Assessing EMI Lecturer Language Proficiency Across Disciplines

Dimova, Slobodanka; Kling, Joyce

Published in:
TESOL Quarterly

Publication date:
2018

Document version
Peer reviewed version

Citation for published version (APA):
Dimova, S., & Kling, J. (2018). Assessing EMI Lecturer Language Proficiency Across Disciplines. *TESOL Quarterly*, *50*, 634-656.

Assessing EMI Lecturer Language Proficiency Across Disciplines

Introduction

Oral English assessment of non-native English speaking (NNESs) teachers in higher education (HE) has been the focus of many discussions in the field of Teaching English to Speakers of Other Languages (TESOL). Over the past 20 years, some European universities have mandated certification of all university teachers, ranging from doctoral students to professors, for teaching in English-medium instruction (EMI) courses and programs at traditionally non-Anglophone universities. A number of states in the United States (US) and Canada have had similar mandates for international teaching assistants' (ITAs) oral English proficiency certification much longer, starting as early as the 1980s and 1990s (Oppenheim, 1998; Thomas & Monoson, 1993).

Nonetheless, debates still exist regarding what type of assessment method is most relevant for assessment of language for teaching purposes. While some universities use standardized English for Academic Purposes (EAP) tests, like the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS), others have developed internal assessment methods based on the language for specific purposes (LSP) model which requires simulated, or actual, teaching (e.g., Douglas, 2000; Freibourg, 2016; RUC, 2015). If the LSP model is adopted, the question remains as to whether teaching behavior (pedagogy) and disciplinary content should be part of the assessment and the scoring rubric or whether assessment should focus only on language. Moreover, when the focus is on language, concerns are raised regarding the degree to which teaching strategies and the field-specific content applied

in the observed teaching, or teaching simulation, affect the assessment if the raters are not familiar with the field.

Given the diversity of classroom settings (large lectures, small seminars, lab sessions, etc.) in higher education and the divergence of academic topics and teaching traditions in different academic disciplines, it seems more appropriate that these certification procedures focus on language and the linguistic aspects of pedagogy (pragmatics, intonation, rhetorical signaling) rather than the behavioral, i.e. classroom-management aspects of pedagogy (student involvement, eye-contact, uses of visuals) (Hoekje & Williams, 1992). However, lack of research evidence exists in relation to whether raters can focus only on language and the linguistic aspects of pedagogy and whether the variation of disciplinary content across different academic fields influences their rating behavior (e.g., raters are more severe when they rate performances in some academic fields than other). Moreover, in terms of disciplinary content, little is known about whether and how teachers' disciplinary language knowledge affects their oral performance in a teaching context granted the argument that lecturers need ESP training with special focus on domain-specific, rather than general, vocabulary when language difficulties are detected.

The present study fills these gaps by investigating whether behavioral pedagogy and disciplinary content affect raters' behavior when assessing lecturer performances in an oral English assessment for university lecturers based on a simulated lecture, as well as the characteristics and the perceptions of their domain-specific vocabulary.

For this purpose, the study was guided by questions related to: whether raters of EMI lecturer oral performances are biased towards or against certain disciplines, what aspects of pedagogy are referred to in lecturer certification formative feedback, what types of vocabulary

uses (general, academic, domain-specific) raters identify in the feedback reports, and what EMI lecturers' perceptions are about their disciplinary knowledge and vocabulary use.

Literature Review

Internationalization has been central in the strategic development of most European universities, especially since the Bologna Declaration in 1999, which promotes student and teacher mobility. To enhance the recruitment of international students and faculty members, universities have rapidly been implementing EMI courses and degree programs in addition to instruction in the national and/or the local language(s). The fast growth of EMI programs has posed a number of challenges for the EMI lecturers who are NNESs. Research suggests that EMI lecturers must make additional efforts to compensate for language-related, pedagogical, and pragmatic difficulties (Airey, 2011; Klaassen, 2008; Tange, 2010; Vinke, 1995; Westbrook & Henriksen, 2011). Apart from lack of language proficiency in some contexts (Campagna & Pulcini, 2014; Dafouz & Camacho-Miñano, 2016), EMI lecturers' speech tends to (1) be restricted in terms of academic and general vocabulary (Tange, 2010) and (2) be formal and dry because it resembles written communication (Thøgersen & Airey, 2011) and lacks sophistication and humor (Tange, 2010; Wilkinson, 2005). Additionally, Björkman (2010; 2011) argues that a high level of proficiency is insufficient for effective EMI classroom communication because it requires pragmatic strategies relevant for its multilingual, multicultural context. Due to these issues, university management teams have become concerned about the quality of teaching and learning in EMI and mandated establishment of quality assurance measures, mostly in the form of English proficiency assessments for lecturers (Ball & Lindsay, 2013; Klaassen & Bos, 2010; Kling &

Hjulmand, 2008; Kling & Stæhr, 2012).

