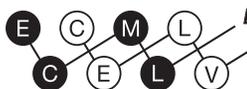# Guidelines for task-based university language testing

Johann Fischer, Catherine Chouissa, Stefania Dugovičová and Anu Virkkunen-Fullenwider

PROMOTING EXCELLENCE IN LANGUAGE EDUCATION

The European Centre for Modern Languages (ECML) is an "Enlarged Partial Agreement" of the Council of Europe to which thirty-four countries subscribe[1]. The institution focuses on bridging the gap between language policy theory and classroom learning practice. In a period of unparalleled change and mobility, the Centre offers concrete approaches to issues and challenges facing Europe's culturally diverse societies.

The Centre seeks to make a positive difference to the language education profession by:

- promoting innovative approaches;
- advancing the quality of teaching and learning languages;
- supporting the implementation of language education policies;
- fostering dialogue between language education practitioners and decision-makers.

ECML activities are complementary to those of the Council of Europe's Language Policy Division, responsible for the development of policies and planning tools in the field of language education and the Secretariat of the European Charter for Regional or Minority Languages.

***

The present series of publications results from the ECML 2008-2011 programme, entitled *Empowering language professionals: Competences – Networks – Impact – Quality*. The programme has taken place against a backdrop of major international political developments in the sphere of education, where increasing demands are placed on the professional skills of teachers. The profession is expected to contribute to national education reform processes and face a wide range of challenges relating, among others, to standard-linked tuition, result-oriented assessment, greater autonomy of educational institutions and increasing ethnic and cultural heterogeneity among students.

The publications illustrate the dedication and active involvement of all those who participated in a series of 24 international projects, particularly the teams who coordinated the projects.

All ECML publications and accompanying materials are available for download: http://www.ecml.at/publications.

[1] The 34 member states of the Enlarged Partial Agreement of the ECML are: Albania, Andorra, Armenia, Austria, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Montenegro, Netherlands, Norway, Poland, Romania, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, "The former Yugoslav Republic of Macedonia", United Kingdom (status 30 June 2011).

Guidelines for task-based university language testing

Johann Fischer, Catherine Chouissa, Stefania Dugovičová and Anu Virkkunen-Fullenwider

with special contribution by Amelia Kreitzer Hope

Table of Contents

Acknowledgments

This publication is the product of an extensive collaborative effort. It combines the work of the project team members with the thoughts and contributions of our workshop and network meeting participants, of language teachers and testers in our project team members' home institutions, and of several colleagues and partners in various networks and professional groups.

First of all, we would like to thank the Scientific Committee of UNIcert, who contributed to the development of a GULT testing structure for proficiency tests.

We are grateful to the European Centre for Modern Languages of the Council of Europe (ECML) for granting us the opportunity to work on the GULT project, and to the ECML staff in Graz for their expert advice and for providing a supportive environment during our meetings.

We value the constructive criticism and feedback from all our workshop participants.

Special thanks go to all the participants in the network meeting in January 2011 for their co-operative spirit, critical thoughts and valuable contributions: Mathilde Anquetil, Andromaqi Haloçi, Ann Elisabeth Laksfoss Hansen, Amelia Kreitzer Hope, Diana Ioniță, Marina Mattheoudakis, Marie-Thérèse Maurer, Cristina Pérez Guillot, Helena Šajgalíková, Joachim Schlabach, Carin Söderberg, Evan Stewart, Andrej Stopar and Rias van den Doel.

In particular, we would also like to thank the four team leaders at our GULT network meeting for their useful comments. We are extremely thankful to Evan Stewart for his insights into the structure of the manual and its language aspects. We are greatly indebted to Amelia Kreitzer Hope for her analytical criticism of testing literature and her theoretical input; in fact, Amelia Hope has played a major role in the development of Chapters 4, 7 and 8 of the GULT guidelines, for which we are extremely grateful.

Last but not least, we want to thank our project consultant Isabel Landsiedler for her continuous support and positive feedback throughout this project.

GULT
Guidelines for task-based university language testing

This publication is the result of a project of the European Centre for Modern Languages entitled "Guidelines for University Language Testing" (GULT).

For further information and materials relating to this publication, visit the website http://gult.ecml.at.

Part A

Guidelines

1.   Aims and objectives

With the integration of project work, case studies, global simulations and webquest activities, a task-based approach to language teaching has become more and more widespread. Too often, however, end-of-course and proficiency exams still measure student achievement with discrete-item tests which do not correspond to what has been done in class. The aim of the GULT Project (Guidelines for Task-Based University Language Testing[1]) has therefore been to develop a structure for a task-based approach to testing languages. The idea is to bring language teaching and testing closer together and to stimulate a task-based approach at both stages. The reason for this is to make language tests more relevant for test takers and to test to what extent they are able to communicate appropriately in a specific communication situation in their field of studies or future workplace.

These guidelines focus on testing languages for specific purposes (LSP) at university at levels B2 and C1 of the *Common European Framework of Reference for Languages (CEFR)[2]* because in most countries LSP teaching takes place at these levels. Nonetheless, the model described below can easily be applied at level C2, and modified for administration at level B1. The GULT Project focuses on LSP teaching and testing, but the practices could be used in testing languages for academic purposes (LAP) and applied to general language testing.

The focus of these guidelines is task-based assessment. Therefore, general aspects of language testing are not covered in the main text, but an introduction to language testing in general is offered in Part B of the publication in Resource document 1.

To complement the guidelines the project website (http://gult.ecml.at) offers model tests at levels B2 and C1 for the languages Spanish, French, German and English.

---

1   Originally named Guidelines for University Language Testing, then changed to Guidelines for Task-Based University Language Testing in order to make it clearer.

2   Council of Europe, 2001.

2. The GULT context

A task-based approach to language teaching has already been implemented in many teaching curricula at university language centres in Europe, but unfortunately not yet in testing. The GULT team thus brought together the competences and results of previous projects, for example:

- the activities of the HERMES group[3] on professional development for teachers at universities and colleges of agriculture and related sciences within the AFANet Project;[4]

- the outcomes of the WebCase Project[5] on the use of case studies in agribusiness teaching programmes at university;

- the LCaS Project[6] aiming at the development of teacher training modules for the use and development of case studies in university language teaching;

- the EXPLICS Project[7] aiming at the development of online case studies and global simulations for eleven languages;

- the competences of the German UNIcert®[8] team with its competence in language testing and in LSP;

- the UNIcert®LUCE[9] team in Slovakia and the Czech Republic;

- the French CLES[10] team with its competence in task-based university language testing; and

- the Finnish network of university language centres with its competence in university LSP testing.

It has thus brought together a number of specialists in task-based university language teaching, in teaching languages for specific purposes and in language testing with the aim of creating a new model for university language testing, based on the principles of task-based teaching and learning and adapted to the specific needs of learners of languages for specific purposes.

---

3 See: HERMES website at www.uni-tuebingen.de/hermes/, and Casey and Fischer, 2004, 2005; Fischer, 2003, 2004; Fischer and Casey, 2005.

4 See: AFANet website at www.afanet.info.

5 See: WebCase website at www.webcase-online.info and Cain and Heath, 2004.

6 See: LCaS website at http://www.ecml.at/mtp2/LCaS, and Fischer et al., 2008.

7 See: EXPLICS website at www.zess.uni-goettingen.de/explics and Fischer et al., 2009.

8 See: UNIcert® website at www.unicert-online.org and Eggensperger and Fischer, 1998; Voss, 2010.

9 See: UNIcert®LUCE website at http://www.casajc.sk/unicert.htm.

10 See: CLES website at http://www.certification-cles.fr/ and Petermann, 2008.

3.   Target audience

These guidelines have been created:

- to provide decision makers in university language teaching (e.g. heads of university language centres or language departments) with a framework of task-based language assessment and the necessary tools to implement the GULT approach in their institutions;

- to help LSP teachers in higher education to develop and administer task-based language tests in order to assess their students' language skills;

- to promote a task-based approach in language teaching and testing at university level by informing stakeholders (e.g. professional associations; regional, national and international networks; national and regional boards and ministries of education; national and international university test providers) about task-based teaching, learning and assessment.

4. Why task-based LSP testing: the logical next step from task-based LSP teaching

Since implementation of task-based models is increasing in university language teaching, the development of task-based language testing is a logical next step.

4.1. Task-based testing in various disciplines

The GULT team does not claim that task-based assessment (TBA) offers better tests, but tests that are closer to real life, and thus prepare students for the future. This link to real-life activities is certainly one of the reasons why several university disciplines, such as medicine, law, social sciences or natural sciences, have implemented a task-based approach both in teaching and assessment.

The original focus of the GULT team was not on general language testing, where some models of task-based exams have already been applied, but on the testing of language skills for specific purposes, as LSP testing is one of the key activities of university language centres in Europe. The GULT models can, of course, be adapted to and used for general language testing.

By using a task-based approach to an LSP test in a specific realistic situation based on facts and authentic documents,[11] the test takers can show to what extent they are able to communicate in a foreign language, which makes the results more useful and relevant for them. This type of test is also more transparent for the test takers, as they themselves can judge more easily how well they managed, linguistically, a realistic communication situation. The overall aim of the GULT Project is therefore to develop guidelines and models for transparent task-based LSP testing at university.

4.2. Task-based / action-based approach to language testing

Definition of "task"[12]

In "task-based" or "action-based" testing, sometimes also called "performance(-based) testing",[13] the examiner gives a task for completion to the test taker who is stimulated by this task to act. "Task" is understood to mean the following:

---

11 Namely, a situation that corresponds to an authentic situation in real life, and documents and facts that have not been created for pedagogical reasons.

12 See also: Resource document 6 "Definition of 'task'".

13 See, for example: Bachman, 2007; Wigglesworth, 2008.

A task is an activity which requires learners to use language, with emphasis on meaning, to attain an objective.

(Skehan, 2003: 3)

Tasks […] are activities which have meaning as their primary focus. Success in tasks is evaluated in terms of achievement of an outcome, and tasks generally bear some resemblance to real-life language use. So task-based instruction takes a fairly strong view of communicative language teaching.

(Skehan, 1996: 20)

In a task-based approach, an overall task describes the general activity that needs to be carried out, with a specific result expected at the end. In order to achieve this outcome, the test takers have to develop specific products, e.g. a written report explaining the problem in question and possible solutions to it, and an oral presentation of these solutions to a specified audience in a specified situation. The various activities and products for the overall task are interrelated. To achieve these results the test takers often have to carry out specific intermediary or build-up tasks.

Terminology

In the literature two terms can be found: "task-based" and "action-based", in French "*approche basée sur les tâches*" and "*approche actionnelle*", and in German "*aufgabenbasiert*" and "*handlungsorientiert*". These terms are strongly interconnected, looking, in one case, more at the teacher, who gives a task to the candidate and, in the other case, concentrating more on the test taker, who has to act. In the English-speaking context the term "task-based" is probably used more frequently, whereas in French it is the term "*approche actionnelle*". In German the expression "*Aufgabenorientierung*" is rather ambiguous, as "*Aufgabe*" is used for everything from a single-item task to a homework activity and a more complex "task". In German it is therefore preferable to use the term "*Handlungsorientierung*" although the word "*Aufgabenorientierung*" has recently been used more frequently.[14]

In fact, the task given to the test takers by the examiner prompts the test takers to act. The relationship between the two terms "task-based" and "action-based" may therefore be illustrated as follows:

---

14 See, for example: Bausch et al., 2006.

task / *tâche* / *Aufgabe* →	**action / *action* / *Handlung***
task-based *entrée par les tâches* *aufgabenorientiert* →	**action-based** *perspective actionnelle /* *approche actionnelle* *handlungsorientiert*
↑ **Input** →	↑ **Output**
teacher's view *perspective de l'enseignant* *Perspektive der Lehrkraft*	**learner's view** *perspective de l'apprenant* *Perspektive des Lerners*

The GULT concept of task-based testing

By "task-based testing" the GULT team understands the testing of language competence based on realistic tasks: the test takers are put into a situation they might encounter in real life and have to work on a major task, normally a problem or a project. With the help of build-up tasks the various language skills can be tested (individually or holistically).

In this approach, the task is used as the input, stimulating the test taker to interact with his or her partner and / or the examiner, while the output (the test taker's production) is assessed at various levels taking into account the linguistic competence, but also the pragmatic competence and content. For the assessment of the productive skills transparent, well-defined criteria and grids are needed.[15]

It is important to remember that, in the GULT case, language teaching, learning and testing take place in a university context, where normally all language skills are needed and where content, because of the use of languages for specific purposes, is of particular importance. While language tests that are not linked to a teaching and learning programme have to limit aspects of content in order to guarantee equal opportunities to every candidate, content is essential in this context.

---

15. See: Chapter 8 and Resource document 5.

The GULT approach tries to combine the experience of different approaches to language testing by setting a specific context and basing the test on an overall task, asking the test takers to interact in a meaningful way about a specific aspect in which content plays a major role, and by assessing the various skills.

Further research is needed to see whether the GULT approach has succeeded in developing a structure for more valid university language tests, and whether this approach has managed to bring together the benefits of different approaches to language testing, as suggested by Bachman (2007: 70f):

> Issues related to language ability and language use contexts and the interaction between these have been addressed, in a dialectic, in language assessment research, and have led to three general approaches to defining the construct, or what we want to assess: (1) ability-focused, (2) task-focused, and (3) interaction-focused. While the different theoretical perspectives that underlie these approaches are not mutually exclusive, they are based on different sets of values and assumptions. […] Because of these differences, the conundrum of ability and context and how they interact in language use and language assessment is, in my view, essentially a straw issue, theoretically, and may not be resolvable at that level.

> Nevertheless, the theoretical issues raised by these different approaches have important implications and present challenging questions for both empirical research in language testing and for practical test design, development, and use. These theoretical issues also provide valuable insights into how we can enrich the ways in which we conceptualize what we assess and how we go about assessing it. For research, they imply the need for a much broader, more catholic methodological approach, involving both so-called quantitative and qualitative perspectives and methodologies. For practice, they imply that *exclusive* focus on any one of these approaches (ability, task, interaction), to the exclusion of the others, will lead to potential weaknesses in the assessment itself, or to limitations on the uses for which the assessment is appropriate. This means that we need to address all three in the design, development, and use of language assessments.

4.3.   Task-based testing and the communicative approach

Task-based language learning and testing can best be seen as a more focused development of the communicative approach to teaching languages. The hallmarks of the communicative approach include a focus on communicating meaning, using authentic materials, integrating skills, and centring instruction on the students and their communicative needs. What the task-based approach adds is an emphasis on embedding holistic communicative acts into a specific context and situation, with a specific aim that mirrors the actual or future communicative aims of the learner.

The relationship between the communicative approach and task-based approach can be illustrated in the following way:
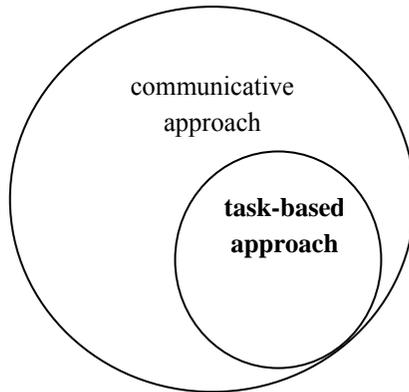
This relationship depends, however, on the definition of "communicative approach", which is seen differently, for example by French researchers. Bourguignon (2006: 58), for example, speaks of a "shift from the 'communicative' era to the 'communic-active' one by referring to an approach which encompasses the communicative one and leads to the 'communic-actional approach'". Bourguignon continues:

> On se rend compte qu'avec la perspective actionnelle, nous changeons de paradigme. Nous passons du paradigme de la connaissance, de la simplification à celui de la compétence, de la complexité qui ne veut pas disjoindre l'objet et le sujet, la réflexion et l'action, l'apprenant et l'usager mais les conjoindre pour que la finalité de l'apprentissage d'une langue ne soit pas la bonne note mais l'utilisation autonome de la langue dans des situations plus ou moins complexes qui peuvent aller de la lecture d'un prospectus à celle d'une œuvre de Shakespeare ! […]

> Au niveau de la perspective actionnelle, la communication est au service de l'action. Il s'agit de passer de l'interaction (= parler avec les autres) à ce que C. Puren appelle la « co-action », et que nous avons choisi d'appeler la « communic-action », c'est-à-dire agir avec les autres, communiquer pour agir. […]

> Avant l'approche communicative, l'enseignement / apprentissage de la langue était totalement coupé de son utilisation sociale. Le savoir de référence était le « savoir savant », la linguistique et la langue s'apprenaient à partir d'exercices de grammaire et de vocabulaire. Avec l'approche communicative, on fait « rentrer » le milieu social dans le milieu scolaire. Il s'agit de simuler des situations de communications « empruntées » au milieu social. Se mettent en place des « tâches communicatives » qui visent l'efficacité dans la transmission de l'information. Pour autant, ces tâches communicatives restent des tâches d'apprentissage qui seront évaluées en tant que telles. Il y a donc bien disjonction entre tâche d'apprentissage et tâche sociale.

> Avec la perspective actionnelle, tâche d'apprentissage et tâche sociale sont liées. Il ne s'agit plus en classe de simuler des situations d'usage en donnant à l'apprenant le rôle de l'usager. L'apprenant sera acteur d'un projet qui ne visera pas seulement la réutilisation ou l'application de connaissances mais aussi sa construction en fonction des objectifs à atteindre et des aléas de la situation.

> (Bourguignon 2006: 63-65)

English translation:

> With the task-based approach we obviously change paradigm. We move from the paradigm of knowledge, of simplification towards the paradigm of competence, of complexity, which does not try to separate object and subject, reflection and action, learner and user, but brings them together so that the final aim of learning a language is not a good grade but the autonomous usage of the language in more or less complex situations, which might go from reading a prospectus to reading a work by Shakespeare ! […]

> In the task-based approach, communication is at the service of the action. The idea is to move from interaction (= speaking with others) to what C. Puren calls "co-action", and what we have called "communic-action", that is acting with others, communicating in order to act. […]

> Before the communicative approach, teaching and learning a language were totally independent of using it in a social context. The point of reference was academic knowledge", and learning linguistics and language were based on grammar and vocabulary exercises. With the communicative approach the social dimension enters the schooling dimension. The idea is to simulate situations of communication taken from the social dimension. Communicative tasks are therefore developed, which aim at the efficiency of transmitting information. These communicative tasks remain, however, learning tasks, which will be assessed as such. There is therefore a gap between the learning task and the social task.

> With the task-based perspective, learning task and social task are linked. The idea is no longer to simulate situations of language use in class by allocating the learner the role of a user. The learner will become the participant in a project which will not only aim at re-using or applying knowledge, but will also aid its development based on the objectives provided and the circumstances of the given situation.

This means that the link between the communicative and the task-based approach can be seen as illustrated here[16]:
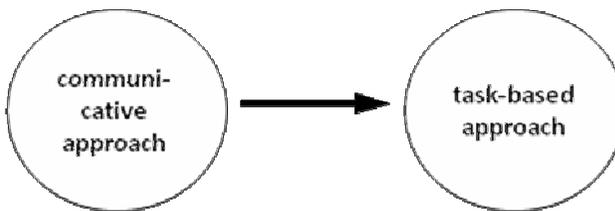
To conclude, according to the GULT team the link between various language teaching and testing approaches could be illustrated as follows, although this illustration is highly simplified:

---

16   See also Puren 2006a, where he explains his distinction between communicative and action-based approach.
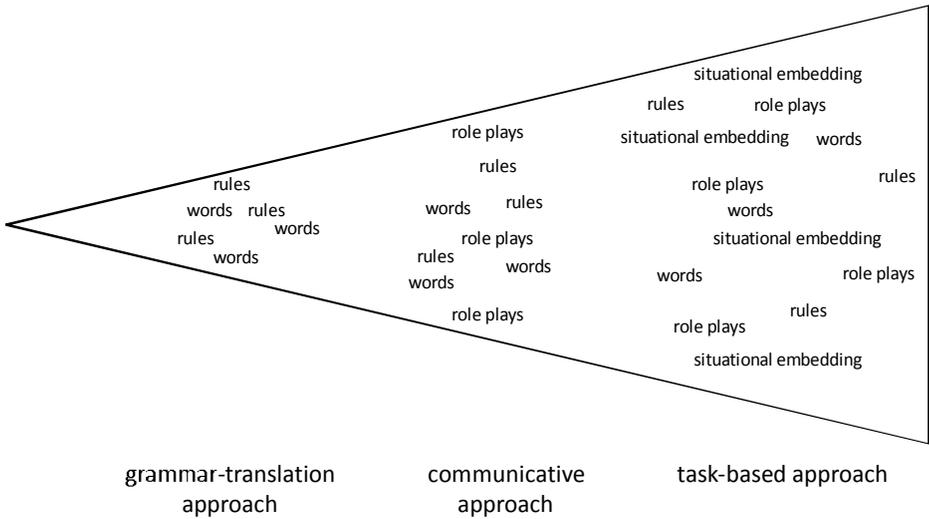
The activities used in task-based teaching and testing should match as closely as possible the communicative tasks the learners face or will face in the world beyond the language classroom. As an example of the difference between communicative tests and task-based tests, a communicative speaking test might involve comparing two pictures or talking about one's opinions on vegetarianism. A task-based test would only use such formats in the unlikely situation where detailed visual descriptions or defending one's views on vegetarianism were communicative contexts test takers would encounter in their professional or academic life outside of the classroom.

