Centre for Internationalisation and Parallel Language Use (CIP),
ENGEROM, Faculty of Humanities, University of Copenhagen

The development of the Test of Oral English Proficiency for Academic Staff (TOEPAS) - Technical Report

Technical Report

Joyce Kling, PhD Scholar

Lars Stenius Stæhr, PhD

Indhold

Acknowledgements....3
Abbreviations....4
1. Background....5
1.1 Resources & constraints....6
2. Analysis of the target language use (TLU) domain....7
2.1 Description of the TLU domain....8
3. Test specifications....11
3.1 Assessment context....11
3.2 Brief description of assessment procedure....12
3.3 Definition of the construct....12
3.4 Tasks specifications....14
4. Analytical criteria for assessment....15
5. Assessment procedure guidelines....18
5.1 Reporting scores and feedback to the participants....18
5.2 Self assessment....20
6. Pilot testing....20
6.1 Modification phase....21
6.2 Post-pilot adjustment phase....21
7. Examiner training and reassessment of criteria....23
7.1 Training/norming....23
8. Finalizing the grid and developing the global scale....25
9. Appendix....26
10. References....27

Acknowledgements

We would like to thank Dr. Renate Klaassen, Technische Universiteit Delft and Virginia Maurer, Harvard University Derek Bok Center for Teaching and Learning for their aid in our research and use of research materials. We would also like to thank the Dr. Alister Cummings, University of Toronto, Ontario Institute for Studies in Education, Dr. Norbert Schmitt, University of Nottingham, School for their expertise and input during the development of this test. In addition, we thank Diane Schmitt, Nottingham Trent University, School of Arts and Humanities for her contribution to both the test development and this report.

We are grateful for all the equipment and assistance we received from the Department IT Media at the Faculty of Humanities for setting up the testing facility and providing video files and backup.

Lastly, our gratitude goes to Dr. Birgit Henriksen and the team at CIP for assisting in the pilot phase of this project both in regard to administration and technical support. In particular, we would like to thank Jimmi Nielsen for his insights, humour and unending attention to detail in the early phases of the testing development and administration.

Abbreviations

ACTFL: American Council on the Teaching of Foreign Languages Proficiency Guidelines C Speaking

COME: Copenhagen Master's of Excellence degree programmes at the University of Copenhagen

CEFR: Common European Framework of Reference for Languages: Learning, Teaching, Assessment

EAP: English for academic purposes

EFL: English as a foreign language

ELT: English language teaching

ESP: English for special purposes

GSI: Graduate teaching assistant

IELTS: International English Language Testing System Speaking band descriptors

ILR: Interagency Language Roundtable Language Skill Level Descriptions

ITA: International teaching assistants

KU: University of Copenhagen

L1: First language (mother tongue)

OPI: Oral proficiency interview

SLA: Second language acquisition

TLU: Target language use

TOEFL: Test of English as a Foreign Language

TOEPAS: Test of Oral English Proficiency for Academic Staff

1. Background

In September 2008, the University of Copenhagen (KU) management team concluded that there should be an assessment procedure that could be used to certify the English language skills of university lecturers teaching at selected graduate programmes at the University of Copenhagen, the Copenhagen Masters of Excellence (COME)[1]. The management team considered a language certification test to be a quality management tool that would ensure that the level of English in the COME programmes would not negatively affect the quality of the teaching. Thus, the overall desired purpose of the certification would be to assess whether the COME teachers have the necessary English language skills to cope with the communicative demands of teaching on these programmes. Moreover, the test should serve a secondary, formative purpose, namely, when teachers do not have the sufficient English language skills to pass the certification, the test should provide some information about the kind of language support or training test takers need to be able to teach at these programmes.

This technical report outlines the development process and the main components of the certification procedure developed at the Centre for Internationalisation and Parallel Language Use (CIP) to meet the KU management team's stated aims. The resulting certification assessment procedure, entitled the Test of Oral English Proficiency for Academic Staff (TOEPAS) is intended for screening or selection purposes and could be regarded as a high-stake test in the sense that the test results have consequences for whether or not test-takers are allowed to teach at the COME programmes. When such high-stakes decisions are made on the basis of information derived from a test, it is important that we can fully justify the inferences about the test-takers ability to cope in the target language use situation which are drawn from the test performance (McNamara, 1996, p. 93–94). Therefore, this report will also touch upon some of the challenges that arise when developing an oral proficiency development tool for highly advanced speakers who operate in a technical academic domain. These challenges include, in particular, the selection of assessment

[1] The COME programmes are elite English-medium graduate degree programmes designed by the University of Copenhagen with the stated aim of attracting the most academically advanced students. (http://come.ku.dk/what_is_come/)

tasks and the determination of the different levels of proficiency in the analytic scale used for assessing this specific group of test takers.

