Principles of Good Practice for ALTE Examinations

Revised Draft
October 2001

Drafted by the ALTE Working Group on the Code of Practice
To be discussed by ALTE Members

Contents

1.0 Introduction

Fairness is an overriding concern in all aspects of assessment and provides a context for the *principles of good practice* discussed below.

Whenever examinations are widely used within educational contexts they affect not only individuals but also institutions and society as a whole. Given the potentially wide-ranging impact of examinations, it is important for both *examination developers* and the *examination users* to follow an established Code of Practice which will ensure that the assessment procedures are of high quality and that all stakeholders are treated fairly. A code of practice of this kind must be based on sound principles of good practice in assessment which allow high standards of quality and fairness to be achieved.

The discussion of what constitutes good practice presented in this paper is an attempt to reflect a concern for accountability in all areas of assessment which are undertaken by ALTE members. It recognizes the importance of *validation* through the collection of data and the role of *research and development* in examination processes. In this respect, it is likely that the principles which are outlined below will continue to evolve over time as research and development programmes expand.

1.1 The ALTE *Code of Practice* (1994)

1.1.1 In the fields of psychological and educational assessment, the USA has a long tradition of setting standards. In 1999 the latest edition of the *Standards for educational and psychological testing* (AERA/APA/NCME) was published, but earlier editions were around since 1954. In the field of language assessment, the *Standards* have influenced both language test developers and test users and are regularly referred to in the language testing literature (see for example Bachman 1990 and Bachman and Palmer 1996). In the USA since the early 80s, ETS have produced their own *Standards for Quality and Fairness* (1981, 1987, 2000) drawing heavily on the AERA/APA/NCME *Standards*.

More recently, ILTA - the International Language Testing Association – conducted a review of international testing standards (1995) and in 2000 published its *Code of Ethics* (2000); this document presents a set of nine principles with annotations which *"draws upon moral philosophy and serves to guide good professional conduct."*

1.1.2 Within the European context the work of ALTE itself has begun to exert an influence in relation to professional standards. In 1994, ALTE published its first *Code of Practice* which set out the standards that members of the association aimed to meet in producing their language exams.

Prior to its publication, this *Code of Practice* was initially drafted and discussed by ALTE members in 1991-93. It drew on *The Code of Fair Testing Practices in Education* produced by the Washington D.C. Joint Committee on Testing Practices (1988) and was intended to be a broad statement of what the users of the examinations should expect and the roles and responsibilities of stakeholders in striving for *fairness*.

1.1.4 The *Code of Practice* identifies three major groups of stakeholders in the testing process:

- the *examination developers,* (i.e. examination boards and other institutions which are members of ALTE)
- the *examination takers,* who takes the examination by choice, direction or necessity
- the *examination users,* who requires the examination for some decision-making or other purpose.

1.1.5 In addition the Code of Practice lays down responsibilities of the stakeholder groups in *four broad areas*:

- developing examinations
- interpreting examination results
- striving for fairness
- informing examination takers

An important feature is that it emphasises the joint responsibility of the stakeholders and focuses on the responsibilities of the *examination users* as well as the *examination developers* in striving for fairness*.*

1.1.6 A supplementary document entitled *Principles of Good Practice for ALTE Examinations* was drafted (Saville and Milanovic, 1991 and 1993) and discussed at ALTE meetings (Alcalà de Henares, 1992, Paris and Munich, 1993). This document was intended to set out in more detail the *principles* which ALTE members should adopt in order to achieve their goals of high professional standards.

The approach to achieving good practice was influenced by a number of sources from within the ALTE membership (e.g. work being carried out by UCLES) and from the field of assessment at large, (e.g. the work of Bachman, Messick and the *AERA/APA/NCME Standards*, 1985).

ALTE members sought feedback on the document from eminent external experts in the field (Spolsky, Bachman, Jan-Mar 1994) and was discussed in Arnhem 1994. While it was not published in its entirety, parts of the document were later incorporated into the *Users Guide for Examiners* produced by ALTE on behalf of the Council of Europe (1997).