To further clarify the difference between communicative tests and task-based tests, consider the ubiquitous role play, the mainstay of many a communicative classroom. In a communicative test, test takers might be asked to assume the roles of doctor and patient and to improvise a conversation about an illness. In a task-based test, however, test takers would never be required to pretend to be doctors unless they were medical students who would in the future need to communicate in their second language with patients. Business students would not be asked to act as managing director, but as students doing a placement, as being a managing director requires the skills gathered over several years of activity in the business sector.

It must be noted that a language test will never be really authentic unless the test takers are tested in real life in a target-language context (direct testing). Nonetheless, a task-based test should attempt to simulate as closely as possible an authentic situation acknowledging that the scenario will be somewhere on a continuum between fictitious and authentic. The more authentic the setting, however, the better:

Test takers are not only judged on whether the phrase they used is linguistically correct, but also on whether they managed to complete the task, for example whether they found the place they wanted to go to (level A), or whether the audience understood the content of the presentation and whether the content had an impact on the group's activities (level C).

In the following illustration, based on Nunan's works (Nunan 1989), Puren (2008: 12) puts the task at the centre of the learning context, as it is linked to objectives, input and activities on the one hand, and the role of the teacher, the role of the learner and the dynamics or situation on the other:
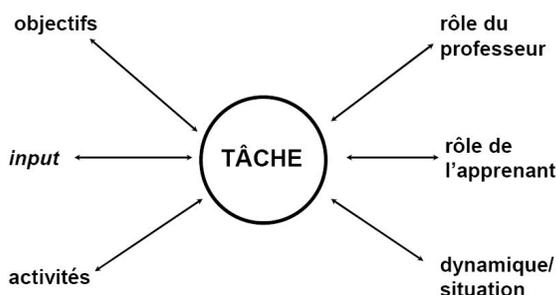
The task is given and carried out with specific objectives, for which specific input is needed and where specific activities have to be completed. Obviously, both the role of the teacher and the role of the learner have to be revisited, as they are very different from more traditional classroom situations.

The following quote by Puren (2008: 4) makes clear the emphasis that the task-based approach places on the holistic nature of tasks:

> La mise en œuvre de la perspective actionnelle suppose logiquement un passage à une « entrée par l'action », l'unité affichée de l'unité didactique (son titre) étant désormais celle d'une action unique..

English translation:

> The implementation of an action-based perspective suggests a move towards an "action-based approach", where the visible unity of the didactic unit (its title) is now the unity based on one single action.

Puren continues (2008: 12):

> […] l' « entrée par l'action », c'est-à-dire un modèle d'unité didactique où toutes les activités dans tous les domaines (CO, CE, PO, PE, lexique, culture, grammaire et graphie-phonie) sont conçues en fonction d'une action unique à partir de laquelle et à propos de laquelle est construite l'unité de l'unité didactique.

English translation:

> […] the action-based approach, that is to say the model of a didactic unit where the activities in all areas (listening, reading, speaking, writing, vocabulary, culture, grammar, spelling and phonetics) are conceived as a single action, on which and towards which the unity of the didactic unit is built.

This holistic view of a task means that all build-up tasks are inextricably linked to the overall task and overall aim. Single tasks in a task-based teaching unit or language exam are not independent of the other activities; their successful completion requires a focus on the overall picture of the project or problem in question. In this way, the skills are more strongly interrelated in a task-based approach, where it is necessary to understand the written and spoken text in order to act in the productive phase of the exam, than in typical discrete-item tests.

4.4.   Task-based approaches

A task-based approach in language teaching may use case studies,[17] project work[18] and global simulations.[19] Whereas global simulations and project work are often more open-ended and creative, and put the focus on the productive skills, case studies are problem-based and start with an important receptive phase that is essential for the outcome of the activity; without a detailed analysis of the problem, the learner or test taker will not be able to develop a solution.[20]

Case studies can be defined as follows:

> Darstellung einer konkreten Situation aus der betrieblichen Praxis oder dem Alltagsleben, die anhand bestimmter Tatsachen, Ansichten und Meinungen dargestellt wird, auf deren Grundlagen eine Entscheidung getroffen werden muss.
>
> (Kaiser 1983: 20)

English translation:

> Presentation of a concrete situation taken from professional or everyday life, which is displayed through specific facts, attitudes and opinions, on the basis of which a decision has to be taken.

---

17   See also: Almagro Esteban and Pérez Cañado, 2004; Daly, 2002; Kaiser, 1983; Kiefer, 2004; Uber Grosse, 1988.

18   See also: Ribé and Vidal, 1993; Schart, 2003.

19   See also: Caré and Debyser, 1995; Debyser, 1996; Yaiche, 1996.

20   See also: Fischer et al., 2008, 2009.

The LCaS team has defined case studies as follows:

> [A]nalysis of a problem / dilemma in a given situation to which no single solution exists.
>
> (LCaS Project team)

"Global simulations" are described by Debyser (1996: IV) as follows:

> Une simulation globale est un protocole ou un scénario cadre qui permet à un groupe d'apprenants pouvant aller jusqu'à une classe entière d'une trentaine d'élèves, de créer un univers de référence – un immeuble, un village, une île, un cirque, un hôtel – de l'animer de personnages en interaction et d'y simuler toutes les fonctions du langage que ce cadre, qui est à la fois un lieu-thème et un univers du discours, est susceptible de requérir.

English translation:

> A global simulation is a script or a scenario framework which allows a group of learners, up to a whole class of about 30 students, to create a universe of reference – an apartment block, a village, an island, a circus, a hotel – to animate it with characters interacting with each other and to simulate all language functions which are probably needed within this framework, which is at the same time a thematic place and a universe of communication.

In language testing, case studies and project work are suitable for proficiency tests and for end-of-course tests. If a course is based on a global simulation, then a specific project activity, integrated into this simulation, can also act as an end-of-course exam. Some subject areas, such as business, economics, law, medicine and social sciences may lend themselves more to case studies, while project work and simulations may be more suitable in areas such as science.

5. Essential features of task-based assessment

5.1. The construct

The general construct of the GULT exam follows an interactionalist perspective,[21] in other words the exam focuses on the interaction in a communication context. The exam does, in general, not follow an integrative approach where all four language skills are tested in an integrated way,[22] but the various tasks assessing the four skills (which may in some cases be combined at local level) are all embedded in an overall context and are built around an overall task. The tasks, which should be as close to real-life activities as possible, serve here as the input, but the language performance of the test taker is assessed according to content-related, pragmatic, (socio-)linguistic and paralinguistic criteria, as specified by the assessment grids available in Resource document 5. Thus, the GULT exam aims at assessing the linguistic ability of the test taker in a specific context.

Whether this is a strong or a weak construct[23] depends on the individual test and the specifications developed by the institution in question. In general, the construct is neither strong nor weak as the emphasis is on both the completion of the task and the language used.

To summarise, the GULT construct can be defined as follows:

- The GULT test is a task-based test.

- The test aims at showing how well learners can express themselves in a real-life situation in their studies or their future workplace, using (where appropriate) LSP language corresponding to the subject area in question, although the results will not necessarily predict the test taker's language performance in a future situation of communication in the workplace.

- The GULT test takes into account and assesses aspects of content, pragmatic competence, linguistic competence and paralinguistic competence. The focus will vary depending on the skill(s) tested, the level and the institutional framework of the test, and the evaluation grids used.

- The GULT test is a criterion-referenced test – and not a norm-referenced test. The criteria are decided upon by the test developers or institution for every test and should be in line with the CEFR criteria / categories.

---

21 See: Bachman, 2007.

22 See: section 5.2.

23 See: Wigglesworth, 2008: 113; see also the distinction by Bachman (2002) between "construct-based" and "task-based" approach.

As the GULT approach offers a testing framework and not concrete test specifications for a single test or exam, this framework then needs to be specified by the individual institution for each test, that is, each institution or testing system needs to define its own construct.

5.2. Integrating the four skills

Usually the four (or five[24]) skills are tested individually and separately. These tests cover up to six different areas: reading, listening, spoken production, spoken interaction, writing and relevant language use (that is, syntax and LSP terminology). Each testlet normally tests one single skill. In some cases it combines two (but rarely more), but in most cases the testlets testing the various language skills are not interconnected in the language test and cover different topics. An exception might be the testing of syntax and LSP terminology, which are often integrated into the testing of productive skills.

By testing each skill separately, the examiner gets a good picture of the strengths and weaknesses of a learner in each skill. Furthermore, testing each skill separately and having a different or new topic for each part of the test will not penalise test takers who are less familiar with one or the other topic. In real life these skills rarely appear separately and are normally interconnected. As with the communicative approach, it is possible to analyse whether test takers are able to understand a written text or to give a presentation and discuss it with the group. It is not possible, however, to say whether they can communicate the key aspects and the underlying ideas of a series of written and spoken texts to a group, extract aspects that are important to them, and use them to develop their own report / presentation or solution to a problem. The individual details can be seen, but not the full picture that a task-based approach is aiming at, in other words one cannot see the wood for the trees.

In a task-based approach, all skills and other language aspects (grammar, vocabulary, LSP, style, register, etc.) are merged through an integrative and holistic approach. For reasons of transparency, of face validity and comparability, but mainly for analytic reasons, the four language skills are tested individually by giving the learners tasks which focus each time on one language skill. These tasks are, however, all integrated into one overall task that looks at a specific situation, as authentic as possible, in a given context.

---

24 In the following, the traditional distinction between four different language skills will be used, as the CEFR makes a distinction between spoken production and spoken interaction for speaking, but not writing. It is also obvious that there are highly important differences between writing a single, isolated text or a text that is part of a longer interaction between two or more partners of communication requiring reading skills (e.g. a chat, a discussion board, a forum, or even an email or letter correspondence).

While the productive tasks in TBA easily correspond to real-life situations, this is not necessarily the case for the testlets focusing on the receptive skills. Here further developments are still needed to make it more authentic in a university teaching and learning environment.[25]

The issue of testing receptive skills

In order to make a test or exam as authentic as possible, ideally only the productive skills are assessed. As the production of texts, both spoken and written, is obviously always based on the analysis of written and aural texts, namely on reception, receptive skills are needed, too. If the candidates do not understand the texts, they cannot present a proper solution to the problem. In fact, in real-life situations in the workplace there are no listening and reading exams, apart from occasional summaries of the problem in question. Therefore, in the most authentic testing situation, a fully holistic approach would be followed. Receptive skills would not be tested separately, but integrated into the parts focussing on the productive skills.

However, a teaching and testing situation is different: if the main purpose is to evaluate the learners' strengths and weaknesses, then their receptive skills should be assessed as well. In particular, competent users of a foreign language have a tendency to underperform in receptive skills. These learners may simply browse through texts (and underperform in the exam part on the receptive skills) and concentrate on the production of their ideas (and perform well in the productive skills part). In fact, they might still have a need to improve their reading or their listening skills, and it is therefore important both to teach and to test their receptive skills.

Furthermore, university language exams are always compared to existing standardised exams on the (international) language testing market. Most high-stakes exams assess each of the four skills separately, often allocating individual points or grades to each skill. The GULT team, therefore, finds it important to test all four skills, too, be it only for reasons of face validity. Nevertheless, the GULT team is looking for innovative ways that make testing of receptive skills even more authentic.

A GULT test or exam that tests all four skills individually, but in an interrelated situation, is therefore a compromise between an authentic linguistic situation in real-life and existing test traditions.

---

25 A higher degree of authenticity can be achieved by asking the learners to use the language in a real-life situation in the country where the target language is spoken. This is, however, normally not possible in a university language exam setting.

5.3.   The structure of a task-based test or exam

The GULT structure

Based on experience with the CLES (Certificat de Compétences en Langues de l'Enseignement Supérieur) and the UNIcert exams the GULT team has developed, in co-operation with the UNIcert Scientific Committee, the following structure for a task-based exam:

- introduction to the topic and the (overall) task
- listening comprehension and reading comprehension[26]
- case study work / case analysis (in pairs or small groups)
- writing
- speaking.

It is, however, important to note that different thematic fields may require different exam structures and different task types or tasks. A problem-based approach can easily be adopted in a business context, where business people develop solutions to existing problems. This might not always be the case, however, in science and engineering, where the focus may be on running projects or developing new machines or products.

The following sections deal specifically with the UNIcert exams as a way of exemplifying how task-based testing might be implemented. In Resource document 2 you can also find the CLES model as another example of implementing a task-based approach in university language testing.

Step 1: Introduction to the topic

It is important to describe the overall situation and overall task to the learners as this will enable them to focus their attention on relevant information. In "traditional" exams the candidates are mainly looking for the specific information asked for in each individual item, and they very often fail to grasp the overall meaning of the text. When using a task-based approach the candidates will, however, look for information they need in order to solve the problem in question. As in real life, attention is much more focused on relevant aspects. Consequently, the introduction to the problem / project is crucial to student understanding.

---

26 If an institution wants to make their exam even more authentic, it may opt for a fully integrative approach and not test the receptive skills individually, but through the outcomes of the productive skills, as described below.

Step 2: Receptive skills

Listening and reading have a new function. Test takers are no longer asked to "understand for the sake of understanding", in other words the overall task is no longer: "Please answer the questions!", but the test takers have to understand the text with a clear aim in mind, namely, they have to manage the (overall) task and will need to use the information provided by the texts in the part of the exam that tests productive skills.

It is important that the listening and the reading texts cover different aspects but they need to be related to the overall topic in order to prevent the test takers from using the information from the listening text to answer the questions in the reading part and vice versa.

The GULT team discussed various possibilities concerning the order in which the receptive skills should be tested. As listening is often considered more difficult by the test takers, it was thought that it might be good to start with the reading (so that the candidates get a general idea of the issue), then interrupt the activity and do the listening, and finally move back to the reading. For practical reasons, this does not seem feasible, however, as students might find it confusing to start with the reading part and then interrupt it. In practice, therefore, the tests always start with the listening test, but put emphasis on the introduction to the task to make sure that the test takers clearly know the overall task and its context before starting with the listening.

Therefore, it was decided to give the candidates both exam parts testing receptive skills, that is, listening and reading, at the same time. Once the recordings have been played, each test taker can decide whether they want to spend more time on answering the listening or the reading tasks. They do not have to wait for their fellow test takers to finish the listening before being able to move on to the reading. Again this corresponds to a real-life situation where people have a certain degree of flexibility in organising their work.

Step 3: Case study work / case analysis

The new element of the GULT test is the independent analysis of the dossier by the candidates. Students are put into pairs (or small groups) that will develop a solution or a proposal together and, in the speaking part of the exam, present it together. Depending on the local circumstances the candidates may be given complete freedom and can use all resources available, for example the library, self-access facilities or the Internet.

An important difference between task-based tests on the one hand and many standardised tests and university exams on the other is the use of an extensive dossier of authentic resources. The aim of this is to prepare the learners and test takers for realistic situations in their future professional life, by improving and testing their ability in using resources – a task they will have to fulfil later on in their career.

Step 4: Productive skills

In the writing and speaking part of the exam, the test takers have to present a specific product in a concrete context, that is, the overall task is embedded in a specific, authentic situation and framework that correspond to real-life activities. For this purpose, they also have to use and process information made available in the other parts of the exam, including the listening and reading part.

Step 4a: Writing

For the writing part of the exam, the learners have to work individually as the aim is to assess each test taker's individual competence in writing. They are, however, allowed to use the dossier and their notes from the case analysis phase.

The writing parts of the exam papers are photocopied after completion, as they will be needed again for the speaking test.

Step 4b: Speaking

The speaking exam does not take place immediately after the other parts of the exams as the learners might already be rather tired and as this would be a disadvantage to those starting first. It is, therefore, better to carry out the speaking exam a couple of days later.

For the speaking part the same candidates who co-operated during the case analysis phase get together in pairs or small groups again. The exam consists of:

- preparation: prior to the exam itself, the candidates are given a copy of their writing exam and are asked to prepare a joint presentation of their solution, that is, a pair (or group) presentation;

- exam – first part: presentation;

- exam – second part: discussion, namely clarifying aspects and answering questions.

Such a situation corresponds to a real-life situation: a team is asked to launch a project or to solve a problem, has to analyse the situation and prepare a written report, which it then presents to the work unit or line manager. It makes the exam situation much more authentic and the test takers put more energy into their proposal as they feel a need to show their competences (in the language and in the thematic domain) and try to convince the audience.

A major concern among language teachers attending GULT workshops or presentations has been the length of the speaking part. Some universities reported that they had dropped the speaking part completely as they were not able to cope with the huge

number of test takers. The GULT team, however, believes that testing speaking should be a key element of each language test and is more important than testing receptive skills or grammar and vocabulary. Communication can only take place if people can express themselves, in spoken and in written form. However, institutions may opt for a shorter duration of the speaking part of the exam. The team believes that the more time that is allocated to the speaking part, the more realistic language teachers' judgments will be, but it is understood that in many cases only the necessary minimum can be implemented.

Timing

As far as timing is concerned, to give an example, the following time allocations have successfully been piloted for a UNIcert III exam at CEFR level C1,[27] although this obviously depends on the length of the texts and the number and complexity of questions or tasks:

- listening comprehension and reading comprehension: 90 minutes for completing the listening and reading tasks after hearing the recordings twice;

- case study work / case analysis (in pairs or small groups): 90 minutes, including a (possible) 30-minute break;

- writing: 90 minutes;

- speaking: 30 minutes per candidate (60 minutes per pair).

Obviously, this is just an example in a given context and can be changed according to the institution's needs and resources.

The exam will be shorter at CEFR level B2.

---

27   This structure has been developed by the Scientific Committee of UNIcert at a meeting in Dötlingen, Germany, and has since been referred to as the "*Dötlinger Modell*". You can find the CLES timing structure in Resource document 2.

6.   The potential benefits and beneficiaries of task-based language assessment

In general, with a task-based approach in language testing there is a purpose in completing the exam and the individual tasks have a concrete function, as they are embedded in the overall task of the exam.

Task-based language testing can be described as authentic, motivating, student-centred and flexible:

- The testing is **authentic** because the materials used in the tests are authentic and the tasks resemble the ones the students may very well face in the near future. In other words, for example in a business English test, students are not asked to take the role of a director but rather that of an intern or an assistant.

- It is **motivating** because the whole assessment procedure is so clearly linked to the students' field of study. It can be said that task-based tests go beyond traditional LSP tests, which are based on field-specific materials but have traditional test items rather than goal-oriented tasks. There is a purpose in completing the exam, because the individual tasks have a concrete function, embedded in the overall task.

- The testing is **student-centred** because 1. the examination is totally based on the language skills and materials needed in the test takers' field of study, and 2. the products created during the assessment are totally student controlled, though the resources and assessment materials have been provided by the examiners.

- It is **flexible,** because the task-based approach suits many kinds of assessment situations, from achievement to proficiency testing. It can be used in continuous assessment or end-of-course examinations and focus on one or more skills. In task-based proficiency tests, all four skills are assessed.

During a task-based examination, several observations were made about the students:

- They appear to be less worried than in traditional exams, which may be due to the social aspect of the test, that is, the pair and group work.

- They seem to concentrate (more) on the task because of its structure and relevance to the outcome, in other words the test does not consist of detached individual test items but is a complete task.

- There is more involvement in the topic / content because of its authenticity and real-life relevance.

- The information received from the listening and reading appears to be successfully integrated into the speaking and writing parts of the exam. In other words, receptive skills are not assessed in a vacuum but are relevant for the assessment of the productive skills.

- Students tend to show better overall test results. Every student can contribute something in the examination situation, which seems to motivate the students to give their best.

- They are better able to show the range of their reading and writing skills, especially students with good language competences. In traditional language testing, every student is expected to answer the same questions in the same way. In task-based testing, students themselves determine the extent and the content of their spoken or written texts, within which they can exhibit their special knowledge and skills, for example comprehending implicit information in a reading task.