1.1 Resources & constraints

Domain specific performance tests like the TOEPAS are relatively resource heavy and time-consuming to develop and administer. From the start, the TOEPAS was heavily under-resourced and subject to a strict time constraint as the test had to be ready approximately seven months after it was commissioned. The management team was initially made aware of this problem, and in the document "*Notat om certificering af COME-underviseres engelskkompetencer*" (September 19, 2008) two models for developing the certification procedure were proposed. The first model outlines the ideal development process of the certification test, involving issues such as domain analysis, development of test specifications and rating scale, piloting, standard setting and training of raters. This model, however, could not realistically be followed within the time frame given and with the resources available. A second model was therefore selected. This model, which is a light version of the first one, involves the same developmental stages but suggests a number of compromises in each of the stages. Moreover, the model proposes that the certification test should only assess test-takers' oral proficiency and not their reading, writing and listening skills – although these skills must be regarded as equally important for coping with the communicative demands of functioning in an academic setting. Due to the time constraint, it was thus decided to focus only on test-takers' oral proficiency when lecturing and interacting with students as this was obviously a key activity for teaching at the COME programmes.

Lack of resources and the time constraints also had an impact on decisions regarding the broad test method. McNamara (McNamara, 1996, p. 96)distinguishes between three overall kinds of performance assessment:

1. Extensive observation of the candidate's performance in the work place or target language use situation (*direct assessment*)
2. Selective observation of the same (*work sample* methods as defined in a narrow sense)
3. Simulation techniques

The ideal solution would have been to observe the teachers in their naturalistic setting, e.g. when lecturing, interacting with students, conducting exams etc. and assess their ability to cope with the communicative demands on the basis of this kind of direct observation. However, observation was not possible for reasons of practicality and the assessment thus had to be some form of simulation technique. Moving away from direct observation will always have some implications for test validity but this is the kind of compromise between practicality and validity that inevitably has to be made (Bachman & Palmer, 1996). However, as will be become evident below, strict efforts have been made to design test tasks that are representative of the commutative tasks of the target language use situation.

A final constraint or challenge that deserves mentioning here is the fact that there is a significant lack of previous research in the area of high proficiency performance assessment, in particular for this particular target group.

2. Analysis of the target language use (TLU) domain

The first step in our development of the test was to analyze the target language use (TLU) domain, that is the "set of specific language use tasks that the test taker is likely to encounter outside of the test itself, and to which we want our inferences about language ability to generalize" (Bachman & Palmer, 1996, p. 44), to identify the communicative tasks facing the teachers. With regard to domain specific tests, it is thus important that the test and the target language use domain share some of the same key characteristics so that the test-takers' performance on the test can be interpreted as evidence of their ability to perform specific language tasks in the target language use domain (Douglas, 2000, p. 47). In other words, a close correspondence between the TLU domain and the test tasks will positively affect the authenticity and the (construct) validity of the test.

When we began developing the test, only four graduate programmes had been awarded COME-programme status. These were *Molecular Biomedicine* (Faculty of Health Sciences and Faculty of Science), *Human Biology* (Faculty of Health Sciences and Faculty of Science), *International, Law, Economics and Management* (Faculty of Law and Copenhagen Business School), *Environmental Chemistry and Health* (Faculty of Life Sciences, Faculty of Health Sciences, Faculty of Science,

Faculty of Pharmaceutical Sciences and Technical University of Denmark). We knew that more programmes would follow in the spring of 2009 but we did not know which. Moreover, one of the programmes, *International, Law, Economics and Management*, was still under development and could not be part of the domain analysis. So, given the time constraint of the project, we had to base our TLU domain analysis on the first three COME-programmes.

The TLU domain analysis involved the following components:

- Interviews with the heads of the study boards who are responsible for the three programmes
- Discussions with the dean of the Faculty of Health Sciences (involved in all three programmes) and the dean of education at the Faculty of Life Sciences (involved in one of the programmes)
- Observation of teaching and short interviews with teachers. Courses in all three programmes were observed and the teachers of these courses were briefly interviewed.
- Literature review: Literature on the following subjects was reviewed:
    - Language tests used for certifying university teachers, e.g. for certifying the language skills of international teaching assistants at American universities
    - What kind of communicative tasks do university teachers face when teaching and what kind of linguistic skills do they need in order to successfully cope with these tasks?
    - Global and analytical scales used in the assessment of English as a foreign language (EFL) learners' oral proficiency
    - Development of oral proficiency tests

2.1 Description of the TLU domain

As noted above, due to a lack of time and resources was decided that the assessment process would only focus on the test takers' oral proficiency. Interviews with heads of the COME

programmes and teachers as well as observation of teaching confirmed that teachers' oral skills must be a top priority when developing a language certification for university teaching. In view of this, our domain analysis was primarily concerned with the kind of oral tasks teachers have to perform as part of their teaching.

