ALTE Quality Assurance Checklists

Unit 4

Test analysis and Post-examination Review

Name(s) of people completing this checklist:

Which examination are the checklists being completed for?

At which ALTE Level is the examination at?

Date of completion:

Instructions for completing checklists.

1. On each sheet you will see various columns. The first one is headed 'Code of Practice' and each page includes one or sometimes two question(s) or point(s) under that column. These are about the main points from the Code of Practice.

2. In the second column are Focus Points, asking for more detailed information about the question under the Code of Practice.

3. For each focus point, please do the following:

a. If the question can be answered by Yes or No, put a cross in the appropriate column.
b. Under 'Comments and Reference' add in further **short** information. This might be a reference to certain documents or as an explanation of why you have ticked Yes or No.
c. In the final column, headed 'Self Evaluation', you will see four boxes for each Focus Point. These are headed as follows:

IMP = In need of improvement
UR = Under review
AD = Adequate
GP = Good practice

For each Focus Point you should tick one of these boxes, depending on whether **in your opinion** this needs to be improved within your organisation (IMP), whether this porcess is being reviewed by your organisation (UR), is adequately dealt with in your organisation (AD), or is good

4. At the end of the Unit you will find questions from the Code of Practice column repeated in Schedule D. Here you can add any longer information there was not room for in the boxes.

5. Please complete the document **electronically** and e-mail or send it on disk to the Secretariat by 1 February

6. **At the moment please do not send any supporting documents, only the questionnaire, even if you have referred to other documents in your answers.**

Example of a completed checklist – this is to give an example of how much information should be added to this part of the checklists.
Please add longer comments in Schedule D at the end of the Unit.

Unit 4. Test analysis and Post-examination Review							
Code of Practice	**Focus Points**			**Explanations and / or references**	**Self evaluation**		
		Yes	No		IMP	UR	AD	GP
K. Data Collection K1. Describe how data are collected.	I. Who is responsible for collecting and analysing data in your organisation?			*Data analysis officer*		x		
	ii. How are these staff recruited and trained?			*By interview. Must have relevant academic qualifications. No further training given.*	x			
	iii. Who decides which data should be collected and analysed?			*Data analysis officer.*		x		
	iv. Do you collect data during **pre-tests**?			*For some papers.*		x		
	v. Do you collect data during or after the live examination?			*Plan to introduce this in 2005.*		x		
	vi. What data are collected routinely or on a sample basis?			*Candidate information (age, L1 etc.)*	x			

Unit 4. Test analysis and Post-examination Review								
Code of Practice	**Focus Points**			**Explanations and / or references**	**Self evaluation**			
		Yes	No		IMP	UR	AD	GP
K. Data Collection **K1. Describe how data are collected.**	I. Who is responsible for collecting and analysing data in your organisation?							
	ii. How are these staff recruited and trained?							
	iii. Who decides which data should be collected and analysed?							
	iv. Do you collect data during pre-tests?							
	v. Do you collect data during or after the live examination?							
	vi. What data are collected routinely or on a sample basis?							

Unit 4. Test analysis and Post-examination Review								
Code of Practice	**Focus Points**			**Explanations and / or references**	**Self evaluation**			
		Yes	No		IMP	UR	AD	GP
	vii. Do you collect item level data?							
	viii. How are **scores** and other data stored and accessed?							
	ix. Who has access to this data?							
	x. Which **populations or population samples** are collected?							
	xi. How do you ensure the representativeness of the samples from which data are gathered?							