- Students are able to overcome their possibly negative initial reaction; they seem happier after the test. Some students may first react negatively to a task-based test, but they all tend to be happier after a task-based test than after a traditional test. One reason may be that, after realising that the student involvement in a task-based test is of a totally different order from that of answering or solving test items in a traditional language test, the students realise what freedom they have in constructing their answers in a task-based test.

- There appears to be less stress, perhaps because of the group work during the preparation phase.

- Less stress is also apparent during the test, because the candidates are working on an authentic problem (their strong involvement in the topic is visible).

- They are surprised at the variety of possible correct solutions.

- Very interesting discussions are initiated and conducted during the assessment.

Some observations have also been made about the examiners:

- There appears to be stronger involvement in and concentration on the topic of the task in the exam, that is, there is a clear increase of interest in the actual test.

- Examiners have more active roles in the discussion, with the second marker also participating. This leads to increased interaction with candidates in the assessment of oral skills.

- The atmosphere in the testing situation appears to be more relaxed: "You don't feel students' stress", as one examiner commented.

- They tend to forget that they are in an examination situation.

- They are surprised at how animated the student discussions are.

- Examiners are also surprised to experience that they have to stop candidates in the test: "Unfortunately the exam is over and we have to stop here".

- They can really see how competent the students are in managing a situation linguistically.

When the feedback given by the test takers and the testers was analysed, the top benefits mentioned were the higher motivation of both candidates and examiners, which led to better results and a higher satisfaction rate.

All in all, it can be concluded that task-based testing:

- increases motivation and involvement (of both candidates and examiners);

- has improved face validity; and

- possibly has also higher content validity, concurrent validity and consequential validity.

Task-based approaches stimulate language competence in all four skills but they also help develop the skills necessary for success in working life, not just in studying languages. These skills include:

- group work skills

- problem solving skills

- presentation skills

- discussion skills

- negotiation skills

- making compromises

- intercultural competence in an international setting

- study skills / "transferable skills"

- learner motivation.

Ideally, in a university context every testing situation should also be a learning situation, and past experiences tell us that in task-based testing every testing situation becomes a learning situation, both for the candidate and the examiner.

Of course, the introduction of the task-based approach in language courses will lead to teaching to the test. The GULT team considers this mainly as a positive "washback effect"[28] as this will lead to the implementation of a task-based approach to teaching and learning, and the learners will move away from focusing on aspects such as grammar and vocabulary without a context and on working with discrete-item testlets. Furthermore, the learners will be stimulated to work on previous task-based tests in self-study while preparing for the test to come.

In fact, with the task-based approach it is not possible to have a simple training for the test without considerable improvement in language skills, in other words there are no

---

28   See also: Norris, 2002: 341.

shortcuts to the test, as might be seen with certain standardised tests where standardised tasks or activities are repeated in every test in one form or another.

Therefore, the individual teacher has to make sure that the task-based approach does not lead to a negative washback effect, which would be the case if the topics and build-up tasks repeated themselves and narrowed down the scope of the subject matter and of task types. This means that test takers cannot prepare for certain parts of the test in advance, for example by learning certain phrases or test parts by heart.

7.   Challenges and limitations

The limitations of task-based testing generally fall into two categories: psychometric issues and practical, real-world concerns.

7.1.   Psychometric concerns

Among the problems cited with task-based tests is the difficulty of developing a truly authentic test. A test by its very nature is not an "authentic" situation. Although authentic materials may be used, test candidates may be asked to interact with them in ways that are not reflective of real-world use. As discussed earlier, authenticity is perhaps best viewed as one end of a continuum, a point test developers should try to approach as closely as possible, while recognising that it is unattainable.

Another criticism of task-based assessment is its lack of generalisability. Bachman, for example, has argued that the complex nature of a task means that a test taker's performance on a test task cannot be used to predict performance in a subsequent real-life situation. In other words, there is practically no generalisability across assessment tasks or extrapolation from tasks to other tasks and real life.[29] While this is true, the GULT team, however, thinks it is equally true (or even more true) for all other language tests and therefore not a special characteristic of task-based assessment. In fact, due to previous needs analyses – which should be carried out for each testing structure – task-based tests correspond more to real-life situations in the test taker's context and may give a better picture of the test taker's language competence.

7.2.   Practical concerns

Challenges related to the implementation of task-based assessment abound at every level: individual, institutional and even national.

University teachers, for example, are justifiably concerned at a personal level about the increased workload that task-based testing demands. In many institutes of higher education, testing has not been given proper attention as part of the university teachers' workloads. It is quite clear that, for teachers, task-based testing means:

- initially more work in test development, and also

- initially more correction work.

---

29   Bachman, 2002: 461f.

It takes time to get used to collecting, for any task-based test, a representative package of materials. Usually, due to the topical nature of the materials and to test security issues, the materials can only be used once in assessment, although they may be reused in teaching. It also takes time to learn to devise the right kinds of tasks, but once learned, the same task structure can be used in the following task-based tests.

Because there are no correct or wrong answers to the tasks in task-based assessment, the raters have to be especially conscious of inter-rater and intra-rater reliability. Also, raters need to meet for every major task-based test to agree on assessment criteria, especially when proficiency is being assessed.

At the institutional level, cost can be an impediment to the implementation of task-based assessment. The increased workload for teachers to develop and mark task-based tests often translates into greater compensation for the teachers. Furthermore, task-based assessments might require more resources such as access to the Internet in testing rooms and / or more photocopies.

At the national level, concerns are raised about the compatibility of task-based assessment with the existing culture of language testing and teaching. Often, task-based assessment conflicts with long-standing traditions and views of education. The following comments were collected from testing experts representing various European countries and their universities:

- The suitability / compatibility of task-based assessment with various local national identities was questioned.

- Any nationwide testing culture is difficult to change, in other words it takes a lot of time.

- The pragmatic aspects of the implementation of task-based assessment would have to be locally analysed and the necessary steps would have to be agreed upon and taken unanimously.

7.3. Other concerns

Testing experts consulted in the development of the GULT project raised several additional points:

- **Quality control:** How can the quality of task-based tests be controlled?

- **Applicability of assessment grids and criteria:** Which assessment grids and criteria can be used in task-based testing?

- **The link between task-based assessment and the CEFR:** How can task-based tests be related to the CEFR where task-based testing is not even mentioned as an example?

- **Absence of (student) self-assessment / reflection:** How can students be involved in the assessment of task-based tests; should there be an extra "reflection subtask"?

- **Student autonomy:** Should students be more involved in finding resource materials for their task-based tests? In the GULT model the testers at present provide the resource material packages but, depending on the local context, tradition and facilities, access to various resources may also be given.

8.   Evaluating test taker performance

It is one thing to devise a method for eliciting a language performance from a test taker; it is another, equally important thing to decide how to evaluate the language sample obtained. Indeed, test scores are useful and useable only if scoring procedures are appropriate. Some task-based assessments include subtests of reading or listening and, in these cases, the scoring usually is quite conventional. For the overall task and productive components of the assessment, however, the evaluation of the performance normally involves subjective judgment, and as such, is a more complicated endeavour. As described in Chapter 7, particular attention should be paid to the scoring grids and the reliability of the raters.

Receptive skills

In the GULT framework, as well as other models of task-based testing, there are test components related to reading and listening comprehension. Although these receptive skills tests are conceptualised as build-up tasks to help the test taker accomplish the overall task, they are normally marked as separate test components. The GULT receptive skills tests are not very different to existing tests, from a structural point of view: test takers listen to and read one or more authentic texts, and then answer questions testing the overall understanding of the text(s), as well as their understanding of specific details. A variety of formats should be used to accommodate the individual strengths and weaknesses of the test takers. As with any test, answer keys need to be prepared in advance for reasons of transparency and objectivity, facilitating at the same time the correction of the tests.

Productive skills

To evaluate productive performance, rating scales / evaluation grids need to be developed which reflect the construct underlying the test design. Practical constraints normally limit the number of aspects of performance that can be evaluated by raters. To decide which aspects are most important, test developers may want to consider the various competences listed in the CEFR (Council of Europe, 2001) and in the *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages* (Council of Europe, 2009), and follow the suggestions presented in the manual, namely to decide on the competence areas on which the test should be based and also on the individual aspects in each of these areas, by looking at the descriptors relevant for each area. The institutional culture or history may also dictate the inclusion of certain aspects, such as grammar, in the scoring grid.

Another key decision in developing the scale is the weight assigned to successful task completion. In some models of task-based testing, the only criterion used in judging test-taker performance is the success accomplishing the task. In most cases, however,

rating scales include criteria related to both language competence and task accomplishment. Many evaluation grids take into account text content, pragmatic skills and linguistic skills for writing, and text content, pragmatic skills, linguistic skills and paralinguistic skills for speaking. These might also include aspects such as complexity, accuracy and fluency.

Once it is determined which criteria will be used to judge test-taker performance, the institution needs to decide on the weighting of each of these elements in the final grade. Linguistic skills will probably be more important and get a higher percentage at a higher level as linguistic correctness becomes more and more important at higher levels of competence. At lower levels learners are expected to be able to simply convey their message (thus the importance of content and pragmatic aspects), at higher levels students are expected to express their ideas in a linguistically correct way. The grids used should guarantee that the number of false positives and false negatives[30] is as small as possible.

In cases where test scores will be used to make important decisions such as exit from a course or a programme, or admission to a course or programme, test developers also need to set standards or cut-off points for those crucial decisions. When doing so, an institution needs to take into account the difficulty of the exam in general and of the individual tasks, and consider the general institutional framework. In many cases these aspects are regulated by general university regulations. This means that the institution needs to define the minimum of marks at which they consider a result as acceptable, and it has to allocate possible marks according to the individual tasks, in compliance with the university regulations. In short, this means that the same standards and procedures apply as for any other well-developed language test.

In order to guarantee the comparability of task-based tests with other language tests, and the comparability of the test results,[31] it is important to make sure that the raters understand the CEFR descriptors in the same way. This makes teacher training events and regular staff meetings necessary and a conscientious utilisation of the *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages* (Council of Europe, 2009).

Well-developed tasks and rating scales will not produce useable scores unless raters are well trained in using the scales. To guarantee inter-rater reliability, that is, consistency among the different raters, it is important to organise training events to make sure that the evaluation grids are understood and used in the same way by the individual examiners. Such events are important, as language testing is not normally an important element in teacher education nor is it a required qualification when employing university language teachers (as the importance of language testing is often neglected

---

30 The concepts of "false negatives" and "false positives" are used to refer to students who either failed when they should have passed the test (false negatives) or passed when they should have failed the test (false positives).

31 Bachman, 2002: 455, 461.

by the institution as a whole and the language teachers' testing skills are taken for granted). Furthermore, these training events will improve the reliability of teachers' judgments of student performance.

Whereas one assessor might be sufficient for a low-stakes exam (for example an end-of-course exam), high-stakes exams (like most proficiency or "prochievement" tests) require two assessors. In this case, the roles need to be defined beforehand, so that each assessor knows what to do during the exam.

9. Potential uses of task-based tests / exams: how to exploit task-based tests

As mentioned at the beginning, task-based teaching goes hand in hand with task-based testing and assessment. Although task-based tests require a certain level of authenticity and need to address current events, they do not have to be thrown away once they serve their purpose as a test. They can be used and exploited again. Task-based tests can easily be used in the classroom and in self-study programmes for other purposes than testing, for example:

- as teaching material
- as learning material in self-study programmes
- as practice test material
- as self-assessment tests
- as model tests.

9.1. Achievement tests: end-of-course exams

Task-based tests can be used as end-of-course tests. In this case, for reasons of time it will in most cases not be possible to follow the integrative approach mentioned above, which tests all four skills in one major exam. An institution or a language teacher may therefore use a case study to test only speaking, only writing, or both speaking and writing.

Clearly, the tasks need to correspond to the teaching carried out during the course, in other words the learners first need to be familiarised with the case study approach by carrying out one or more case studies in class if they have to work on a case study in the final exam.

Over the past few years, institutions have moved towards a "portfolio" approach in teaching and testing languages in their course syllabi. In this case, the learners work on one or several projects, a global simulation and / or case studies throughout the course, and results of these activities will then compose the final grade. Certain aspects or text types can be tested during the course; others will be tested in the final exam. If, for example, a teacher uses a global simulation in class, he may ask the learners to write a business letter in relation to their global simulation project during the course, and in the final exam ask them to write a report on a case study based on their project. By doing so the teacher can test the test takers' competence in writing different text types they might need in their future professional life.

9.2. Proficiency and prochievement tests

A task-based approach can also be used for proficiency (or prochievement) testing. Both the CLES exam and the UNIcert exam are existing models of proficiency / prochievement tests based on tasks. They are not directly linked to a course, but test the overall language competence of the test taker at the end of a teaching and / or learning process.

"Prochievement tests" are understood to be a combination of proficiency tests and achievement tests, in other words they are related to a specific syllabus, but go beyond it, while the evaluation is criteria-based:

> In general, the profession defines achievement tests as those limited to a particular body of material just covered in class(es) and proficiency instruments as those testing the total range of skills and contexts a learner may be able to handle – regardless of where and when they may have been learned – and testing them through actual interaction in realistic situations. Prochievement *tests* are a combination of the preceding two types, testing students' ability to perform in only the contexts and situations that have been practiced in class.
>
> (Gonzalez Pino, 1998: 120)

In the UNIcert context, the learners first have to attend at least 112 hours of face-to-face teaching (plus an extensive element of self-study), and succeed in end-of-course exams, before being allowed to register for a UNIcert exam. It can therefore be considered as a "prochievement test". The CLES, on the other hand, does not necessarily require previous participation in a specific course or learning programme, although students will normally first do so before registering for the exam. It is thus a proficiency test testing linguistic competence in a specific situation.

If the UNIcert and the CLES exams are compared, UNIcert allocates grades for each exam as it is fully integrated into the university teaching and evaluation system (the grades are those used in that particular institution), whereas the CLES does not allocate grades but works on "pass" / "fail" basis. The CLES may, however, also be used as an end-of-course test (for example at the University of Strasbourg), in which case grades are given. This demonstrates that GULT tests and exams can either opt for a "pass" / "fail" system or a detailed grading system.

See Resource document 2 for details of how UNIcert and CLES have been used in these ways.

10. How to develop a task-based test

10.1. Test design

10.1.1. The problem

When developing a task-based language exam, the test developer first has to look for an authentic problem or project. It needs to be a problem (or project) the test takers are not (or only partly) familiar with, as otherwise those who are competent in the area have an advantage compared to those for whom everything is new. The problem or project needs to be authentic and correspond to the test takers' field of research or interest, in other words it must be relevant to them.

10.1.2. Authenticity of the task

The test developer then has to think about an overall task that corresponds to the test takers' situation. This includes the authenticity of their role, for example students at university will not be put into the role of managing director of a multinational company, as in real life they will not start their professional career after graduation as a managing director or immediately become a manager. They will start at a far lower level and will have to climb up the hierarchical ladder. If the role and the task are authentic, the proposals developed by the test takers will be far more authentic, too.

10.1.3. The dossier

Next, the test developer has to collect relevant documents, both oral and written, that deal with the problem or project in question. This material then needs to be sorted:

- Which audio file(s) / video(s) can be used for the listening part of the exam?

- Which written text(s) are most appropriate for the reading part?

- Do the texts for listening and reading complement each other (or can the candidate use the information from the listening to answer the questions in the reading part)?

- Which documents are useful for the dossier? In which order should they be presented? Which documents can be left out?

10.1.4. The situation

The test developer now describes the background situation of the proposed task in writing. This description should be fairly detailed in order to guarantee that the test takers have fully understood the situation.

10.1.5. The task

In the following step the test developer has to develop a realistic overall task in more detail. This task needs to specify the general aim of the activity, namely the case study work or the project work. In other words, the test takers are told: "This is the situation, this is your role, this is the problem / project, and now you have to develop a solution / proposal."

10.1.6. The build-up tasks for the individual skills

The final step of the test design phase consists in the writing of the individual build-up tasks for the four individual skills. These all need to be integrated into the overall task and complement each other.

For each part of the exam, once the test developer has prepared the task, he or she has to look back at the overall task and check whether each individual task is part of the whole picture, helps to reach the overall aim (the solution of the problem or the development of a project), and is relevant for the overall task. Following a holistic approach, the test developer puts all the individual tasks together to make it "a whole".

When drafting the individual tasks for the individual parts of the exam the test developer has to think how to integrate "relevant language use" (syntax and terminology) into the individual tasks, in other words how the "relevant language use" is made (or can be made) inherent in the receptive or productive tasks.

In general, repeating the overall task in each section of the test and explaining in a very detailed way the tasks and products the test takers have to develop can be very useful as repetition helps to clarify the situation, roles and tasks.

10.2. Next steps in test development: trying out

As with any other language test that follows quality assurance measures, task-based tests need to be piloted, pretested and / or trialled, as described in the *Manual for Language Test Development and Examining* (2011). In a university context, and in particular for a GULT test, which is based on current issues, this can mean:

Piloting involves asking a small number of people to complete items as if in a test. This may be quite informal and could, for example, involve work colleagues if no one else can be found. Their RESPONSES are analysed and, together with their comments ..., may be used to improve items further.

(*Manual for Language Test Development and Examining*, 2011: 32)

Similarly, a pre-testing with a small number of potential test takers can take place to avoid unexpected responses by the test takers in the actual delivering of the test. Trialling, that is, a small-scale statistical analysis of subjectively marked tasks, will probably only be possible in larger institutions or in networks of institutions.

11. Bibliography[32]

Abendroth-Timmer, D. and Breidbach, S. (eds) (2000). *Handlungsorientierung und Mehrsprachigkeit*. Frankfurt am Main: Peter Lang.

AFANet website: http://www.afanet.info.

Almagro Esteban, A. and Pérez Cañado, M. L. (2004). "Making the Case Method Work in Teaching Business English: a Case Study". In: *English for Specific Purposes* 23, 137-161.

ALTE (1998). *Multilingual Glossary of Language Testing Terms: Studies in Language Testing 6*. Cambridge: Cambridge University Press.

*Bachman, L. (2002). "Some Reflections on Task-Based Language Performance Assessment". In: *Language Testing* 19(4), 453-476.

*Bachman, L. (2007). "What is the Construct? The Dialectic of Abilities and Contexts in Defining Constructs in Language Assessment". In: Fox, J., Wesche, M., Bayliss, D., Cheng, L., Turner, C.E. and Doe, C. (eds). *Language Testing Reconsidered*. Ottawa: University of Ottawa Press, 41-71.

Bausch, K.-R., Burwitz-Melzer, E., Königs, F. G. and Krumm, H.-J. (eds) (2006). *Aufgabenorientierung als Aufgabe*. Tübingen: Narr.

Bourguignon, C. (2006). "De l'approche communicative à l' « approche communic-actionnelle » : une rupture épistémologique en didactique des langues-cultures". In: *Synergies Europe* 1, 58-73.

Bourguignon, C. (2007). "Apprendre et enseigner les langues dans la perspective actionnelle: le scénario d'apprentissage-action". Paper presented on 7 March 2007, and online publication in: *Les Langues Modernes,* available at: http://www.aplv-languesmodernes.org/spip.php?article865

*Bourguignon, C., Delahaye, P. and Puren, C. (2007). Évaluer dans une perspective actionnelle: L'exemple du « Diplôme de Compétence en Langue » [available from C. Bourguignon].

Brown, J. D. (1998). *New Ways of Classroom Assessment*. Alexandria, VA: Teachers of English to Speakers of Other Languages.

*Brown, J. D., Hudson, T., Norris, J. and Bonk, W.J. (2002). *An Investigation of Second Language Task-Based Performance Assessments*. Technical report No. 24, Honolulu: University of Hawaii Press.

---

32 Publications that cover particular aspects of task-based (or performance-based) language testing and that the GULT team recommends for reading are marked with a preceding asterisk: *.

Cain, P. and Heath, S. (eds) (2004). *Handbook for Lecturers: Design, Development and Delivery of Web-based Case Study Resources Supporting Student Centred Learning*. Newcastle: University of Newcastle.

Caré, J.-M. and Debyser, F. (1995). *Simulations globales*. Sèvres: CIEP.

Casey, E. and Fischer, J. (2004). "Improving Interaction Between Subject Area Specialists and Language Teachers at UCARS (Universities and Colleges of Agricultural Sciences)". In: Satchell, R. and Chenik, N. (eds). *University Language Centres: Forging the Learning Environments of the Future. Papers from 7th CercleS Conference. Paris, 19-21 September 2002*. Paris / Dublin: CercleS, 431-443.