Similar concerns about the difficulties in teaching in English exist in relation to ITAs in North America (Kang, Rubin, & Lindemann, 2015). Two major differences, however, exist between the EMI lecturers and the North American ITAs. In EMI, communication occurs primarily between NNESs, as both local and most international students and lecturers have L1s other than English (Kling, 2017). Therefore, the European EMI context in higher education represents English as the lingua franca (ELF) in the academic context, which has distinct characteristics when compared to the Englishes used in traditionally Anglophone universities (Mauranen, Hynninen, & Ranta, 2010). Another difference between the European EMI lecturer context and that of the North American ITAs is that EMI lecturers are already experts in a certain disciplinary field with teaching experience, be that in their L1, another language, or English, while ITAs are still in the process of their academic training. Despite these differences, the development and use of EMI lecturer certification can be supported by the discussions in the ITA literature, especially those related to the selection of the lecturer assessment methods and models.

A number of universities have initiated design of local certification methods for university teachers, most of which are based either on simulated teaching or direct classroom observation (Farnsworth, 2004; Freibourg, 2016; Kling & Hjulmand, 2008; RUC, 2015; Saif, 2002). The development of some of these assessment methods has been governed by findings from research that identifies the specific communicative and interactional demands across different academic settings (Bailey, 1984, 1985; Byrd & Constantinides, 1992; Hoekje & Williams, 1992; Madden & Myers, 1994; Rounds, 1987; Williams, Inscoe, & Tasker, 1997).

The simulated teaching or classroom observation model, also known as the LSP model,

has been popular because it is supposed to elicit the aspects of language use relevant to the target language use (TLU) domain of teaching (Douglas, 2000; Farnsworth, 2013; Wagner, 2016). While some universities practice the strong performance hypothesis of the model by including teaching style assessment, others use just the context of teaching to assess only language while teaching competence is not part of the scoring rubric (Douglas, 2000; McNamara, 1996). The supporters of the strong version of the LSP model argue that pedagogy should be included in the assessment because it cannot be separated from language and is an essential element in ITA training programs (Wagner, 2016). Pedagogy has also gained importance in the EMI lecturer screening because EMI in the ELFA context requires different strategies from pedagogy applied in less linguistically diverse classrooms (Ball & Lindsay, 2013).

The critics of the strong performance hypothesis in the LSP model argue that inclusion of teaching style assessment in ITA and EMI lecturer screening appears discriminatory against NNESs because NESs are not usually subjected to teaching-related certification (Bailey, 1985; Farnsworth, 2013; Hoekje & Williams, 1992; Saif, 2002). In addition, as early as in the 1980s, issues about teaching style generalizations across disciplines were raised because huge variations in what was considered appropriate teaching style were identified (Rounds, 1987). This becomes even more problematic when the raters of oral English tests for university instructors are English as a Second or Foreign Language (ESL/EFL) teachers because their pedagogical values diverge significantly from those of the teachers in the sciences (Byrd & Constantinides, 1988). For instance, research has shown that mathematicians apply different teaching approaches to teach math even when compared to math instruction approaches in closely related fields, such as engineering and computer technology (Byrd & Constantinides,

1992). However, research on whether and how teaching style influences raters' perceptions of test-takers' language remains scarce.

Another important aspect worth considering when designing oral English assessments for university teachers is the actual needs of the local university stakeholders (students, teachers, administrators). In relation to ITAs, some department administrators may need only a language screening procedure because they believe their graduate students are academically prepared to take on the teaching responsibilities in the department and that teaching experience is not a prerequisite for the job (Saif, 2002). The same situation is evident in the European EMI context where all EMI lecturers have been hired by the university because of their disciplinary expertise, and many of them have had rich teaching experience in the local language(s) (Dimova, 2017; House & Lévy-Tödter, 2010); assessment of topical knowledge and pedagogical skills becomes, therefore, superfluous.

The pending question regarding whether or not it is better to assess teaching styles (pedagogy) in addition to language in oral assessments for teachers can be answered by examining the needs of the local HE contexts. From the review of literature on ITAs and EMI lecturers, several scenarios are possible. If the HE institution offers teacher training for all prospective teachers, be they NESs or NNESs, then assessing candidates' teaching skills is redundant as the assessment goal is to assess only whether teachers have adequate English language proficiency to teach EMI courses (Saif, 2002). For example, some programs offer courses or mentoring for all new teaching assistants [e.g., mentoring programs for freshman composition teaching assistants (Pytlik & Liggett, 2002)] so prospective ITAs only need proof of oral English proficiency. Similarly, junior teaching staff at the University of Copenhagen are

required to complete a year-long program in university pedagogy (Teaching and Learning in Higher Education Programme – *Universitetspædigogikum*).

