Life after oral English certification
The consequences of the Test of Oral English Proficiency for Academic Staff for EMI lecturers
Dimova, Slobodanka

Published in:
English for Specific Purposes

DOI:
10.1016/j.esp.2016.12.004

Publication date:
2017

Document version
Peer reviewed version

Document license:
CC BY-NC-ND

Citation for published version (APA):
Life after Oral English Certification: The Consequences of the Test of Oral English Proficiency for Academic Staff for EMI Lecturers

1. Introduction

The increased implementation of English-medium instruction (EMI) programs in non-Anglophone countries in Europe has raised concerns about the quality of teaching and learning. EMI related research has not only focused on students’ preparedness to learn in English but also on lecturers’ language proficiency and preparedness to teach field-specific content in English (Airey, 2011; Jensen, Denver, Mees, & Werther, 2011; Tange, 2010). These concerns have resulted in university mandates to introduce quality assurance measures in the form of language proficiency requirements for incoming students and locally-developed assessments for EMI lecturers (Klaassen & Bos, 2010; Kling & Hjulmand, 2008; Kling & Stæhr, 2011).

The Test of Oral English Proficiency for Academic Staff (TOEPAS) is an oral performance test based on a simulated lecture, which was locally-developed to address such a university mandate for quality assurance at the University of Copenhagen (UCPH) (Kling & Stæhr, 2011). The university management and the union representatives, however, required that TOEPAS did not threaten lecturers’ professional status in their units (e.g., institutes, departments, research centers). Hence, TOEPAS was designed to assist unit leaders’ (e.g., program leaders, department heads) decision-making related to EMI course assignments and to raise lecturers’ awareness about their own English skill strengths and weaknesses. For that purpose, in addition to the holistic score on a 5-point scale ranging from 1 to 5, lecturers receive a video recording of their performance and extensive written formative feedback on their language skills.

The present study applies an embedded case study design to examine to what degree the required “non-threatening” nature of TOEPAS is maintained and how the formative feedback is utilized by lecturers. More specifically, the investigation focuses on lecturers’ (n=24) interpretations and uses of the written feedback, as well as their perceived implications of the TOEPAS results for their professional status. To gain extended understanding of lecturers’ feedback interpretations and uses, quantitative analysis of the linguistic descriptors found in the feedback report was performed (n=400).

To situate the case of the University of Copenhagen within the broader discussion on lecturer EMI preparation in European higher education institutions (HEIs), this paper starts with an overview of research related to EMI teaching and lecturer certification, which is presented in Section 2. In Section 3, the methodology is described through detailed information about participants, instruments, and procedures. The main results are given in Section 4, and Sections 5 and 6 present the discussion and conclusions.

2. English-medium instruction in Europe

Research reports on EMI in European HEIs suggest that non-Anglophone lecturers experience an additional effort when they teach in EMI programs (Airey, 2011; Crawford Camiciottoli, 2004; Dafouz & Núñez, 2009; Morell, 2007; Tange, 2010). In addition to the language-related limitations, EMI lecturer challenges are related to the increased student population diversity, the need for adaptation of pedagogical skills, and the growing
attention to intercultural communicative competence (Airey, 2011; Klaassen, 2008; Tange, 2010; Vinke, 1995; Westbrook & Henriksen, 2011).

Language-related limitations have been ascribed to EMI lecturers’ overall English proficiency, vocabulary range, and style. In terms of English proficiency, research has pointed to English proficiency disparities among the European countries, especially the north-south proficiency divide, where southern Europeans tend to have lower academic English competence due to various socio-educational and linguistic factors (Campagna & Pulcini, 2014; Dafouz & Camacho-Miñano, 2016). One aspect of English proficiency, vocabulary range, has received attention in EMI research. Findings suggest that EMI lecturers tend to have a good command of their domain specific vocabulary, so their communication suffers mostly due to their restricted academic and general vocabulary ranges (Tange, 2010). Stylistically, EMI lectures have a tendency to be formal and dry because lecturers’ speech resembles written communication norms (Thøgersen & Airey, 2011) and lacks nuancing and humor (Tange, 2010; Wilkinson, 2005). Because of these limitations, lecturers need more time and effort for EMI lesson preparation than for lessons in their L1s (Vinke, 1995; Vinke, Snippe, & Jochems, 1998). Concerns have also been raised regarding the amount of material covered in EMI lectures as findings suggest that lecturers’ speech rate in their L2 is slower than when speaking in L1s (Thøgersen & Airey, 2011; Vinke, 1995).

The language-related issues are less acute with lecturers who have stronger English skills and experience teaching in international settings (Vinke, 1995). Tange (2010) found that these lecturers experienced more issues with the cultural aspects of the EMI classroom. Similarly, Björkman (2010; 2011) argues that pragmatic ability, rather than proficiency, has importance in the EMI classroom, which represents an English as a lingua franca (ELF) setting. In other words, highly proficient speakers might not be effective communicators if they lack pragmatic strategies relevant for the multi-lingual, multi-cultural setting.

Given the linguistic, pragmatic, and pedagogical issues associated with EMI, concerns about the instructional quality have been raised by the universities, which led to the development and implementation of language policies and quality assurance measures across Europe (Ball & Lindsay, 2013; Klaassen & Bos, 2010; Kling & Hjulmand, 2008; Kling & Stæhr, 2011). Despite the challenges with intercultural communication and pragmatics, the emphasis of these quality measures remains on language proficiency.