1.1.7 This revised version of *Principles of Good Practice for ALTE Examinations* (2001) is based on the earlier version but has been thoroughly updated and reworked in many parts. It addresses in more detail the central issues of *validity* and *reliability* and looks at the related issues surrounding the impact of examinations on individuals and on society.

1.1.8 This version, like the earlier drafts, draws heavily on the revised *AERA/APA/NCME Standards* document (1999) - especially in the sections on validity and reliability - as well as the work of Bachman, 1990 and Bachman and Palmer, 1996.

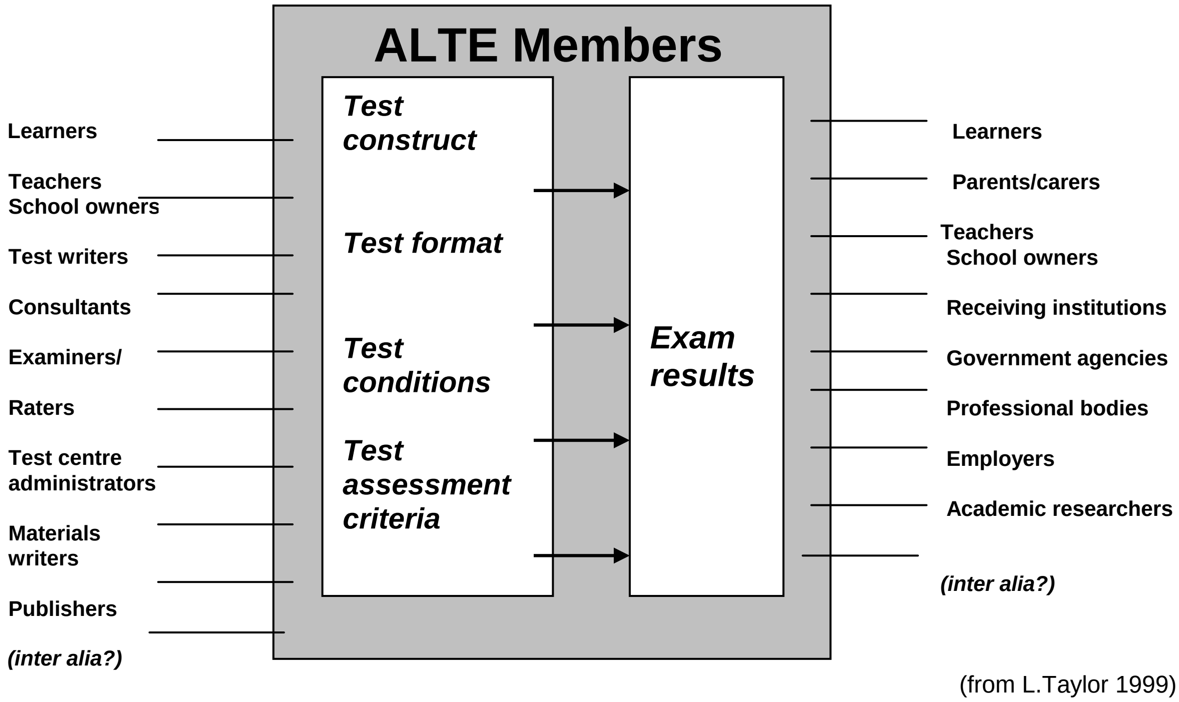
1.2 Participants in the Examination Process

1.2.1 Examinations affect not only individuals but institutions and society as a whole as noted in the ALTE *Code of Practice*. However, The three broad categories of stakeholder identified in the *Code of Practice* do not represent the full range of participants in the examination processes surrounding the ALTE examinations.

The individuals who are affected by the exams include the takers and their sponsors (students, parents, teachers, job applicants, employees etc). In addition there are the professionals and academics involved in the process of designing, writing, administering and validating the exams themselves (i.e. teachers, item writers, consultants, examiners, school owners, test centre administrators, supervisors, text-book writers etc.).

The institutions affected may include schools, universities and colleges, government agencies, publishers, businesses and industry.

Individuals and institutions benefit when testing helps them achieve their goals, and society benefits when the achievement of these goals contributes to the general good.



