Principles and Practice in Language Testing: Compliance or Conflict?

J Charles Alderson,
Department of Linguistics and English Language,
Lancaster University

LTRG
Language Testing Research Group

LANCASTER
UNIVERSITY

Outline

- The Past
- The Past becoming Present – Present Perfect?
- The Future?

Standards?

Shorter OED:

- Standard of comparison or judgement
- Definite level of excellence or attainment
- A degree of quality
- Recognised degree of proficiency
- Authoritative exemplar of perfection
- The measure of what is adequate for a purpose
- A principle of honesty and integrity

Standards?

Report of the Testing Standards Task Force, ILTA 1995 (International Language Testing Association)

http://www.iltaonline.com/ILTA_pubs.htm

1. Levels to be achieved
2. Principles to follow

Standards as Levels

- Foreign Service Institute
- Interagency Language Round Table
- American Council for the Teaching of Foreign Languages
- Australian Second Language Proficiency Ratings

Standards as Levels

European traditions?

- 1-5? 4-10? F-A*?
- Beginner/ False Beginner/ Intermediate/Post Intermediate/Advanced
- How defined?

Standards as Principles

- Validity
- Reliability

- Authenticity?
- Washback?
- Practicality?

Psychometric tradition: Pluperfect?

- Tests externally developed and administered
- National or regional agencies responsible for development, following accepted standards
- Tests centrally constructed, piloted and revised
- Difficulty levels empirically determined
- Externally trained assessors
- Empirical equating to known standards or levels of proficiency

Standards as Principles: The Simple Past?

In Europe:

- Teacher knows best
- Having a degree in a language means you are an ‘Expert’
- Experience is all
- But 20 years experience may be one year repeated twenty times and is never checked

Past (?) European tradition

- Quality of important examinations not monitored
- No obligation to show that exams are relevant, fair, unbiased, reliable, and measure relevant skills
- University degree in a foreign language qualifies one to examine language competence, despite lack of training in language testing
- In many circumstances merely being a native speaker qualifies one to assess language competence.
- Teachers assess students’ ability without having been trained.

Past (?) European tradition

- Teacher-centred
- Teacher develops the questions
- Teacher's opinion the only one that counts
- Teacher-examiners are not standardised
- Assumption that by virtue of being a teacher, and having taught the student being examined, teacher-examiner makes reliable and valid judgements
- Authority, professionalism, reliability and validity of teacher rarely questioned
- Rare for students to fail

Past becoming Present: Levels

- Threshold 1975/ Threshold 1990
- Waystage/ Vantage
- Breakthrough/ Effective Operational / Mastery
- CEFR 2001
- A1 – C2

LTRG
Language Testing Research Group

LANCASTER UNIVERSITY

Past becoming Present: Levels

- CEFR translated into 37 languages so far:
- *Arabic*, Albanian, Armenian, Basque, Bulgarian, Catalan, *Chinese*, Croatian, Czech, Danish, Dutch, English, Esperanto, Estonian, Finnish, French, Friulian, Galician, Georgian, German, Greek, Hungarian, Italian, *Japanese, Korean*, Lithuanian, Moldovan, Norwegian, Polish, Portuguese, Russian, Serbian (Iekavian version), Slovak, Slovenian, Spanish, Swedish and Ukrainian.
- The CEFR is also currently being translated into Macedonian and Romanian.

Past becoming Present: Levels

- CEFR enormous influence since 2001
- ELP contributes to spread
- Claims abound
- Not just exams but also curricula/ textbooks
- Familiarity with CEFR claimed, but evidence suggests that this is extremely superficial.
- Claims of levels are made without accompanying evidence – by exam boards, textbook publishers and universities

Manual for linking exams to CEFR

- Familiarisation – essential, even for ‘experts’ – Knowledge is usually superficial
- Specification
- Standard setting
- Empirical validation

Manual for linking exams to CEFR

BUT FIRST

- If an exam is not valid or reliable, it is meaningless to link it to the CEFR

Language Testing Research Group

Standards as Principles: Validity

- Rational, empirical, construct
- Internal and external validity
- Face, content, construct
- Concurrent, predictive
- Construct

How can validity be established?

- My parents think the test looks good.
- The test measures what I have been taught.
- My teachers tell me that the test is communicative and authentic.
- If I take the X test instead of the FCE, I will get the same result.
- I got a good English test result, and I had no difficulty studying in English at university.