2.1. Oral English certification methods for EMI

In the initial stages of EMI in Europe, in the early 2000s, oral English proficiency of lecturers was not considered a priority. EMI teaching assignments were distributed based on lecturers’ expertise and experiences abroad, rather than their language skills (House & Lévy-Tödter, 2010). In the following years, fearing complaints from students, university leaders put lecturers’ ability to teach in EMI under scrutiny (see for example, Hellekjær, 2010; Vinke, Snippe, & Jochems, 1998). In the beginning, the requirement for minimum lecturers’ oral English proficiency at some universities was an IELTS score (ranging from 6.0 to 8.0) (Klaassen & Bos, 2010).

However, soon universities initiated development of local assessment methods which were founded primarily on the EMI literature suggesting a threshold proficiency of
C1 on the Common European Framework of Reference (CEFR) (Klaassen, 2001; Klaassen & De Graaff, 2001; Verhelst, Van Avermaet, Takala, Figueras, & North, 2009). For instance, both the Copenhagen Business School and the University of the Basque Country designed assessment methods drawn directly from the CEFR holistic and analytic descriptors, i.e. the Project in Language Assessment for Teaching in English (PLATE) (Kling & Hjulmand, 2008) and the Test of Performance for Teaching at University Level through the Medium of English (TOPTULTE) respectively (Ball & Lindsay, 2013). Some universities utilize other assessments in conjunction to the CEFR criteria (e.g., the Delft University of Technology used the Oxford Quick Placement Test, QTP) (Klaassen & Bos, 2010). Though the literature lacks studies describing the standardization procedures applied to the adaptation of the oral English certification scales to CEFR, universities continue to adopt the CEFR levels for their scales, and level C1 as the cut-score, claiming international transparency and recognition of the results (LACS: Language Communication and Services, 2015).

Face validity is often at the core of lecturer certification assessment procedures, sometimes at the expense of assessment administration and score reliability. Universities emphasize the importance of authentic classroom observation and student evaluations as important elements in the certification procedure, disregarding the construct irrelevant variability these elements may bring (e.g. contextual, didactics, physical, field-specific language variation). Feedback based on the certification results and follow-up language training of lecturers who are not certified represent important elements of lecturer certification (Roskilde University, 2015; University of Freiburg, 2016). Since EMI-related research falls short of investigating the assessment consequences, including the effects of the feedback procedures, the present study sets out to elucidate what kind of impact, if any, EMI certification procedures and results have on university lecturers.

### 2.2. The case of TOEPAS (Test of English Proficiency for Academic Staff)

In 2007, as part of its internationalization strategy “Destination 2012,” the University of Copenhagen introduced a parallel language policy, which required that all “relevant materials and courses will be available in English” (University of Copenhagen, 2007, p. 31). The implementation of this strategy required EMI quality assurance, so the management team at University of Copenhagen (UCPH) commissioned the development of the Test of oral English Proficiency for Academic Staff (TOEPAS). The purpose of the test was to assess whether lecturers had the necessary English skills to cope with the communicative demands of EMI. During test development, various stakeholders at UCPH, including management, head of departments, lecturers, and union representatives, suggested that formative feedback should be included with the assessment result. Thus, in addition to the overall score, the results of the assessment were to include performance description and recommendation for further language training.

The TOEPAS test procedure utilizes simulated teaching performances in a controlled setting. Each TOEPAS test administration lasts roughly two hours and involves the assessment of three lecturers from the same program or area of expertise. Test-takers take turns giving a prepared mini-lecture and participating in a role-play as ‘students’ in order to simulate a graduate classroom setting. As ‘students’, the participants are asked to both interrupt the lecturer (test-taker) during the lecture and to ask spontaneous questions.
after the lecture. The shared expertise of the lecturers allows them to engage in meaningful interaction about the specialized topic. The test-takers' lectures are based on a subject or topic that they typically teach.

All TOEPAS administrations are digitally video recorded for use in both assessment and feedback. Two trained raters rate the test takers’ performance on a 5-point holistic scale based on their observation of the live performance and, in some cases, review of the video. In the period between 2009 and 2012, when the first version of the TOEPAS scale (TOEPAS 1.0) was in use, a result was reported as an overall holistic score on a scale from 1 to 5, which was derived from analytic descriptors of the following criteria: fluency, pronunciation, vocabulary, grammar and interaction skills. The overall score was not a sum or an average of scores received for each criterion, but provided general, global evaluation of the performance (Table 1).

Table 1: TOEPAS 1.0 Assessment Scale

<table>
<thead>
<tr>
<th>TOEPAS Global Assessment 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</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
</tr>
<tr>
<td>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.</td>
</tr>
</tbody>
</table>

Once assessed, each examinee receives a detailed feedback report, which provides a description of their language performance characteristics in relation to the five criteria, fluency, pronunciation, vocabulary, grammar, and interaction. In addition to the written feedback, examinees receive a copy of the digital video recording of their performance, and an invitation to a follow-up meeting with one of their examiners. TOEPAS test takers
ultimately receive a certificate with their holistic result, alongside a detailed feedback report that discusses the strengths and weaknesses of their performances. The report is accompanied by a digital link to test-takers' recorded performances.

2.3. Formative feedback

Formative feedback report is a central feature of TOEPAS although it is generally associated with classroom, rather than standardized assessment. Shute (2008) defines formative feedback as, “information communicated to the learner that is intended to modify the learner’s thinking or behavior for the purpose of improving learning” (p.1). The challenge, though, is to understand what feedback features and delivery methods would yield success in learner’s thinking and behavior because research findings have been inconsistent and contradictory (Azevedo & Bernard, 1995; Kluger & DeNisi, 1996).