(from L.Taylor 1999)

This diagram shows the wide range of stakeholders who can be considered *participants in the examination processes*; they include those who contribute to the production and administration of ALTE examinations and those who make use of the test scores and certificates.

ALTE members must be prepared to monitor the views and attitudes of this constituency and to review/change what they do in light of the way these stakeholders use the exams and what they think about them.

1.3 Achieving Good Practice

1.3.1 The appropriate involvement of the stakeholder groups (outlined above) in examination processes is an important principle in achieving good practice.

1.3.2 Achieving good practice also depends on two fundamental principles:

a) the rational planning and management of resources relating to the development, administration and validation of examinations;

b) the collection, storage and use of data/information about all aspects of the examining process.

Failure to capture adequate data means that *evidence* of standards being reached and maintained cannot be provided (e.g. regarding validity, reliability, impact and practicality).

1.3.3 Good practice and thus high quality examinations can only be achieved if appropriate procedures are implemented for *managing* all aspects of the examination process. It is therefore necessary to adopt a rational approach (that incorporates the notion of iterative cycles) to examination development, administration and validation.

1.3.4 The first stage of this approach must involve *planning* including a detailed situational analysis (i.e. a feasibility study) which looks at the perceived need for a new examination within a given educational context. The aim is to identify the *considerations and constraints* which will be relevant to the examination development project and which will determine how examination usefulness will be achieved.

1.3.5 Whenever it is decided that a new examination development should go ahead, there should be agreed procedures which address at least the following areas:

- the management structure for the development project;
- a clear and integrated assignment of roles and responsibilities;
- a means of monitoring progress in terms of development schedules and resources;
- a methodology for managing the examination production process when the examination becomes operational (i.e. item writing, vetting, moderation, pre-testing, item banking, question paper construction).

1.3.6 Once an examination becomes operational, information must be collected regarding the production of materials and administration in order to judge whether the procedures are reaching expectations regarding aspects of practicality such as cost and efficiency. The cyclical processes which follow the initial planning involve on-going monitoring of the examination development itself and the subsequent live administrations of the examination. Careful record keeping and data collection is required to monitor all activities. The techniques of monitoring include all kinds of records which serve to establish a documented history of the examination; these in turn serve for both *formative* and *summative* evaluation. The data which is collected can be both "hard data" (empirically collected facts and figures) and "soft data" (feelings, impressions, attitudes, etc.).

1.3.7 A key aspect of this approach is that validation is an *integral* part of the process. In order for this to occur, the procedures which are implemented for the on-going production and administration of the examination should be designed so that adequate data can be collected, stored and retrieved as required.

1.3.8 As a principle, it will be the *overall usefulness* of an examination that must be maximised. This means that it is inevitable that evidence collected regarding one aspect of an examination will be relevant to the others. For example, data collected on examination reliability will not only be used in a narrow way to address the question of reliability, but will also be used to address questions of validity and impact (as defined below).

1.4 The Concept of Usefulness in Examinations

1.4.1 The principles of good practice proposed here are aimed at ensuring that ALTE examinations can be shown to meet explicit criteria in terms of the following *examination qualities*:

- Validity
- Reliability
- Impact
- Practicality

- Quality of Service

By addressing these aspects of their examinations in a principled way, the members of ALTE ensure that their commitments in their *Code of Practice* are met and sufficiently high standards can be maintained.

1.4.2 Not surprisingly the qualities of *validity and reliability* have been discussed extensively in the literature on measurement and language testing. For example, the *Standards for educational and psychological testing* (both the 1985 and 1999 editions) provide extensive discussions and Bachman 1990 dedicates a chapter to each. The other two qualities have always been important considerations for examination developers but have only recently emerged in the language testing literature. It is now broadly recognised that the individual examination qualities cannot be evaluated independently and that the relative importance of the qualities must be determined in order to maximize the *overall usefulness* of the examination (see for example in Bachman and Palmer, 1996).