If prospective teachers come from different educational and teaching cultures and are not offered teacher training, then evaluating teaching alongside language skills may be needed to identify those who need support for language, teaching, or both. Moreover, teaching skills could also be evaluated if the assessment is used for hiring purposes, similar to the Occupational English Test (OET), which migrant healthcare professionals take to demonstrate their English proficiency when they apply for jobs in Australia, New Zealand, and Singapore (Lumley, 1998; McNamara, 1997). In other words, those who fail the test do not fulfill the hiring requirements.

In the European EMI courses, the central change for teachers is in the selection of language, a shift in medium. The teachers continue to teach the same content (and at times the same courses that are now taught in English as opposed to the local language) in the same educational context that values the same teaching approaches. English is a transplanted language in the local university settings. In other words, EMI lecturers do not need to adjust to the local educational culture as some of them may have been part of it for years. Unlike the ITA context where ITAs teach mostly NES students whose expectations are that the ITAs will adjust to the local English varieties and educational context, EMI lecturers need to accommodate students that are mostly NNESs with different proficiency levels (Dimova, 2017). These differences need to be reflected upon in the assessment and training procedures universities employ for teacher certification and support.

Methodology

To address issues raised in the field regarding whether and to what degree teaching styles and specific disciplinary content, which may lead to bias in language assessment based on a simulated teaching performance (Hoekje & Williams, 1992; Saif, 2002), data from the Test of Oral English Proficiency for Academic Staff (TOEPAS) were used to answer the following research questions:

(1) To what degree do TOEPAS raters maintain a uniform level of severity across examinees from different departments?

(2) What aspects of pedagogy, if any, do TOEPAS raters refer to in their formative feedback?

Given the discussion about the role of topical knowledge and the argument that lecturers need LSP training with focus on specialized vocabulary when language difficulties are detected, we also examined how disciplinary language knowledge affects lecturers' oral performance in a teaching context:

(3) What types of vocabulary uses (general, academic, domain-specific) TOEPAS raters identify in their formative feedback reports?

(4) What are lecturers' perceptions about the role of their disciplinary knowledge and domain-specific vocabulary use when teaching in English?

Three main types of data were used to answer the research questions, TOEPAS scores, TOEPAS feedback reports, and observation of and interviews with lecturers. A mixed-method approach was employed (Creswell & Clark, 2007), which allowed for a complementary analysis of the different data types. The generalizable findings supported by details from the local context contributed to deeper understanding of the characteristics of the oral English assessment of EMI

lecturers.

Test of Oral English Proficiency for Academic Staff (TOEPAS)

TOEPAS is a locally-developed, performance-based, oral English proficiency test used for certification of academic lecturers for teaching EMI courses at the UCPH. TOEPAS, which became operational in 2009, is based on simulated teaching performances where three lecturers from the same program or area of expertise take turns giving a self-selected prepared mini-lecture and participate in a role-play as "students" in order to simulate a graduate classroom setting. The mini-lecture is based on a subject or topic the lecturer typically teaches. It consists of three main parts: explanation of terms/concepts/processes, giving homework assignment, and answering questions from "students". The test format was designed on the basis of a thorough analysis of the target language use (TLU) domain, which included interviews with the heads of the study boards, discussions with deans, and observation of teaching and short interviews with teachers (see technical report, Kling & Stæhr, 2012).

The TOEPAS scoring rubric was constructed to focus on language, including verbal elements of pedagogy identified in the needs analysis, such as signposting, pragmatics, intonation, and emphasis. In the first version of TOEPAS (2009-2012), each test takers' performance was rated by at least two trained raters (exact agreement 70%, adjacent agreement 99.9%) on a five-point holistic scale. The TOEPAS has been validated through a number of studies analyzing the scale, the rater behavior, the feedback effectiveness, and the test score uses (Dimova, 2017; Dimova & Kling, 2015; Kling & Dimova, 2015). Based on the validation analyses, the TOEPAS scale was revised between 2012 and 2015, and it now has six levels. TOEPAS performances are digitally video recorded for use in both assessment and feedback. Lecturers

receive a certificate with their holistic result, as well as a detailed written and oral feedback report accompanied by a digital link to their recorded performance.

