my first encounter and at the time a rather curious example of the use of a language test as a proxy for failure in another area but in fact analogous to my example of the Chinese candidate presenting for the Australian dictation test. I should add that Caroline and I were far too junior to exercise any refocusing of the press releases. I feel that with such a lapse of time I can speak openly: the GMC also refused permission for me to base my doctoral research on the development of these assessment procedures which were then really new in their approach to assessing communicative English language proficiency in the workplace. All the medical examiners at that time were male, of a certain age, including some specifically appointed to the Queen, with several most eminent and titled. How could a lowly applied linguistic researcher and female to boot hope to ‘put the record straight’!!
Code of Ethics for the International Language Testing Association, 2002; American Psychological Association, 2005, USA). Thus, to conclude this first part of my presentation, we see a shift from a technicist approach, narrowly focused on the internal statistical properties of a test, to one that is concerned with how assessment procedures are used and the ethicality associated with such use in terms of consequences, i.e. consequential validity, for individuals or groups. As language assessors, we are not solely interested in the statistical properties of an assessment. We are also concerned about how assessments are used, by whom, for which purposes, in which contexts and with which consequences: this is the core of our ethical positioning.

So far, then, I have analysed some historical perspectives on the ways in which the quality of assessment procedures have been evidenced. I have also summarised some of the current thinking about how to gauge quality in the domain of language assessment. With this background in place, I wish, next, to share with you some examples of language assessment practices.

**Uses, Misuses and Abuses of Language Assessment Data**

**Context: Assessment of Overseas Doctors' Language Proficiency**

**Vignette 2: The ‘Language’ Test as ‘Proxy’**

An early example is from language assessment in a medical context (see Rea-Dickins, 1987). I was working at the University of Lancaster funded by the General Medical Council to develop with a colleague, Dr Caroline Clapham, the first language assessment system in the UK for overseas doctors wishing to work within the National Health Service. We were also involved in the development of the interview, providing guidance for ‘medical-examiners’ on how to assess a doctor’s skill in

4. State policies generally provided a minimum of guidance to schools. (Rivera et al. 2000:xii)

Here we see that in spite of an Act to provide equitable educational opportunity for all students, accommodation policies and their implementation were shown to fall very short of the mark. Whilst the ELLs were now mostly included in the statutory assessments – a positive position - they may not be provided with adequate opportunities to show their potential and demonstrate fully their abilities and, thus, this is another example of a lack of fairness to the individual test taker that might have significant social and educational consequences, for reasons I have given above. I have shown in my own research with learners using English as an Additional language in mainstream classrooms in England (e.g. Rea-Dickins & Gardner, 2000) that if a child is classified as a grade higher than actual ability, one implication might be that additional language support may be either reduced or even withdrawn. On the other hand, where a child is assessed as being in need of additional language support, and in fact they are not, these learners may lose out in terms of being provided with more interaction with the curriculum and teaching that challenges rather than supports.

Further, I invite you at this point to reflect on how many school children throughout the world are learning their school subjects through a language that is not their first language. This will include the majority of children in Africa south of the Sahara where either English is the medium of instruction from middle primary school onwards, or an indigenous language – a lingua franca and not a child’s first language – is used. In some cases, children who have not yet fully developed their first language literacy start their learning careers in a language other than their first, with potentially stark consequences. Such educational
consequences may be that they fail their primary school leaving examinations and do not even have an opportunity to enter secondary education. Can this be called ‘fair’?

My penultimate example is a pan-European one.

**Context: An International Systems Approach to Language Assessment**

**Vignette 7: Uniformity, Conformity or Celebrating Diversity?**

The Council of Europe has always promoted life-long language learning amongst its citizens with the aim of facilitating communication for social or work purposes in different first languages. Three aims within European Policy are to achieve:

a) *quality*, in relation to improving the quality of language teaching and learning  
b) *quantity*, in relation to an increasing number of citizens learning more languages more of the time  
c) *diversity*, in relation to the promotion and maintenance of linguistic diversity.

In this example I want to consider the extent to which recent innovations in assessment in the European context in the form of the *Common European Framework for Assessment* (Council of Europe, 2001) promotes linguistic diversity (see Rea-Dickins, 2004). This assessment framework focuses on communicative language proficiency exemplified in terms of:

- 5 language skills: Reading, writing, listening, oral interaction and oral presentation, at
- 6 different levels of proficiency, from A1= Breakthrough, A2; B1, B2; C1, C2 = Mastery

The intention is to be able to identify an individual’s language proficiency

I now turn to some current perspectives on quality in assessment.