The effects of formative feedback have been attributed to a number of variables such as functions, specificity, complexity/length, timing, and cognitive load. Regarding its functions, feedback can be directive or facilitative (Black & Wiliam, 1998). Directive feedback provides specific information about the parts of the task that need improvement. Facilitative feedback, though, is less specific than directive feedback because it gives general comments and guidelines related to the need for revision of the task. Research suggests that specific feedback reduces uncertainty and minimizes the cognitive load needed to perform the task, especially if the task is conceptual and procedural (Phye & Sanders, 1994). However, the assumption that the levels of detail and feedback complexity tend to correlate negatively with learners’ motivation to use the feedback has not been consistently confirmed (Dempsey, Driscoll, & Swindell, 1993; Mason & Bruning, 2001; Narciss & Huth, 2004). Findings regarding timing are also mixed. Some research findings support immediate feedback, i.e. feedback occurring immediately after the task or test performance because it prevents erroneous encoding in memory (Kippel, 1974; Newman, Williams, & Hiller, 1974; Phye & Baller, 1970). Others argue that delayed feedback, i.e. feedback occurring later (minutes, hours, days) relative to task or test performance, has stronger effects because it reduces error interference; initial error is forgotten by the time feedback is received, so the feedback, rather than the error, is preserved in the memory (Kulhavy & Anderson, 1972; Surber & Anderson, 1975). Despite the dichotomous description of the above feedback variables, all of them can be analyzed as part of a continuum.

When it comes to the content of formative feedback, research corroborates the assumption that feedback is effective if it includes elements of both verification, i.e. confirmation of correctness of response, and elaboration, i.e. explanation why the response is correct or not and how it could be improved (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Mason & Bruning, 2001). Though explanation can be directed to the topic, the response, specific errors, examples, or guidance, research asserts that response-specific feedback is far superior with regard to learning (Corbett & Anderson, 2001; Gilman, 1969; Mory, 2004).

Using the descriptive variables above, the TOEPAS formative feedback can be characterized as facilitative due to the general descriptors and suggestions regarding lecturers’ oral English skills, though supported with specific examples and quotations directly taken from the test. It is quite complex, covering detailed linguistic information, which may be cognitively taxing. Due to the established administrative procedures of
TOEPAS which require high quality and precision, the formative feedback is sent to test-takers (lecturers) two weeks after the test, which means it is delayed.

3. The study

The main objective of this study was to examine UCPH lecturers' interpretations and uses of the TOEPAS written feedback report, as well as their perceived implications of the TOEPAS results for their professional status. Research questions 1 and 2 guided the investigation of the extent to which the TOEPAS feedback was exploited, while 3 elicited the implications of the TOEPAS results for the lecturers' status:

1. To what extent do lecturers use the video and the TOEPAS written formative feedback?

2. How extensive is the use of linguistic terminology in the written report descriptors? How does the use of linguistic terminology affect lecturers' comprehension of the feedback?

3. How are the TOEPAS results used by lecturers and their institutions? How do these uses affect lecturers' personal and professional status at their units?

A case study approach (Merriam, 2007) was used to investigate the research questions. This approach allows for framing the interviews within the specific context so that lecturers' perceptions and interpretations could be understood with respect to the local policies and historical circumstances (Merriam, 2007; Yin, 2013). More specifically, an embedded case design was utilized, where several units of analysis (EMI lecturers) were examined within a single case (University of Copenhagen oral English certification) (Yin, 2013). This design allows for a broad but focused analysis of the case, where the units of analysis are not holistic but still context dependent (Yin, 2013). The embedded case design also lends itself to an instrumental rather than intrinsic approach. While an intrinsic case study concentrates completely on the particular case, an instrumental approach is applied to shed light on broader concerns extending beyond the focal context (Stake, 1995). Two main data collection procedures were undertaken: analysis of TOEPAS reports with written feedback and semi-structured interviews with test-takers. Though quantitative text analyses are not traditionally applied in case studies, they are deemed important to understand the feedback characteristic for the particular case.

The aim is not simply to describe perceptions and interpretations within the selected case but to gain insights from the particular context in order to form more general assumptions about the consequences and uses of EMI lecture certification at HEIs.

3.1. Participants

Participants were 24 teaching staff members (assistant, associate, and full professors, as well as post-docs and Phd students) who took part in the certification process across different faculties and departments at the University of Copenhagen (6% of the total tested). They were randomly selected from each scale level in the TOEPAS
database, which was designed in MS Access. All participants were telephoned for invitation and scheduling of the interview time. Table 2 shows the background characteristics of the participants.

Table 2: Participant background information

<table>
<thead>
<tr>
<th>Years at UCPH</th>
<th>range=1--36 years, median=11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>F=5; M=19</td>
</tr>
<tr>
<td>L1</td>
<td>Danish=19; Swedish=2; Portuguese=1; Dutch=1; German=1;</td>
</tr>
<tr>
<td>Department</td>
<td>Computer Science</td>
</tr>
<tr>
<td></td>
<td>Department of Forest and Landscape</td>
</tr>
<tr>
<td></td>
<td>Institute of Food and Resource Economics</td>
</tr>
<tr>
<td></td>
<td>Department of Agriculture and Ecology</td>
</tr>
<tr>
<td></td>
<td>Department of Food Science</td>
</tr>
<tr>
<td></td>
<td>Department of Large Animal Science</td>
</tr>
<tr>
<td></td>
<td>Department of Human Nutrition</td>
</tr>
<tr>
<td></td>
<td>Law</td>
</tr>
<tr>
<td>Position</td>
<td>Full Professor (n=6); Associate Professor (n=13); Assistant Professor (n=2); Post-doc (n=1); PhD student (n=2)</td>
</tr>
<tr>
<td>Score</td>
<td>5 (n=4); 4 (n=10); 3 (n=9); 2 (n=1)</td>
</tr>
</tbody>
</table>

3.2. Interviews and written feedback reports

Two data types were collected to answer the questions related to the consequences of the TOEPAS, i.e. lecturers’ feedback comprehension and their perceived personal and professional effects. The main data were gathered through in-depth, semi-structured interviews (Seidman, 2013) with randomly selected academic staff members who had taken the TOEPAS. These were supplemented with written report data to gain an improved understanding of the feedback report content.