Casey, E. and Fischer, J. (2005). "The Use of Case Studies in Language Teaching". In: Moravčíková, A., Taylor Torsello, C. and Vogel, T. (eds). *University Language Centres: Broadening Horizons, Expanding Networks. Proceedings of the 8th CercleS Conference. Bratislava, 9-11 September 2004*. Bratislava: CercleS, 163-175.

*Chalhoub-Deville, M. (2001). "Task based assessments: Characteristics and validity evidence". In: Bygate, M., Skehan, P. and Swain M. (eds). *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*. Harlow: Longman.

Clapham, C. and Corson, D. (eds). *Encyclopedia of Language and Education: Vol. 7: Language testing and assessment*. Dordrecht: Kluwer Academic.

CLES website: http://www.certification-cles.fr/.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe / Cambridge: Cambridge University Press.

Council of Europe (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). A Manual*. Strasbourg: Council of Europe, Language Policy Division; available at: http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp.

Daly, P. (2002). "Methodology for Using Case Studies in the Business English Language Classroom". In: *The Internet TESL Journal – For Teachers of English as a Second Language*, Vol. VIII, No. 11. http://iteslj.org/Techniques/Daly-CaseStudies/ [Internet publication].

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999). *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

Debyser, F. (1996). *L'immeuble*. Paris: Hachette.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

Eggensperger, K.-H. and Fischer, J. (eds) (1998). *Handbuch UNICERT®*. Bochum: AKS-Verlag.

Elder, C. (1993). "How Do Subject Specialists Construe Classroom Language Proficiency". In: *Language Testing* 10(3), 235-254.

*Elder, C. and Brown, A. (1997). "Performance testing for the professions: Language proficiency or strategic competence?". In: *Melbourne Papers in Language Testing* 6(1), 68-78.

Elder, C., Iwashita, N. and McNamara, T. (2002). "Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer?". In: *Language Testing* 19(4), 47-368.

*Ellis, R. (2003). *Task Based Language Learning and Teaching*. Oxford: Oxford University Press.

EXPLICS website: www.zess.uni-goettingen.de/explics.

*Finch, A. (2002). "Authentic Assessment: Implications for EFL Performance Testing in Korea". In: *Secondary Education Research* 49, 89-122. Available at: http://www.finchpark.com/arts/Authentic_Assessment_Implications.pdf.

Fischer, J. (2003). "AFANet – HERMES: A Bridge between Scientists and Language Teachers: Goals and Perspectives". In: *Języki obce: narzędzie komunikacji naukowej. Ogólnopolska Konferencja Naukowa z okazji 50-lecia Studium Języków Obcych Akademii Rolniczej im. Augusta Cieszkowskiego w Poznaniu.* Poznań: Wydawnictwo Akademii Rolniczej, 16-30.

Fischer, J. (2004). "HERMES – The AFANet Language Policy Group: Strengthening language teaching for students of agriculture and related sciences". In: Heath, S. (ed.). *Capitalising on Innovation in the Curriculum in European Higher Education. Practice and Promise in the Disciplines of Agriculture, Forestry, Aquaculture and Environmental Sciences. Proceedings of the AFANet – ICA Conference held at the University of Gent, Belgium, 8 to 10 January 2004.* Nancy: Institut National Polytechnique de Lorraine (INPL), 113-131.

Fischer, J. and Casey, E. (2005). "Der Einsatz von Fallstudien im Fremdsprachenunterricht an Hochschulen". In: Gebert, D. (ed.). *Innovation aus Tradition. Dokumentation der 23. Arbeitstagung 2004.* Bochum: AKS-Verlag, 175-184.

Fischer, J., Casey, E., Abrantes, A. M., Gigl, E. and Lešnik, M. (eds) (2008). *LCaS: Language Case Studies. Teacher training modules on the use of case studies in language teaching at secondary and university level. A handbook.* Strasbourg: Council of Europe.

Fischer, J., Musacchio, M. T. and Standring, A. (eds) (2009). *EXPLICS – Exploiting Internet Case Studies and Simulation Templates for Language Teaching and Learning. A Handbook.* Göttingen: Cuvillier Verlag.

Foster, P. and Skehan, P. (1996). "The influence of planning and task type on second language performance". In: *Studies in Second Language Acquisition* 18, 299-323.

*Foster, P. and Skehan, P. (1999). "The influence of source of planning and focus of planning on task-based performance". In: *Language Teaching Research* 3, 299-324.

Gonzalez Pino, B. (1998). "Prochievement testing of speaking: Matching instructor expectations, learner proficiency level, and task type". In: *Texas Papers in Foreign Language Education* 3.3, 119-133.

HERMES website: www.uni-tuebingen.de/hermes/ .

Hill, B. C. and Ruptic, C. (1994). *Practical aspects of authentic assessment: Putting the pieces together*. Norwood, MA: Christopher-Gordon Publishers.

Hughes, A. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Kaiser, F. J. (ed.) (1983). *Die Fallstudien. Theorie und Praxis der Fallstudiendidaktik*. Bad Heilbrunn: Klinkhardt.

Kiefer, K.-H. (2004). "Fallstudien – zum Umgang mit einer erfolgreichen Lernmethode im Fachsprachenunterricht Wirtschaftsdeutsch". In: *Info DaF* 1, 68-98.

Kohonen, V. (1997). "Authentic Assessment as an integration of language learning, teaching, evaluation and the teacher's professional growth". In: Huhta, A., Kohonen, V., Kurki-Suonio, L. and Luoma, S. (eds), *Current Developments and Alternatives in Language Assessment: Proceedings of LTRC 1996*. Jyväskylä: University of Jyväskylä, 7-22.

Kohonen, V. (1999). "Authentic assessment in affective foreign language education". In: Arnold, J. (ed.) *Affect in Language Learning*. Cambridge: Cambridge University Press.

LCaS website: http://www.ecml.at/mtp2/LCaS.

Lewkowicz, J. (2000). "Authenticity in language testing". In: *Language Testing* 17(1), 43-64.

Manual for Language Test Development and Examining. For use with the CEFR. Produced by the Members of ALTE on behalf of the Language Policy Division, Council of Europe. (2011). Strasbourg: Council of Europe (Language Policy Division), available at: http://www.coe.int/t/dg4/linguistic/ManualtLangageTest-Alte2011_EN.pdf.

McNamara, T. (1996). *Measuring Second Language Performance*. London: Longman.

McNamara, T. (2000). *Language Testing*. Oxford et al.: Oxford University Press.

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

Morton, J., Wigglesworth, G. and Williams, D. (1997). "Approaches to validation: Evaluating interviewer performance in oral interaction tests". In: Brindley, G. and Wigglesworth, G. (eds). *Access: Issues in Language Test Design and Delivery*. Sydney: NCELTR, 175-196.

*Norris, J. (2002). "Interpretations, Intended Uses and Designs in Task-Based Language Assessment". In: *Language Testing* 19 (4), 337-346.

Norris, J. M., Brown, J. D., Hudson, T. and Yoshioka, J. (1998). *Designing Second Language Performance Assessments*. Honolulu: University of Hawaii Press.

Norris, J. M., Brown, T. D. and Bonk, W. (2002). "Examinee Abilities and Task Difficulty in Task-Based Second Language Performance Assessment". In: *Language Testing* 19(4), 395-418.

Nunan, D. (1989). *Designing Tasks for the Communicative Classroom*. Cambridge: Cambridge University Press.

O'Malley, J. M. (1996). Authentic Assessment for English Language Learners: Practical Approaches for Teachers. Reading, Mass.: Addison-Wesley Pub. Co.

*Petermann, A. (2008). "CLES – Certificat de Compétences en langues de l'enseignement supérieur. Zertifizierung der Sprachkompetenzen im französischen Hochschulbereich". In: *Fremdsprachen und Hochschule* 79/80, 33-54.

Puren C. (2002). "Perspectives actionnelles et perspectives culturelles en didactique des langues : vers une perspective co-actionnelle co-culturelle". In: *Les langues modernes* 3, 55-71.

Puren C. (2006a). "De l'approche communicative à la perspective actionnelle". In: *Le Français dans le Monde* 347, 37-40.

Puren C. (2006b). "Les tâches dans la logique actionnelle". In: *Le Français dans le Monde* 347, 80-81.

Puren, C. (2008). "Formes pratiques de combinaison entre perspective actionnelle et approche communicative : analyse comparative de trois manuels". In: *Langues Modernes* 5 January 2008, 1-19, available at: http://www.aplv-languesmodernes.org/spip.php?article1409#.

Ribé, R. and Vidal, N. (1993). *Project Work*. Oxford: Heinemann.

Schart, M. (2003). *Projektunterricht – subjektiv betrachtet. Eine qualitative Studie mit Lehrenden für Deutsch als Fremdsprache.* Hohengehren: Schneider Verlag.

Skehan, P. (1996). "Second language acquisition research and task-based instruction". In Willis, J. and Willis, D. (eds). *Challenge and Change in Language Teaching*, Oxford: Heinemann, 17-30.

Skehan, P. (2001). "Tasks and language performance". In: Bygate, M., Skehan, P. and Swain, M. (eds). *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing.* Harlow: Longman, 167-187.

Skehan, P. (2003). "Task-based instruction". In: *Language Teaching* 36, 1-14.

Spence-Brown, R. (2001). "The eye of the beholder: Authenticity in an embedded assessment task". In: *Language Testing* 18(4), 463-481.

*Standring, A. (2009). "Language Assessment". In: Fischer, J., Musacchio, M. T. and Standring, A. (eds). *EXPLICS – Exploiting Internet Case Studies and Simulation Templates for Language Teaching and Learning. A Handbook*. Göttingen: Cuvillier Verlag, 45-65.

Teasdale, A. (1996). "Authenticity, validity, and task design for tests of well-defined LSP domains". In: *Thames Valley University Working Papers in English Language Teaching* 2, 137-150.

Uber Grosse, C. (1988). "The Case Study Approach to Teaching Business English". In: *English for Specific Purposes* 7, 131-136.

UNIcert® website: www.unicert-online.org.

UNIcert®LUCE website: http://www.casajc.sk/unicert.htm.

Van den Branden, K., Bygate, M. and Norris, J. M. (eds) (2009). *Task-Based Language Teaching. A reader*. Amsterdam / Philadelphia: Benjamins.

Vaughn, C. (1991). "Holistic assessment: What goes on in the raters' minds?" In: Hamp-Lyons, L. (ed.). *Assessing Second Language Writing in Academic Contexts*. Norwood, NJ: Ablex: 111-126.

Voss, B. (2010). *Handbuch UNIcert 2*. Bochum: AKS-Verlag.

WebCase website: http://webcase-online.info.

Weideman, A. (2006). "Assessing academic literacy in a task-based approach". In: *Language matters* 37 (1), 81-101.

Weir, C. J. (1990). *Communicative language testing*. Hemel Hempstead: Prentice Hall.

Weir, C. J. (2005). *Language Testing and Evaluation*. Basingstoke: Palgrave Macmillan.

Wesche, M.B. and Skehan, P. (2002). "Communicative, task-based, and content-based language instruction". In: Kaplan, R. B. (ed.). *The Oxford handbook of applied linguistics*. Oxford: Oxford University Press, 207-228.

*Wigglesworth, G. (2001). "Influences on performance in task-based oral assessments". In: Bygate, M., Skehan, P. and Swain, M. (eds). *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*. Harlow: Longman, 186-209.

*Wigglesworth, G. (2008). "Task and Performance Based Assessment". In: Shohamy, E. and Hornberger, N. H. (eds). *Encyclopedia of Language and Education, 2nd Edition, Volume 7: Language Testing and Assessment*. 111-122.

*Wu, W. and Stansfield, C. (2001). "Toward authenticity of task in test development". In: *Language Testing* 18(2), 187-206.

Yaiche, F. (1996*). Les simulations globales – mode d'emploi.* Paris: Hachette.

For further reading on language testing in general, see also detailed list available at: http://languagetesting.info/video/reading.html.

Part B

Resource documents

Resource document 1: About tests and assessment

The following provides background information on language testing in general. Although this text does not specifically relate to task-based assessment, it provides a context and basic testing information for those who are less familiar with testing literature.

Pre-university foreign language instruction aims at giving learners the necessary tools to be effective communicators in the target language in everyday situations. This is language for general purposes (LGP). Very few of these foreign language programmes go beyond the difficulty of the language in newspapers. In other words, pre-university language learners should be able to discuss matters that come up in newspapers at the level of the language used in newspapers. This level of language proficiency is, however, not enough to meet university language requirements.

Universities and other academic institutions often emphasise their responsibility to improve their students' language skills so that they can study in a foreign language and, after graduation, are able to function as professionals internationally, using the foreign language/s in their future work. It is also becoming more and more common for university students to complete part of their degrees in some foreign language either as exchange students or, in their home institutions, by participating in degree programmes given in a foreign language. This is in addition to the fact that most universities have foreign language requirements in their degrees, whether or not their students have to use foreign language materials in their studies.

University foreign language instruction differs from pre-university foreign language programmes in emphasising or concentrating on languages for academic purposes (LAPs) or languages for specific purposes (LSPs), also called "professional languages". LAP courses cover the foreign language material students need if they study in a university / degree programme using that foreign language. Thus, in addition to some subject specific language material, the courses cover, for example, language functions the students have to master in the situations they face in a university bureaucracy.

Languages for specific purposes differ from both LGPs and LAPs in many ways. First of all, LSP users can be clearly defined according to their occupations, professions or academic orientations. Each LSP has its own lexical, syntactic and stylistic characteristics. For instance, words that are otherwise very infrequent may occur quite frequently (e.g. "catkin", "irruption"), the passive voice can be much more commonly used than in LGPs or LAPs, and otherwise stylistically awkward long compounds are needed for the sake of accuracy ("skin cancer vaccine research").

Learning an LSP has proven to be very motivating for university students, many of whom plan on working in international jobs. In addition, the possibilities of successful exchanges in foreign universities are increased when the students know the language used by foreign professionals in their area of study.

LSP instruction in the universities is based on previous research and needs analyses. LSPs were a very fashionable subject of research in the 1980s, and the most popular LSPs in Europe then were, and perhaps still are today, English, French and German. Typical LSP courses would be Business English, French for medical students and German for biology students.

When the importance of a language test is discussed, the terms "high stakes test" and "low stakes test" are used. A high stakes test is a very important examination the outcome of which is the basis of far-reaching decisions such as admission to a university. Also, high stakes tests can usually be taken only a set number of times, and a fee may be required. Low stakes tests are minor examinations which can be retaken or, if a series of tests is taken, poor results in one test can be compensated for by good results in another test. Typical low stakes tests are quizzes or smaller language tests.

Every test can be placed on a continuum with **"**speed tests" at one end and the so-called "power tests" at the other end**.** In speed tests, the decisive factor is how fast the test taker can do the test. In a true power test, the test takers can have as much time as they need to complete the test, that is, to show their power in the area to be tested. It is quite clear that no pure speed or power tests exist in university foreign language instruction, but when language tests are being planned, test designers must have a very clear idea of how much time will be allocated for each item and for the whole test. One has to depend on estimated averages: there will never be an occasion when all students will need the same amount of time to complete a test.

The terms "achievement tests" and "performance tests" are used in relation to the material to be tested. Achievement tests show how much students have learned of the language material they have been taught. Therefore, achievement tests have to be based on the course materials and very little else, and they are typically final tests in language courses. In contrast, performance tests are not based on any definite body of language but they show how well the testee can perform in certain situations. Students know what to study for an achievement test while performance tests may cover all kinds of competences, not taking into account the testee's previous language studies. Many internationally available commercial language tests are performance tests.

Language examinations can be used to predict a student's future success in language studies, and these are "prognostic tests"**.** The results of a "diagnostic test" show what the student's language skills are at the time of the examination. In universities, the most common test type of these two is, without doubt, the diagnostic test.

In the design phase of a language examination, one of the important decisions to be made is the extent to which the test is to measure the testees' language skills (the other decisions being the amount of time allowed for completing the test, the selection of test materials and the types of items or tasks used in the test). "Formative tests" concentrate on some aspect of the foreign language, such as the use of articles, while **"**summative tests" have one or more tasks where language skills are observed as a whole, for example an essay. The concept of a "formative test" is also used to describe a test which is not graded. If a grade is given or the test result is a pass or a fail, students are

told the grades or the levels for their information, so that they can learn where to improve their skills. In assessment for learning (AFL), formative tests are used as teaching tools.

"Objective tests" and "subjective tests" should actually be called "objectively scored tests" and "subjectively scored tests". There are no truly objective language tests because every examination is based on some test maker's idea of what language material is important and what competences are needed to complete a language examination successfully. Objectively scored tests are typically multiple-choice or true-false tests with a key of correct answers which every rater uses in the same way. The prerequisite is that there is only one correct answer for each item. The rater's personality and personal preferences play a more important role in subjectively scored tests because there is no key, and written instructions can often be interpreted in many ways. That is why in high stakes examinations, raters have training sessions before they start grading summative tests.

"Scoring" means, first of all, deciding how many points are granted for each item, task and section and what is required for receiving full credit. The most important point in any test score is the "cut-off point", in other words, the point where the line is drawn between those who pass and those who fail the test. The optimal cut-off point is based not only on tester intuition but also on research.

In universities, many tests are not graded on a scale but are either passed or failed ("pass / fail tests"). In a language examination, the cut-off point and the borders between grades can be decided on the basis of either a "norm-referenced" or "criterion-referenced" evaluation. In norm-referenced evaluation, the grades are determined by the previous experiences of the tested population, and the testees receive their grades according to the percentiles they are placed in. For example, it is already known what percentage of the test takers will get the best grade or fail even before the first test is graded. This means that the raw score used as the cut-off point may differ from test to test but not so much that it would make the test unreliable. Norm-referenced testing is not appropriate in assessing small populations.

Criterion-referenced evaluation means that there are certain criteria that the test taker must fulfil in order to pass the test or get a certain grade. Most university language tests are criterion-referenced, which means that it is known beforehand what is needed for achieving maximum points, but it is not known how many students will achieve which grades or will fail the test.

With regard to language examinations, one can tell whether the test is a **"discrete-item test"** ("discrete-point test") or an "integrative" test. Discrete-item tests have sections that focus on clearly identifiable linguistic phenomena and that are totally separate from one another, with their own sets of points; the score is the sum of the points given for the various items. An integrative test may test several skills at the same time or language in general without an emphasis on, for example grammar or vocabulary. A "cloze test" is an integrative test because in order to fill a gap, the test taker must understand both the contextual and syntactic cues.

"Direct testing" means, for example, that speaking is tested by letting the test takers speak and then their speech is rated. "Indirect testing", when taken literally, covers all testing of receptive skills because reading comprehension or listening comprehension cannot be tested as such – a teacher has to resort to speaking or writing to find out what has been understood in a reading or listening test. A more concrete example of an indirect test is when one half of a discussion is given on paper to the test taker, who has to provide the other half in writing. In other words, speaking skills are tested through writing.

For approximately a thousand years, the main foreign language instruction method was the so-called "grammar-translation method". Testing meant mainly translating texts from and into the target language. By the middle of the last century, the "psychometric-structuralist method" started gaining ground, and "multiple-choice tests" became popular. This method and these kinds of tests were, however, severely criticised because they were felt to be so far removed from the reality of using languages; ticking alternatives was thought to have very little to do with actually using the foreign language to communicate. After a couple of decades, the era of the "communicative method" brought about "communicative tests", in which the test takers were really communicating in the target language, conveying and exchanging relevant information. The extreme form of communicative testing is **"authentic tests"**, which should consist of authentic situations of foreign language use to be assessed as performance tests. Such a test might be to send students to a library where they have to get a certain book using the target language; the tester would video the exchanges between the students and the librarian, and then assess the students' speech.

A language is such a complex entity that a foreign language test must concentrate only on part of it. When testing language skills, these are usually divided into two kinds: the "receptive" skills (listening and reading) and "productive skills" (speaking and writing). The term "modalities", often used in testing literature, refers also to listening, speaking, reading and writing. The *Common European Framework of Reference for Languages* divides languages into 57 "competences", which, however, are not to be tested one by one. Rather, in a writing test for example, the tester has to keep in mind what competences are actually being tested.

Every language test should have "validity" and "reliability". A test can be shown to be valid through "a priori validation" or "a posteriori validation". A priori validation, which is done after the test has been constructed but before it is used, is based on studying and weighing various testing theories and using previous research on similar tests and testing situations. A posteriori validation, which takes place after the test has been administered, uses all kinds of measures and analyses.