Unit 4. Test analysis and Post-examination Review								
Code of Practice	**Focus Points**			**Explanations and / or references**	**Self evaluation**			
		Yes	No		IMP	UR	AD	GP
K2. Describe how each test or test component is analysed in terms of its psychometric characteristics.	I. Do you carry out analysis to address the following features of the examination;							
	Difficulty?							
	Discrimination?							
	Reliability?							
	ii. If so, how do you do it?							
	iii. If not, how do you address these issues? For example: how do you calculate the difficulty of the items / tasks used in the sub-tests in relation to the ALTE Framework / Common European Framework of Reference.							
	iv. Do you use **internal consistency estimates** for item-based tests (such as **Cronbach's alpha**)							

Unit 4. Test analysis and Post-examination Review								
Code of Practice	**Focus Points**			**Explanations and / or references**	**Self evaluation**			
		Yes	No		IMP	UR	AD	GP
	v. Do you use measures of **reliability**, such as **test - retest** methods?							
	vi. Do you estimate **rater agreement** for tests of writing, speaking or constructed response items?							
	vii. If so, what data are used and what kind of analyses are carried out?							
	viii. Do you establish targets or minimum standards of **reliability** for your sub-tests?							
	ix. If so, what are these standards and how are they monitored?							

Unit 4. Test analysis and Post-examination Review								
Code of Practice	**Focus Points**			**Explanations and / or references**	**Self evaluation**			
		Yes	No		IMP	UR	AD	GP
K3. Describe how the performances of candidates from different groups are analysed.	I. Do you compare the results of different groups in terms of psychometric characteristics in order to detect unwanted differences such as test or item **bias**?							
	ii. What kinds of analysis are carried out?							
	iii. Is this done routinely or on a sample basis?							
	iv. How is this information used?							

Unit 4. Test analysis and Post-examination Review								
Code of Practice	**Focus Points**			**Explanations and / or references**	**Self evaluation**			
		Yes	No		IMP	UR	AD	GP
L. Confidentiality L1. Describe the procedures to protect the confidentiality of all data (raw or processed).	I. How is the security and confidentiality of all materials guaranteed when the materials are at the test centres or in transit?							
	ii. Are exam papers and scripts stored so that access is limited to authorised personnel only? For example: do you have secure storage facilities and password protected databases?							

Unit 4. Test analysis and Post-examination Review							
Code of Practice	**Focus Points**			**Explanations and / or references**	**Self evaluation**		
		Yes	No		IMP	UR	AD	GP
M. Monitor and report on the impact of examinations. M1. Describe how the results of validation and research are made available to internal colleagues, external stakeholders and to the public in suitable forms.	I. How do you collect the views and attitudes of the different stakeholder groups, for example to investigate test **impact**?							
	ii. How is this information analysed and reported?							
	iii. How is the information used to inform the routine test development process or to identify the need for change?							

Unit 4. Test analysis and Post-examination Review								
Code of Practice	**Focus Points**			**Explanations and / or references**	**Self evaluation**			
		Yes	No		IMP	UR	AD	GP
M2. Describe how the relevant measurement concepts are explained for users.	i. Do you produce regular reports on the examination which are provided to user groups? For example: on a quarterly or yearly basis.							
	ii. If so, what information is included in these reports?							
	iii. How are they distributed to ensure access to them by users?							
	iv. Do you produce articles in scientific and educational magazines for a range of stakeholders?							

Unit 4. Test analysis and Post-examination Review							
Code of Practice	**Focus Points**			**Explanations and / or references**	**Self evaluation**		
		Yes	No		IMP	UR	AD	GP
	v. If so, which ones?							
	vi. Do you carry out presentations at seminars and national / international conferences?							
	vii. Do you produce brochures, handbooks, leaflets for different stakeholder groups which attempt to explain issues to do with language assessment? For example: do these publications describe what your examination is measuring and how the results should be interpreted by users?							

Schedule D

Please add below any further information you have in answer to the questions:

K Data Collection

K1 Describe how data are collected.

K2 Describe how each test or test component is analysed in terms of its psychometric characteristics.

K3. Describe how performances of candidates from different groups are analysed.

L Confidentiality

L1 Describe the procedures to protect the confidentiality of all data (raw or processed).

M Monitor and report on the impact of examinations.