1.4.3 The concept of **examination usefulness** requires that, for any specific assessment situation, an appropriate *balance* must be achieved between the 4 main examination **qualities:** validity, reliability, impact and practicality. In addition, for ALTE members as providers of examinations to users around the world, *quality of service* is an important consideration. It is recognized that members of ALTE have a responsibility to be held accountable for all matters related to use of their examinations; this involves providing a high quality service to the users of their examinations which meets the principles of good practice as outlined in this document.

1.4.4 All examinations are context specific and this means that practical *considerations and constraints* must be taken into account regarding examination development and examination administration so that the appropriate balance between the examination qualities is achieved for any given situation (e.g. educational context, group of examination takers and examination purpose). The relative importance of the qualities must be determined in order to maximize the *overall usefulness* of the examination. Successful examinations cannot be developed, however, without due consideration being given to all qualities.

1.4.5 In considering the context in which an examination is to be developed and used, it is necessary to take into account the *specific* considerations and constraints which characterize that situation. These will not be the same for all ALTE examinations and will determine whether an examination is feasible and can be produced and administered with the *resources* available. With regard to resources, this applies to the resources which are available *internally* to the ALTE institutions and also to the resources which are available in the contexts where the ALTE examinations will be administered. In particular, costs for both development and administration must be controlled and managed.

2.0 Examination Qualities

ALTE members should ensure good practice in relation to the four qualities of their examinations which were noted above, namely, Validity, Reliability, Impact and Practicality.

2.1 Validity

2.1.1 Aspects of Validity

Validity is generally considered to be the most important examination quality; it concerns the appropriateness and meaningfulness of an examination in a specific educational context and the specific inferences made from examination results. The validation process then is the process of accumulating evidence to support such inferences.

Validity is normally taken to be the extent to which a test can be shown to produce scores which are an accurate reflection of the candidate’s true level of language ability. While validity is now accepted as a *unitary concept* (c.f. Messick’s chapter on Validity in Linn 1988) it is convenient to describe the validity of an examination in relation to a number of related concepts for which evidence of validity can be provided:

- Construct-related evidence – the extent to which the test results conform to the model of communicative language ability underlying the test
- Content-related evidence – the extent to which the test covers the full range of knowledge and skills relevant and useful to real world situations and authentic language use.
- Criterion-related evidence (predictive and concurrent validity) – the extent to which test scores correlate with a recognised external criterion which measures the same area of knowledge or ability (e.g. with reference to a system of levels such as the ALTE Framework).

2.1.1 Construct-related evidence

The focus in construct-related validation is primarily on the examination score or grade as a measure of a trait or "construct". The examination developer defines traits of ability for the purpose of measurement and it is these definitions which are the constructs.

A model of *communicative language ability* represents a construct in the context of language testing (cf. Canale and Swain 1980, Canale 1983, Bachman 1990).

The process of compiling construct-related evidence for examination validity starts with examination development and continues when the examination ‘goes live’ and is used under operational conditions.

Validating inferences about a construct requires paying great attention to many aspects of measurement such as examination format, administration conditions, or level of ability which may affect examination meaning and interpretation.

Construct validity is seen by many as the “unifying concept” within test validation that incorporates content and criterion considerations (Messick 1980). As a process, construct validation seeks evidence from a variety of sources in order to provide information on construct interpretation. The choice of which approach to be used in gathering evidence for the interpretation of constructs depends on the particular validation problem and the importance of the role of given constructs within the investigation. In the literature a wide-range of statistical techniques have been used, largely based on correlations and often using experimental designs to collect data.

2.1.2 Content-related evidence

Content-related validation investigates the degree to which the sample of items, tasks, or questions on an examination are representative of a defined *domain of content*. It is concerned with both *relevance* and *coverage*.

A wide range of methods for testing the representativeness of the sample are available; major characteristics of the domain can be specified through *a model* (e.g. the Waystage and Threshold specifications) and *experts* in the field can be asked to assign examination items to the categories defined by these characteristics; in this way the representativeness of the content can be judged.

The specification of the domain of content that an examination is intended to represent is very important for the ALTE exams; the degree to which the format of items or tasks in an examination are representative of the domain is crucial and the involvement of stakeholder groups and relevant experts is a key element in the process of test development.