TOEPAS scores

Holistic scores assigned to lecturer performances from 15 different departments (n=340) were analyzed. These scores were assigned by six raters (female=4; male=2), out of the total of nine TOEPAS raters. The scores from the other three raters could not be included in the study because they performed an insufficient number of assessments for the MFRM analyses (<29 out of 400). The six raters were experienced EAP and ESP teachers (a requirement for all TOEPAS raters), four of whom had master's degrees and two had PhD degrees in TESOL, applied, or theoretical linguistics. The raters had high proficiency in English (and other languages) although they had different L1 backgrounds (English=2; Danish=2; Flemish=1; Polish=1). The 15 departments represented in the score sample ranged from Archeology and Ethnology to Plant Biology and Biotechnology, Veterinary Studies, Human Nutrition, Forestry, Math, and Computer Science.

TOEPAS formative feedback reports

All formative feedback reports (n=400) from the TOEPAS database were examined to answer the research questions related to raters' references to pedagogy and vocabulary uses. The feedback reports included general description of lecturers' oral performances in relation to five different aspects of oral production: fluency, pronunciation, grammar, vocabulary, and interaction. An important feature of these feedback reports was a list of transcribed quotes from the lecturer's performance that exemplify and support the performance descriptions. For example, if the description stated that the lecturer failed to use certain words and expressions

effectively, examples of this ineffective vocabulary use from the actual performance would follow.

Interviews with lecturers

To contextualize the quantitative TOEPAS data, interviews were held with 10 Danish L1 lecturers who all have expertise in applied natural sciences and had each obtained the minimum required score for certification (i.e., 3) because difficulties with vocabulary were not expected at the highest scalar levels. All the lecturers were tenured academic staff, with extensive teaching experience both in English and/or Danish. As a group, they had been teaching for an average of 17 years, with an average of 8.7 years of EMI experience.

Data collection took place approximately one to two years post TOEPAS-certification. The process consisted of three one-on-one meetings with each lecturer. Each lecturer was observed and digitally recorded teaching in their own graduate level classroom setting. Following observation, they had a chance to comment and reflect on this teaching event through stimulated recall. Here the lecturers were invited to openly reflect on their performance and comment on aspects they found salient, e.g., language, pedagogy, student interaction. These teacher cognition responses served as a foundation for subsequent semi-structured interviews.

Individual interviews took place, in English or Danish, in the privacy of the lecturers' offices. The semi-structured interview schedule included both scripted questions as well as two forms of card sorting activities that were used to elicit response (Spencer & Warfel, 2004) from the lecturers regarding their reflections on the use of English as a medium of instruction. The process involved sorting and commenting on cards, each marked with some type of content or information, into groups that made sense to the lecturer, particularly in relation to teaching in Danish and English.

At the end of each interview, the lecturers' TOEPAS formative feedback report was reviewed and discussed with focus on the feedback and considerations for competence development. Here the lecturers had the opportunity to comment on the TOEPAS certification administration overall and the effect(s) (if any) of this experience in relation to their teaching.

Analyses

MFRM analysis of TOEPAS scores

TOEPAS data were examined using multifacet Rasch measurement (MFRM) where the facets analyzed were raters, lectures, departments, and scale (Linacre, 2010). MFRM is an extension of the basic Rasch model in that it allows for addition of more than two facets that are not necessarily dichotomous (e.g., the TOEPAS scale is polytomous with five levels of ordered categories) (De Ayala, 2013; Embretson & Reise, 2000; Ostini & Nering, 2006). A separate parameter value represents each observation within each facet, which means the parameters indicate the proficiency of each lecturer, the effect of department, the severity of each rater, and the difficulty of scalar levels.

We selected MFRM for the analysis because we could analyze all facets in the model simultaneously by calibrating them onto a single linear scale with equal-interval points, referred to as log-odds units or logits. In other words, we could directly compare all facets by measuring rater severity on the same scale as examinee proficiency, department effect, and scalar level difficulty. More importantly, with MFRM, we could also explore the possible interaction effects between raters and departments, i.e. we could investigate whether raters were biased against certain departments. Therefore, in our MFRM model, we included an interaction term that

represented to what degree each rater-department pairwise combination may differ from each of their average parameter estimates (Engelhard, 2002). The following is the model:

$$ln\,[\frac{P_{ndjk}}{P_{ndjk-1}}] = \Theta_n - \alpha_j - \gamma_d - \varphi_{jd} - \tau_k$$