How can validity be established?

- Does the test match the curriculum, or its specifications?
- Is the test based adequately on a relevant and acceptable theory?
- Does the test yield results similar to those from a test known to be valid for the same audience and purpose?
- Does the test predict a learner’s future achievements?

How can validity be established?

Note: a test that is not reliable cannot, by definition, be valid

- All tests should be piloted, and the results analysed to see if the test performed as predicted
- A test's items should work well: they should be of suitable difficulty, and good students should get them right, whilst weak students are expected to get them

Factors affecting validity

- Unclear or non-existent theory
- Lack of specifications
- Lack of training of item/ test writers
- Lack of / unclear criteria for marking
- Lack of piloting/ pre-testing
- Lack of detailed analysis of items/ tasks
- Lack of standard setting
- Lack of feedback to candidates and teachers

Standards as Principles: Reliability

- Over time: test – re-test
- Over different forms: parallel
- Over different samples: homogeneity
- Over different markers: inter-rater
- Within one rater over time: intra-rater

Standards as Principles: Reliability

- If I take the test again tomorrow, will I get the same result?
- If I take a different version of the test, will I get the same result?
- If the test had had different items, would I have got the same result?
- Do all markers agree on the mark I got?
- If the same marker marks my test paper again tomorrow, will I get the same result?

Factors affecting reliability

- Poor administration conditions – noise, lighting, cheating
- Lack of information beforehand
- Lack of specifications
- Lack of marker training
- Lack of standardisation
- Lack of monitoring

Present Perfect?

Present Tense and Tension:
Practice vs. Principles

- Teacher-based assessment vs central development
- Internal vs external assessment
- Quality control of exams vs. no quality control
- Piloting or not
- Test analysis and the role of the expert
- The existence of test specifications – or not
- Guidance and training for test developers and markers – or not

Exam Reform in Europe
(mainly school-leaving exams)

- Slovenia
- The Baltic States
- Hungary
- Russia
- Slovakia
- Czech Republic
- Poland
- Germany
- Austria

Hungarian *English* Exams Reform Teacher Support Project

- Project philosophy:

  "The ultimate goal of examination reform is to encourage, to foster and to bring about change in the way language is taught and learned in Hungary."

Achievements of *English* Exam Reform Teacher Support Project

- Trained item writers, including class teachers
- Trained teacher trainers and disseminators
- Developed, refined and published Item Writer Guidelines and Test Specifications
- Developed a sophisticated item production system

Achievements of *English* Exam Reform Teacher Support Project

- In-service courses for teachers in modern test philosophy and exam preparation
  - Modern Examinations Teacher Training (60 hrs)
  - Assessing Speaking at A2/B1 (30 hrs)
  - Assessing Speaking at B2 (30 hrs)
  - Assessing Writing at A2/B1 (30 hrs)
  - Assessing Writing at B2 (30 hrs)
  - Assessing Receptive Skills (30hrs)

Achievements of *English* Exam Reform Teacher Support Project

- Developed sets of rating scales and trained markers
- Developed Interlocutor Frame for speaking tests and trained interlocutors
- Items / tasks piloted, IRT-calibrated and standard set to CEFR using DIALANG/ Kaftandjieva procedures

Achievements of *English* Exam Reform Teacher Support Project

- Into Europe series: textbook series for test preparation:
  - many calibrated tasks
  - explanations of rationale for task design
  - explanations of correct answers
  - CDs of listening tasks
  - DVDs of speaking performances

Into Europe

Reading + Use of English

Writing Handbook

Listening + CDs

Speaking Handbook + DVD

All downloadable for free from

http://www.lancs.ac.uk/fass/projects/examreform

LTRG

Language Testing Research Group

LANCASTER UNIVERSITY

But what happened since?