Often systematic observations of behaviour in the ‘real world’ can be used to identify distinctive features or characteristics of the *criterial situation* (cf Bachman and Palmer’s *Target Language Use – TLU - domain,* 1996 pp 44-45). These observations may be combined with expert judgements to build up a representative sample of the content domain.

A concern for the *authenticity* of test content and tasks and the relationship between the input” and the expected response or “output” is an important feature of content validation. The authenticity of the tasks and materials in the ALTE examinations can be considered a major strength of the approach to assessment they adopt. The examination content must be designed to provide sufficient evidence of the underlying abilities (i.e. construct) through the way the test taker responds to this input. The responses to the test input (tasks, items, etc.) occur as a result of an interaction between the test taker and the test content. The authenticity of test content and the authenticity of the candidate's interaction with that content are important considerations for the examination developer in achieving high validity.

(See Widdowson 1978, 1983 on *situational* and *interactional authenticity* and Bachman and Palmer 1996 for the application of these concepts to language tests).

In summary, content-related validation is linked to examination construction as well as to establishing evidence of validity after the examination has been through the developmental phase and is considered "live".

2.1.3 Criterion-related evidence

The *criterion-related* aspect of validity is of particular importance to ALTE in that ALTE examinations are *criterion-referenced.* That is, the five level ALTE Framework represents a series of criterion levels to which the examinations are linked. This approach has implications for other aspects of validity, (such as test content) and for the estimation and reporting of reliability

Criterion-related validation aims at demonstrating that test scores or examination grades are systematically related to an external criterion or criteria (e.g. another indicator of the ability tested). It is the criterion, therefore, that is of central interest. This criterion may be defined in different ways; for example, by group membership, by performance on another examination of the same ability, or by success in performing a real world task involving the same ability.

The five-level system provides the external criterion and the interpretative frame of reference for all the ALTE exams.

There are two specific kinds of criterion-related evidence which are discussed in the literature - *concurrent* and *predictive.*

- Concurrent validity involves obtaining information on the accuracy with which examination data can be used to estimate or predict criterion behaviour. The most common information is based on correlations between various measures which are made concurrently (e.g. to show the relationship between scores on two different tests of the same ability). In the case of performance tests, qualitative comparisons can be made between the criterion norms and samples of the output from the test (e.g. in writing or speaking).

- Predictive validity serves a similar purpose but obtains predictive information in relation to the future such as future examination results, performance in higher education or performance in a future job. Evidence of this kind of validity is particularly important where the examination or test results are used for screening or placement purposes.

2.1.4 In providing evidence of validity, good practice should involve at least the following:

a) A description of the constructs to be measured and the domain of content covered by the examination.

b) Evidence related to the use of examination results, including a description of how the evidence provided is appropriate for the inferences that are drawn and the actions that will result from examination results.

c) A description of the validation procedures used and their results including as appropriate:

- logical and empirical analyses of processes underlying performance in examinations;
- the relationship between examination results and other variables, including likely sources of variance not related to the construct;
- how the examination questions/items were derived and are related to the domain of knowledge or skill appropriate to the intended inferences to be made;
- logical and empirical evidence supporting discriminant validity sub-scores;
- the number and the qualifications of any experts who made judgements which are pertinent to the validation process;
- procedures used to arrive at judgements, which are pertinent to the validation process;
- the rationale and procedures used in designing the examination specifications (including range of materials surveyed, etc.);
- the rationale and procedures for determining criterion relevance;
- information relative to the interpretation of quantitative evidence.

d) The carrying out of new studies on validity whenever there is a substantial change in the examination, the mode of administration, the characteristics of intended examination takers, or the domain of content to be sampled.

e) The provision of information to examination users to help them interpret validation studies with respect to intended examination results, such as pass/fail decisions, selection or placement.

(*cf. AERA/APA/NCME Standards*, 1999, pp 17-24)

2.2 Reliability

2.2.1 Aspects of reliability

Reliability is a key concept in any form of measurement and contributes to overall validity.

In language assessment, reliability concerns the extent to which test results are *stable, consistent,* and *free from errors of measurement.* In other words it concerns the degree to which examination marks can be depended on for making decisions about the candidate. Estimates of reliability should not only consider relevant sources of error, but the types of decision which are likely to be based on examination marks.