M1 Describe how the results of validation and research are made available to internal colleagues, external stakeholders and to the public in suitable forms.

M2 Describe how the relevant measurement concepts are explained for users.

Term	Definition
Assessment	In language testing, the measurement of one or more aspects of language proficiency, by means of some for of test or procedure.
Bias	A test or item can be considered to be biased if one particular section of the candidates population is disadvantaged by some particular aspect of the test or item which is not relevant to what is being measured. Sources of bias may be connected with gender, age,
Calibrate	In item response theory, to estimate the difficulty of a set of test items
Certificates	A document stating that a names person has taken a test or component of a test and had achieved a particular grade, usually at least a pass. See also *Diploma*
Clerical markers	A method of marking in which markers do not need to exercise any special expertise or subjective judgement. The mark by following a mark scheme which specifies all acceptable responses to each test item.
Communicative competence	The ability to use language appropriately in a variety of situations and settings.
Construct	A hypothesized ability or mental trait which cannot necessarily be directly observed or measured, for example, in language testing, listening ability. Language tests attempt to measure different constructs which underlie language ability. In addition to language ability itself, motivation, attitude and acculturation are all relevant constructs.
Construct validity	A test is said to have construct validity of the scores can be shown to reflect the theory about the nature of a construct or its relation to other constructs. It could be predicted, for example, that two valid tests of listening comprehension would rank learners in the same way, but each would have a weaker relationship with the scores on a test of grammatical competence.
Constructed response	A form of written response to a test item that involves active production, rather than just choosing from a number of options.
Content validity	A test is said to have content validity if the items or tasks of which it is made up constitute a representative sample of items for the area of knowledge or ability to be tested. These are often related to a syllabus or course.
Co-ordination session	For the assessment of Speaking and Writing human markers (raters / examiners)are required. Six aspects of the process of ensuring that the markers can mark in a reliable and valid way can be identified: RITCME - **R**ecruitment; **I**nduction and Initial Training; **T**raining for the specific exam; **C**o-ordination (before each exam takes place or at least regularly); **M**onitoring of their conduct; **E**valuation of their conduct. A co-ordination session is the event to ensure that all examiners have been co-ordinated effectively before they examine.
Criterion-related validity	A test is said to have criterion-related validity if a relationship can be established between test scores and some external criterion which is believed to be a measure of the same ability. Information on criterion-relatedness is also used in determining how well a test predicts future behaviour.
Cronbach's alpha	A reliability estimate, measuring the internal consistency of a test. It ranges in value from 0 to 1. It is often used for tests with rating scales as opposed to tests with dichotomous items, although it may be used for both. Also referred to as coefficient alpha.

Curriculum	An overall description of the aims, content, organisation, methods and evaluation of an educational course.
Cut score	The minimum score a candidate has to achieve in order to get a given grade in a test or examination. In mastery testing, the score on a test which is considered to be the level required in order to be considered minimally competent or at 'mastery' level.
Difficulty (index)	In classical test theory, the difficulty of an item is the proportion (p) of candidates responding to it correctly. This means that the difficulty estimate of an item is sample dependent, and changes according the level of ability of candidates.
Diploma	A document stating that a names person has taken a test or component of a test and had achieved a particular grade, usually at least a pass. Often interpreted as being of a higher level qualification than a certificate. See also *Certificate*
Examiner	Someone who assigns a score to a candidate's responses in a test, using subjective judgement to do so. Examiners are usually qualified in the relevant field and are required to undergo a process of training and standardization. In oral testing the roles of examiner and interlocutor are sometimes distinguished. Also referred to as assessor or rater.
Grading	The process of converting test scores or marks into grades.
Impact	The effect created by a test, both in terms of influence on general education process, and in terms of the individuals who are affected by the results.
Internal consistency (sample / estimate)	A feature of a test, represented by the degree to which candidates' scores on individual items in a test are consistent with their total score. Estimates of internal consistent can be used as indices of test reliability, various indices can be computed, for example KR-20 alpha. See also **Cronbach's alpha**
Invigilator	A person of authority employed at an examination centre to ensure that the exam is conducted according to the established procedures.
Marker	Someone who assigns a score to a candidate's responses to a written test. This may involve the use of expert judgement, or in the case of a clerical marker, the relatively unskilled application of a mark scheme.
Marking	Assigning a mark to a candidate's responses to a test. This may involve professional judgement, or application of a mark scheme which lists all acceptable responses.
Optical mark reader (OMR)	An electronic device used for reading information directly from answer sheets or mark sheets. Candidates or examiners can mark item responses or tasks on a mark sheet and this information can be read directly into a computer. Also referred to as scanner.
Performance	The act of producing language by speaking or writing. Performance, in terms of language actually produced by people, is often contrasted with competence, which is the underlying knowledge of a language.
Population sample	A selection of a sub-set of elements from a population.
Pretesting	A stage in the development of test materials at which items are tried out with representative samples from the target population in order to determine their difficulty. Following statistical analysis, those items that are considered to be satisfactory can be used in live tests.