The TLU analysis revealed three main teaching formats:

- Lecture: The teacher gives a lecture, typically supported by a visual presentation such as a PowerPoint slide show, explaining text book material, figures, graphs, pictures etc. The lectures are given to between 20-35 students, with what appears to be a relatively high degree of interaction between teacher and students. Interaction occurs when the teacher asks comprehension questions to test the students' understanding of the material and when students interrupt and ask questions.

- Group work: The students work in groups of two to four, solving a specific task on a computer or on paper or discussing a case. The role of the teacher is to help the groups with their different questions.

- Laboratory work: The students work on an experiment in groups in the lab and the teacher supervises them.

In these different TLU situations the teachers were faced with a number of different communicative tasks. Our domain analysis, observations and interviews, indicated that the tasks outlined below were (some of) the most significant ones:

- Presenting highly complex content material to students, on the basis PPT slides or other visual aids – but without a manuscript
- Explaining domain-specific terms and concepts
- Presenting a case or assignment, describing administrative details
- Clarifying, paraphrasing or restating concepts and main points
- Asking questions to students
- Understanding student questions
- Responding to student questions

- Dealing with unclear questions or misunderstandings and negotiating meaning

As will become evident from the test specifications below, we attempted to develop a testing procedure that included these communicative tasks.

Whereas the communicative tasks outlined above must be regarded as central to most university teaching involving lecturing and teacher-student interaction regardless of subject, the content of the teaching naturally varies greatly from program to program. This is evident when looking at the variety of graduate programs which have now become COME programmes:

- MSc in Molecular Biomedicine
- MSc in Human Biology
- MSc in Environmental Chemistry and Health
- MSc in International Law, Economics and Management
- MA in Applied Cultural Analysis
- MA in the Religious Roots of Europe
- MSc in Food Science and Technology
- MSc in Computer Science

In view of this, we decided to construct a testing procedure in which the test takers should select the content themselves based on their field of expertise. In other words, to strengthen the content validity and the authenticity of the test procedure, test takers would have to demonstrate ability to carry out the relevant communicative tasks with reference to the content they are familiar with. Although laboratory work was a part of all the three COME programmes analyzed here, it is not a TLU situation that we want to directly simulate in the certification procedure. Laboratory work is limited to only the natural and health sciences and future COME programmes will come from all eight faculties at the university. In addition, the variety of groupwork activities across the various programmes makes it difficult to standardize a specific type of assignment that lends itself to assessment. We thus decided that it would not make sense to directly simulate these two kinds of teaching formats in the certification. However, the interaction between teacher and students which takes place in the laboratory and groupwork might resemble the student-

teacher interaction taking place in the teacher-fronted lecture. Therefore, the communicative task of interacting with students is a significant task that needed to be part of the certification.

The assessment procedure is described in the test specifications below.

3. Test specifications

3.1 Assessment context

The TOEPAS is given to university teachers who lecture in elite English-medium graduate degree programs. The overall purpose of the test is to certify the lecturers' English language skills by assessing whether they have the necessary skills to cope with the communicative demands of teaching at the COME programmes. More specifically, the test aims to assess whether the teachers have an adequate level of oral proficiency for lecturing and interaction with graduate students in a university setting. Moreover, when teachers do not have the sufficient English language skills to pass the certification, the test provides some information about the kind of language training they need to be able to teach at these programmes.

Wanting to ensure that the teachers' level of English proficiency does not have a negative effect on the quality of teaching, the University of Copenhagen management team commissioned a test whose results can assist the heads of study boards, heads of departments and deans in determining who can and cannot teach on the COME programmes. In addition, the test results provide information for the administration about the type of language training or support teachers need to be able to teach on the COME programmes. The test results also provide the test takers themselves with a tool for getting specific feedback on their speaking skills for teaching in English.

The test takers are primarily associate professors and full professors who are experts in their field of expertise and they have wide variety of different EFL learning backgrounds. The majority of test takers have Danish as their first language (L1), but teachers with a variety of other L1s take the test as well. Teachers with English as their L1 are exempt.

The testing procedure is conducted at the CIP and examiners are English language teaching (ELT) specialists in the fields of second language acquisition (SLA), language testing and pronunciation.

3.2 Brief description of assessment procedure

Based on the TLU analysis described above, we decided to develop a test procedure that could simulate two main tasks found in the TLU domain: 1) lecturing to students on the basis of visual aids but without a manuscript; 2) interacting with students in the classroom about the content of the lecture or related issues. The two main tasks are thus designed to elicit whether test takers can handle a range of more specific communicative tasks which the TLU domain analysis showed to be of importance for university teaching.

The test procedure lasts approximately two hours and involves the assessment of three teachers from the same programme or area of expertise. Each participant has to give a prepared mini-lecture and participate in a role-play as a *'student'* in order to simulate a graduate classroom setting. Hence, each of the test takers gives a lecture on a specialized topic within his/her area of expertise and discusses aspects of this topic with his/her colleagues who act as students. This means that the test takers select the content themselves, i.e. the subjects that they normally teach. In order to assess the test taker's ability to interact with students about the specialized topic, it was necessary to have three test takers from the same programme/area of expertise participate in the same testing procedure as the two examiners would not be able to engage in interaction with the lecturer about his/her selected topic.