Semi-structured interviews seemed most appropriate for this study because they were based on a framework of themes that needed to be covered (language policies, feedback effects, certification impacts), but also allowed for exploration of new ideas and themes which might come up during the interview (Mason, 2004).

The semi-structured interview consisted of 23 basic questions to four different themes: background information, video from test performance, written report with formative feedback, and perceived test impact. Additional questions were raised during interviews to further explain certain statements or to obtain more details regarding specific cases.

The background information questions sought information related to participant’s experience with teaching in EMI programs, international students, communication practices in and outside the classroom, and understandings of the university’s quality assurance policy. In addition, participants were asked about the existence of local language policies, i.e. language-related decisions made within their research group, lab, or department.
The video and written report questions related mostly to whether and how participants employed them to improve their language skills for the classroom. In particular, the questions elicited information about whether participants sought language instruction, self-learning, or self-monitoring. Here, participants were also asked to what degree they understood the technical terminology used to describe their performance in the written report. This part of the interview also served to identify the specific terminology which became focus of the written report analysis.

The last set of questions inquired about the impact of TOEPAS on the individual participants. More specifically, participants were questioned about how certification results were exploited and the degree to which the certification affected their teaching assignments and status in the department.

Once the problematic performance descriptors were identified during the interviews, content analysis of all written reports from the TOEPAS database was conducted to examine the extensiveness of the linguistic term presence (n=400). The written reports contained general description of test-takers’ oral performances in relation to five different aspects of oral production, namely fluency, pronunciation, grammar, vocabulary, and interaction. The descriptions were supported with examples from test-takers’ videos. The examples were transcribed quotes that contained the feature discussed in the description. The following example shows a description and examples of the test-takers’ use of discourse markers:

<table>
<thead>
<tr>
<th>He produces a broad range of discourse markers to show the structure and organisation of his presentation and to highlight important points. Moreover, he uses rhetorical questions to good effect. Some examples observed in the presentation include:</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Let’s talk about...” <em>(introducing topic)</em></td>
</tr>
<tr>
<td>“Let me give you an example...” <em>(giving example)</em></td>
</tr>
<tr>
<td>“So what is ‘I’?” <em>(rhetorical question to introduce point)</em></td>
</tr>
<tr>
<td>“Let’s imagine that...” <em>(setting the scene for a particular point)</em></td>
</tr>
</tbody>
</table>

The written formative feedback report examination focused on frequency analyses of 1) words in performance descriptions, 2) definitions or explanations of four linguistic terms identified during interviews (discourse markers, sound contrasts, hesitation, and pragmatic meaning), and 3) examples from actual test performances to exemplify the four linguistic terms in descriptions.

The formative feedback analysis followed the interviews with lecturers to confirm the extent of the written feedback effects found during the interviews.

### 3.3. Procedures

The study was conducted in four stages: (1) interview question design, (2) procedures and questions piloting, (3) interview administration, and (4) written formative feedback report analysis.

#### 3.3.1. Interview procedures
The interview questions were piloted with randomly selected participants. During this stage, the participant recruitment and interview procedures were finalized and the interview questions were trialled.

Interviews, which lasted around 30 minutes, were conducted in participants’ offices at their own convenience. They were audio recorded on a MacBook laptop using GarageBand™ as a recording application. The recordings were subsequently transcribed, coded, and analyzed with NVivo 10™. The coding procedure began with data segmentation based on turns as individual segments because participants rarely shifted topic without being prompted by a question. Since interpretation of some categories was difficult without the context provided in other, more than one code was applied in certain segments. Three broad categories, and several subcategories, were teased out from the data:

Language policy
- Participants’ familiarity with university language policy
- Existence of local (departmental) language policies
- TOEPAS in relation to the policy

Effects of formative feedback
- Uses of video
- Understanding and uses of written report
- Type of language support sought based on TOEPAS recommendation

Impact of TOEPAS
- TOEPAS result uses
- Consequences at personal level
- Consequences at professional level
- Consequences at university level

3.3.2. Formative feedback report analysis

During the interviews, four language-related terms were identified as difficult to understand by lecturers, namely “pragmatic meaning(s),” “sound/phonological contrasts,” “hesitation,” and “discourse markers.” Therefore, the written formative feedback reports (n=400) were analyzed in NVivo 10™ to investigate the frequency and the applications of these terms. The analysis started by inspecting the TOEPAS rating scale grid for linguistic terms used to describe performances at different levels to examine whether and how those four terms were used. The inspection confirmed that these terms were applied as performance descriptors at different levels of the TOEPAS scale, so the degree to which these terms were also included in the written reports needed further examination.