A test is valid if it measures what it is supposed to measure and convinces the stakeholders of that. There are many kinds of validity. "External validity" depends on the representability of the test and the generalisability of the test results. "Internal validity" refers to the content of the test and, specifically, to the fact that the test itself does not contain any irrelevant factors that would disturb the test taker. "Predictive

validity" means that the test can be used to predict the examinee's future success in learning that or some other language. "Content validity" shows whether the test has measured what it was supposed to measure, for example that the content of an achievement test does not extend beyond what was taught before the test. "Concurrent validity" is calculated when the test is being compared to other tests of approximately the same ilk or given at the same time. "Construct validity" has to do with the theoretical framework underlying the test, that is, whether the test is really that of a construct, such as an LSP reading comprehension. "Face validity", sometimes jokingly also called "faith validity", deals with the acceptance by the stakeholders of the test as a measure of what it promises to measure. For instance, some students find the face validity of multiple-choice tests as reading comprehension tests somewhat questionable because they find the alternatives too limiting or naïve.

"Reliability" in testing means that if a test is administered to similar populations in similar testing situations, it will yield similar results. "Intra-rater reliability" refers to the ability of one rater to grade tests systematically. Intra-rater reliability is calculated in cases when the rater's judgments may have been influenced by illness or exhaustion. Even the time of the day and the amount of work done before grading may affect the results. "Inter-rater reliability" is calculated when there are several raters grading the same test, and the testers need to know that all the raters rate the tests in the same way. It is a student's right to know that grading is fair.

When the test has been administered and graded, it is time to take a good look at the test again. Every teacher-tester should complete a simple "item analysis", which consists of calculating the "difficulty and discrimination indices". Difficulty (marked "p" in tables) is calculated by dividing the number of correct answers by the number of test takers. The difficulty index varies from 0 to 1. The larger the number is, the easier the item. It is still customary to approve of higher difficulty indices for the very first and last items so that the examinees are not terrified by the test when they start, and they finish the test feeling that they were able to contribute. However, the test designers must in advance know the target population well enough to have some idea of how difficult the test items will be.

Even if the test items have the accepted number of correct answers, the test designers must find out whether the test items really discriminate between the good and the bad language learners. In other words, if the predicted difficulty index is .5 (= half the test population has got the item right), the item is of poor quality if those who are not good at the language had the right answers while the good language learners have answered the item incorrectly. The discrimination index ("D" in tables) can be calculated using a simple formula: The tested population is divided into a good half and a poor half according to their total "raw scores" (the total scores before any manipulation or adjustment). Then the number of correct answers in the poorer half is deducted from the number of correct answers in the better half. The resulting number of correct answers is divided by the total of all correct answers. The discrimination index varies between -1 (all the correct answers are in the poorer half) and 1 (all the correct answers are in the better half).

"Transparency" in testing means that the tests are described in such detail that all those handling the test results know what was tested and how, and what the test results mean, that is, what the test taker with a certain grade or score can do in the target language. That is why European universities take great pains to unify the language used in test descriptors and criteria, the most important tool now being the *Common European Framework of Reference for Languages.*

Resource document 2: Examples of the implementation of task-based testing in university proficiency testing systems: "Make it real"

The CLES exam

The French CLES exam, "Certificat de compétences en langues de l'enseignement supérieur", was developed in 2000 by an order of the French National Ministry of Education. A working group of teachers from different universities, teaching languages to students of all disciplines, after a trial period developed the exact examination format for the *Common European Framework of Reference for Languages (CEFR)* levels B1 (CLES 1), B2 (CLES 2) and C1 (CLES 3). This group still co-ordinates the development of the tasks[1] and the administration of the exam sessions organised at the various university language centres, and provides the evaluation grids.

In a university context, a language certificate needs to show that the candidates have the necessary linguistic and pragmatic competences which allow them to:

▪ handle information correctly: selecting, weighing and prioritising facts, identifying and comparing different points of view;

▪ convey this information: building an argumentation, relating different arguments, presenting an opinion, debating, offering a solution and reaching an agreement.

In order to be able to assess whether a candidate has acquired these competences, the student has to be put into a situation that is taken from "real life" and asked to work on a complex task which requires him or her to use, and combine, multiple competences and skills (linguistic, sociolinguistic, pragmatic, etc.). For a university student it is normal to read and / or to listen in order to acquire knowledge, to find information and to use it in writing and speaking. The option of organising the whole test around one single overall task is therefore completely in line with the task-based approach of the CEFR.

The topics covered in the tests are built on scenarios: the candidates are put into a specific situation and are given a dossier with various documents. They then have to work on tasks and produce documents within the context of the scenario, while the receptive build-up tasks allow them to integrate elements into their writing and speaking.

---

1   The test developers, teachers at university language centres, normally working in teams of two, have to follow the test specifications, which are revised annually. The tests developed at one centre are proof-read and tested at two other centres before being accepted and entered into the national exam database.

CLES 2, level B2: language and topic of general academic interest

The exams follow the GULT structure as described in section 5.3, as all four skills are evaluated. The case study work in small teams is, however, not carried out. The exam specifications set the format of the exams:

- listening: two or three authentic spoken documents of 5 minutes in total, consisting of three different passages;

- reading: a dossier of three to five authentic texts of approximately 9 000 characters in total;

- writing: a summary, a report (described in more detail in the description of the situational setting) of approximately 250 to 300 words.

These three first parts constitute the first 2 hours 45 minutes of the exam, in which the candidates can freely manage the time once the 30 minutes of listening comprehension is over. The last part of the exam follows immediately afterwards:

- oral interaction: of two or three students, 10 to 15 minutes; following the initial situational embedding, the students take their allocated roles and try to reach a compromise while exploiting the information previously acquired.

CLES 3, level C1: topic linked to field of study and LSP

This competence level targets mainly students who have reached a higher level in their university studies (graduate programmes) and who have to demonstrate that they are able to communicate about topics that are linked to their field of studies or research. It therefore appeared more realistic not to separate receptive and productive skills as mentioned in section 5.2. Hence only the productive skills are evaluated explicitly. The task cannot, however, be carried out if the candidates lack the necessary receptive skills, which are, therefore, tested implicitly.

The test takers need to analyse the dossier which consists of:

- one or two authentic listening documents of 10 minutes in total;

- written documents of approximately 15 000 characters, which may also include pictures, graphs, tables, statistics, etc., taken from academic publications.

The candidates have 3 hours at their disposal to acquire the information and to prepare a support document (PowerPoint presentation or OHP slides) to accompany their oral presentation. They work on their own and manage their time independently. The students must then present their topic:

- oral production: 10 minutes of presentation (the candidate presents his / her conclusions to the evaluators) followed by 10 minutes of interaction with the evaluators;

- written production: after the speaking exam the candidate is then given 60 minutes to write a summary (or an article, an abstract, etc., depending on the situational context) of approximately 600 words, which also includes the aspects discussed with the examiners during the speaking part.

The test specifications require that a subject-area specialist approves the topic of the exam and the documents, and takes part in the examination panel.

Evaluation

As the grids in Resource document 5 show, it is not only the linguistic competence (vocabulary and grammar) that is evaluated in the exam. Whether a candidate accomplishes the task and is able to communicate are criteria that are as important as linguistic accuracy.

Furthermore, in order for the student to pass the exam and receive a certificate, each competence tested must be found to be at an acceptable level. However, no grades are given, and the result is either "pass" or "fail". In order to pass the receptive skills parts (CLES 2), a minimum of 60% of the items must be correct. In order to pass the productive skills parts of the exam, each criterion in the grids must be fulfilled. In fact, the CLES considers that all criteria are essential both at level B2 and at level C1.

Use of CLES

As the CLES is a proficiency test, it is independent of the teaching programmes carried out at language centres, and could theoretically be passed without prior participation in a teaching programme. Furthermore, preparation courses for the CLES do not exist: due to the format of the test, any work on the four skills aiming at understanding or producing authentic language prepares for the CLES.

The CLES can, however, also be used as an end-of-course exam or at the end of a degree programme. At the University of Strasbourg, for example, the students studying science have to reach level B2 by the end of their Bachelor level studies, and therefore all students have to take the CLES 2 exam at the end of their 5th semester. It is, however, not required that they pass the CLES. If the CLES is used as an exam, each part of the exam is graded according to the local exam modalities; the average is then calculated, which may allow a compensation between the results in the various skills. This may allow a student to pass the exam and get a pass grade without being eligible for a CLES certificate (see above).

The UNIcert exam

UNIcert in general[2]

UNIcert is a university language learning and testing system developed by members of the German association of university language centres (Arbeitskreis der Sprachenzentren – AKS) at the beginning of the 1990s. This system, initiated by Bernd Voss and Christine Klein-Braley and then adopted by AKS, aims to enhance language learning at university at all levels. It is a framework system which allows member institutions to develop their own teaching and testing profiles following common regulations and guidelines. Teaching and testing programmes are normally accredited by the Scientific Committee for a period of three years (or five years in case of an extensive *in situ* evaluation of the language centre).

UNIcert distinguishes four UNIcert levels: UNIcert I is oriented towards CEFR level B1, UNIcert II towards CEFR level B2, UNIcert III towards CEFR level C1 and UNIcert IV towards CEFR level C2. A preparatory level "UNIcert Basis", introduced in 2011, particularly for structurally more complex languages, is oriented towards CEFR level A2. As UNIcert focuses on university-specific language competences, the system concentrates on the higher language levels, where academic language is more likely to be encountered.

UNIcert member institutions develop a course programme for their centre that is in line with the UNIcert regulations. These courses may be electives in Bachelor's and Master's programmes or form an integral part of the studies, depending on the institutional context. These teaching and testing programmes are then checked by the UNIcert offices at the Technical University Dresden and the Ruhr University Bochum, and by members of the Scientific Committee, which accredits these programmes.

In 2011 over 50 institutions offered UNIcert teaching programmes and exams in approximately 25 different languages. The highest number of certificates is issued in English (particularly at levels II and III), followed by Spanish (mainly levels I and II) and French (levels I to III), although in recent years Swedish on the one hand and Asian languages, as well as Arabic and Turkish on the other, have encountered a considerable increase.

In 2003 a franchise system was set up in Slovakia and the Czech Republic, called UNIcert®LUCE, with its office at the University of Economics in Bratislava. Since then more than 60 teaching and testing programmes at eleven institutions have been accredited in both countries.[3]

---

2   For further information on UNIcert see: B. Voss (2010). *Handbuch UNIcert 2*. Bochum:AKS-Verlag; and the UNIcert website at: www.unicert-online.org.

3   Further information on UNIcert®LUCE is available on the Internet at: http://www.casajc.sk/unicert.htm.

The UNIcert committee has recently adopted the task-based approach to language testing, as described in the GULT guidelines, and encourages its members to introduce task-based approaches in teaching and testing, particularly at the higher language levels.

UNIcert at the University of Göttingen

At the Centre for Languages and Transferable Skills (Zentrum für Sprachen und Schlüsselqualifikationen – ZESS) of the University of Göttingen a task-based approach has been introduced for the UNIcert III exam (approximately CEFR level C1), which is a (high stakes) proficiency test at the end of a series of courses preparing the learners for level C1. In each course, however, it is not possible to run such a complex and time-consuming exam, but each skill is tested individually during and at the end of the course. While listening and reading are tested "traditionally" in an exam situation, the learners are asked to carry out fairly authentic real-life tasks for the productive skills and to put together a kind of portfolio, for example writing minutes, negotiating a contract, writing a report, presenting a solution to a case study, in oral and / or written form. This way, the learners are familiarised with a task-based approach to language testing and already have a good competence in project work and case study activities when they take the UNIcert III exam.

In other words, when introducing a task-based approach to language testing, an institution does not have to immediately do "the whole thing" and carry out a task-based exam that comprises all four skills, as described above, but, instead, it can introduce certain elements at different stages of the teaching / learning and testing programme.

As far as the UNIcert exam is then concerned, it will test all four skills. The structure of a UNIcert exam, in this case a UNIcert III exam (approximately CEFR level C1) at the University of Göttingen, is described above in section 5.3 "Timing".

Resource document 3: Examples of task-based tests and exams[1]

Test 1: CLES 2 Spanish – Level B2

As can be seen from the sample on the website (http://gult.ecml.at), the build-up tasks in the CLES provide material to be used for performing the overall tasks, namely written production and oral production, which is based on interaction between two test takers.

From this point of view and through the use of scenarios that aim at replicating real-world activities, the CLES draws heavily on the task-based performance approach whereby "the simulation of real-world tasks, and associated situational and interactional characteristics, wherein communication plays a central role (see Bachman & Palmer, 1996)." (Brown et al., 2002: 10).

As mentioned above, the receptive skills are not assessed separately as CLES 3 follows a holistic approach to testing: like in real-life situations, the test takers have to convey, discuss and evaluate relevant information from the documents provided. Receptive skills are thus needed in order to be able to accomplish the tasks focused on the productive skills.

Test 2: CLES 3 German – Level C1

For test description see above: Test 1: "CLES 2 Spanish – Level B2".

Test 3: UNIcert III Business English – Level C1

The UNIcert exam presented here was developed at the University of Göttingen and given to students who had followed a teaching programme of approximately 120 contact hours (and another 240 learning hours outside the classroom) in Business English. In this exam, the test takers are put in an authentic situation that they might encounter in real life (doing a placement at an international company) and have to develop a solution to a problem in a small team after a detailed analysis of a comprehensive set of documents. They then have to submit a written document (a report) and present their case orally (presentation of their solution, followed by a detailed discussion of their approach).

---

1   The sample exams described in the following are available at: http://gult.ecml.at

Each of the four skills is then tested separately, for which individual grades are given. For the evaluation of the productive skills the UNIcert grids available in Resource document 5 are used.

Test 4: UNIcert III French – Level C1

This UNIcert exam in French, developed by Claudie Bréhinier at ZESS at the University of Göttingen, follows the same approach as Test 3 (Business English), but was taken at the end of a UNIcert III teaching programme in Academic French by students of different university disciplines.

Test 5: GULT exam English – Level B2

This model exam was developed by the English-speaking working group at the GULT Network Meeting, which took place in January 2011 at the European Centre for Modern Languages in Graz, and is based on a first draft developed during the GULT Event at the ECML in May 2010. It aims at students at *Common European Framework of Reference for Languages (CEFR)* level B2. The individual teams at the network meeting developed different variations of this exam, focusing either on students studying environmental sciences or on students coming from a variety of academic disciplines.

Test 6: GULT exam French – Level B2

This model exam was developed by the French-speaking working group at the GULT Network Meeting, which took place in January 2011 at the European Centre for Modern Languages in Graz. It is aimed at students in engineering at CEFR level B2.

References

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

Brown, J.D., Hudson, T, Norris, J. and Bonk, W. J. (2002). *An Investigation of Second Language Task-Based Performance Assessments*. Honolulu: University of Hawaii Press.

Resource document 4: Examples of the implementation of task-based assessment in university language teaching

Example of foreign language learning and evaluation in a Master's programme at the University of Strasbourg

Various factors have over several years led the teachers to rethink the learning and evaluation setting for the students taking part in a Master's programme:

- A reduced number of hours has been allocated to these groups (16 hours per semester).

- They want to prepare these students for the start of their professional life after their studies.

- The idea is to co-operate more closely with lecturers in the individual academic disciplines.

- It is hoped to make the students autonomous in their learning of languages, starting at Bachelor level, and to avoid offering ready-made language courses.

- They recognise the influence of the CLES examination and want to organise the students' work around one single "task" they have to carry out during the semester.

A very fruitful co-operation with the lecturers in earth sciences and environmental studies has been set up, which has led to the following structure:

- Students are given the task at the beginning of the semester. The task is to prepare an oral presentation on a research topic that is directly linked to their field of study, including the analysis and exploitation of at least two publications (pseudoscientific in the first year of their Master's programme, scientific in the second year) in a foreign language (normally English). The students in their first year can carry out the task in pairs, but in their second year they have to work individually.

- The students decide themselves which topic they want to work on and which sources they want to use, but these need to be approved by an academic specialist in the specific field of studies.

- The students regularly report to the foreign language teacher and to the whole student group on their progress, which allows them to practice their oral skills (giving a presentation and oral interaction), to work on phonological and lexical aspects, and to work on their linguistic imperfections.

- Once they have analysed the documents sufficiently, the students write an abstract for their future presentation, which is then reviewed with the teacher (activity on improving their writing skills), but not evaluated.

- They then prepare their PowerPoint presentation, which is also reviewed by the language teacher (activity on the methodology of the language).

- At the end of the semester, they give their presentation in front of the group and the two teachers, who together evaluate the outcome. The presentation is followed by a discussion based on questions from the teachers and / or the group of students.

The complexity of the task they must carry out and the type of sources to be used correspond certainly to level C1 of the *Common European Framework of Reference for Languages (CEFR),* or even higher. The evaluation takes into account the students' communicative competences (operational and pragmatic competences) and the quality of the language used, but it also analyses to what extent the students have understood the documents and to what extent the scientific content presented is correct. The students are students just as in their real life, they act within the reality of their academic discipline, and the task, inspired by the traditions of the scientific community, is not only realistic, but actually real for them.

Resource document 5: Assessment grids

A. Existing grids used within the context of CLES and UNIcert

Assessment grids CEFR level B2

Speaking

Student number:

Criteria	Level B2	passed	failed
Pragmatic competence	**1. Situation:** Manages the role given in oral interaction adequately.	☐	☐
	2. Content: Uses varied and relevant arguments, based on the documents provided, maybe adding personal ideas.	☐	☐
	3. Interaction: Is able to interact. Takes his / her turn and the initiative when suitable; knows how to restart an oral exchange when necessary.	☐	☐
	4. Fluency: Is able to express his / her ideas fluently without longer pauses (some hesitations acceptable).	☐	☐
	5. Pronunciation: His / her pronunciation and intonation are sufficiently clear in order to be understood.	☐	☐

Linguistic competence	**6. Grammatical accuracy:** Masters the grammar sufficiently to avoid errors that might cause misunderstandings (some non-systematic errors tolerated).	☐	☐
	7. Coherence: Is able to use the necessary linguistic devices in order to link, accentuate and adapt his / her speech.	☐	☐
	8. Vocabulary: Uses a varied vocabulary appropriate to the task (some gaps acceptable).	☐	☐

RESULT: B2 is <u>only given</u> if each criterion has been passed	
YES ☐	NO ☐

Grid 1: CLES assessment grid speaking CEFR level B2

Exam Protocol (B2.1, B2.2)

Language / Course: _____   Surname: _____
Semester: _____   Name: _____
Teacher: _____   Student No: _____

Topic:_____

Speaking Task: Date: _____

The student is able to express him/herself orally in an appropriate manner concerning a variety of cultural and technical topics and to actively participate in discussions, in which he / she uses a complex sentence structure and technical vocabulary.

According to the specific level standards:	excellent	good	satisfactory	poor	insufficient
Content of the presentation / pragmatic competence	1	2	3	4	5
Task accomplishment: - The student identifies the problem / the task and is able to deal with it correctly. - The student shows the ability to discuss the issue. - The content is presented clearly and convincingly. - The student's own opinion and outlooks are presented in a detailed and convincing manner. - The student is able to apply general knowledge to the chosen subject.					
Correctness of content: - The design of the content matches the task. - The content is relevant and presented correctly. - Certain aspects are discussed or analysed critically. - The presented approach to the solution is reasonable and convincing. - Good examples are given.					
Logic and coherence of the presentation: - The presentation is structured clearly. - The introduction is good. - The arguments are structured and linked in a felicitous way. - The major points are emphasised and transitions are used effectively. - The conclusion and the last transition are used effectively and are well formulated and evident. - The student successfully initiates a debate.					
Evaluation of content and pragmatic competence:					

Discussion and debating of arguments	1	2	3	4	5
Ability to scrutinise the opinions of others: - The student asks critical questions about the opinion of his / her fellow students. - The student's questions are clear and precise.					
Ability to defend one's own opinion: - The student reacts adequately to the comments of his / her fellow students and answers their questions clearly and convincingly. - The student is able to defend his / her position, to specify it and to avoid misunderstandings. - The student is able to pick up a previous debate and extend it.					
Evaluation of the ability to debate:					

According to the specific level standards:	excellent	good	satisfactory	poor	insufficient
Linguistic competence	1	2	3	4	5
Phonetics / prosody / fluency: - The content is presented fluently. - The pronunciation is correct. - Intonation and stress are correct. - The volume is good.					
Grammar / morphosyntax: - Morphosyntax and syntax are correct. - Contextual constructions are used correctly. - The student shows grammatical flexibility. - The student is able to spontaneously react to questions and comments in a grammatically correct manner.					
Vocabulary: - The student makes use of a broad range of vocabulary. - The student makes use of a variety of technical terms. - Technical terminology is used correctly.					
Speech competence: - The student shows his / her ability to speak fluently and to make use of the complexity of the language (variability, diversity of useful phrases). - Modal particles and discourse particles are used competently. - Weaknesses in speaking are overcome and false starts are adequately corrected. - Pauses are used consciously and reasonably for planning, emphasising and correcting.					
Register / style: - The presentation and contributions to the discussion match the register of corresponding professional and university situations. - The style matches the corresponding professional or university situation.					
Evaluation of linguistic competence:					

Paralinguistic aspects and presentation techniques:	1	2	3	4	5
Gestures / facial expressions: - Gestures and facial expressions are used according to the target language. - Gestures and facial expressions appear natural and convincing. - Gestures and facial expressions emphasise the statements.					
Body language: - The student uses a body language that is appropriate to the situation. - The student is able to emphasise statements through his / her body language when necessary.					