The testing procedure is digitally recorded and two examiners rate the test takers' performance based on their observation of the live performance and on the recording. The assessment is given as a global score from 1 to 5 and as an analytic profile based on the following criteria: fluency, pronunciation, vocabulary, grammar and interaction skill.

3.3 Definition of the construct

This is a test of spoken production and interaction in English. More specifically, it is assesses test takers' ability to lecture and interact with students in an academic context. The test tasks are designed to elicit whether the test taker can handle a range of communicative tasks which are central to university teaching at graduate level, namely present highly complex content material; explain domain-specific terms and concepts; clarify, paraphrase and restate concepts and main

points; present and explain an assignment; ask, understand and respond to student questions; deal with unclear questions and misunderstandings and negotiate meaning when necessary.

The important subskills involved in successfully coping with these communicative tasks are related to the test taker's fluency, pronunciation, vocabulary, grammar and interaction skill. These subskills also comprise the assessment criteria on which the analytic profile and the global assessment are based. This means that a good performance on the test would reflect the test taker's ability to speak smoothly, effortlessly and coherently at an appropriate pace and without unnatural language-related pauses or hesitations. Moreover, the test taker's pronunciation would be intelligible and precise and would not cause strain for listener or impede effective communication. In terms of vocabulary the test taker would demonstrate a correct use of a broad range of academic and domain-specific vocabulary for effective communication and would show a good command of idiomatic expressions and collections. In a good performance the test taker would also consistently display a high degree of grammatical accuracy in both simple and complex structures. Finally, in terms of interaction skill, the test taker would understand questions and comments and respond appropriately and effectively and would be fully capable of dealing with unclear questions or misunderstandings when necessary, e.g. through comprehension checks, clarification requests and confirmation checks.

With regard to Bachman and Palmer's (1996) model of language ability, the test directly assesses grammatical knowledge and directly or indirectly covers some aspects of textual, functional and sociolinguistic knowledge. Grammatical knowledge, as defined by Bachman and Palmer, is directly assessed through the criteria of vocabulary, grammar and pronunciation. Textual knowledge reflects the test taker's ability to structure ideas and to produce coherent and cohesive speech, and this ability is mainly covered by the fluency criteria. However, it is also assessed through vocabulary and grammar. Functional knowledge is only covered indirectly in the test. When describing, explaining, exemplifying and interpreting information, when expressing views and attitudes, when requesting something from students and when interacting with students, the test taker will perform a range of different functions. However, the ability to express different functions is only assessed indirectly through the criteria of fluency, grammar, vocabulary, pronunciation and interaction skill. To some extent, this is also the case for sociolinguistic

knowledge. This aspect of Bachman and Palmer's model of language ability is not tested directly but will obviously be involved in any kind of communicative language task. However, it might be argued that sociolinguistic knowledge is semi-directly assessed in the criteria of *interaction skill* as this involves the ability to respond *appropriately* to questions and comments, and that it is also assessed in the criteria of *vocabulary* as this involves correct and *appropriate* use of vocabulary.

3.4 Tasks specifications

The testing procedure is divided into three parts: 1) a warm up; 2) a mini-lecture; and 3) a question and answer session. Only parts 2 and 3 are assessed. These three tasks are described in turn below.

Part 1

Part 1 consists of a warm-up session which is not assessed. The session lasts approximately 10 minutes and aims to allow participants to interact with each other and with the examiners in English before the assessment. The underlying purpose is to get the participants to relax and give them the opportunity to speak English immediately before they are assessed and to get used to the variants of English spoken by the examiners and fellow participants.

The examiners ask the participants basic background questions about their professional interests, work, and areas of research and attempt to initiate an open discussion among the participants and examiners.

Part 2

In Part 2 each test taker gives a prepared mini-lecture of 20 minutes to an assumed audience of graduate students in his/her programme. This can be drawn from a lesson the test taker has taught in the past. As part of the lecture, the test taker should briefly give his/her *students* instructions for a group work assignment to be completed at a later time. During the course of the lecture, the other participants will interrupt and ask a few questions related to the presentation. The test taker is allowed to use visual aids (e.g. PowerPoint presentation, whiteboard, etc.).

The purpose of this task is to allow test takers to demonstrate oral proficiency in English when giving a mini-lecture in an academic setting. Moreover, the task aims to elicit whether the test taker can use the administrative language needed for giving instructions for a group assignment and whether the test taker has the required language ability to deal with questions from the audience.