Proficiency	Knowledge of a language and a degree of skill in using it.
Rater	See definition for examiner
Rater agreement	The degree of agreement between two assessments of the same sample of performance made at different times by the same assessor. This has particular relevance to the assessment of speaking and writing skills in tests where subjective judgements by examiners are required.
Regulations	An official document provided by the examination board which states the conditions under which enrolment for the exams, the conduct of the exams and the issue of results will be made. Candidates need to be aware of the regulations before the take the exam, including the rights and obligations they are signing up to.
Reliability	The consistency or stability of measures from a test. The more reliable a test is, the less random error it contains. A test which contains systematic error, e.g. bias against a certain group, may be reliable, but not valid. See also **Test - Retest**
Results	The outcome of a test, as reported to a test taker or test user.
Rubrics	The instructions given to candidates to guide their responses to a particular test task.
Score	A) The total number of points someone achieves in a test, either before scaling (raw score) or after scaling (scaled score). B) To assign numerical values to observed performance.
Standard error of measurement (SEM)	In classical true score test theory, the SEM is an indication of the imprecision of a measurement. The size of the standard error of measurement depends of the reliability and the standard deviation of the test scores.
Standardisation	The process of ensuring that assessor adhere to an agreed procedure and applying rating scales in an appropriate way.
Supervisor	A senior invigilator who is responsible for the conduct of an examination at an examination centre or in the examination room.
Test-retest	An estimate of reliability obtained by administering the same test to the same candidates in the same conditions, and correlating the scores on two sittings. It is concerned with the stability of scores over time, and is also appropriately used where estimates of internal consistency are not possible.
Validity	The extent to which scores on a test enable inferences to be made which are appropriate, meaningful and useful, given the purpose of a test. Different aspects of validity are identified, such as content, criterion and construct validity; these provide different kinds of evidence for judging the overall validity of a test for a given purpose. See also: *Construct validity, content validity, criterion related validity*
Discrimination	The power of an item to discriminate between weaker and stronger candidate. Various indices of discrimination are used. Some (e.g. biserial, point-biserial) are based on the correlation between the score on the item and a criterion, such as the total score on the test or some external measure of proficiency. Others are based on the difference in the item's difficulty for high and low ability groups. In item response theory the 2, and 3, parameter models estimate item

K1(iv) Do you collect data during pretests?

Revision

- delete bad items / tasks
- selection anchor items / tasks (minimum 20 items)
- revision next year

Supplementary research

- to determine cut off score of PMT, PPT en PAT:external scoring by professionals + anchoring with exams 'old-style'
- anchoring to state examinations NT2

- quality of examiners and test effect results

- in depth research on rating and rater reliability

- research on use of open tasks as anchor tasks

- comparing task outsome with task construction

K1 Describe how data are collected.

Question (e.g. C2i)	Answer