A word frequency analysis was also performed to rank all content words used in the feedback reports. Content words were defined as lexical words (nouns, verbs, adjectives, adverbs) while all function words (prepositions, articles, conjunctions) were eliminated. The word count was based on all occurrences which shared the same stem. For example, the word “sound” occurred in several different forms (sound, sounds, sounded, sounding) for a total of 1,207 times and represented 0.9% of all words the data set. The main reason for this frequency analysis was to detect the number of linguistic terms, including elements focal terms above, among the 50 most frequent words found in the reports.
4. Results

4.1. Uses of the TOEPAS formative feedback

As mentioned earlier, lecturers receive a score, a video recording, and a written formative feedback report after the TOEPAS certification session. While the purpose of the TOEPAS score is reported to the lecturers' department or institute leaders, the intended outcome of the video recording and the written feedback is to raise awareness about lecturers' oral English skills in relation to the EMI teaching context. To gain an improved understanding of the uses of the video and written feedback, lecturers were asked about whether and how they used them in their personal and professional development.

4.1.1. Uses of the TOEPAS videos

Based on lecturers' reports, only 13 (54%) of the interviewed lecturers kept the video recording of their own TOEPAS session, and only 5 (21%) actually watched it. Lecturers felt uncomfortable watching their own video because they were overly self-conscious. They focused more on their gestures, body language, and presence, rather than the characteristics of their speaking performance, which was the main focus of the feedback. For example, one lecturer claimed that he watched the video “not so much about English but more about my appearance and my gesture. Concerning the language, it didn’t really affect me much” (P11, Veterinary Disease Biology).

Those who did not watch the video stated they were not used to being “filmed” and they felt “awkward.” For example, one lecturer stated,

I found…the feedback report sufficient…. watching yourself on the video …I think it would be embarrassing…Perhaps you get aware of some body language or something that … you think it’s funny, or perhaps you think it’s awkward…so better not be aware of it. (P04, Geosciences and Natural Resource Management)

And in a similar manner, another lecturer mentioned, “I don’t know. Maybe I would...be uncomfortable to a certain extent…Well because I’m not as good looking as I have imagined” (P02, Plant and Environmental Science).

Another reason lecturers provided to justify not watching the videos was that the written formative feedback report was more informative and sufficient.

4.1.2. Uses of the written formative feedback

Indeed, results suggest that the interviewed lecturers did keep and read the written feedback report (n=19), and 16 remembered in general terms what the report stated about their oral English performance. This was confirmed by checking the performance descriptions lecturers provided against their actual written reports.

According to the lecturers, the specific examples from their own performances in support of the performance descriptions were very useful. For instance, if they did not know the term “pronunciation errors”, they would check the examples of pronunciation errors from their performance, and that is what they remembered.
However, they experienced problems understanding some of the terminology used in the description. During the interviews, four terms were found particularly difficult to understand: *hesitation(s)*, *discourse markers*, *pragmatic meaning*, and *sound/phonological contrasts*. While 18 of the interviewed lecturers believed they understood *hesitations*, a much lower number understood the rest of the terms.

When asked to define these terms, the lecturers confirmed their understanding of *hesitation(s)* by providing reasonable explanations, mostly referring to filled and unfilled pauses. For example, one lecturer stated that *hesitation* occurs "[i]f you make uncalled breaks or pauses to search your memory for words" (P12, Food Science).

Those lecturers (n=5) who claimed to understand the term *discourse markers* also provided cogent definitions and explanations, which were mostly associated with directions in speaking and writing. For example, lecturers stated that, "[D]iscourse markers are different words or phrases that indicate which direction you want to go" (P13, Food and Resource Economics) and that, "[I]t is … words that somehow point in directions" (P14, Plant and Environmental Sciences).

However, the lecturers were unfamiliar with the terms *pragmatic meaning* and *sound contrasts* even though some tried to define them based on the general meaning of the words. For example, while one of the lecturers tried to interpret the term as, “[p]ragmatic means practical—if you are pragmatic you adapt to the situations by the means available or something like that” (P09, Department of Resource Economic and Food Policy), another lecturer admitted that he was familiar with the words, but not the term saying, “I know what ‘pragmatic’ means and I know ‘meaning,’ but that constellation is not immediately meaningful for me” (P12, Food Science).

The term *sound contrasts* was even less transparent for the lecturers, so all their attempts to define it revolved generally around pronunciation. As one of the lecturers said, "I suppose it relates to pronunciation, but I don’t know exactly what sound contrasts are" (P02, Plant and Environmental Science), and another thought that it “must have to do with intonation… or something like that… or playing with words...” (P18, Food Science).

Due to the finding that the terminology found in the written formative feedback report challenged the lecturers’ comprehension, while examples from their own performances were more useful, the feedback reports were subsequently examined in

![Fig. 1. Comprehension of linguistic terms](image-url)
relation to what terms were frequently utilized, whether the difficult terms were defined, and whether these terms were accompanied by examples from the test performances.

The word frequency analysis in NVivo 10™ revealed the following language-related terminology among the top 50 words presented in Table 3.

Table 3: Word frequency of linguistic terms

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
<th>Weighted Percentage (%)</th>
<th>Similar Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>interaction</td>
<td>1222</td>
<td>0.91</td>
<td>interact, interacting, interaction, interactional, interactions, interactive, interactively, interacts</td>
</tr>
<tr>
<td>sound</td>
<td>1207</td>
<td>0.90</td>
<td>sound, sounded, sounding, sounds</td>
</tr>
<tr>
<td>vocabulary</td>
<td>1063</td>
<td>0.79</td>
<td>vocabulary</td>
</tr>
<tr>
<td>pronunciation</td>
<td>1048</td>
<td>0.78</td>
<td>pronounce, pronounced, pronounces, pronouncing</td>
</tr>
<tr>
<td>pronunciation</td>
<td>924</td>
<td>0.69</td>
<td>pronunciation, pronunciations</td>
</tr>
<tr>
<td>meaning</td>
<td>859</td>
<td>0.64</td>
<td>mean, meaning, meanings, means</td>
</tr>
<tr>
<td>hesitation</td>
<td>670</td>
<td>0.50</td>
<td>hesitate, hesitates, hesitation, hesitations</td>
</tr>
<tr>
<td>stress</td>
<td>570</td>
<td>0.42</td>
<td>stress, stressed, stresses</td>
</tr>
<tr>
<td>domain</td>
<td>523</td>
<td>0.39</td>
<td>domain</td>
</tr>
<tr>
<td>discourse</td>
<td>504</td>
<td>0.38</td>
<td>discourse</td>
</tr>
<tr>
<td>markers</td>
<td>500</td>
<td>0.37</td>
<td>marker, markers</td>
</tr>
<tr>
<td>context</td>
<td>447</td>
<td>0.33</td>
<td>context, contexts</td>
</tr>
<tr>
<td>fluency</td>
<td>445</td>
<td>0.33</td>
<td>fluency</td>
</tr>
</tbody>
</table>