Part 3

Part 3 consists of a question-and-answer session lasting approximately 5-7 minutes. After each lecture the two test takers taking on the role as students are required to ask questions about the lecture and are told they their goal is to initiate an open dialogue on the topic of the mini-lecture and to engage in a discussion on a relevant point they find interesting. The purpose of this task is to simulate student/teacher interaction in an academic setting and the test takers are assessed their English interaction skills, both when asking and answering questions.

As in evident from the above, Part 2 and 3 involve simulation of student/teacher interaction to different degrees. The following role-play instructions are given to the test takers:

> *In order to simulate 'student/teacher' interaction during this assessment, you are to take on the role of a graduate student. Below are guidelines for this role.*
>
> *DURING THE MINI-LECTURE*
>
> *Find an opening or interrupt the lecturer one time during the course of the mini-lecture to ask any question you find relevant (e.g. ask for clarification of a specific term, a concept or a graph (any visual aid), the assignment, etc.)*
>
> *AFTER THE LECTURE (QUESTION & ANSWER SESSION)*
>
> *Ask questions about the mini-lecture. Your goal here is to initiate an open dialogue on the topic of the mini-lecture and to engage in a discussion on a point you find interesting.*

4. Analytical criteria for assessment

Once the test specifications were in place, designating a desired level of English proficiency for teaching university courses, specifically at the graduate level, was necessary. Unlike some of the

other universities currently implementing English certification procedures for academic/scientific staff, the University of Copenhagen does not have a specific language policy requiring a particular level of language proficiency, i.e, on a standard internationally recognized commercial test. Therefore, we needed to determine an acceptable and transparent proficiency scale for this playing field and a range of levels that would be suitable for our needs. We approached this task using combined intuitive, quantitative and qualitative approaches (Council of Europe, 2001, p. 226).

Beginning with an intuitive approach to developing proficiency descriptors for this certification test, a number of existing scales for measuring language competency were evaluated, most specifically the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR), the Interagency Language Roundtable (ILR) Language Skill Level Descriptions for speaking and the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines C Speaking, the International English Language Testing System (IELTS) Speaking band descriptors (public version), as well as variety of American university rating scales for assessing graduate teaching assistant (GSI) / international teaching assistants (ITA)[2].

In general, a number of universities running certification programmes[3] have selected criteria directly from the CEFR. These universities all maintain the policy that lecturers must have a proficiency at a minimum level of C1. This level is a realistic expectation for academics working in English-medium settings given that once having completed their studies, students should have acquired this same level of proficiency (Klaassen & De Graaff, 2001). In fact, while most degree programs at the University of Copenhagen require a minimum iBT TOEFL result of between 79[4] (B1 upper level) and 100[5] (B2), in 2009 iBT test takers with Danish as their L1 averaged 101, a strong B2 result(ETS/TOEFL, 2010).

[2] University of Michigan Graduate Student Instructor Oral English Test (GSI-OET), University of Pennsylvania Interactive Performance Test, University of California Los Angeles Test of Oral Proficiency (TOP),
[3] E.g., Delft University of Technology (DUT), Technische Universiteit Eindhoven (TU/e), and Copenhagen Business School (CBS)
[4] University of Copenhagen, Faculty of Science
[5] COME degree program in International Law, Economics and Management

With a CEFR equivalent level of C1 identified as the starting point as an acceptable level for certification, we decided on a 5-point assessment scale, loosely linked to the relative CEFR levels (5/C2+, 4/C2, 3/C1, 2/B2, 1/B1). Participants receiving a result of 3, 4 or 5 would be certified to teach in English-medium programs. An assessment of 1 or 2 would not be an acceptable proficiency level and the participant would not be certified. We thus proceeded to analyze the CEFR descriptors for levels B1 to C2 to determine if they were suitable for our assessment purposes. The use of the criteria directly from the CEFR provides a desirable level of transparency and reliability. The actual descriptors at the C1 level also described minimum general proficiency levels skills that we believed were necessary for academic work. Unfortunately, these descriptors did not completely meet our analytic demands in regard to the assessment of high level proficiency in English for specific purposes. Consequently, we set out developing unique descriptors, drawing from the documents mentioned above. Throughout the entire development process, we kept both the global and discrete CEFR criteria on hand in an attempt to maintain a parallel between these levels and those of the TOEPAS. Thus, after sifting through the assessment criteria of CEFR, as well as the above mentioned tests and scales, we drafted criteria descriptors at five levels for six categories of competence of the CEFR: fluency, pronunciation, vocabulary, grammar, coherence and cohesion and interaction. The decision to have a 5x6 scale was supported by the literature (McNamara, 2000).