Where,

P_{ndjk} is the probability of lecturer *n* from department *d* receiving a rating *k* when rated by rater *j;* P_{ndjk-1} is the probability of lecturer *n* from department *d* receiving a rating *k-1* when rated by rater *j;* Θ_n is the proficiency of lecturer *n,* α_j is the severity of rater *j;* γ_d is the department facet term, φ_{jd} is rater by department interaction term; and τ_k is the difficulty to receive rating of *k* relative to rating of *k-1.*

MFRM provides a bias measure, which is based on the estimate of the interaction parameter for each rater and each department. If the bias estimates are above 0, the observed scores are higher than expected, while if estimates are below 0, the observed scores that are lower than expected according to the model. Dividing the MFRM bias measure by the standard error yields the value of the bias statistic t. A significant t-statistic (i.e., p<= 0.01), or a t statistic with an absolute value greater than 2, denotes substantial rater bias (Engelhard & Myford, 2003).

NVivo analyses of TOEPAS formative written feedback

The TOEPAS formative feedback written reports were analyzed in NVivo to identify references to pedagogy and references to lecturers' vocabulary usage.

To understand the results from the score analyses, all formative feedback reports were searched for specific references to pedagogy, teaching, and classroom behavior (eye contact, use of board, use of slides, etc.) but also for references to particular language tools that are used in

teaching (pragmatics, coherence, appropriateness, etc.).

In terms of lecturers' vocabulary analysis, all references in the vocabulary section of the formative feedback were coded according to whether they were related to 1) general, 2) academic, or 3) domain-specific vocabulary, and according to whether the identified problems were in relation to a) collocation structure, b) morphology, or c) word choice, hence yielding nine possible coding categories. For example, problems could be identified in relation to general collocations (1a), problems with morphology with general vocabulary words (1b), or choice of general vocabulary (1c). Suggestions for precision in vocabulary usage were included as the tenth category.

Collocations were identified as academic if they were found in the *Pearson Academic Collocations* list (Ackermann & Chen, 2013) and general if they were listed in the *Oxford Collocation Dictionary* (Deuter, 2008). If the collocations were not listed in either of these two sources, they were considered domain-specific collocations. Individual words were considered academic vocabulary if they were listed in the *Academic Word List* (Coxhead, 2000) and general if they were found in general English dictionaries. Some domain-specific words did occur in the general English, but they were still considered pre-technical domain-specific words, i.e. they were recognizable by people outside the field (e.g., *chronic disease*).

NVivo analyses of interviews

The stimulated recall and semi-structured interviews with the lecturers were transcribed and translated into English (when necessary) using a denaturalized transcription process (Bucholtz, 2000). After two rounds of open coding in Nvivo, thematic analysis was utilized to identify, analyze, and report patterns and themes (Saldaña, 2015).

The data were then coded using focus on the informant lecturers' reflections to the shift in language of instruction (from Danish to English), their teaching experience and their thoughts about the specific formative feedback they had received from their TOEPAS certification. Several rounds of coding produced three themes focused on the lecturers relationship to the use of English for teaching, their concerns regarding multi-disciplinary, multi-cultural student populations (regardless of language), and their ability to draw on their previous teaching experience. The overarching themes contained several sub-themes, including meta-cognitive reflections on vocabulary use, pedagogy, and compensatory strategies.

Results

TOEPAS facets and interactional bias

As mentioned earlier, the MFRM analysis enabled us to calibrate the parameters for all facets (lecturer, rater, department, scale) on one linear logit scale. The output of the MFRM analysis, referred to as a Wright map, comparatively presents the facets positioned on the logit scale (the first column in Figure 1

Figure 1. Wright map of TOEPAS facets.

According to the map, the lecturers (examinees column) and the departments that occur lower on the scale have lower proficiency than those found at the upper end of the scale. Similarly, the raters (judges column) at the lower end are less severe than those at the higher end, and level 2 of the scale is less difficult than level 5. As the map shows, while almost no variation is present among the departments, some variation is found among the raters, where rater 7 is the most severe and raters 4, 6, and 9 are the least severe. The fixed model chi-square

(5)=36.7, p<.0001 indicates that the raters had different levels of severity after allowing for measurement error. Nevertheless, all raters fit in the model because the Infit Mean Square indices are within the recommended range of .7-1.3 (McNamara, 1996), which means that the raters' behavior is consistent, i.e. not too random or too predictable (see Table 1).