Elements of the four difficult terms (hesitation, discourse markers, pragmatic meaning, sound/phonological contrasts) were all found among the top 50 words. However, except for hesitation, which seemed to be most frequent with 670 references found in 383 written reports, the other three were two-word terms, so additional frequency analyses were performed. According to these, discourse markers closely followed hesitation(s) with 475 in 397 reports. Unlike discourse markers and hesitation(s) which could be found more than once in a given written report, pragmatic meaning(s) (n=267) and sound/phonological contrasts (n=128) occurred only once.

The next step was to investigate whether these four terms were used together with definitions or explanations to ensure feedback comprehensibility for lecturers. Based on the NVivo™ frequency queries, definitions of these four terms were rarely found in the written reports, though sometimes they co-occurred with related, more commonly used, terms. The term hesitation(s) was most commonly used in the expression “language-related hesitation” (n=368), which was directly taken from the TOEPAS grid with descriptors. Other descriptors qualifying hesitation were “undue,” “natural,” “unnatural,” “occasional,” “slight,” “some,” and “moderate.” No instances were found where the meaning of “language-related hesitation” was defined, though this expression was usually used when describing “speech rate,” “false starts,” and “pauses,” as well as “smoothly,” “coherently,” and “effortlessly.”
Like hesitation(s), the term discourse markers was also commonly used as part of longer expressions, i.e. “a wide range of discourse markers” (n=182) and “a variety of discourse markers” (n=135). Discourse analysis was also used with “connectors,” as in “discourse markers and connectors” (n=250). Although only one explicit attempt for clarification of the term was found in the written reports (“transition markers”), sometimes the purpose of discourse markers was defined, hence explaining the term to a certain extent.

The most common expressions in which the term pragmatic meaning (n=267) was found were, “convey basic pragmatic meaning” (n=143) or “a variety/range of pragmatic meanings” (n=83). This term is even less frequently defined and explained than the terms hesitation and discourse markers. The word that is most frequently used with this term is “intonation,” as in “use intonation to convey basic pragmatic meaning” (n=102). No definitions or explanations of this term were found in the written reports.

Examples from the performances to support statements related to pragmatic meaning were also rare. Only four examples (1%) of all written reports had examples provided in parenthesis, e.g., “(such as new topic; continuation of utterance; utterance finished),” “(such as end of utterance; new topic; utterance continued).”

The term sound/phonological contrasts (n=127, sound n=65, phonological n=62), was the least frequent term of the four. Unlike the other three terms, this one was not commonly found in regular set expressions. The most recurrent expression with this term was “produces English sounds and sound/phonological contrasts...” with only 18 occurrences. No definitions of the term were found in the written reports though examples from performances were common in almost all reports where the term was found. For example, the lecturer

“produces many English sound contrasts with reasonable accuracy. Exceptions are between the ‘b’ and ‘v’ sounds (so that, for example, ‘biotic’ sounds like ‘viotic’), between ‘i’ and ‘ee’ (so that the words ‘this’ and ‘these’ are pronounced the same), and between the ‘th’ and ‘t’ sounds (so that ‘three’ sounds like ‘tree’).”

The written formative feedback analysis suggests that the performance descriptions contained unfamiliar terms for the lecturers, and that these terms were not always defined and supported with examples from the performance. Table 4 summarizes how the four terms were exploited in the written feedback reports.

Table 4: Uses of hesitation, discourse markers, pragmatic meaning, and sound contrasts in written feedback reports

<table>
<thead>
<tr>
<th>Term</th>
<th>Modifiers</th>
<th>Explanations/Descriptions</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hesitation</td>
<td>• Undue</td>
<td>• Smoothly</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>• Unnatural</td>
<td>• Coherently</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Occasional</td>
<td>• False starts</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Slight</td>
<td>• Effortlessly</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Moderate</td>
<td>• Speech rate</td>
<td></td>
</tr>
<tr>
<td>Discourse markers</td>
<td></td>
<td>• Signal structure and emphasis</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Show organization and emphasis</td>
<td></td>
</tr>
</tbody>
</table>
4.2. Uses of TOEPAS results

The intended use of the TOEPAS scores was making internal decisions about teaching assignments in EMI courses at the different departments. The interviews revealed that the department management reportedly used the scores consistently with the intended use, though the management (department heads) did not explicitly discuss the results, or possible follow-up support, with any of the lectures who were certified. The few lecturers who sought English support post certification received it upon request from the management.

Almost half of the interviewed lecturers either used, or thought of using, the TOEPAS results for other purposes: CV, promotion, job applications, grant applications, and personal websites. Nine lecturers reported the score on their CVs when applying for promotion at the UCPH or other jobs or for project funding from granting agencies.