Paralinguistic aspects and presentation techniques:	1	2	3	4	5
Presentation techniques: - The content is presented freely. - Eye contact with the audience is existent and supports the attention of the audience.					
Use of visual aids: - The student uses PowerPoint effectively. - The use of graphics and tables visualises the content of the presentation. - The student uses other aids convincingly. - The student engages with the audience through the use of visual aids.					
Evaluation of paralinguistic aspects / presentation techniques:					

	calculation	grade
Content of the presentation / pragmatic competence **(30%)**	_____ x 30%	
Discussion and debating of arguments **(20%)**	_____ x 20%	
Linguistic competence **(40%)**	_____ x 40%	
Paralinguistic aspects and presentation techniques **(10%)**	_____ x 10%	
Final grade for the speaking task (100%)		

Signatures: _____

Grid 2: University of Göttingen (UNIcert) assessment grid speaking CEFR level B2

Writing

<div align="center">

CERTIFICATION NATIONALE CLES
Evaluation grid for writing CLES2

</div>

Student number:

Criteria	**Level B2**	passed	failed
Pragmatic competence	1. Completes the task defined by the situation. Respects: • the form (report), • the content (neutral), • the given length.	☐	☐
	2. Uses in general a language register suitable for the context (and adequate to the respective reader).	☐	☐
	3. Uses a broad variety of relevant information presented in the documents.	☐	☐
	4. Structures and formulates the text coherently.		
Linguistic competence	5. Masters the syntax of simple sentences and is also able to use complex sentences corresponding to the completion of the task.	☐	☐
	6. Masters the grammar sufficiently to make reading the text easy (some non-systematic errors tolerated).	☐	☐
	7. Uses a varied vocabulary appropriate to the task.	☐	☐

RESULT: B2 is only given if each criterion has been passed	
YES ☐	NO ☐

<div align="center">

Grid 3: CLES assessment grid writing CEFR level B2

</div>

Exam Protocol (B2.1, B2.2)

Language / Course: _____ Semester: _____ Teacher: _____	Surname: _____ Name: _____ Student No: _____

Writing Task: Date: _____

The student is able to communicate appropriately in written form about a variety of cultural and technical topics, using a certain amount of complex sentence structures and technical subject-related vocabulary.

According to the specific level standards:	excellent	good	satisfactory	poor	insufficient
Content of the text	1	2	3	4	5
Task accomplishment (integrity of the text): - The student understands the task and is able to accomplish it. - The student shows the ability to address the question in a thematically correct manner. - The content is presented clearly and convincingly. - The student is able to write clearly and in detail about a variety of familiar topics.					
Correctness of content: - The content corresponds with the task. - The content is factually correct. - Certain aspects are challenged when necessary. - The student is able to critically analyse causes, consequences and hypothetical situations.					
Relevance of content: - The content is relevant and convincing and is enriched with appropriate examples.					
Originality: - The student wins the reader's attention with an original composition of the text.					
Evaluation of content:					

Pragmatic competence	1	2	3	4	5
Functionality: effect on the recipient: - The student is able to write different types of texts. - The student is able to write official documents such as letters, e-mails, and reports. - The student is able to write letters in which he / she describes his / her feelings sophisticatedly, stresses the personal meaning of events and experiences, and comments on the recipient's opinions and news. - The student is able to express news, opinions and feelings and to comment on those of others' in writing. - The student is able to discuss a topic in an essay or a report and to give reasons for and against a certain opinion and to calculate advantages and disadvantages of different options. - The student is able to write a short review on a film or a book.					
Organisation / coherence / layout of the text: - The text is clearly structured. - The introduction is well written. - The structure and the linkage of arguments and transitions are well done. - Main points are expressed clearly. - The student is able to weigh different ideas and solutions for a problem against each other. - The student is able to summarise information and arguments from different sources. - The student is able to write clear, detailed and coherent descriptions of real or fictitious events and experiences which are easy to read. The student makes the connection of different ideas obvious. - The ending is well written. - The student is able to reproduce information in an essay or a report and to find arguments for or against a certain point of view. - Layout and paragraphs are correct and support the organisation of the text.					
Evaluation of pragmatic competence:					

According to the specific level standards:	excellent	good	satisfactory	poor	insufficient
Linguistic competence	1	2	3	4	5
Correctness of language:					
Grammar: - The syntax is correct. - Contextual constructions are used correctly.					
Vocabulary: - The student uses a broad range of vocabulary. - The student is able to use conjunctions correctly to emphasise logical connections.					

According to the specific level standards:	excellent	good	satisfactory	poor	insufficient
Linguistic competence	1	2	3	4	5
Spelling / punctuation: - Spelling and punctuation are correct.					
Variability of linguistic devices:					
Grammar: - The student shows an extensive grammatical flexibility.					
Vocabulary: - The student is able to express opinions, feelings, and experiences in a nuanced manner. - The student is able to convey news and opinions successfully. - The student is able to write a variety of cohesive texts about personal, study or work related fields of interest.					
Register / style: - The student is able to summarise texts about personal, study or work related topics as well as technical information from different sources and media. - The student is able to write standardised official letters, using the appropriate register and considering conventional expressions.					
Evaluation of linguistic competence:					

	Calculation	Grade
Content of the text (25%)	_____ x 25%	
Pragmatic competence (25%)	_____ x 25%	
Linguistic competence (50%)	_____ x 50%	
Final grade for the writing task (100%)		

Signatures: _____

Grid 4: University of Göttingen (UNIcert) assessment grid writing CEFR level B2

Assessment grids CEFR level C1

Speaking

CERTIFICATION NATIONALE CLES
Evaluation grid of the oral exam CLES 3

Student number:

Level C1	Production		Interaction	
	yes	no	yes	no
1. Manages the task given by using one or several support documents adequately.	☐	☐		
2. Takes into account the questions of the board and reacts appropriately in the interaction.			☐	☐
3. Manages the time allocated correctly.	☐	☐		
4. Presents the information and arguments described in the documents correctly and with precision.	☐	☐		
5. Structures his / her presentation clearly and coherently.	☐	☐		
6. Speaks freely without concentrating on his / her notes; looks at the interlocutors.	☐	☐		
7. Speaks fluently, spontaneously and with ease.	☐	☐	☐	☐
8. Uses a language register suitable for the task.	☐	☐	☐	☐
9. Manages the syntax adapted to oral argumentation.	☐	☐	☐	☐
10. Maintains a high level of grammatical correctness (occasional errors).	☐	☐	☐	☐

11. Uses extensive and precise vocabulary, appropriate for the task.	☐	☐	☐	☐
12. Manifests a satisfactory competence of the phonological system of the language (pronunciation, intonation, stress).	☐	☐	☐	☐

RESULT: C1 is <u>only given if</u> each criterion has been passed	
YES ☐	NO ☐

Grid 5: CLES assessment grid speaking CEFR level C1

Exam Protocol (C1.1, C1.2)

Language / Course: _____   Surname: _____
Semester: _____   Name: _____
Teacher: _____   Student No: _____

Topic:_____

Speaking Task: Date: _____

The student possesses a high level of general and work-related language skills, which enable him / her to use different oral speaking techniques to talk effectively about the chosen topics. He / she can express him/herself fluently and adequately about complex topics relating to his / her subject by using both general and technical vocabulary. Additionally, he / she can choose and argue a certain position logically, within the scope of the context and articulating him/herself at an advanced stylistic level.

According to the specific level standards:	excellent	good	satisfactory	poor	insufficient
Content of the presentation / pragmatic competence	1	2	3	4	5
Task accomplishment: - The student identifies the problem / task and is able to deal with it correctly. - The student shows the ability of discussing the issue. - The content is presented clearly and convincingly. - The student's own opinion and outlooks are presented in a detailed and convincing manner. - The student is able to apply general knowledge to the chosen subject.					
Correctness of content: - The design of the content matches the task. - The content is relevant and presented correctly. - Certain aspects are discussed or analysed critically. - The presented approach to solution is reasonable and convincing. - Good examples are given.					
Logic and coherence of the presentation: - The presentation is structured clearly. - The introduction is good. - The arguments are structured and linked in a felicitous way. - The major points are emphasised and transitions are used effectively. - The conclusion and the last transition are used effectively and are well formulated and evident. - The student successfully initiates a debate.					
Evaluation of content and pragmatic competence:					

Discussion and debating of arguments	1	2	3	4	5
Ability to scrutinise the opinions of others: - The student asks critical questions about the opinion of his / her fellow students. - The student's questions are clear and precise.					
Ability to defend one's own opinion: - The student reacts adequately to the comments of his / her fellow students and answers their questions clearly and convincingly. - The student is able to defend his / her position, to specify it and to avoid misunderstandings. - The student is able to pick up a previous debate and extend it.					
Evaluation of the ability to debate:					

According to the specific level standards:	excellent	good	satisfactory	poor	insufficient
Linguistic competence	1	2	3	4	5
Phonetics / prosody / fluency: - The content is presented fluently. - The pronunciation is correct. - Intonation and stress are correct. - The volume is good.					
Grammar / morphosyntax: - Morphosyntax and syntax are correct. - Contextual constructions are used correctly. - The student shows grammatical flexibility. - The student is able to spontaneously react to questions and comments in a grammatically correct manner.					
Vocabulary: - The student makes use of a broad range of vocabulary. - The student makes use of a variety of technical terms. - Technical terminology is used correctly.					
Speech competence: - The student shows his / her ability to speak fluently and to make use of the complexity of the language (variability, diversity of useful phrases). - Modal particles and discourse particles are used competently. - Weaknesses in speaking are overcome and false starts are adequately corrected. - Pauses are used consciously and reasonably for planning, emphasising and correcting.					
Register / style: - The presentation and contributions to the discussion match the register of corresponding professional and university situations. - The style matches the corresponding professional or university situation.					
Evaluation of linguistic competence:					

Paralinguistic aspects and presentation techniques:	1	2	3	4	5
Gestures / facial expressions: - Gestures and facial expressions are used according to the target language. - Gestures and facial expressions appear natural and convincing. - Gestures and facial expressions emphasise the statements.					

Paralinguistic aspects and presentation techniques:	1	2	3	4	5
Body language: - The student uses a body language that is appropriate to the situation. - The student is able to emphasise statements through his / her body language when necessary.					
Presentation techniques: - The content is presented freely. - Eye contact with the audience exists and draws the attention of the audience.					
Use of visual aids: - The student uses PowerPoint effectively. - The use of graphics and tables visualises the content of the presentation. - The student uses other aids convincingly. - The student engages with the audience through the use of visual aids.					
Evaluation of paralinguistic aspects / presentation techniques:					

	calculation	grade
Content of the presentation / pragmatic competence **(30%)**	_____ x 30%	
Discussion and debating of arguments **(20%)**	_____ x 20%	
Linguistic competence **(40%)**	_____ x 40%	
Paralinguistic aspects and presentation techniques **(10%)**	_____ x 10%	
Final grade for the speaking task (100%)		

Signatures: _____

Grid 6: University of Göttingen (UNIcert) assessment grid speaking CEFR level C1

CERTIFICATION NATIONALE CLES – SESSION _____
Evaluation grid for written production CLES3

Student number:

Level C1	passed	failed
1. Manages the given task within the context of the scenario.	☐	☐
2. Respects the required length (600 words +/-10%).	☐	☐
3. Structures the text coherently (convincingly using connectors and other discourse markers).	☐	☐
4. Uses the relevant information presented in the documents.	☐	☐
5. Integrates in his / her writing the comments made during the conversation with the board.	☐	☐
6. Takes into account different points of view and weighs the arguments in order to come to a conclusion suitable for the task.	☐	☐
7. Uses a language register suitable for the task.	☐	☐
8. Manages complex sentences.	☐	☐
9. Sustains a high level of correctness (grammar, spelling etc.).	☐	☐
10. Uses extensive and precise vocabulary.	☐	☐

RESULT: C1 is only given if each criterion has been passed	
YES ☐	NO ☐

Grid 7: CLES assessment grid writing CEFR level C1

Exam Protocol (C1.1, C1.2)

Language / Course: _____	Surname: _____
Semester: _____	Name: _____
Teacher: _____	Student No: _____

<u>Writing Task</u>:   Date: _____

He / she has command of a high level of general and occupational language skills, which enable him / her to communicate about a chosen topic by using a variety of language devices. He / she can express him/herself in writing fluidly and effectively, using complex grammar structures and a broad general and technical vocabulary. Furthermore, he / she can present his / her point of view coherently, logically and in a stylistically suitable manner about a variety of complex topics relating to his / her subject or that are relevant for work and study abroad situations.

According to the specific level standards:	excellent	good	satisfactory	poor	insufficient
Content of the text	1	2	3	4	5
Task accomplishment (integrity of the text): - The student understands the task and is able to accomplish it. - The student shows the ability to address the question in a thematically correct manner. - The content is presented clearly and convincingly. - The student's own opinion and outlooks are presented in a detailed and convincing manner. - The student is able to apply general knowledge to the chosen subject. - The student is able to state complex issues clearly and in detail in written form.					
Correctness of content: - The content corresponds with the task. - The content is factually correct. - Certain aspects are challenged when necessary. - The presented approach to the solution is reasonable and convincing. - Good examples are given. - The student is able to express his / her opinion clearly and effectively in complex formal letters (e.g. to write a letter of complaint, to take a controversial stand on a problematic issue).					

According to the specific level standards:	excellent	good	satisfactory	poor	insufficient
Content of the text	1	2	3	4	5
Relevance of content: - The content is relevant and convincing - The student is able to write about complex issues in letters, essays or reports, and is able to stress those aspects that are important to him / her. - The student is able to explain and reinforce his / her point of view through additional points, motivations and meaningful examples.					
Originality / essence: - The student is able to keep the reader's attention though his / her creative presentation of the content. - The student uses a variety of relevant information in his / her text.					
Evaluation of content:					

Pragmatic competence	1	2	3	4	5
Functionality: effects on the recipient: - The student is able to write different kinds of texts. - The student is able to write official letters, e-mails and reports. - The student is able to write protocols. - The student is able to write standardised letters, in which he / she is able to convey and ask for detailed information (e.g. to reply to an advertisement, to apply for a job). - The student is able to write clearly structured and detailed descriptions or fictional texts in a reader-friendly, convincing, personal and natural way.					
Text structure / coherence / cohesion / layout: - The text is clearly structured. - The introduction is well written. - The structure, the linkage of arguments and transitions are well done. - Main points are expressed clearly and are supported by details. - The student is able to write clearly and fluently. - The student is able to present his / her point of view in detail. - The ending is well written. - The layout and paragraphs are correct and support the organisation of the text.					
Evaluation of pragmatic competence:					

Linguistic competence	1	2	3	4	5
Correctness of language:					
Grammar: - The syntax is correct. - Contextual constructions are used correctly.					
Vocabulary: - The student uses a broad range of vocabulary. - Technical terminology is used correctly. - The student is able to use conjunctions correctly to emphasise logical connections.					
Spelling / punctuation: - Spelling is correct. - Punctuation is correct.					
Variability of linguistic devices:					
Grammar: - The student shows an extensive grammatical flexibility. - The student uses complex structures.					
Vocabulary: - The student makes use of a broad range of general and technical vocabulary. - The student is able to express him/herself in a precise and sophisticated manner. - The student is able to use synonyms properly. - The student is able to formulate gradual differences in opinions and statements, e.g. concerning certainty / uncertainty, persuasion / doubt and probability. - The student is able to express opinions, feelings and experiences in a nuanced manner.					
Register / style: - The student is able to decide on a style, which is appropriate to the respective reader, by using the appropriate register. - The student is able to adjust his / her vocabulary and style to fit the context. - The student appropriately uses the register for professional and university contexts.					
Evaluation of linguistic competence:					

	Calculation	Grade
Content of the text (25%)	_____ x 25%	
Pragmatic competence (25%)	_____ x 25%	
Language competence (50%)	_____ x 50%	
Final grade for writing task (100%)		

Signatures: _____

Grid 8: University of Göttingen (UNIcert) assessment grid writing CEFR level C1

B.   Evaluation grid developed by participants at GULT Network Meeting in Graz, 13 to 14 January 2011

Evaluation grid adapted to task-based approach for written and spoken production

Context

Workshop during the GULT Network Meeting at the ECML: compare existing grids used by CLES and UNIcert and develop suitable grid for GULT guidelines.

Aspects to be assessed

After analysing the different existing models the working group has developed a three part structure of the aspects to be assessed:

1.   the efficient accomplishment of the task in a task-based perspective: 30 % of final grade

2.   interaction skills shown during performance: 30 % of final grade

3.   mastery of linguistic competence: 40 % of final grade.

Allocation of a grade

The majority of the working group participants underlined the necessity to allocate a grade for the performance shown, in addition to the general level acquired, in order to facilitate the integration of certificates into evaluation and credit point systems used in European university degree programmes (such as the ECTS).

The different categories of the grid

The following grid gives only general categories which need to be further developed according to the characteristics of the individual tasks given to the candidates.

1. Task Accomplishment
(Does the candidate interact successfully in the given context?)

1.1.　Respecting task criteria

1.2.　Interacting according to role given

1.3.　Using relevant information and arguments according to given thematic and socio-cultural context.

2. Interaction
(Does the candidate structure his / her speech according to the socially accepted patterns of interaction?)
(Please specify here descriptors depending on type of interaction defined by task.)

2.1.　Competence to manage presentation or narrative text

2.2.　Competence to structure argumentation successfully

2.3.　Competence to manage interaction linguistically.

3. Linguistic competence
(Does the candidate show the linguistic competence according to the level defined?)

3.1.　Syntax

3.2.　Morphology

3.3.　Vocabulary

3.4.　Pronunciation, prosody (oral production) / spelling (written production).

Resource document 6: Definition of "task"

According to *The Multilingual Glossary of Language Testing Terms*, originally developed by ALTE members, a task is defined as: "A combination of rubric, input and response. For example, a reading text with several multiple choice items, all of which can be responded to by referring to a single rubric." (ALTE / UCLES, 1998: 165).

GULT team members understand "task" differently. Definitions provided below (in English, French and German) best represent GULT's understanding of "task / tasks":

Tasks are real world activities "that people do in everyday life and which require language for their accomplishment" (Norris, 1998: 33).

A task is "an activity that involves individuals in using language for the purpose of achieving a particular goal or objective in a particular situation" (Bachman and Palmer, 1996: 44).

> A task is an activity promoting language learning in which meaning is primary, and through which communication must take place that has a link with real-life activities and its completion has priority, because the success of its completion is assessed by outcomes.
>
> (Wesche and Skehan, 2002)

> Throughout the paper, the notion of "task" is used in a broader sense to refer to any structural language learning endeavour which has a particular objective, appropriate content, a specified working procedure, and a range of outcomes for those who undertake the task. "Task" is therefore assumed to refer to a range of workplans which have the overall purpose of facilitating language learning – from the simple and brief exercise type to more complex and lengthy activities such as group problem-solving simulations and decision-making.
>
> (Breen, 1987: 23)

> [O]ne of a set of differentiated, sequenceable problem-posing activities involving learners and teachers in some joint selection from a range of varied cognitive and communicative procedures applied to existing and new knowledge in the collective exploration and pursuance of foreseen or emergent goals within a social milieu.
>
> (Candlin, 1987: 10)

> A task is a "workplan"; that is, it takes the form of materials for researching or teaching language. A workplan typically involves the following: (1) some input (i.e., information that learners are required to process and use); and (2) some instructions relating to what outcome the learners are supposed to achieve. As Breen (1989) has pointed out, the task-as-workplan is to be distinguished from the task-as-process (i.e., the activity that transpires when particular learners in a particular setting perform the task). As we will see, the activity predicted by the task-as-workplan may or may not accord with the

activity that arises from the task-as-process. Definitions of "task" typically relate to task-as-workplan.