For a wide variety of reasons individuals may score differently on two forms of an examination which are intended to be parallel; when these differences cannot be accounted for, they are called errors of measurement. Measurement errors reduce reliability (and thus the generalizability) of marks obtained for an individual from a single measurement.

Reliability is generally estimated and reported in terms of *reliability coefficients*. Since this is a generic term, the information about error it conveys varies with the specific estimation method used, and since not all sources of error will be relevant to every examination, it is the responsibility of the examination developer to decide on appropriate forms of reporting error variance. This may involve reporting standard errors of measurement, confidence intervals, dependability indices etc.

Within language testing, much of the literature to do with computing the reliability of language tests has been based on work in educational and psychological testing more generally, e.g. the APA *Standards* between 1954 and 1985. In the new volume of *Standards* (1999) the revised chapter on *Reliability and Errors of Measurement* (Part 1, Section 2) still identifies the *three broad categories of reliability* which have traditionally been recognised in the field:

a) alternate-form coefficients (derived from the administration of parallel forms in independent sessions)
b) test-retest coefficients
c) internal consistency coefficients

Of these three, the use of internal consistency coefficients, such as Cronbach's alpha or KR20 to estimate the reliability of objective tests is common (e.g. for multiple-choice reading or listening tests). The fact that these coefficients are relatively easy to calculate mean that other, perhaps more appropriate estimates, are not used as commonly (e.g. test-retest estimates are less often reported because adequate data is difficult to obtain under operational conditions).

For tests of Speaking and Writing the APA *Standards* make it clear that when the scoring of a test involves judgement by examiners or raters, it is important to consider reliability in terms of the *accuracy and consistency* of the ratings which are made. The tests of Speaking and Writing found in many of the ALTE exams fall into this category because the assessments are made by examiners.

The reliability of subjective assessments (using examiners) is usually estimated using correlations, e.g. *intra*- and *inter-rater* correlations.

2.2.2 In providing evidence of reliability, good practice should involve at least the following:

a) Serious efforts to identify and quantify major sources of measurement error, including:

- the degree of reliability expected between pairs of marks in particular contexts (e.g. marks achieved by a candidate on two different tasks which are intended to be of equivalent difficulty);
- the generalizability of results across tasks and items, different forms of the same exam, examiners, different administrations, etc.

b) An assurance that examination marks, including sub-scores and combinations of marks, are sufficiently reliable for their intended use.

c) Provision of information on reliabilities, standard errors of measurement, or other equivalent information so that examination users can also judge whether reported examination marks are sufficiently reliable for their intended use.

d) Provision of information for examination users about sources of variation and other sources of error considered significant for score interpretation.

e) Estimates of the reliability or consistency of reported examination marks by methods that are appropriate to the nature and intended use of the examination marks and that take into account sources of variance considered significant for score interpretation.

f) Documentation of the reliability analysis, including:

- a description of the methods used to assess the reliability or consistency of the examination marks and the rationale for using them, the major sources of variance accounted for in the reliability analysis and the formula used and/or appropriate references;
- a reliability coefficient, an overall error of measurement, an index of classification consistency, or other equivalent information about the consistency of examination marks;
- standard errors of measurement or other measures of mark consistency for mark regions within which decisions about individuals are made on the basis of examination marks;
- the degree of agreement between independent markings when judgemental processes are used;
- correlations among reported sub-scores within the same examination or the marks within an examination battery.

g) Descriptions of the conditions under which the reliability estimates were obtained, including:

- a description of the population involved, e.g. demographic information, education level, employment status;
- a description of the selection procedure for, and the appropriateness of, the analysis sample, including the number of observations, means, and standard deviations for the analysis samples and any group for which reliability is established;

- when marks are based on judgements, the basis for marking, including selecting and training markers and the procedures for allocating papers to examiners and adjusting discrepancies;
- the time intervals between examinations, the rationale for the time interval and the order in which the forms were administered if alternate-form or test-retest methods were used.