Table 1

Measurement results for the rater facet

Rater	Measure logit	Model S.E.	Infit Mnsq	severity
Rater 9	-1.35	.66	.87	Least
Rater 6	-.79	.36	1.00	
Rater 4	-.78	.39	1.02	
Rater 3	.36	.28	.98	
Rater 5	1.03	.38	.88	
Rater 7	1.53	.38	.87	Most

In terms of rater bias, the MFRM results do not indicate significant interaction effect between raters and departments (see Table 2). In the 74 interactions that were identified, no significant or large t-values (>=2) were found (see Table 2). This means that the scores the raters assigned were based on the consistent use of the language rating criteria rather than chance or criteria not included in the scale (e.g., content, pedagogy).

Table 2

Summary statistics for interaction analysis

Statistic	Department x Rater
N combinations	74
% large t-values	0
Minimum t	-2.31
Maximum t	1.58
M	-0.6
SD	.85

Results from analysis of TOEPAS written feedback report

The statistical analyses failed to identify rater bias against any of the included departments, which means the diversity in teaching strategies and disciplinary content did not affect the rating procedure. To provide further confirmation of this finding, content analysis of the written formative feedback was performed to identify references to pedagogy and different types of vocabulary use (general, academic, and domain-specific). No references to teaching competences and behaviors (eye contact, body language, black-/whiteboard use, or slides) were found in the written reports. However, the raters used performance descriptors, which could be referred to as linguistic aspects of pedagogy, including the pedagogical roles of intonation, vocabulary, interaction strategies, and signposting. These references were grouped into four

main categories:

- Utilization of stress and intonation to convey pragmatic meaning
- Application of varied vocabulary (e.g. synonyms) to explain and emphasize terms
- Employment of interaction strategies to clarify, explain, and/or confirm information in order to make sure that points have been clearly understood
- Use of signposting to organize lecture and to direct students' attention.

The following is a quote taken from an actual written report.

> He generally uses stress and intonation to convey basic **pragmatic meaning**.
>
> …he generally uses correct and appropriate words and can vary vocabulary for **emphasis and meaning**.
>
> When interacting with students, he **restates, clarifies and confirms** information to make sure that points have been clearly understood.
>
> Student: Is this a large scale synthesis [Lecturer: "yes"] that is carried out…?
>
> Lecturer: Yes. You start out by extracting chlorophyll using organic solvents …
>
> Student: so what is the yearly production of this?
>
> Lecturer: I don't know, tons?
>
> Student question: tons scale, [lecturer: "yeah"] wow
>
> (direct response including **back-channeling** followed by **clarification** and **explanation**)

As for vocabulary references, analysis revealed that the identified problems with general vocabulary (n=686) occurred much more frequently than problems with academic (n=58) or domain specific vocabulary (n=49). The most frequent type of problems identified were mis-constructed collocations, inappropriate word choice, and word morphology.

Table 3

Vocabulary use: Written reports

	general	**academic**	**domain-specific**
Collocations	219	14	4
Word Choice	391	25	34
Morphology	76	19	11
Total	686 (87%)	58 (7%)	49(6%)
Suggestions	240		

Results from interview analysis

The interview data corroborate the findings from the TOEPAS feedback report analysis in that it is not the domain-specific vocabulary, but the nuanced use of general vocabulary that is reportedly problematic for lecturers. Lectures associated their disciplinary fields with English, so they did not think disciplinary knowledge and domain-specific vocabulary were problems. They acknowledged that they do experience difficulties with non-disciplinary vocabulary, but they overcome these difficulties by application of compensatory strategies.

The lecturers believed that using English domain-specific terminology was natural when teaching natural science courses because English is the language of science. They reported rare uses of domain-specific terminology in Danish in academic circles, be they in teaching, publishing, or presenting at conferences. Therefore, explaining terminology in English may be

easier than in Danish. For example, some lecturers stated,

> […] I guess, to explain new terminology. That could be relatively challenging. Well, actually, I think that, um, often explaining new terminology might be easier in English because the words are often derived from English literature, and they make sense in English, whereas they may not always make as much sense in Danish. So it could be actually a little more challenging to explain it in Danish than in English. (Informant no. 9)

and

> No, I think that it is completely natural to use English at the university level because it has been the language of science, language of publication for years. In that regard, it is completely natural … it is all in English. (Informant no. 10)

Although the lecturers tended to associate English with their disciplinary fields, some reported occasional lapses in spontaneous domain-specific term retrieval. However, in these situations, they were able to draw on compensatory strategies to find solutions for their linguistic limitations. For example, they used visuals and circumlocution to convey meaning. The statements below illustrate some of these compensatory strategies.