The analytical descriptors were subsequently revised through a qualitative approach (Council of Europe, 2001, p. 209). Small workshops with groups of informants provided specific feedback on both the overall construction of the grid as well as he natural progression of the working of the specific descriptors. Qualitative input was collected from the examiners, as well as a group of 19 graduate students of English at the University of Copenhagen through a jigsaw exercise to determine whether there was a natural progression in the wording of the descriptors[6]. Overall, the scale was deemed clear. However, two categories continued to cause issue throughout the

[6] The informants received the grid in pieces and asked to a) put the descriptors in the correct order, b) explain their rationale for placing the pieces, and c) identify any key points that aided or confused them (CEFR 2000: 209)

development and examiner training process, namely coherence & cohesion and pronunciation (see section 7.1 for discussion).

With the analytical scale in place, we were able to quickly draw up a global scale that clearly informed all parties of the five levels of assessment.

5. Assessment procedure guidelines

Using the analytic criteria, the two examiners (Examiner A & Examiner B) independently assess each of the three participants immediately following the assessment session and assign a global result. The examiners also rate the participants from 1-5 for each of the individual criteria areas. After each examiner has reached an independent rating, the examiners discuss these ratings and must reach an agreement as to the overall global assessment as well as the discrete categories. In cases of lack of agreement for a global result between the two examiners, a third examiner assesses the digital video performance of the participant. The three examiners then discuss their positions and award a global result. In all cases of a result of 2 or below, or a result of 5, a third examiner must assess the video performance to confirm the result. Prior to viewing, the third examiner is not informed of the result, but should independently rate the performance. To support this system and to alleviate rater bias, random samples of participants' video performances are distributed over time to all examiners for their assessment. This provides the examiners with a quality check of their assessment as well as a training/norming exercise for the third examiner.

5.1 Reporting scores and feedback to the participants

When a global assessment has been agreed upon, Examiners A & B record this result, as well as the results of the five linguistic categories, on a separate assessment form. The assessment is supported by documentation in the form of specific examples from the participant's performance. Two sets of documents are created at this time: one for record keeping and a second that is sent directly to the participant. The first document records not only the narrative feedback with quotes from the participant' performance to support the assessment, it also records the grades awarded for each linguistic category. On the second document, the one sent directly to the participants, the

individual results for each analytic criterion are removed. The participants receive a global result and the narrative feedback.

The results of the certification assessment are distributed to the participants, their department head and/or faculty dean. Feedback is only sent directly to the participant for their own personal consideration. They are free to share the specific feedback with their colleagues, but CIP does not provide this information to any administrative units at the university.

All participants receive this written feedback, regardless of their results. This means both positive and negative examples of language related performance in an EAP/ESP situation. The rationale for providing this explicit written feedback stems from the need to make the assessment process as transparent and comprehensible as possible and to develop a positive rapport with the participants and their respective departments. This is a mandatory, high-stakes test for a very specific population. The participants are sophisticated professional educators who want to understand how we arrived at the assessment result. Some of the participants might resent the imposition of this type of assessment since many of them have been teaching in English for years, sometimes in groundbreaking programmes. So, to lower their resentment towards the testing situation and simultaneously create a learning experience, we provide detailed written feedback.

In conjunction with this written feedback, the participants also receive a digital link to their video recorded performance. The participants find this feature a natural accompaniment to the written feedback. Having access to the videos gives the participants the opportunity to review their performance in a productive manner. In the privacy of their own offices, they can consider the detailed feedback and pursue training if necessary. Having the video recording allows them to review their performance in a more productive manner. The video also allows the examiners to provide detailed explanations and justifications of the assessment in cases of discontent.

From an assessment perspective, the video recording naturally support the assessment, both globally and analytically, and allows us to provide detailed feedback on the participants' performances supported by specific personal examples. However, questions as to the extent to which our access to the video affects our assessment are still a concern. One could argue that we

may alter our assessment based on review of the recorded performances, which of course leads to questions regarding the affect on the validity and reliability of the results.

Lastly, the participants are invited to contact CIP at any time for a face-to-face feedback session with one of the examiners. The element allows the participants to review any aspects of the feedback that they are unsure of. It provides a first step toward training and competence development. In this session, the examiner can review the areas that the test taker should focus on in order to improve their language proficiency for teaching English-medium courses.

5.2 Self assessment

In addition to the live assessment, we have included a participant self assessment task drawn from the 'can-do' statements from the CERF. This self assessment provides the examiners with a baseline from which to provide feedback to the participants. The self assessment helps to identify if the participants have a realistic perception of their language skills, in comparison with the observed performance. If the self assessment and the TOEPAS result do not correlate, it can be noted in the feedback. In addition, this information provides us with data to investigate the relationship between the participants' self assessment on the CEFR (general) and the TOEPAS assessment (domain specific).

6. Pilot testing

To collect qualitative information and prepare for the pilot operational testing, a pre-pilot field triail was conducted. Using internal staff at CIP as participants, we ran an assessment session under operational conditions. Feedback from this field trial allowed us to redraft the tasks required of the participants and clarify instructions and administrative procedures. No changes were made to the assessment grid based on this activity.