(*cf. AERA/APA/NCME Standards*, 1999, pp 31-36)

2.3 Impact

2.3.1 Aspects of Impact

It is recognized that, as providers of examinations, members of ALTE have a major impact on educational processes and on society in general. This impact operates on at least two levels:

a) a *macro level* in terms of general educational processes;
b) a *micro level* in terms of the individuals (stakeholders) who are affected by examination results.

One area of general impact concerns the role of ALTE in promoting the public understanding of assessment and related pedagogical issues within Europe and world-wide. This can be achieved by providing public information, research and advisory services. The aim should be to achieve greater understanding of the purposes and procedures of testing and the proper uses of examination information (results, grades, etc.).

In terms of impact on individuals, it is necessary to establish that the examination is fair and not biased.

Positive impact on *teaching and learning is* an important aspect of impact which operates on both levels (macro and micro). It is in this context that the notions of "face validity" (or test appeal) and washback are considered. It is important to be able to investigate the educational impact that examinations have within the contexts in which they are used. As a point of principle, examination developers must operate with the aim that their examinations will not have a negative impact and, as far as possible, strive to achieve positive impact.

2.3.2 In providing evidence of impact, good practice should involve at least the following:

a) the development and presentation of examination specifications and detailed syllabus designs;

b) provision of professional support programmes for institutions and individual teachers/students who use the examinations.

Positive educational impact can also be achieved through the following practices:

- the identification of suitable experts within any given field to work on all aspects of examination development;

- the training and employment of suitable experts to act as question/item writers in examination production;
- the training and employment of suitable experts to act as examiners.

Procedures also need to be put into place when an examination becomes operational in order to collect information which allows impact to be estimated.

This should involve collecting data on the following:

- who is taking the examination (i.e. a profile of the candidates);
- who is using the examination results and for what purpose; who is teaching towards the examination and under what circumstances;
- what kinds of courses and materials are being designed and used to prepare candidates;
- what effect the examination has on public perceptions generally (e.g. regarding educational standards);
- how the examination is viewed by those directly involved in educational processes (e.g. by students, examination takers, teachers, parents, etc.);
- how the examination is viewed by members of society outside education (e.g. by politicians, businesspeople, etc.).

This information should be made available within the ALTE organizations, for example in the form of written reports, and suitable versions of such reports should be made available to the other stakeholders.

From the evidence collected, it should be possible to demonstrate that the examination is sufficiently valid and reliable for the context in which it is used. This in itself is a way of ensuring that positive impact is achieved.

2.4 Practicality

2.4.1 Aspects of Practicality

Practicality is an integral part of the concept of test usefulness and affects many different aspects of an examination.

It can be defined as the extent to which an examination is practicable in terms of the resources necessary to produce and administer it in its intended context of use.

The *practicality* of any examination, involves 2 factors:

a) the resources that are required to produce an operational examination that has the appropriate balance of qualities (validity, reliability and impact) for the context in which the examination will be used;
b) the resources that are available.

A practical examination is one that does not place an unreasonable demand on available resources. If available resources are exceeded, then the examination is not practical. In this case, the examination developer must either modify the design of the examination or make a case for an increase or reallocation of resources. If practical constraints make the second option impossible, the first option must be chosen.

Before the examination development can proceed, it must be established whether or not the examination will still be *useful* if the changes to the specifications are implemented. If this cannot be established the examination development should not proceed.

2.4.2 Good Practice in ensuring practicality should include the following:

Good practice can only be achieved if appropriate procedures are implemented for *managing all aspects of the examination process,* including the development, administration and validation of the examinations.

The development of practical examinations requires that an *explicit model of test development* be adopted – see for example the *User Guide for Examiners.*

Whenever a new or revised examination is to be developed, there should be procedures in place to address the management structure for the development project with a clear and integrated assignment of roles and responsibilities. There needs to be a means of monitoring progress in terms of development schedules and resources, and a methodology for managing the examination production process when the examination becomes operational (item writing, vetting, moderation, pre-testing, item banking, question paper construction)

The process of development should begin with a *feasibility study* dealing with at least the following:

- The purpose of the new examination
- The level of difficulty for the intended examination takers (e.g. in relation to the ALTE Framework)
- External factors
  - the market place
  - the competition - provided by existing exams of a similar kind
  - societal demands or requirements (e.g. from parents, Ministries of Education, etc)
- Intrinsic factors
  - theories related to the examination constructs and content
  - advances in technology
  - fixed institutional parameters
- The predicted relevance and acceptability of the new examination to intended takers and users;
- The cost of the new examination to the taker;
- The resources available for
  - development,
  - administration,
  - reporting of results,
  - replication (future administrations of the examination).