(Ellis, 2000: 195)

From a psycholinguistic perspective a task is a device that guides learners to engage in certain types of information-processing that are believed to be important for effective language use and / or for language acquisition from some theoretical standpoint. This perspective is predictive, and, in some cases, deterministic. That is, it assumes that there are properties in a task that will predispose, even induce, learners to engage in certain types of language use and mental processing that are beneficial to acquisition. As Skehan, Foster and Mehnert (1998) put it "task properties have a significant impact on the nature of performance" (p.245).The claim is, therefore, that there is a close correlation between the task-as-workplan and the task-as-process because the activity that results from the task-as-workplan is predictable from the design features of the task.

(Ellis, 2000: 197-198)

Il y a « tâche » dans la mesure où l'action est le fait d'un (ou de plusieurs) sujet(s) qui y mobilise(nt) stratégiquement les compétences dont il(s) dispose(nt) en vue de parvenir à un résultat déterminé.

(Council of Europe, 2001: 15)

Als Merkmale einer *task* werden deshalb unter anderem die folgenden beschrieben: Eine *task* nennt den Zweck und das erwartete Ergebnis einer Aktivität, sie legt den Schwerpunkt auf die Bedeutung dessen, was gesagt wird, und nicht auf die Verwendung einer bestimmten Form (z. B. die Anwendung einer grammatischen Struktur), und sie versucht, die Sprache so zu verwenden, wie sie im Alltag vorkommen könnte (*real or authentic language use*).

(Müller-Hartmann and Schocker-v. Ditfurth, 2005: 2)

A task is an activity which requires learners to use language, with emphasis on meaning, to attain an objective.

(Skehan, 2003: 3)

A task is intended to result in language use that bears a resemblance, direct or indirect, to the way language is used in the real world.

(Ellis, 2003: 16)

Im TBL [*Task-Based Learning*] sieht man die Aufgabe demgegenüber weniger als Zielpunkt einer Unterrichtssequenz, sondern als deren Ausgangs- und Angelpunkt. Das befrachtet die Aufgabe mit dem Anspruch, genuine Lernsituation zu sein, in der neues Wissen und Können erworben werden können. Zuweilen erhofft man sich in den konkreten Anforderungen der Aufgabenstellung auch ein erhebliches

Motivationspotential, das bei den Schülern selbstgesteuertes Lernen auslöst, damit diese die Aufgabe zufriedenstellend bewältigen können.

(Klippel, 2006: 111)

Börner *sieht* im Aufgabenbegriff die Möglichkeit, die Implikationen eines starren Übungsbegriffs zu überwinden, der „zum Kampfbegriff für unkommunikative Sprachlerntätigkeiten mutierte" (1999, 212) und an dessen Stelle nunmehr eine differenzierter Aufgabenbegriff treten zu lassen, der sowohl die lehrerseitige Initiierung sprachlicher Aktivitäten des Lerners umfasse, die als problemhaft antizipiert werden, sowie die Betrachtung lernerseitiger sprachlicher und kognitiver Handlungen (Lösen einer Aufgabe) als auch das Produkt eben dieser Handlungen (Lösung einer Aufgabe).

(Königs, 2006: 115)

Est définie comme tâche toute visée actionnelle que l'acteur se représente comme devant parvenir à un résultat donné en fonction d'un problème à résoudre, d'une obligation à remplir, d'un but qu'on s'est fixé. Il peut s'agir tout aussi bien, suivant cette définition, de déplacer une armoire, d'écrire un livre, d'emporter la décision dans la négociation d'un contrat, de faire une partie de cartes, de commander un repas dans un restaurant, de traduire un texte en langue étrangère ou de préparer en groupe un journal de classe.

(Council of Europe, 2001: 16)

Aufgaben bedeuten für die Lehrenden mehr und anderes als für die Lernenden, *Aufgaben* sind mehr und etwas anderes als *Aufgaben-Formulierungen.* Aufgaben-Formulierungen sind für die Lernenden da. Sie verlangen obligatorisch eine Fragestellung bzw. Aufforderung (die eigentliche Aufgabe), Angaben zu Themen und Materialien, die benutzt werden sollen, Angaben zu erwarteten Aktivitäten der Lernenden und zu den Resultaten, die aus diesen erwachsen sollen, und optimalerweise enthalten sie auch Angaben, die die Aufmerksamkeit auf Sprache sowie die Fähigkeit zur Selbststeuerung und Selbstbeurteilung der Lernenden fördern. Ihr Zweck ist erfüllt, wenn sie ihre Adressaten richtig informieren und orientieren.

(Portmann-Tselikas, 2006: 185)

A *task* is defined as any purposeful action considered by an individual as necessary in order to achieve a given result in the context of a problem to be solved, an obligation to fulfil or an objective to be achieved. This definition would cover a wide range of actions such as moving a wardrobe, writing a book, obtaining certain conditions in the negotiation of a contract, playing a game of cards, ordering a meal in a restaurant, translating a foreign language text or preparing a class newspaper through group work.

(Council of Europe, 2001: 10)

References

Bachmann, Lyle F. and Palmer, Adrian S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Breen, M. (1987). "Learner contributions to task design". In: Candlin, C. and Murphy D. (eds). *Language Learning Tasks*. Englewood Cliffs NJ: Prentice-Hall, 23-46.

Candlin, C.N. (1987). "Towards task-based learning". In: Candlin, C. N. and Murphy, D. (eds). *Lancaster Practical Papers in English Language Education. Vol. 7. Language Learning Tasks*. Englewood Cliffs, NJ: Prentice Hall, 5-22.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe / Cambridge: Cambridge University Press.

Ellis, R. (2000). "Task-based research and language pedagogy". In: *Language Teaching Research* 4(3), 193-220.

Ellis, R. (2003). *Task Based Language Learning and Teaching*. Oxford: Oxford University Press.

Ellis, R. (2009). "Task-based language teaching: sorting out the misunderstandings". In: *International Journal of Applied Linguistics* 19, 221–246.

Klippel, F. (2006). "Aufgabenorientierung im Fremdsprachenunterricht. Wirklich ein neues Paradigma?". In: Bausch, K.-R., Burwitz-Melzer, E., Königs, F. G. and Krumm, H.-J. (eds). *Aufgabenorientierung als Aufgabe. Arbeitspapiere der 26. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts*. Tübingen: Gunter Narr Verlag, 109-115.

Königs, F. G. (2006). "Aufgabenorientierung als Aufgabe. Überlegungen zur aktuellen Diskussion um ein 'neues' Konzept für den Fremdsprachenunterricht". In: Bausch, K.-R., Burwitz-Melzer, E., Königs, F. G. and Krumm, H.-J. (eds). *Aufgabenorientierung als Aufgabe. Arbeitspapiere der 26. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts*. Tübingen: Gunter Narr Verlag, 115-122.

Müller-Hartmann, A. and Schocker-v. Ditfurth, M. (eds.) (2005). *Aufgabenorientierung im Fremdsprachenunterricht. Task-Based Language Learning and Teaching*. Tübingen: Narr.

Norris, J. M. et al. (1998). *Designing Second Language Performance Assessments*. Honululu: University of Hawaii Press.

ALTE (1998). *Multilingual Glossary of Language Testing Terms: Studies in Language Testing 6*. Cambridge: Cambridge University Press.

Portmann-Tselikas, P. R. (2006). "Was müssen Lehrende über Aufgaben wissen?" In: Bausch, K.-R., Burwitz-Melzer, E., Königs, F. G. and Krumm, H.-J. (eds) *Aufgabenorientierung als Aufgabe. Arbeitspapiere der 26. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts.* Tübingen: Gunter Narr Verlag, 182-191.

Skehan, P. (2003). "Task-based instruction". In: *Language Teaching* 36, 1-14.

Wesche, M. B. and Skehan, P. (2002). "Communicative, task-based, and content-based language instruction". In: Kaplan, R. B. (ed.). *The Oxford handbook of applied linguistics.* Oxford: Oxford University Press, 207-228.

Resource document 7: Definition of "authenticity"

"Authenticity"

Authenticity: the degree to which test materials and test conditions succeed in replicating those in the target use situation.

(McNamara, 2000: 131)

On a pu affirmer que l'authenticité est la résultante de l'interaction entre le lecteur et le texte et pas seulement une caractéristique du texte […]. L'identité du lecteur, son projet de lecture, l'intention du scripteur et le degré de proximité sociale et culturelle entre le lecteur et le texte ont une incidence sur la nature de l'interaction entre un lecteur et un texte donné. (*Evaluation de compétences en langues et conception de tests*, 2002: 24)

L'authenticité interactionnelle peut se définir comme l'interaction entre l'activité d'évaluation (la tâche) et le candidat; elle suppose que les rédacteurs et les concepteurs de tests devraient:

- proposer des textes, des situations et des tâches qui simulent la "vraie vie" sans essayer de la reproduire à l'identique;
- essayer de proposer des situations et des tâches qui ont des chances d'être pertinentes pour le candidat potentiel à un niveau donné;
- clarifier la *finalité* de chaque tâche ainsi que le *public cible* en mettant en contexte adéquat;
- expliciter les *critères de réussite* de la tâche.

(Evaluation de compétences en langues et conception de tests, 2002: 24)

"Authentic materials"

There are a number of definitions related to authentic materials and three are presented here:

An authentic text is a stretch of real language, produced by a real speaker or writer for a real audience and designed to convey a real message of some sort. (Morrow 1977: 13)

Authentic texts (either written or spoken) are those which are designed for native speakers: they are real texts designed not for language students, but for the speakers of the language in question. (Harmer 1983: 146)

A rule of thumb for authentic here is any material which has not been specifically produced for the purposes of language teaching. (Nunan 1989: 54)

(Teaching Functional English Through Authentic Materials, 2011)

The definitions of authentic materials are slightly different in literature. What is common in these definitions is "exposure to real language and its use in its own community". Rogers (1988) defines it as "appropriate" and "quality" in terms of goals, objectives, learner needs and interest and "natural" in terms of real life and meaningful communication (p. 467). Harmer (1991), cited in Matsuta (n.d., para. 1) defines authentic texts as materials which are designed for native speakers; they are real text; designed not for language students, but for the speakers of the language. Jordan (1997, p. 113) refers to authentic texts as texts that are not written for language teaching purposes. Authentic material is significant since it increases students' motivation for learning, makes the learner be exposed to the "real" language as discussed by Guariento & Morley (2001, p. 347). The main advantages of using authentic materials are (Philips and Shettlesworth 1978; Clarke 1989; Peacock 1997, cited in Richards 2001):

1.   They have a positive effect on learner motivation.

2.   They provide authentic cultural information.

3.   They provide exposure to real language.

4.   They relate more closely to learners' needs.

They support a more creative approach to teaching.

We can claim that learners are being exposed to real language and they feel that they are learning the "real" language. These are what make us excited and willing to use authentic materials in our classrooms, but while using them, it is inevitable that we face some problems.

(Kilickaya, 2004)

Peacock (1977) says that authentic materials are "materials that have been produced to fulfil some social purpose in the language community".

Widdowson's (1990) differentiation of the terms "authentic" and "genuine material" should be mentioned here:

Authentic would be material designed for native speakers of English used in the classroom in a way similar to the one it was designed for. For example, a radio news report brought into the class so students discuss the report on pollution in the city where learners live. Most of the time, though, this material is used in a *genuine* way, in other words, not in the way it was intended, but in a somewhat artificial way. For example, news articles where the paragraphs are cut up and jumbled so students have to put them back together in the correct order.

"Authenticity of text and task"

Authentic texts or materials have been defined by Villegas Rogers and Medley (1988) as "[...] language samples – both oral and written – that reflect a naturalness of form, and an

appropriateness of cultural and situational context that would be found in the language as used by native speakers" (p. 468). Texts that are prepared *for* native speakers *by* native speakers reflect the culture and societal values of everyday life. "No textbook culture note on the Hispanic family, for example, can replace the study of authentic birth or christening, wedding and death announcements, where, under the observable linguistic conventions, lie the rituals of events, the connotations of rites of passage, the meaning of 'family', and the dynamic nature of culture"

(Galloway & Labarca, 1990, p. 139).

For our purposes, any text that is *purposeful*, *meaningful*, and has a real *communicative intent* for a real audience can be considered to be authentic. In other words, it is authentic in the sense that it was not originally produced for language-teaching purposes but rather for the purpose of communicating meaning (Brinton et al., 1989, p. 17). This means that an e-mail message sent via the Internet by a student of German to another student of German is "authentic" as long as the message is meaningful (even though the message was not written by a native speaker for another native speaker). Furthermore, authenticity in a deeper sense does not reside in the text itself but rather is determined by how that text is used (Hutchinson & Waters, 1987), i.e., the authenticity of the task. For example, if a teacher uses an article from a target culture magazine for the sole purpose of having the students underline all of the instances in which the subjunctive appears, the authenticity of the task disappears.

Let's examine a task and consider ways in which it can be slightly altered to become more authentic. Imagine that students are engaged in a unit on Costa Rica (or any other target country). As a culminating activity at the end of the unit, the teacher decides to have students create travel brochures in the target language to demonstrate their knowledge of what they have learned. Such a task asks that the students pretend to act as native speakers, which they clearly are not. Kramsch (1993) would argue that authenticity involves having students be who they are prelearners of the target language. To revise the task somewhat with an eye toward greater authenticity, the teacher can have students create travel itineraries for a group of students who will be traveling to Costa Rica, the intent being to demonstrate their knowledge of what they have learned by communicating it to other students.

Another example would involve having students at the beginning of the unit write letters in the target language to various travel agencies, tourist bureaus, and "Chamber of Commerce" equivalents to indicate that they (1) are students of Spanish, (2) are studying about Costa Rica, and (3) are interested in receiving travel information in Spanish. Such a task has a real purpose and a real audience. The added benefit is that it will also lead to additional authentic materials for classroom use!

A final example of an authentic task for this instructional setting is to have students write to Costa Rican students about Minnesota (i.e., their home state), given what they have learned about Costa Rica. A letter written for this task might include, for example, a comparison between Minnesota's Boundary Waters and Costa Rica's Tortuguero National Park in terms of their environmental restrictions.

These suggestions highlight the importance of creating tasks that involve students in using language for real communicative purposes and for real audiences. For example, a

teacher might have high school students write children's stories that are then shared with an elementary language program in the same district. It is important to note, however, that it is not possible to make every task or text authentic in the language classroom. Sometimes students need to pretend to be native speakers for a role play; sometimes they need to write for a hypothetical audience; sometimes they need to read a text that has been adapted for nonnative speakers of the language. Such activities are valuable and certainly have a place in the language curriculum. What is important (and possible!), however, is for teachers to find a good balance in their curriculum between tasks and texts that are less authentic and those that represent the principles of authenticity as described above. Teachers should also make sure that some of the texts they use in the curriculum contain language as used by native speakers so as to incorporate cultural and linguistic authenticity. A number of authentic texts (i.e., written by native speakers for native speakers of the target language) are used in CoBaLTT lessons / units found at the Web Resource Center.

(Tedick, 2003)

"Authentic assessment"

Authentic assessment: a form of assessment in which students are asked to perform real-world tasks that demonstrate meaningful application of essential knowledge and skills.

(Mueller, 2011)

Authentic assessment aims to evaluate students' abilities in "real-world" contexts. In other words, students learn how to apply their skills to authentic tasks and projects. Authentic assessment does not encourage rote learning and passive test-taking. Instead, it focuses on students' analytical skills; ability to integrate what they learn; creativity; ability to work collaboratively; and written and oral expression skills. It values the learning process as much as the finished product.

In authentic assessment, students:

- do science experiments

- conduct social-science research

- write stories and reports

- read and interpret literature

- solve math problems that have real-world applications.

(TeacherVision, 2011)

References

Evaluation de compétences en langues et conception de tests (2002). Edited by members of ALTE. Available at

http://www.coe.int/t/DG4/Portfolio/documents/Guide%20pour%20l'Evaluation%20-%20rev%20octobre%202002.doc [accessed 9 September 2011].

Kilickaya, F. (2004). *Authentic Materials and Cultural Content in EFL Classrooms*, from: http://iteslj.org/Techniques/Kilickaya-AutenticMaterial.html [accessed 9 September 2011].

McNamara, T. (2000). *Language Testing*. Oxford et al.: Oxford University Press.

Mueller, J. (2100) *What is Authentic Assessment?*, from: http://jonathan.mueller.faculty.noctrl.edu/toolbox/whatisit.htm [accessed 9 September 2011].

Peacock, M. (1997) "The Effect of Authentic Materials on the Motivation of EFL Learners". In: *English Language Teaching Journal* 51, pp. 2, cited in: http://www3.telus.net/linguisticsissues/authenticmaterials.html [accessed 9 September 2011].

Teaching Functional English Through Authentic Materials (2011), from: *http://www.teachingstylesonline.com/teaching_functional_english_through_authentic_materials.html* [accessed 9 September 2011].

TeacherVision® (2011). *Authentic Assessment Overview*. From: http://www.teachervision.fen.com/teaching-methods-and-management/educational-testing/4911.html#ixzz1ERN2vV00, [accessed 9 September 2011].

Tedick, D. J. (2003). *CAPRII: Key Concepts to Support Standards-Based and Content-Based Second Language Instruction*. Minneapolis, MN: University of Minnesota, Center for Advanced Research on Language Acquisition (CARLA). The original publication can be found on the CARLA website at : http://www.carla.umn.edu/cobaltt/modules/strategies/CAPRII/READING1/caprii.htm. [accessed 19 September 2011]. Used with permission.

Widdowson, H. (1990), *Aspects of Language Teaching*. Oxford: Oxford University Press, cited in: http://www3.telus.net/linguisticsissues/authenticmaterials.html [accessed 9 September 2011].

Further reading:

In his article "Authentic Assessment: Implications for EFL Performance Testing in Korea", published in 2002 in *Secondary Education Research* 49, 89-122, Andrew Finch offers a thorough description of authentic assessment and compares it to standardised testing. Further information is available at: http://www.finchpark.com/arts/Authentic_Assessment_Implications.pdf

See also:

Wiggins, G. (1990). The case for authentic assessment. In: *Practical Assessment, Research & Evaluation*, 2(2). Available at: <u>http://PAREonline.net/getvn.asp?v=2&n=2</u>

Resource document 8: Steps in developing a task-based language test

When preparing your own task-based language test, the following procedure can be helpful:

Developing the scenario:

1. Think of your candidates. What is their level of competence and what is their field of study? Have they any special interests?

2. Think of an authentic situation in which they would need their foreign language skills in their domain (in their studies or their future profession). What situation could serve as the framework for your test?

3. Think of a problem or a project they should be able to handle in the target language in their studies or future profession. What overall task should they carry out to come up with a specific outcome which serves the specific need in the given situation?

4. Think of the individual spoken and written outputs they have to produce in order to fulfil the overall task. What do they have to prepare in writing and to present orally?

Collecting materials:

5. Now collect the required materials for listening and reading: Start with searching for listening material, as this will probably be more time-consuming and more complicated.

Test writing: Now prepare your test!

6. Describe the situation.

7. Develop a (realistic) overall task (general aim).

8. Decide on how you want to test the receptive skills, choose the (oral and written) texts and develop the task sheets.

9. Prepare an answer key for the receptive part of the test.

10. Think of a logical link between the receptive part and the productive part of the test.

11. For the productive part of the test, decide whether the test takers have to work individually or (partially or entirely) in pairs or groups.

12. For the productive part of the test, develop build-up tasks for the individual skills, and clearly define the expected outcomes (both for speaking and writing).

13. Define the roles of the individual test takers / team members and their respective tasks and expected outcomes.

14. Then define what outcomes you would expect and adapt your evaluation grids. In this context, decide on the minimum requirements for passing the test.

15. Think of the logical order of the individual tasks. Put them together according to an authentic situation in real life (e.g. first prepare a written report and then present it orally to an audience, for example the team).

16. While adapting the test to the needs of your institution pay attention to the respective weight of the various build-up tasks (productive and receptive skills) in the assessment of the overall task.

GULT test template

Title:	
CEFR level:	
Scenario (problem):	
Problem / simulation / project:	
Background:	
Situation:	
Role(s):	
Tasks:	
Overall task:	
Build-up tasks:	
Listening task / product:	
Reading task / product:	
Writing task / product:	
Speaking task / product:	
Listening text(s):	
Reading text(s):	
Documentation material:	
Answer key for evaluators:	
Grids for evaluators:	
Notes for evaluators:	

Checklist for developing a GULT test – version 1

1.	Is there an overall task?	☐ yes	☐ no
2.	Is the test problem-based or simulation-/project-based? The problem is: The simulation / project is:		
3.	Is this task embedded into a specific situation? Situation:	☐ yes	☐ no
4.	Are the roles of the test takers clearly defined? Roles:	☐ yes	☐ no

5.	Are there build-up tasks?
	
	Please specify: a)
	                     b)
	                     c)
	                     d)
	                     e)
	                     f)
		☐ yes	☐ no

6.	In what respect is the overall task authentic? and the build-up tasks? and the problem or the simulation / project? and the situation? and the testing materials?		