**Current Perspectives**

Reliability is still important. How could it not be? And of course we still need to know that what is being assessed is what we want to assess. But other considerations come into play that have to do with the use and interpretation of assessment results. I mention three here.

1. The notion of **consequences** from an assessment. The concept of *consequential validity* derive from the writings of Messick (1989) who was concerned with the social consequences of test use and how test data are interpreted. As Lynch observes:

   “Part of the goodness of a test, part of its validity, would be a consideration for the consequences of its use. If those consequences entail unethical or immoral outcomes, then the test and those associated with it are accountable, albeit to greater or lesser degrees.” (Lynch, 1997:317).

   This links with what Hamp-Lyons (1989:13, cited in Lynch, 1997) has identified as a core principle of ethical language testing that:

   “no test taker shall be harmed by the test”.

2. The notion of **fairness** and the need to give the individual a ‘fair chance’ to demonstrate his or her ability

3. The use of very different methodological approaches to establishing test validity, for example the analysis of qualitative data (i.e. this does not involve counting and tests of significance!) and, in particular, allowing the voices of different stakeholders in assessment processes to be heard. This could involve, for example, gathering information from both those who are assessed (i.e. individuals of all ages) and those who examine (i.e. examiners, raters, moderators) (see Rea-Dickins, 1997).

In summary, there is now a strong drive towards assuring, i.e. demonstrating, ethical assessment practice, and those involved in my profession adhere to a socially responsible code of ethics (e.g. The
In brief, any procedure to assess a set of skills (language or otherwise) had to show that it was both reliable and valid and the way that this was done was by analysing test results statistically, with the statistics informing on whether the test was ‘good’ or ‘bad’. Reliability referred to the consistency of the assessment procedure in measuring what it was supposed to measure. You would get a very different result if you used a metre stick on one occasion to measure someone’s height and a piece of elastic on another. So clearly we would chose the metre stick to get a more reliable result. The other concept, validity, was concerned with whether what was supposed to be assessed was actually being assessed. You would therefore want to show that an assessment was indeed measuring height – as in my example - and not hair length or boot size!

The most important point to retain here is that a good assessment tool was one that showed itself to be both reliable and valid through rigorous statistical procedures. Thus, statistical goodness of fit was the primary quality indicator: an assessment procedure was good to the extent that the statistics were right (e.g. good correlations, appropriate factor loadings and so forth). The assessment of language was, thus, sometimes treated as something that could be done with some certainty, accompanied by the belief that it is possible to achieve a ‘true’ score, i.e. a ‘true’ account of an individual’s language abilities. As my colleague Richard Kiely suggested when he read an early draft of this talk: “A factor in constructing this notion of certainty is the demand for it by users for making decisions – a test is a technology that has been constructed as certain and safe, especially in the context where decisions that affect people in negative ways are challenged and have to be defended” (personal communication).

level, irrespective of which language, that is informed by this ‘common’ / global scale. There is a growing literature relating to the use of this assessment framework about which I have two observations here.

1. This assessment system is gathering considerable momentum, especially across higher income earning countries in Europe with the effect of affirming the ‘spread of English’. Although other languages - such as Spanish, Dutch, Greek, German, French, Italian and Portuguese - are assessed, these cannot be considered representative of linguistic diversity in the EU.

   In 2001, for example, Extra and Gorter (2001, 12-13) wrote:

   “the prognosis is that immigration populations will continue to grow as a consequence of the increasing number of political refugees, the opening of the internal European borders, and political and economic developments in Central and Eastern Europe and other regions of the world. It has been estimated that in the year 2000 about one third of the population under the age of 35 in urbanised Western Europe had an immigration background”.

   Further, a survey of Roma in EU countries (reported in Bakker 2001) estimated that the official numbers in Greece to be between 150,000 and 300,000 and that approximately 90% of the Roma speak Romani. But, nowhere, can I find an example of the application of the CEF in the context of Romani language skills. This same comment would apply to a significant number of other languages spoken within the EU.