- No coordination between English, German, French, Spanish etc: Clash of testing cultures
- German Model preferred by Ministry official.
- English tasks not handed over, but published
- No piloting of any tasks in current examination
- Speaking tasks left up to teachers to design and administer, typically without any training in task design
- Administering speaking tasks to one candidate whilst four or more others are preparing their performance in the same room

Lack of professionalism

- No training of markers
- No double marking
- No monitoring of marking
- No comparability of results across schools, across markers/towns/ regions or across years (test equating)
- No guidance on how to use centrally devised scales, how to resolve differences, how to weight different components, no guidance on what is an “adequate” performance

Lack of professionalism

- Pre-setting cut scores without knowledge of test difficulty
- No understanding that the difficulty of a task item or test will affect the appropriacy of a given cut-score
- Belief that a ‘good teacher’ can write good test items: that training, moderation, revision, discussion, is not needed
- Lack of provision of feedback to item writers on how their items performed, either in piloting, or in live exam

Lack of professionalism

- Failure to accept that a ‘good test’ can be ruined by inadequate application of suitable administrative conditions, lack of or inadequate training of markers, lack of monitoring of marking, lack of double / triple marking.
- Lack of political will to fail students.
- Virtually all candidates pass -> Exam results worthless

Present Perfect? Negative features

- Political interference
- Politicians want instant results, not aware of how complex test development is
- Politicians afraid of public opinion as drummed up by newspapers
- Poor communication with teachers and public
- Resistance from some quarters, especially university ‘experts’, who feel threatened by and who disdain secondary teachers

Present Perfect? Negative features

- Assessment not seen as a specialised field: “anybody can design a test”
- Decisions taken by people who know nothing about testing
- Lack of openness and consultation before decisions are taken
- Urge to please everybody – the political is more important than the professional

The Austrian Matura

"There are no external examiners: Candidates are set tasks both for their written and oral finals by their own (former) teachers. Formally, however, there is an examination board consisting of a candidate's teachers/examiners, the headmaster/headmistress and a Vorsitzende(r) (head), usually a high-ranking school official or the head of another school. All oral exams are public, but attendance by anyone other than a candidate's former schoolmates is legally possible but not encouraged, and indeed rare."

(continued)

The Austrian Matura

“Criticism of the Austrian Matura has been persistent. In particular, it has been argued that the current system encourages rote learning, hinders candidates’ creativity and obscures the fact that the body of knowledge is constantly changing.“

(Wikipedia, last accessed 12 Sept. 2011)

The Reform

- Began in 2007, to be implemented across Austria in 2014/15
- Parallel reforms, coordinated by University of Innsbruck, in English, French, Spanish, Italian, Latin, Greek and Spanish.
- Other, non-standardised foreign languages can also be chosen.
- First foreign language (English) started in 2007, aiming at CEFR B2 in Listening, Reading and Language Use (The Written Examination)
- Second foreign languages (French, Italian, Spanish) 6-year and 4-year courses, targeted next (for 6-year courses, B2 except for Listening and Writing = B1. For 4-year courses, target is B1).