> I use that word, the phrase for parameter. I probably wouldn't do that in Danish. [...] I am lacking some nuance in what I want to say. I would have said it in a more substantive fashion. That word, parameter, is more a technical word. [...] I am sure that I am not always satisfied with my own formulations. (Informant no. 2)

And,

> I have particular problems with terminology, explaining new terminology in anatomy. … I write them up on my slides to help myself! So, this I would not have done in Danish, of

course. (Informant no. 10)

In addition, they claimed that it is not only the language proficiency in Danish or English, but it is their topical knowledge and preparation that facilitates their language use. One of the lecturers commented on topic familiarity and the importance of preparation for teaching,

> When I am searching for a word in English, I can become nervous. But on the other hand, last time I taught in Danish, when it was a new area for me, I was also nervous. So it can be for both languages. (...) It basically depends on if I am well prepared and have thought it through. But there are some times (laugh), honestly, when I suddenly stop and think, which way am I going now. […] the more I have prepared, the less that happens. And that is the same in Danish and in English. (Informant no. 3)

and another stressed the efficiency of developing materials in English,

> Well, actually I would say the last 5 years it [teaching] has been almost all English. To the extent, at least that all materials and all, I mean, I would never make slides in Danish because I can't be bothered. I will, I have so many times realized, OK, I have made a slide or a note in Danish and 3 weeks later I find myself wishing I had done it in English. (Informant no. 8)

While they expressed confidence in their ability to use domain-specific vocabulary because of their disciplinary knowledge, they indicated difficulties to express themselves in a sophisticated and nuanced manner. For instance, a lecturer pointed to his difficulties finding synonyms,

> Sometimes there are pauses that are a little longer than they would be in Danish because I just can't find the word and I have to find another. So I think that my vocabulary isn't that

> large in English, as it could be [...] but it is good enough that we can sit down and hold a conversation [...] it just flows on its own and every now and then there is a little obstacle. (Informant no. 7)

but another noted that this can also happen in Danish,

> [...] is missing some nuances. There are some variations in the language, expressions that every now and then you miss when you are a little unsure about the most precise meaning of a word. Because I think I experience with English that there are words that sometimes are used in certain contexts when I think, oh, hmm, I could use that here. It of course happens with Danish [...] in some contexts you would use a slightly different word. There are so many synonyms in English that I, of course, don't have breadth. (Informant no. 10)

Regardless of the challenges they experience with non-domain specific terminology, the lecturers seemed to maintain their identity as disciplinary experts. One lecturer emphasizes the importance of his content knowledge over his language production,

> […] and it happens in English that I am just missing the word - so I search for some other word to compensate […] to a large extent it is very important for my identity that the students can see that I can also say things that might be wrong, that I might be fumbling for the words - I don't want to appear stupid, but for me it isn't a question of proving anything to the students, showing them how great I am. (Informant no. 6)

Discussion

The findings addressed some of the current concerns regarding whether raters would show bias when rating simulated, or actual, teaching performances across different disciplines--the assumption being that bias could occur due to variation in disciplinary content and pedagogy.

The results from the MFRM rejected the assumption of bias because the raters used the language-related criteria consistently across the departments. The lack of discrimination against performances from any department suggested that external variables, i.e., variables not represented in the measured construct, such as disciplinary content or teaching traditions, did not interact with the raters' behavior. Raters were successfully trained to rate the language that was grounded in the teaching context, which means they applied the rating criteria in the same manner regardless of whether the performance dealt with content from veterinary medicine, math, computer science, human nutrition, or archeology.

The analysis of the TOEPAS written feedback reports supported the finding that raters maintained the focus on the linguistic aspects of the performance, especially those related to classroom-related communication. Although the raters seemed to avoid comments on the teaching materials lecturers used (e.g., slides, board use) or their classroom presence (e.g., body language, eye-contact), they still managed to provide feedback that was relevant for the teaching context. Alongside the descriptions of the linguistic characteristics of the performance, the raters offered feedback on how certain aspects of language use (e.g., pragmatics, emphasis, intonation) could contribute to effective teaching. All written reports contained quotes from the performance that exemplified how lecturers used, or could use, language to explain or clarify difficult material as well as how they interacted with students. Arguably, the observed type of written feedback could be qualified as feedback on the linguistic aspects of pedagogy in the broader sense.

Results from written feedback analysis and lecturer interviews answered the research question related to EMI lecturers' range of vocabulary use. The analysis of references to lecturers'

vocabulary use in the feedback reports confirmed previous assumptions (Tange, 2010) that it was general vocabulary, rather than domain-specific vocabulary, that lecturers needed when they experienced language-related difficulties in their teaching. For example, the raters frequently pointed instances where lecturers overused vague vocabulary words like "thing," "stuff," and "make" in their explanations due to their inability to retrieve more precise general vocabulary words (e.g., "the *things* I find interesting").