7.	Are the problem / simulation / project, the situation, the tasks and the roles relevant to the test taker and his or her interests?		
		☐ yes	☐ no
8.	Is the test interesting and stimulating?	☐ yes	☐ no
9.	How far are the learners already familiar with the topic and the content of the materials?		
10.	In what respect does the topic allow for a controversial discussion?		
11.	What is the complexity of the topic?		
12.	Are the test takers' roles clear?	☐ yes	☐ no
13.	Are the individual steps the test takers have to take in completing the test presented in a clear, structured way?	☐ yes	☐ no
14.	Is there enough material available?	☐ yes	☐ no
	for each individual topic?	☐ yes	☐ no
15.	Which products have to be developed and presented? written products: oral products:		
16.	Is it possible to manage the task without using the materials provided (e.g. the test takers using their own background knowledge only)?	☐ yes	☐ no

Checklist for developing a GULT test – version 2

	Yes	No
1. Are the learners already familiar with a task-based approach?		
2. Is there an overall task?		
3. Is this task embedded in a specific situation?		
4. Are the roles of the test takers clearly defined?		
5. Are there appropriate build-up tasks?		
6. Is the overall task sufficiently realistic?		
7. Are the build-up tasks sufficiently realistic?		
8. Is the problem / simulation / project sufficiently realistic?		
9. Are the input (testing) materials authentic?		
10. Is the problem / simulation / project relevant to the test takers' needs and interests?		
11. Are the situation and tasks relevant to the test takers' needs and interests?		
12. Does the topic allow for argumentation?		
13. Is the complexity of the topic at an appropriate level for the learners?		
14. Are the test takers' roles both clear and relevant to the test takers' needs and interests?		
15. Are the instructions clear for every component of the test?		
16. Is there enough material for test takers to complete the task?		
17. Do the test takers need to use the material provided to complete the task?		

Resource document 9: Glossary[1]

Achievement tests: show how much students have learned from the language material they have been taught. Therefore, achievement tests are based on, for example, the course materials and very little else, and they are typically final tests in language courses. They are also called "attainment tests".

Analytic rating: the rating of each aspect of a performance separately, as in writing, grammar, organization, content, et**c**. (McNamara, 2000: 131).

Analytic scoring: method of scoring which can be used in tests of productive language use, such as speaking and writing. The assessor makes an assessment with the aid of a list of specific points. For example, in a test of writing the analytic scale may include a focus on grammar, vocabulary, use of linking devices, etc. (ALTE, 1998: 135).

A posteriori validation: takes place after the test has been administered, using all kinds of measures and analyses.

A priori validation: is done after the test has been constructed but before it is used. It is based on studying and weighing various testing theories and using previous research on similar tests and testing situations.

Aptitude tests: are designed to measure capability or potential, whether a candidate is able to succeed in an academic programme, to learn a foreign language, to acquire a specific vocation, or some other capability (Testing Glossary, 2011).

Attainment test: see "achievement test".

Assessment for learning: see "formative assessment".

Authentic tests: are the extreme forms of communicative testing and should consist of authentic situations of foreign language use to be assessed as performance tests. Such a test might be to send students to a library where they would have to obtain a certain book using the target language. The tester would video the exchanges between the students and the librarian and then assess the students' speech.

Authentic text: text used in a test which consists of materials originally produced for a non-language testing purpose, and not specially produced for the test (ALTE, 1998: 135).

Authenticity: the degree to which test materials and test conditions succeed in replicating those in the target use situation (McNamara, 2000: 131).

Cloze tests: a type of gap-filling task in which whole words are deleted from a text. In a traditional cloze, deletion is every nth word. Other gap-filling tasks where

---

1   If no reference is given, the entry was prepared by the project team.

short phrases are deleted from a text, or where the item writer chooses the words to be deleted, are commonly referred to as cloze tests, for example "rational cloze". Candidates may have to supply the missing words (open cloze), or choose from a set of options (multiple-choice or banked cloze). Marking of open cloze may be either "exact word" (only the word deleted from the original text is taken as the correct response) or "acceptable word" (a list of acceptable responses is given to markers) (ALTE, 1998: 138).

Communicative competence**:** the ability to use language appropriately in a variety of situations and settings (ALTE, 1998: 138).

Communicative method / communicative tests: are tests in which the test takers are really communicating in the target language, conveying and exchanging relevant information.

Concurrent validity: A test is said to have concurrent validity if the scores it gives correlate highly with a recognized external criterion which measures the same area of knowledge or ability (ALTE, 1998: 139).

Consequential validity: the way in which the implementation of a test can affect the interpretability of test scores the practical consequences of the introduction of a test (McNamara, 2000: 132).

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 the different constructs which underlie language ability. In addition to language ability itself, motivation, attitude and acculturation are all relevant constructs (ALTE, 1998: 139).

Construct validity: 1) a test if scores can be shown to reflect a 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 scores on a test of grammatical competence (ALTE, 1998: 139). 2) It concerns the theoretical framework underlying the test, that is, whether the test is really that of a construct, such as LSP (languages for specific purposes) reading comprehension.

Content validity: shows whether the test has measured what it was supposed to measure, for example, that the content of an achievement test does not extend beyond what was taught before the test.

Criterion: 1) is the domain of behaviour relevant to test design; or 2) is an aspect of performance which is evaluated in test scoring, for example fluency, accuracy, etc.

Criterion-referenced evaluation: means that there are certain criteria that the test taker must fulfil in order to pass the test or get a certain grade. Most university language tests are criterion-referenced, which means that it is known beforehand

what is needed for achieving, for example, full marks, but it is not known how many students will be getting which grades or failing the test.

Criterion-referenced tests: test[s] in which the candidate's performance is interpreted in relation to predetermined criteria. Emphasis is on attainment of objectives rather than on candidates' scores as a reflection of their ranking within the group (ALTE, 1998: 140).

Cut point / cut score / cut-off point / cut-off score: is a score which marks the border between the passing scores and the failing scores.

Diagnostic test: test which is used for the purpose of discovering a learner's specific strengths or weaknesses. The results may be used in making decisions on future training, learning or teaching (ALTE, 1998: 142).

Difficulty index: in classical test theory, relates to the proportion (p) of candidates responding to an item correctly. This means that the difficulty index is sample-dependent, and changes according to the ability level of the candidates (ALTE, 1998: 142). The difficulty index (marked "p" in tables) is calculated by dividing the number of correct answers by the number of test takers. The difficulty index varies from 0 to 1. The larger the number is, the easier the item. It is still customary to approve of higher difficulty indices for the very first and last items so that the examinees are not terrified by the test when they start, and they finish the test feeling that they were able to contribute. However, the test designers must in advance know the target population well enough to have some idea of how difficult the test items will be.

Direct tests: test[s] which measure the productive skills of speaking or writing, in which performance of the skill itself is directly measured. An example is testing writing ability by requiring a candidate to write a letter (ALTE, 1998: 142).

Discrete-item / discrete-point tests: have sections that focus on clearly identifiable linguistic phenomena and are totally separate from one another, with their own sets of points, and the score is the sum of the points given for the various items.

Discrimination index: is a tool to help the test designers find out whether the test items really discriminate between the good and the bad language learners, even if the test items have the accepted number of correct answers. In other words, if the predicted difficulty index is .5 (= half the test population has got the item right), the item is of poor quality if those who are not good at the language had the right answers while the good language learners have answered the item incorrectly. The discrimination index ("D" in tables) can be calculated using a simple formula. The tested population is divided into a good half and a poor half according to their total raw scores or other information of their language skills. The formula is as follows: (the number of correct answers in the better half minus the number of correct answers in the poorer half) divided by the total number of correct answers. The D varies between +1 (all the correct answers in the better half) to -1 (all the correct answers in the poorer half).

Domain: the defined area of content and / or ability which is to be tested by a specific task or component of an examination (ALTE, 1998: 143).

External validity: depends on the representability of the test and the generalisability of the test results**.**

Face validity: the extent to which a test appears to candidates, or those choosing it on behalf of candidates, to be an acceptable measure of the ability they wish to measure. This is a subjective judgment rather than one based on any objective analysis of the test, and face validity is often considered not to be a true form of validity. It is sometimes referred to as "test appeal" (ALTE, 1998: 145).

Formative assessment: testing which takes place during, rather than at the end of, a course or programme of instruction. The results may enable the teacher to give remedial help at an early stage, or change the emphasis of a course if required. Results may also help a student to identify and focus on areas of weakness (ALTE, 1998: 146). It is also called "assessment for learning".

High stakes tests: are very important examinations the outcome of which is the basis of far-reaching decisions such as admission to a university. Also, high stakes tests can usually be taken only a set number of times and a fee may be required.

Indirect tests (tasks): test[s] or task[s] which attempt to measure the abilities underlying a language skill, rather than testing performance of the skill itself. An example is testing writing ability by requiring the candidate to mark structures used incorrectly in a test (ALTE, 1998: 147).

Indirect testing: when taken literally, covers all testing of receptive skills because reading comprehension or listening comprehension cannot be tested as such — the teacher has to resort to speaking or writing to find out what has been understood in a reading or listening test. A more concrete example of an indirect test is when one half of a discussion is given on paper to the test taker, who has to provide the other half in writing. In other words, speaking skills are tested through writing.

Integrative items / tasks: used to refer to items or tasks which require more than one skill or subskill for their completion. Examples are the items in a cloze test, an oral interview, reading a letter and writing a response to it (ALTE, 1998: 148).

Integrative tests: may test several skills at the same time or language in general without an emphasis on, for example, grammar or vocabulary. A cloze test is an integrative test because in order to fill a gap, the test taker must understand both the contextual and syntactic cues.

Inter-rater reliability: is calculated when there are several raters grading the same test, and the testers need to know that all the raters rate the tests in the same way.

Intra-rater reliability: an estimate of the reliability of assessment, based on the degree to which the same assessor scores the same performance similarly on different occasions (ALTE, 1998: 149).

Internal validity: refers to the content of the test and, specifically, to the fact that the test itself does not contain any irrelevant factors that would disturb the test taker.

Item analysis: occurs when the test has been administered and graded and it is time to take a good look at the test again. Every teacher-tester should complete a simple item analysis, which consists of calculating the difficulty and discrimination indices.

Item difficulty: concerns the proportion of correct responses to total responses on a test item; for example, if 20 out of 30 students get an item right, the item difficulty is 66% (20/30), and the difficulty index (p) is 0.66.

Language ability: the competences which together define an individual's capacity to use language for a variety of communicative purposes (ALTE, 1998: 150).

Level descriptors: in rating scales, statements describing the character of a minimally acceptable performance at a given level (McNamara, 2000: 134).

Low stakes tests: are minor examinations which can be retaken or, if a series of tests is taken, poor results in one test can be compensated for with good results in another test. Typical low stakes tests are quizzes or shorter language tests.

Multiple-choice format: is a format for test questions in which candidates have to choose from a number of presented alternatives, only one of which is correct (McNamara, 2000: 135).

Multiple-choice item: a type of test item which consists of a question or incomplete sentence (stem), with a choice of answers or ways of completing the sentence (options). The candidate's task is to choose the correct option (key) from a set of three, four or five possibilities, and no production of language is involved. For this reason, multiple choice items are normally used in tests of reading and listening. They may be discrete or text-based (ALTE, 1998: 153).

Norm: 1) an empirically derived distribution of scores on a test, which provides reference data for appropriate groups of examinees, e.g., students' results on the Test of English as a Foreign Language (TOEFL) are reported with reference to norms so that students can see where they stand in comparison with the general population of foreign students (Testing Glossary, 2011); 2) a standard of performance. In a standardized test the norm is determined by recording the scores of a large group. The norm or standards based on the performance of that group are used in assessing the performance of subsequent groups of candidates of a similar type (ALTE, 1998: 154).

Norm-referenced evaluation / measurement: means that grades that are determined by the previous experiences of the tested population, and the testees receive

their grades according to the percentiles they are placed in. For example, it is already known what percentage of the test takers will get the best grade or fail even before the first test is graded. This means that the raw score used as the cut-off point may differ from test to test but not so much that it would make the test unreliable. Norm-referenced testing is not appropriate in assessing small populations.

Norm-referenced tests: test[s] where scores are interpreted with reference to the performance of a given group, consisting of people comparable to the individuals taking the test. The term tends to be used of tests whose interpretation focuses on ranking individuals relative to the norm group or to each other (ALTE, 1998: 155).

Objective / objectively scored tests: are typically multiple-choice or true-false tests with a key of correct answers, normally one, which every rater uses in the same way.
(Note: There are no truly objective language tests because every examination is based on some test maker's idea of what language material is important and what competences are needed to complete a language examination successfully.)

Pass / fail tests: often occur in university instruction, where many tests are not graded on a scale but are either passed or failed. In a language examination, the cut-off point and the borders between grades can be decided on the basis of either a norm-referenced or criterion-referenced evaluation.

Performance tests: are not based on any definite body of language but they show how well the testee can perform in certain situations. Students know what to study for an achievement test while performance tests may cover all kinds of competences, not taking into account the testee's previous language studies. Many internationally available commercial language tests are performance tests.

Power test: test which allow sufficient time for almost all candidates to finish it, but contain some tasks or items of a degree of difficulty which makes it unlikely that the majority of candidates will get every item correct (ALTE, 1998: 157).
(Note: It is quite clear that no pure speed or power tests exist in university foreign language instruction, but when language tests are being planned, test designers must have a very clear idea of how much time will be allocated for each item and for the whole test. One has to depend on estimated averages: there will never be an occasion when all students will need the same amount of time to complete a test.)

Predictive validity: means that the test can be used to predict the examinee's future success in using a language and / or learning that or some other language.

Productive skills: are speaking and writing.

Proficiency tests: test[s] which measure general ability or skill, without reference to any specific course of study or set of materials (ALTE, 1998: 158).

Prognostic tests: are language examinations which can be used to predict a student's future success in language studies, and these are prognostic tests.

Raw scores: are the total scores before any manipulation or adjustment.

Receptive skills: are listening and reading.

Reliability: the consistency or stability of the 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 (ALTE, 1998: 160).

Scoring: is the decision as to how many points are granted for each item, task and section and what is required for receiving full credit. The most important point in any test score is the cut-off point, in other words the point where the line is drawn between those who pass and those who fail the test. The optimal cut-off point is based not only on tester intuition but also on research.

Speeded test: test with a time limit for completion. Slower candidates achieve a lower score because they do not reach the final questions. Normally in a speeded test the difficulty of questions is such that candidates would generally respond correctly, were it not for the time constraint. Also known as a speed test (ALTE, 1998: 164).

Standardisation: the process of ensuring that assessors adhere to an agreed procedure and apply rating scales in an appropriate way (ALTE, 1998: 165).

Standardised tests: a measure that has been piloted (usually on a large sample, representing different types of respondents) and for which interpretive data, such as norms, reliability, and validity coefficients, have been provided; has been administered to a large group of examinees from a target population, often more than 1 000 persons, and has been analyzed and normed for use with other samples from that population (Testing Glossary, 2011).

Subjective tests / subjectively scored tests: are tests where the rater's personality and personal preferences play a more important role because there is no key, and written instructions can often be interpreted in many ways. That is why in high stakes examinations, raters have training sessions before they start grading tests.

Summative evaluation: evaluation that comes at the conclusion of an educational programme or instructional sequence (Testing Glossary, 2011).

Task / tasks: 1) are "an activity which requires learners to use language, with emphasis on meaning, to attain an objective" (Skehan, 2003: 3); or 2) are "an activity promoting language learning in which meaning is primary, and through which communication must take place …; that has a link with real-life activities; and its completion has priority, i.a. because the success of its completion is assessed

by outcome(s)" (Weideman, 2006: 87); or 3) "are activities which have meaning as their primary focus. Success in tasks is evaluated in terms of achievement of an outcome, and tasks generally bear some resemblance to real-life language use. So task-based instruction takes a fairly strong view of communicative language teaching" (Skehan, 1996: 20).

Task-based / action-based approaches to language testing: are where the examiner gives a task for completion to the test taker, who is stimulated by this task to act. This means that in a task-based approach an overall task describes the general activity that needs to be carried out, with a specific result to be expected at the end. In order to achieve this outcome, the test taker has to develop specific products, for example a written report, explaining the problem in question and possible solutions to it, and an oral presentation of these solutions in a specific context (situation and setting). All these activities and products are interrelated. To help the test taker to achieve these results he or she has to carry out specific build-up tasks.

Task-based language learning (and testing): follows a communicative approach, but goes beyond it. The aim is no longer to carry out a role play in which learners can show what they have learned, but to embed it into a specific context and situation, with a specific aim. The candidates will not be judged on whether the phrase they used is (linguistically) correct (e.g. whether they were able to ask for directions at CEFR level A or to give a presentation on a specific topic at level C), but on whether they managed to reach the aim of the task. In other words, in the examples given above, they will be judged on whether they found the place they wanted to go to (level A) or whether the audience understood the content of the presentation and whether the content had an impact on the group's activities (level C). In the GULT approach to language testing, the assessment of student performance is not only based on the outcome of the activity, but also on descriptors taking into account content as well as linguistic and pragmatic competences, following an interactionalist approach.

Tests: formal measure of skill, announced in advance and requiring a substantial amount of time to complete (an hour or more). For example, a test might include a reading passage with questions, a grammar section and a series of words to use in sentences (Testing Glossary, 2011). They are also a procedure for testing specific aspects of proficiency or knowledge.

1. A set of components which together constitute an assessment procedure, often used to mean the same as examination.

2. A single task or component for assessing an area of skill or knowledge, e.g. speaking or writing. In this sense a test may also form part of a complete examination as a component (e.g. the speaking test) or as a single task (e.g. cloze test).

3. An assessment procedure which is relatively short and easy to administer often devised and administered within an institution (e.g. a progress test)

or used as a part of a research programme or for validation purposes (e.g. anchor test) (ALTE, 1998: 166).

Test item: one entry or question on a test or quiz, e.g. of an item: "Write out a sentence using the word 'salubrious'" (Testing Glossary, 2011).

Transparency: means that the tests are described in such detail that all those handling the test results know what was tested and how, and what the test results mean, that is, what the test taker with a certain grade or score can do in the target language.

Validity: a) means that a test can be shown to be valid through a priori validation or a posterior validation. A test is valid if it measures what it is supposed to measure and convinces the stakeholders of that. There are many kinds of validity. b) the extent to which scores on a test enable inferences to be made which are appropriate, meaningful and useful, given the purpose of the 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 (ALTE, 1998: 168).

Weighting: the assignment of a different number of maximum points to a test item, task or component in order to change its relative contribution in relation to other parts of the same test. For example, if double marks are given to all the items in Task One of a test, Task One will account for a greater proportion of the total score than other tasks (ALTE, 1998: 169).

References

ALTE (1998). *Multilingual Glossary of Language Testing Terms: Studies in Language Testing 6.* Cambridge: CUP.

McNamara, T. (2000). *Language Testing*. Oxford et al.: Oxford University Press.

Testing Glossary (2011). Available at:
https://facultystaff.richmond.edu/~rterry/Middlebury/testgloss.htm
[accessed 9 September 2011].

Skehan, P. (1996). "Second language acquisition research and task-based instruction". In Willis, J. and Willis, D. (eds). *Challenge and Change in Language Teaching*, Oxford: Heinemann, 17-30.

Skehan, P. (2003). "Task-based instruction". In: *Language Teaching* 36, 1-14.

Weideman, A. (2006). "Assessing academic literacy in a task-based approach". In: *Language matters* 37 (1), 81-101.

Guidelines for task-based university language testing

Johann Fischer, Catherine Chouissa, Stefania Dugovičová and Anu Virkkunen-Fullenwider

This publication is targeted at:

- decision-makers in university language teaching and testing (e.g. heads of university language centres or language departments);
- teachers and testers of languages for specific purposes in higher education;
- language teacher educators;
- other stakeholders in university-level language instruction and assessment.

Guidelines for task-based university language testing is a practical manual for those language teachers and testers who are looking for a valid tool to measure their students' language skills in a meaningful way. It shows how to link the language skills taught with those needed in studies and later in working life. It helps language instructors, who already conduct task-based language courses, to design corresponding tests and to evaluate their students' language performance. The publication also highlights the benefits of task-based language testing for all the stakeholders.

For further information and materials relating to this publication, visit the website: http://gult.ecml.at.