...if you want to shift from one position to another, then it’s nice to have it on your CV, I think... I appreciate it they have done this at this university because it really shows commitment to internationalization of the university, and at least they are guarantee that at the University of Copenhagen we want to have certain standards when we deal with international courses and education. (P18, Food Science)

Some interpreted their scores as insufficient oral English proficiency even though they were able to meet the cut-score and become certified, “No I think I got a 3, and that’s kind of medium, so I don’t think that’s OK” (P14, Plant and Environmental Sciences).

Standardization of the TOEPAS scores emerged as an issue pertaining to score use beyond UCPH. One lecturer stressed the challenges to understand the meaning of the scores as the TOEPAS scale is localized and limited to the UCPH uses. He mentioned that in an unofficial discussion,
...in the Netherlands in Amsterdam with a professor in Philosophy... who had gone through a similar certification where a score system they used was different, so we tried to figure out what the connection between these was I think this one [TOPEAS] was 1 to 5 or some a five-tiered system, and he used something that was supposedly a European standard. (P01, Computer Science)

Although explicit reference to a particular European standard was not offered, based on the EMI literature, the assumption is that the lecturer meant the Common European Framework of Reference for Languages (CEFR), a framework which has gained a wide range of applications in Europe and beyond (see Section 2.1.).

4.3. Professional and personal consequences of TOEPAS

The professional and personal consequences depended on the TOEPAS result uses discussed above. The term consequences applied to the identified changes in lecturers’ teaching assignments and practices and individual emotional and attitudinal effects. At professional level, though the TOEPAS results raised lecturers’ awareness about the characteristics of their performances, they did not affect significantly the lecturers’ teaching assignments. Lecturers’ self-perceptions of their own English speaking ability equaled the TOEPAS results. Eight lecturers explicitly stated that the test confirmed their self-perceptions, which increased their confidence levels,

...as I said before my expectations about my ability were quite similar to what they [raters] judged... so...it was just a reconfirmation of my ability...I would have got disappointed obviously if it was lower than 4. And if it was 5, I would be surprised because obviously I know I’m not fluent in English. (P04, Geosciences and Natural Resource Management)

Though ten lecturers asserted that the score and the feedback report increased their awareness about their oral English skills, only four paid more attention to some of the problems pointed in the written report when teaching.

When it comes to teaching assignments, no adjustments to lecturers’ involvement in EMI courses was found after TOEPAS (100%). Most certified teachers had taught EMI courses before TOEPAS, and they continued teaching the same courses post certification. Similarly, those lecturers who failed to become certified had already limited participation in EMI programs, and that remained unchanged until they were certified.

Most lecturers were confident that their teaching career at the university was not compromised due to their TOEPAS results because they had been hired as experts in their fields, and they were the most knowledgeable about the course material. When asked about possible fears about their teaching status, one lecturer claimed,

No, because we are teaching according to profession...they [management] just can’t go out and find somebody else who can do what I can do, so they’ll have to manage what I can do. (P03, Food Science)
The lecturers believed that their departments would support, rather than replace, them if their TOEPAS results suggested that their English proficiency was not adequate for teaching EMI courses. One lecturer mentioned,

If I didn’t pass then, well, I don’t know how it would be in practice, but I think the institute’s policy is to offer help. I seriously doubt that they’d actually say ‘you didn’t pass so you cannot teach this course. We’ll find someone else’. (P10, Department of Nutrition, Exercise and Sports)

From a personal perspective, results suggested that the greatest impact of failing to become certified was embarrassment and disappointment. However, the interviews also revealed certain categories of lecturers who were particularly vulnerable. Though initially oral English certification was mandatory for those who had permanent positions, i.e. associate professorship and above, in some departments certification was required also for assistant professors, postdocs, and PhD students who taught EMI courses. Due to the temporary status of their positions (assistant professors are hired on a three-year contract), three lecturers expressed fear of failing to become certified. The unknown consequences of failing was one of the reasons lecturers mentioned,

...obviously you are judged on something that...you are going to use it in your professional career. There is a kind of...at least for me, it was a kind of nervous atmosphere...was a little bit anxious to see the test results because if you had bad test results, what do you do then? (P04, Geosciences and Natural Resource Management)

Another unaccounted group of lecturers appeared to be international lecturers whose Danish proficiency was insufficient. The consequences for this group were potentially much more serious because they lacked the option to teach in Danish if not certified for English. These consequences were even more exacerbated for international lecturers at postdoc or assistant professor level,

Lecturer: I guess they would have fired me. I don’t know. My Danish is much worse than my English so then, I don’t know what would have happened. Interviewer: Were you worried? Lecturer: Yea, a little. I had another problem--they wouldn’t have extended my contract if I was not certified... (P13, Food and Resource Economics)

5. Discussion

Formative feedback has gained popularity as part of oral English assessment procedures and EMI training and support programs at many European HEIs even though very little is known about its effectiveness. The present study contributes to our understanding of formative feedback revealing that the convergence of the expected and the actual formative feedback uses is not always warranted, especially if it does not take into account users’ background. For instance, in the present case, the TOEPAS’ performance video and written report were supposed to provide extensive formative feedback with the
purpose of raising lecturers’ awareness of their oral English skills needed to teach in the EMI classroom, but they remained largely inaccessible due to the linguistics-specific terminology and the style of the performance descriptions. Given lecturers’ unfamiliarity with the terminology, they seemed to obtain a fragmented picture of their performances, mostly construed from the specific performance examples, if such examples accompanied the description. These findings substantiate previous research on feedback which delegates feedback effectiveness to the type of the informative aspects of its content and the presentation of its components, as well as to the learner content knowledge and motivation (Shute, 2008). If the content of the feedback departs from learners’/users’ knowledge base, the feedback effectiveness will suffer, as is the case with TOEPAS. Nonetheless, results provide further evidence about the effectiveness of feedback specificity (e.g., reference to specific elements of the performance and inclusion of detailed examples), since lecturers remembered the examples and quotations taken from their own simulated-lecture performances (Phye & Sanders, 1994).