Following the field trial, we proceeded to operational testing and administered the test to 19 volunteer participants from LIFE. In return for volunteering, the participants received written feedback on their performance. This operational testing phase allowed us to focus on three issues: the test takers' language abilities, the usefulness of the analytic descriptors and the assessment

grid as a whole, and the testing and administrative procedures. The information drawn from this pre-testing help to determine which modifications were necessary to improve the usefulness of the test (Bachman & Palmer, 1996).

6.1 Modification phase

It was vital to test the assessment grid to confirm that the two examiners conducting the pretesting were able to assess the participants' English proficiency in the TLU context as well and interpret the descriptors the same. Through a course of assessment and negotiation, the examiners were able to identify areas which needed to be refined so that the descriptors would be more effective as tools for assessment and examiners could reach the same result more consistently.

As for data collection in regard to the testing and administrative procedure, a debriefing session was included in the operational testing. This debriefing session was purposefully designed to allow the participants to relax and openly share with us their reflections on the assessment session they had just participated in. With this in mind, the language of the debriefing session switched from English (the language of the test) to Danish (the L1 for the participants). Participants were asked open ended questions on each aspect of the assessment procedure in order to elicit a broad range of responses (Bachman & Palmer, 1996). Some of the developmental concerns here included clarity of written communication, task appropriateness, assessment construction (warm up, task, interaction), number of participants, use of technology, etc. During this debriefing session, we were also able to explain the feedback procedure to the participants and get feedback on this aspect of the test as well.

Additionally, the operational testing also provided us with an opportunity to develop appropriate formulations and phrasing for the assessment feedback forms. During this phase we discussed at length the best method by which to provide detailed feedback with appropriate specificity in a timely and efficient manner. It was also during this phase that our learning curve regarding access to the video files was quite steep.

6.2 Post-pilot adjustment phase

Based on the feedback from operational testing and on our own assessment experience, we felt no need to make any significant changes the testing procedure or administration. However, the debriefing sessions with the participants did provide us with some insights and allowed us to make some minor adjustments. During the debriefing session, the participants were asked to reflect on 10 aspects of the certification session:

1. the instructions (administration prior to the session)
2. length of the presentation session
3. atmosphere
4. the warm-up
5. the (student) role-play,
6. assessment in group setting with colleagues
7. authenticity of procedure
8. use of video recording equipment
9. feedback process
10. their own self assessment (using the CEFR self assessment tool)

In regard to the instructions and communications sent out prior to the certification session (1), as well as the length of the presentation session (2), the majority of the participants were satisfied, however a few of the participants expressed some confusion regarding what precisely was required of them and how much they would be able to cover in the allotted period of time. To be more specific, the participants were not sure how much of an existing lecture they could pare down to the time allotted for their presentation. Given this feedback, the test instructions were clarified and the length of time allotted for the presentation was expanded from 15 to 20 minutes.

The participants all expressed complete satisfaction in terms of the atmosphere (3) of the testing session, expressing that they actually found it to be a comfortable and relaxed setting. They unanimously agreed that the warm up session (4) helped them to loosen up and “get into the language”.

As we are not aware of any other task-based ESP OPI such as this one, where colleagues are group in pairs and groups of three and are responsible for all input and interaction (no input/interaction

on the part of the examiner(s)), we were most concerned with the participants acceptance of this type of assessment session and the reliability and validity of assessing this type of OPI. The feedback we received from the participants in regard to the authenticity of the situation, both in terms of task and role (5), (6) & (7), was positive. The participants noted that they are "accustomed having colleagues present at lectures". They felt that the role play and question & answer sessions were appropriate and were happy to have colleagues from the same discipline present, as it promoted authentic questions and interaction.

The participants alleviated any concerns we had regarding the recording equipment and microphones distracting focus during the assessment (9). All those who commented noted that they forgot about the camera immediately and focused on the task at hand.

The last point, self assessment (10), was not taken as seriously as other aspects of this procedure. Many of the participants did not complete the self assessment prior to arriving at the center and simply filled it out during one of the breaks or after the session. Some commented that they found the CEFR self assessment difficult to fill out due to the fact the they themselves understand their own strengths and weaknesses in relation to daily language versus domain specific language. Overall, the participants considered the activity appropriate and thought it would be interesting to see if their self perception would be the same as their test result.

7. Examiner training and reassessment of criteria

As the two test developers had conducted all pilot sessions, new examiners were rotated into the examination schedule only after training/norming and observation of three to four certification sessions. As a constant, one of the test developers administered all the sessions during each new examiners official certification sessions.

7.1 Training/norming

Following a review of the procedure based on the feedback from the participants and some focus on the grid, a training / norming session for two new examiners was conducted.