2. Whilst it is asserted that this framework may be used both to assess language achievement and to diagnose learner needs as the basis for nurturing language development, this latter function is far less well researched and articulated. There is thus a strong possibility of
unwanted washback on classroom language assessment practices, with undue teacher emphasis on the monitoring of language levels to the detriment of providing their students with opportunities to develop their language skills. This is something that is very obvious in my own classroom research in England. In my view, a restricted view of assessment that places systems to the fore has a tendency to obscure and stifle diversity in assessment contexts. Assessment procedures associated with identifying levels of achievement and standards have a certain homogeneity about them: not only do they tend to conform in terms of which languages are assessed, they have a tendency to conform in terms of approach, format, the skills assessed and so forth. Where, then, is the diversity we are requested to celebrate?

I therefore invite you to consider the impact of an assessment policy – which affirms the ‘spread of English’ and the use of other ‘highly valued’ foreign languages on:

1) the marginalisation of other languages
2) the devaluing of other languages which may, in turn, have a consequence for
3) the suppression of those languages which are perceived as unimportant (by an elite) relative to others which have more recognition, in that they are formally assessed. Is it a case of a language gaining its ‘cachet of approval’ through assessment.

We need, I suggest, to reflect on the appropriateness of our assessment mechanisms provided for our citizens and the diverse languages that they speak and how these might on the one hand inhibit, and on the other promote linguistic diversity. If the Council of Europe/EU is really science. We cannot evaluate an individual’s performance in a language with absolute precision, in the way we can measure the temperature of a liquid. One consequence of this is that in order to capture a ‘true’ reflection of language abilities, it is important to have different ways of capturing an individual’s language abilities through a range of different assessment procedures. In everyday life, if you wanted to know what someone looked like, you wouldn’t rely on just one photograph to show the different facets of a person, especially if it was taken first thing in the morning or late at night (or even at a lectern as I am now!): you would need a selection of photographs. In this respect language assessment is no different. We should be basing our decisions about an individual’s language abilities on several different types of evidence, preferably gathered over time.

I now turn briefly to how language assessors have approached achieving quality in their assessments: how do you know what constitutes quality in assessment?

How Good are the Assessment Procedures? An Early Perspective

I am not going to start – for obvious reasons – as far back as the Han Dynasty for my analysis but will highlight one significant influence on language assessment practices as a consequence of the pursuit of so-called ‘objective measurement’. In the early part of the 20th century, there was a move in the US towards standardised and ‘objective’ language testing. This was made possible by advances in the field of psychometrics more generally and educational measurement specifically. These had an enormous impact – that you could call both invasive and pervasive - in terms of the effects on assessment practices generally as well as, indirectly, on the field of language assessment.
interview. Assessment, however, is a broader concept, of which tests are a part. It includes the function of certification (e.g. GCSEs, A levels) and documenting achievement (e.g. at the National Curriculum Key Stages) within a school context. There are also examination boards, such as Cambridge Assessment (England) and Educational Testing Service (ETS, US), that provide certification for proficiency in English as a Foreign Language but this form of assessment is divorced entirely from specific learning and teaching contexts. By way of contrast, assessment is also central to effective teaching and learning – as an integral part of classroom instruction. Much of this ongoing assessment will be informal and take place through interaction between teachers and their learners. In this way, teachers develop keen insights into the strengths and weaknesses of the language abilities of their class and plan instruction accordingly; and the students themselves learn from the feedback provided by their teachers or their peers. In this last function of assessment, I am referring to ongoing formative classroom assessment, the focus of much of my own recent research, which aims to nurture rather than measure language development (e.g. Rea-Dickins 2003a, 2003b, 2001). This dimension of assessment links directly to research in second language acquisition (e.g. Ellis 2003, 1997, 1994) and in general educational assessment (for example, Assessment Reform Group 2001; Broadfoot & Black, 2004). This university is also in the process of developing its policy and practice in assessment, teaching and learning to ensure a balance between the development of learning on the one hand and the measurement of learning on the other. Thus, a teacher’s observation of a learner’s language use in the classroom - known as observation-driven assessment - is an important means to assess linguistic progress.

Secondly, I wish to signal that language assessment is not an exact about life-long language learning and celebrating linguistic diversity then the stranglehold of the prestige languages needs to be loosened. Assessment mechanisms such as the CEF have the potential to do this but I am very sceptical about whether they will succeed in so doing.

For my final example, I return to the immigration context.

Vignette 9: Language Assessment for Deportation Purposes

Language assessments are now used around the world as a forensic tool to detect illegal immigrants. In other words, assessment is being used to determine nationality and ethnicity rather than proficiency in a language. Eades and colleagues (2003) have identified significant problems in current practices and have shown that:

- A person’s nationality cannot always be determined by the language he or she speaks
- A few key words (for example on a tape and not even in a face to face interview) and their pronunciation cannot normally reveal a person’s nationality or ethnicity – this is reminiscent of the shibboleth test in the Book of Judges I mentioned at the outset
- Everyday perceptions about pronunciation differences among groups of people cannot be relied upon
- Any analysis of pronunciation must be based on thorough knowledge of the language and region in question and must involve detailed phonetic analysis.