AHS Teams: 3 x 5 days training per year for 3 years

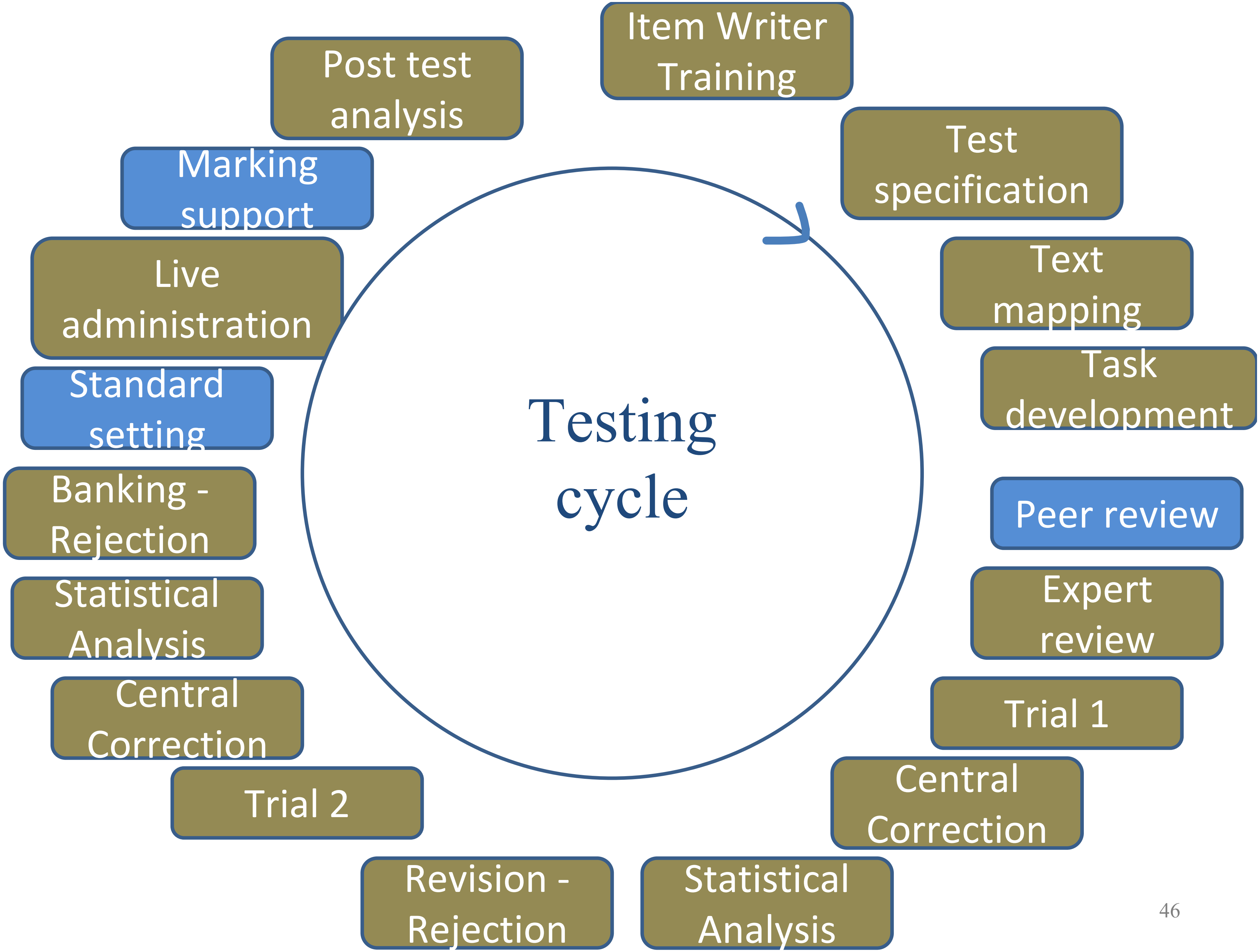
Team	N	Skill	Language	Training
A	15	Receptive	E, F	2007-2010
B	20	Receptive	E, F, I, S, R	2008-2011
C	15	Writing	E, F,	2009-2012
D	20	Receptive	E, F, I, S	2010-2012
E	20	Writing	E, F, I, S	2010-2012

BHS Team Training

Team	N	Skill	Language	Training
BHS 1	20	L, R Writing	E	2010-2013
BHS 2	16	L, R	F, I, S	2011-2014
BHS 3	12	Writing	F, I, S	2011-2014

The Reform

- A rolling reform, first with 59 selected pilot schools in 2008, then gradually spreading as new schools or individual teachers volunteer to take the new standardised Written Exam tasks.
- In Spring 2011, 300+ gymnasia volunteered to take tests in Reading, Listening and Language in Use in English, French, Italian or Spanish
- The Standardised Written Exam will be obligatory for all gymnasia in 2014 and for all vocational schools in 2015
- See **http://.uibk.ac.at/srp/**



Testing cycle

Item Writer Training

Test specification

Text mapping

Task development

Peer review

Expert review

Trial 1

Central Correction

Statistical Analysis

Revision - Rejection

Trial 2

Central Correction

Statistical Analysis

Banking - Rejection

Standard setting

Live administration

Marking support

Post test analysis

Some statistics

Students	2009	2010
English	14, 559	15, 335
French	818	930
Italian	123	155
Spanish	32	81
Schools	294	308
Teachers	1,457	1,146

Some more statistics

Students	2009	2010
English LC/RC	1,701	2,281
English LC/RC/LU	2,285	2,794
English LC only	10,573	10,260

LTRG
Language Testing Research Group

LANCASTER
UNIVERSITY

Some more statistics

Students	2009	2010
French 4-year LC/RC	16	45
French 6-year LC/RC	165	54
French 6-year LC	637	831

Some more statistics

Students	2009	2010
Italian 4-year LC/RC		3
Italian 6-year LC/RC	123	9
Italian 6-year LC		143