The intended use of the video was to ground the written report in the teaching context and offer supportive evidence for the report descriptions; nevertheless, the lecturers seemed to find it obsolete and uninformative. The absence of information about the performance video utility and the lack of reference to the video in the written reports resulted in its underutilization. As literature suggests, the purpose of the formative feedback components needs to be explicated and goal-oriented to achieve positive effects and incite meaningful learning (Shute, 2008). Though the TOEPAS written report delineates the difference between the current level of performance and the desired performance goals, it fails to establish a connection with the video.

The broader implications of these findings, both for EMI certification and EMI support programs, suggest that successful language-oriented feedback demands detailed non-technical description of the language performance supported by language examples from the specific teaching context of the academic discipline. The functions of all feedback elements (performance descriptors, examples, video), however, need to be brought to lecturers’ attention in order to maximize their effectiveness.

Given the HEIs increased interest in development of oral English certification procedures, it should be highlighted that implementation of such procedures needs careful exploration of possible misuses, as well as negative consequences for the involved stakeholders. Of special importance is provision of explicit information about the intended test-taker groups and the expected uses for which the certification is designed. In the UCPH case, for instance, the TOEPAS test users, management and lecturers, largely complied with the proposed inferences regarding the lecturer population at UCPH. Lecturers who met the cut-score were involved in EMI courses, while those who failed maintained limited EMI participation. Most lecturers interpreted the passing results as sufficient oral English proficiency to teach in EMI courses, so they opted not to seek language support despite the recommendations.

Nevertheless, findings also exposed sub-populations of lecturers who might experience negative outcomes as a consequence of the TOEPAS results. The intended population for which TOEPAS was designed was Danish lecturers with permanent positions at the university. The expected consequence of failing to obtain certification, i.e. to meet the cut-score, was that lecturers continue teaching in Danish while receiving English language support. However, results revealed sub-populations of test-takers
(lectures) who were not accounted for in the intended outcomes of TOEPAS. The test stakes seem higher for a sub-population of lecturers who held temporary positions at the university because the results could influence decisions about their contract renewal or promotion. Another affected sub-population of lecturers comprised international lecturers who had low L2 Danish proficiency. If they fail to meet the required TOEPAS cuts-core, these lecturers will not be able to teach at all. The stakes might be even higher for international lecturers with temporary positions at the university, for which TOEPAS might have adverse effects on their ability to keep their jobs.

Unexpected was the finding that lecturers used their results beyond the purpose and the context for which they were intended. The novel, trans-institutional, trans-national uses of lecturers’ oral English certification results require standardization of the scores to offer greater transparency for the users. If recognition of certification results beyond the local institution, in this case UCPH, is anticipated (e.g., lecturers do not need to be re-certified if they move to a different university), then alignment with more widely known standards, such as the CEFR, is immanent. To gain international and cross-institutional recognition of certification results, the scale alignment requires undertaking established standardization procedures (Figueras, North, Takala, Verhelst, & Van Avermaet, 2005) rather than simple application of the CEFR scale level names (A1, A2, B1, B2, C1, C2), which has been a common practice so far among European HEIs. Despite the detailed scalar descriptions of competences required for effective communication in an L2, the main purpose of the CEFR scale is to provide a framework of reference, so any applications in particular contexts stipulate cautious interpretation and adaptation (Verhelst et al., 2009).

6. Conclusion

Local test design for oral English certification of tertiary education lecturers has amplified in the past decades as a result of the rapid growth of EMI at European HEIs. The implementation of such procedures is expected to have personal and professional impact on lecturers because results are used for decision-making and training purposes regarding teaching EMI courses. To minimize the negative consequences of their implementation, the responsibility of local test-developers is to discover the best way to communicate the meaning and the intended uses of the results to test users (lecturers and decision-makers). Lecturers, then, need to consider the effects of their own decisions to use the scores for purposes other than intended (e.g., lecturers’ use of TOEPAS results for job or grant application) and for the broader educational planning effects in their local units and the university.

Written formative feedback can be a powerful instrument for raising lecturers’ awareness about their own oral English strengths and weaknesses. To ensure its effectiveness, in the TOEPAS case, the feedback format needs to be redesigned to increase lecturers’ comprehensibility. According to the results, linguistic terminology should be avoided, or defined, specific examples from the performance should accompany descriptions, and references to the video should be apparent. The positive feedback outcomes could be intensified with an oral feedback component where the written feedback report descriptions are explained while watching the video of the performance. Test-developers can also contribute to the alignment of TOEPAS scores with more-widely known scales of reference, like the CEFR, to mediate a more transparent decision-making
beyond the local institution. This alignment, though, must follow suggested standardization procedures outlined by the Council of Europe (Figueras et al., 2005), rather than simply borrowing descriptors and scale point labels.

Finally, “EMI teaching preparedness” remains part of EMI-related discussions though certifications continue to focus on language proficiency. The available materials on oral English tests for EMI certification tend to concentrate on descriptions of the test method, emphasizing considerations of face validity during the developmental stages (e.g., authentic classroom observations, student involvement). Findings of this study indicate that attention should also fall on the consequences after the method has been implemented, which will improve understanding of how results are interpreted and used within and beyond their local context.

References