(Eades et al., 2003:7-8)

In an analysis of 58 cases where linguistic identification led to a decision not to grant the right to remain in a country, Eades et al. found evidence that:

- the linguistic identification was based on insufficient data
- the qualifications of the analysts were not provided
- the linguistic identification ignored the fact that languages, of both
groups and individuals, can change through language contact and spread, and

- the possibility of the applicant accommodating to the interpreter’s dialect was not considered: in the case they report, the interpreter’s dialect was different from the applicant and the applicant appeared to alter his own pronunciation and choice of words in order to be understood by the Farsi-speaking interpreter.

(Consider within this context my own lexical shifts when I am in America: from lift to elevator, from pavement to sidewalk, from tomato to tomayto or from chips to French fries: this does not make me American).

Evidence of the kind reported above suggests an abuse of linguistic assessment practices where individuals are denied basic human rights on untenable grounds. I am not saying here that a linguistic assessment should not be done. What I am saying is that any such assessment should be theoretically grounded, and demonstrate its validity. Returning to the question I asked at the beginning of this presentation and whether we could find any examples analogous to the Australian dictation test used as a tool for exclusion, we can. Indeed, it is deeply concerning that language assessment can serve as an agent against the legitimacy of individuals to reside in countries in safety.

And now, finally, to some conclusions about the contemporary realities of language assessment practices.

**Contemporary Realities of Assessment: What are the Threats?**

The problems I have documented in my presentation reside not with the assessments themselves but, rather, in how they are used and I am guided in these conclusions by Messick’s (1989) construction of validity that has at its core notions of educational and social consequences in England and the use of some questions that have appeared too challenging for members of the general public surveyed!}

What follows is some advice for a prospective administrator of the assessment in respect of a specific candidate:

**Sir**

_I have the honour to inform you that a Chinese named Kang Yu Wei contemplates visiting Australia shortly. He was formerly a man of considerable standing in China. For various reasons some of which have regard to his moral reputation the Government have refused permission for him to land in Australia. I shall be glad, therefore, if you will issue instructions that in the event of this man arriving at Freemantle he is subjected to the dictation test which should be applied in such a manner as to ensure its efficiency. It is probable that Kang Yu Wei understands English. Inquiries should be made on that point, and your officers should be in readiness to apply the test say in Spanish or Italian._

Here we have a blatant example of an assessment used to prevent immigration – not just a misuse but an abuse of a language assessment system. But this wouldn’t happen now, would it? Or, would it? I shall come back to this question towards the end of my lecture.

**Background**

**What is Assessment?**

I want to make two general points here. Firstly, the term ‘assessment’ in my title, as opposed to ‘testing’, has been a deliberate choice. These terms are not synonymous, although both language testing and assessment have the aim of obtaining information about an individual’s language abilities. Tests are usually formal procedures - of the pencil and paper variety that could include a dictation procedure - but may also be in the form of speaking tests, either a 1-1 interview or a paired
Curriculum assessments and the way it uses these results to inform on literacy and other targets and policy delivery through the League Tables in the national press and, in the case of language assessment, the rather poor performances nationally in French. Before I move on to say something about how we may judge the quality of language assessment, I shall introduce my first example, from Australia in 1910, and to what McNamara calls “the notorious Australian Dictation Test” (McNamara, 2005:5).

**Vignette 1: Assessment as a Tool for Exclusion**

This assessment was used from 1901 until 1957 and nobody passed it from 1910 onwards. The dictation was not to be less than 50 words and could be in any language. The following is an English example of the dictated test (from McNamara, 2005:5):

> The tiger is sleeker, and so lithe and graceful that he does not show to the same appalling advantage as his cousin, the lion, with the roar that shakes the earth. Both are cats, cousins of our amiable purring friend of the hearthrug, but the tiger is king of the family.

McNamara also provides some other interesting details about this procedure, in particular a letter dated 4th March 1927 from the Commonwealth Home and Territories Department to the Collector of Customs at Freemantle:

> … The test when applied to an immigrant is intended to serve as an absolute bar to such a person’s entry into Australia, or as a means of depriving him of the right to remain in the Commonwealth if he has landed. The test should therefore be applied in a language with which the immigrant is not sufficiently acquainted to be able to write out at dictation.