LTRG
Language Testing Research Group

LANCASTER
UNIVERSITY

Some more statistics

Students	2009	2010
Spanish 4-year LC/RC		27
Spanish 6-year LC/RC	32	3
Spanish 6-year LC		51

Number of test sessions

Test administrations	2009	2012
1. Main admin	5 +7 May	X
2. Main admin	12 +14 May	
3. Main admin	19 +20 May	
1st Re-sit	September	X
2nd Re-sit	January 2010	X

Negative features of the Reform, so far

- Work has been very demanding and hand-to-mouth
- 2007 First pilot May, 2nd pilot September, Standard setting December
- First test administration May 2008, thereafter 2 main sessions and 2 re-sits in 2008
- Small entries for second languages but equal amount of quality control work: sustainable?
- Inefficiency of multiple test administrations

Negative features of the Reform, so far

- Speaking ability not part of the reform
- Teachers will continue to set whatever test of speaking they wish
- No central marking of the Matura: teachers mark and determine the grades
- Teachers are said to want central marking but the unions are likely to oppose it.
- National elections in 2013: will Reform be a political football?
- No item bank yet.
- Disruption to schools when teachers attend workshops and when piloting takes place.

Positive features of the Reform

- Competence tested, not mere knowledge
- Fairness: all students tested on same standardised examination
- Students not subject to their teacher's idiosyncracies or biases
- Central correction of Written exam
- Hot line and Help desk

Positive features of the Reform

- Teachers and students enthusiastically positive
- Legal framework matching Project’s aims and design features.
- BiFIE very supportive
- Adequate resourcing of finance and personnel
- School-leaving exam now in sync with 2004 CEFR-based curriculum

Future Perfect?

- Sustainability – what happens when a project becomes a programme?
- End 2012 responsibility transfers from University of Innsbruck to BiFIE
- How will quality be guaranteed in future?
- Who will monitor?
- Who will do standard-setting?

Future Perfect?

- Will piloting continue?
- Who will train item writers? markers?
- Constant updating, development and research essential. Who?
- Validation process needs to be ongoing
- Stakeholders' reactions – employers, universities, media, unions?
- Washback and currency?

Lessons to be learned from Hungary and Austria

- Beware language rivalries and cultures
- Political support is essential
- Without full Ministry trust and cooperation things will go wrong
- Assessment literacy of officials is crucial
- Adequate continuing resourcing and staffing indispensable
- Quality and standards cannot be compromised

Language Testing Research Group

Lessons to be learned from Hungary and Austria

- Can a project change an assessment culture?
- Need to identify change drivers and resisters
- BUT agendas and personal ambitions often well hidden
- Xenophobia and Anglophobia?
- Grass roots support less important than support of those in power.
- How to win friends and influence: Hungary lost it. Austria?

Alderson, J.C. (Ed.) (2009) *The Politics of Language Education: Individuals and Institutions*. Bristol: Multilingual Matters

The Future: More Generally

Quis custodiat custodies?

The Future

- Validation of claims of links: Self-regulation acceptable? Role of ALTE? Role of EALTA? Role of IATEFL TEA SIG?
- Validation is not rubber stamping
- Claims of links will need rigorous inspection
- Codes of Practice? Not just for exams but also for classroom assessment

The Past and the Future

- Alderson and Buck (1993). Standards in testing: A study of the practice of British examination boards in EFL/ESL testing. *Language Testing*, 10/1 1-26
- Alderson et al (1995) *Language Test Construction and Evaluation*. CUP
- Alderson (2010). A survey of aviation English testing*, Language Testing*, 27 (1) 51 - 72

LANCASTER
UNIVERSITY

LTRG
Language Testing Research Group

The Past and the Future

ILTA Code of Ethics

http://www.iltaonline.com/code.pdf

EALTA Guidelines for Good Practice in Language Testing and Assessment

(in 35 languages)

http://www.ealta.eu.org/guidelines.htm

IATEFL TEA SIG Code of Practice?

LANCASTER UNIVERSITY

L T R G
Language Testing Research Group

The Present and the Future

- An IATEFL TEA SIG survey of the quality of national school-leaving examinations in English?

Why?

Good tests and assessment, following European standards, cost money and time

But

Bad tests and assessment, ignoring European standards, waste money, time and LIVES

Thank you for your attention!

c.alderson@lancaster.ac.uk

LTRG
Language Testing Research Group

LANCASTER
UNIVERSITY