The initial activity for the session was the completion of the jigsaw activity described above (see section 4). This went quite quickly, as the examiners had little difficulty placing the descriptors in

the correct categories and levels. However, this exercise did lead to a great deal of focused discussion on the working of the descriptor in all categories and levels. To support the categories and the wording of the descriptors, the examiner-trainees also viewed video recordings from the operational pilot testing to gain an understanding of the construct of each category and agree on the level of proficiency required to achieve each level. With the examination team in the initial stages of this certification program consisting of only four examiners (the exam development team plus two trainees), we were able to work efficiently to reach agreement.

In order to apply the criteria, the examiner-trainees were shown three video recordings of pilot participant presentations, each representative of a specific level of proficiency. After each video, all four examiners discussed the ratings for each category. At this point, the inter-rater reliability across categories was still low. The wordings of the descriptors for coherence & cohesion, grammar and pronunciation caused disagreement between the examiners. Therefore, the team time adjusted the wording of the descriptors to meet the perceptions of the examiners.

It became apparent that the analytic category of coherence & cohesion did not stand well on its own. We realized from our discussions that although coherence & cohesion could be interpreted quite broadly, in the pilot testing we had operationalized it very narrowly as linguistic connectedness. Given this construct, we decided to merge this category with fluency. The pilot data supported this merge, as the results from the two categories highly related. In reviewing the literature, we found that this definition has been implemented as one element of fluency in other existing oral proficiency tests, e.g. IELTS and Test of English for Aviation (TEA). Therefore, we decided to broaden the construct of fluency to include connectedness.

Based on our assessment experience with the pilots, the pilot data (results + debriefing), a meeting with consultants and sorting tasks with other examiners and graduate students, the grid was adjusted to make it more user friendly and transparent. Changes were implemented in all categories. The most significant change being the merging coherence & cohesion into the category of fluency.

The examiners continue to struggle with the analytic descriptors for pronunciation. The descriptors in this category are more specific than in the other rubrics available to us as resources.

For example, on the IELTS oral assessment scale, pronunciation is a shorter scale. However, pronunciation and intelligibility is a very important category for our population, both because of the face validity in regard to Danish lecturers lecturing to Danish students, and because of the international student body represented in the classroom. Non-Scandinavian students must be able to understand a 'new accent' when they come to an English-medium program in Denmark and Danish students must be able to understand lecturers coming from other non-English L1s.

In regard to vocabulary, the testing of participants led us to consider the level at which to assess a strong command of formulaic language and alter the descriptors appropriately. Lastly, the interaction category received a bit of an overhaul. The wording here had caused problems since it appeared that too much was trying to be covered the category (understanding of clear questions, understanding of unclear questions, response to both of these, etc.). After the pilot testing we agreed that the overarching category had to do with negotiation of meaning and the ability to clarify and rephrase in unclear situations.

8. Finalizing the grid and developing the global scale

From the start, we had a clear understanding regarding the holistic ratings and the level required to achieve certification (level 3). This was based on the original postulation that we wanted to link the new scale to the CEFR and have a positive result of approximately C1. Once the individual analytical descriptors were in place, we were able to draw up a global scale that combined the scores for the separate aspects for reporting purposes (McNamara, 2000). Similar to the IELTS global ratings, the TOEPAS global scale does not weave together statements directly from the descriptors. Instead, the scale gives general, overarching, transparent statements for use by the participants and the appropriate stakeholders (i.e., heads of department, deans, administration, etc.). Once the 5-point scale was agreed upon, it was translated into Danish so that all stakeholders would be satisfied. Several experts were consulted to find the most appropriate wording in Danish to complement the English global scale.

The actual official certification program commenced approximately 16 weeks from the start of the test construction.

9. Appendix

Global Scale

The overall certification result is based on a combined assessment of the lecturer's fluency, pronunciation, vocabulary, grammar and interaction skills in English for university teaching

5: The lecturer has demonstrated English language proficiency for university teaching equivalent to that of a highly articulate, well-educated native speaker of English. The lecturer has been certified to teach English-medium courses. No training is required.

4: The lecturer has demonstrated excellent English language proficiency for university teaching. The lecturer has been certified to teach English-medium courses. No training is required.

3: The lecturer has demonstrated good English language proficiency for university teaching. The lecturer has been certified to teach English-medium courses. No training is required, but training may be beneficial in one or more of the assessed areas.

2: The lecturer has demonstrated less than sufficient English language proficiency for university teaching. The lecturer has not been certified to teach English-medium courses. Training is required.

1: The lecturer has demonstrated limited English language proficiency for university teaching. The lecturer has not been certified to teach English-medium courses. Significant training is required.

10. References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. Oxford University Press.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Douglas, D. (2000). *Assessing Language for Specific Purposes*. Cambridge Language Assessment. Cambridge University Press.

ETS/TOEFL. (2010). *Test and Score Data Summary for TOEFL Internet-based and Paper-based Tests: January 2009-December 2009 test data*.

McNamara, T. (2000). *Language testing*. Wiley Online Library.

McNamara, T. F. (1996). *Measuring second language performance*. Longman London.