Not all assessment is ‘bad’ (but that would be for another talk!). Many assessments are, by definition, meant to discriminate but we do need to show that this is done fairly. The boundaries of language assessment now extend well beyond educational and language proficiency certification contexts into the domains of ‘language for citizenship’ and immigration policy. However, the ways in which the language assessment profession attempts to assure quality within its traditional boundaries have not yet been extended to those new domains in which language data on individuals or groups are captured. What concerns me is what happens to language assessment procedures and/or results once these are out of the hands of the ethical and responsible language assessor? As I have highlighted, there are social and educational consequences for learners whose linguistic needs are inappropriately assessed and I have not even mentioned the considerable damage caused by some assessment practices, and the ensuing loss of self-esteem that can stay with an individual for a lifetime. I have shown that language assessment is not limited to the classroom, with social and also political consequences. In a discussion of where responsibilities lie (i.e. begin and end) within the assessment profession, Davies (1997) has questioned whether you would put the blame on an architect if the building he had designed was being used to house ugly racist meetings. The inference we draw from his comment is clear. However, as language assessment researchers, we are in a position to ‘out’ unethical and questionable practice, and to ask difficult questions. Indeed, it is our responsibility to do so. As I observed at the outset, language assessment is not an exact science yet, on the basis of the pronunciation of a few words on a tape, decisions are taken to deport an individual.

In conclusion, therefore, unlike Hans Blix who failed to find weapons of
mass destruction in Iraq, there are – I suggest - in the area of language assessment too many examples where assessment results are both misused and abused in a thoroughly destructive manner: to promote political careers, to exclude, to repatriate and to deny significant numbers of school learners with educational opportunities that would allow them to compete on equal terms with their peers, both nationally and globally. Assessment can be destructive in other ways, too. Assessment cultures – and England is a prime but not exclusive example – have become obsessed with summative assessment of achievement or certification, encouraging continuous measurement of language knowledge. Jumping on and off the bathroom scales every minute does not lead to weight loss; by the same token the continuous measurement of student achievement does not necessarily lead to learning.

In 1974, Jacobovits wrote of language testing as the ‘reactionary wing of Applied Linguistics’. This has alarming currency some 31 years later and I doubt whether he could have conjectured at the time of his comment just how reactionary the broader domain of language assessment might become.

NOTES
1. This is not a ‘new’ phenomenon. For example, it is reported that the Chinese imperial examination lasted between 24 and 72 hours and was conducted in cell like rooms. “The personal suffering that individuals underwent both in the preparation and in the taking of these exams has become part of Chinese lore. .... Some committed suicide because of the disgrace that these failures brought to their families. Others continued taking exams even as very old, grey-haired men.” (http://www.china.org)

The Weaponry of Language Assessment: what are the threats?

Introduction
In this presentation, there will be a short introduction in which I explain what I mean by assessment and I provide my first example. Secondly, I will briefly consider ways in which the language assessment profession has tried to achieve quality in its assessment practices. This will be followed by a range of examples on the basis of which – in my concluding remarks - we can determine whether language assessment practices warrant being equated with weapons, and whether in fact they pose any kind of threat.

Historically, we ‘go back a long way’ in language assessment, to the Bible and the Book of Judges in fact. Spolsky writes of the ‘shibboleth test’ that used dialectical differences to recognise and dispose of enemies (Judges, 12, 4-6 cited in Spolsky (1995:242). The Chinese imperial examination system – dating from the Han Dynasty, (206BC-220) - is an interesting example on account of its meritocratic nature: whereby appointment to the civil service was gained through ability and not “through special or inherited privilege” (http://www.china.org). This was a examination system that lasted for 2000 years and must have been the first attempt to introduce ‘objectivity’ (for example, by “sealing examinees’ names on paper, blocking the examination site during the exam, and separate evaluation of each answer sheet by three different examiners.” (http://www.china.org). There were the Jesuits in the seventeenth century who sought to control school curricula through an examination system (Madaus, 1990, cited in Spolsky 1995). More recently, indeed currently in England, we are able to recognise government control of our education system through the use of National
REFERENCES
Assessment Quarterly.


Canberra: National Languages and Literacy Institute of Australia.

McNamara, T. 2005. 21st Century Shibboleth: language Tests, Identity and Intergroup Contact. (Mimeo, provided by author)