The determination of examination usefulness is both cyclical and iterative; considerations of *practicality* affect decisions at all phases of the examination development process. When operational considerations are taken into account, it is necessary to consider to what extent it is possible to achieve this balance with the resources that are available (e.g. in terms of people, equipment, time and money).

2.5 Quality of Service (to be updated following further discussion at meetings)

2.5.1 Aspects of Quality of Service

Quality of service concerns the examination developer's ability to meet specific commitments to the examination takers and users.

The includes the provision of secure examination materials, the confidentiality of examination data and results, and procedures to handle enquires about results and appeals procedures.

2.5.2 Good Practice in achieving high quality of service should include the following:

- Making available the results of research, and seeking peer review of such activities.
- Asking for information about individuals and institutions only when it is potentially useful to them by way of furthering the research on products, and thereby improving them. The purpose of gathering such information should be made clear to everyone concerned.
- Protecting the confidentiality of all data (raw or processed) held by ALTE members on any institutions or individuals, and encouraging any group or institution to or from which data is transferred to adopt the same policy.
- Making realistic delivery commitments and subsequently making every effort to meet these commitments.
- Using adequate quality controls to ensure that any products and services offered by ALTE members are on delivery accurate and within the time spans promised.
- Accepting the responsibility for informing those negatively affected if, subsequent to its release, information is found to be inaccurate.
- Informing those negatively affected if there is likely to be a substantial departure from scheduled commitments.
- Reviewing and revising examination questions and related materials in order to avoid potentially insensitive content and language.
- When feasible, making appropriately modified forms of examinations or administrative procedures available for handicapped candidates.
- Providing candidates with the information they need in order to be familiar with the coverage of the examination, the types of question formats, rubrics and appropriate test-taking strategies. Striving to make such information equally available to all candidates.
- Telling candidates how long their results will be kept on file and indicating to whom and under what circumstances examination results will or will not be released.
- Describing the procedures that the candidates may use to register a complaint or an appeal and have their problems resolved.

Sources and references

AERA/APA/NCME (1985/1999). *Standards for educational and psychological testing.* Washington: AERA.

ALTE (1994) Code of Practice

ALTE/Council of Europe (1997) Users Guide for Examiners

ALTE (1998) Handbook of Language Examinations and Examination Systems

Bachman L.F. (1990). *Fundamental considerations in language testing*. Oxford: OUP.

Bachman, L. F. and Palmer, A. (1996). *Language Testing in Practice*. Oxford: OUP.

Canale, M. (1983) On some dimensions of language proficiency, in Oller 1993

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

Council of Europe

Common European Framework of Reference (1997/2001)

Threshold Level (1990)

Waystage Level (1990)

ETS (1981, 1987, 2000) *Standards for Quality and Fairness*

ILTA (1995) Task Force on Testing Standards

ILTA (2000) Code of Ethics

Linn, R.L. ed. (1989) *Educational Measurement.* Third edition. New York: ACE/Macmillan

Messick, S.A. (1980) Test validity and the ethics of assessment. *American Psychologist,* 35.

Messick, S.A. (1989) Validity, in Linn 1989

Saville, N. and Milanovic, M. (1991, 1993), *Principles of Good Practice for ALTE Examinations,* Unpublished documents

Taylor, L. (1999) Constituency matters. Paper presented at Language Testing Forum, Edinburgh, November 1999

Washington D.C. Joint Committee on Testing Practices (1988) *The Code of Fair Testing Practices in Education*

Widdowson, H.G. (1978) *Teaching Language as Communication*. Oxford: OUP

Widdowson, H.G. (1983) *Language Purpose and Language Use.* Oxford: OUP