This finding was supported by data from the interviews, in which the lecturers exhibited a certain level of confidence related to their disciplinary knowledge and domain-specific vocabulary, while the problems they referred to were primarily of linguistic nature (grammar and vocabulary). Lecturers' open concession to weaknesses in their oral English production, usually associated with lack of ability to speak with sophistication and precision, came from their lack of adequate general vocabulary and idiomatic language use. While confident with their domain-specific vocabulary, in most cases, EMI lecturers seemed to need expansion of their general vocabulary to gain tools for explaining domain-specific terms and concepts to students.

In fact, lecturers claimed to rely on their disciplinary expertise and knowledge of domain-specific vocabulary when teaching, which they tended to value more than the range of their general vocabulary. This echoes the findings of Pecorari, Shaw, Irvine, and Malmström (2011) whose survey of Swedish academics across disciplines found a general tendency for EMI lecturers to place greater importance on domain specific terminology than on general English vocabulary.

The lecturers did not declare experiencing difficulties in teaching because EMI courses maintain the same educational culture while the only change in this educational setting was the

shift of instructional medium. They seemed to preserve the same teacher identity as when they taught in their L1 by applying the same teaching approaches. Unlike ITAs, EMI lecturers are already experts in their fields who frequently present, publish, and participate in international projects and networks in English. In fact, some of the interviewed lecturers claimed that it was even more difficult to discuss some disciplinary terminology in their L1 than in English as they had been exposed to the English terminology throughout their careers (Kling, 2013).

Conclusion

Given EMI lecturers' teaching experience and disciplinary expertise, as well as the different teaching traditions across the disciplinary fields, we argue that evaluation of pedagogical skills and topical, i.e. disciplinary, knowledge is redundant, but the weak LSP assessment model remains relevant because it elicits the language used in the TLU domain. With proper training, raters can focus on the linguistic aspects of the performance even if it is elicited through a simulated lecture, which can be concluded from the finding that raters can rate consistently despite disciplinary and pedagogical variation across the different academic fields. However, the study does not provide clear evidence regarding whether good pedagogues compensate for lack of language skills, i.e. whether raters are more likely to award higher language scores to the lecturers with good pedagogical practice, and vice versa.

Additional research is needed to improve our understanding about the interaction between pedagogy and language. The success of such research depends on a precise operationalization of the term *pedagogy,* which seems overlooked in the current discussions on EMI. It could be differentiated between linguistic pedagogy (pragmatics, intonation, rhetorical signaling) and behavioral, i.e. classroom-management aspects of pedagogy (e.g., student

involvement, teaching activities, eye-contact, uses of visuals). Since the linguistic aspects of pedagogy are already embedded in the TOEPAS rubrics, the study suggests consistency of their application across the disciplines. A lack of evidence exists about the influence of the classroom-management aspects of pedagogy on the performance ratings.

Regarding disciplinary content and vocabulary knowledge, it can be concluded that lecturers are confident in their disciplinary domains, and the problems they experience tend to be related to general vocabulary, which is needed to explain new and/or difficult disciplinary material. Although these findings are based only on perceptions, i.e., lecturer interviews and rater feedback reports, their importance is in that they challenge the traditional assumptions that EMI lecturer language support must be grounded in ESP. In other words, the implication is that focus on general English vocabulary development may be more beneficial for lecturers who need to improve their classroom communication. To confirm these findings, however, analysis of the vocabulary in lecturer performances is needed.

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235-247.

Airey, J. (2011). Talking about teaching in English: Swedish university lecturers' experiences of changing teaching language. *Ibérica, 22*, 35-54.

Anderson-Hsieh, J. (1990). Teaching suprasegmentals to international teaching assistants using field-specific materials. *English for Specific Purposes, 9*(3), 195-214.

Bailey, K. M. (1984). A typology of teaching assistants. In K. Bailey, F. Pialorsi J., & Zukowski/Faust (Eds.), *Foreign teaching assistants in US universities* (pp. 110-125). Washington, D.C.: National Association for Student Affairs.

Bailey, K. M. (1985). If I'd Known What I Know Now: Performance Testing of Foreign Teaching Assistants. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second Language Performance Testing = L'Évaluation de la performance en langue seconde* (pp.153-180). Ottawa, CA: University of Ottawa Press.

Ball, P., & Lindsay, D. (2013). Language demands and support for English-medium instruction in tertiary education. Learning from a specific context. *English-medium instruction at universities: Global challenges (pp.44-64)*.Bristol, UK: Multilingual Matters.

Björkman, B. (2010). So you think you can ELF: English as a lingua franca as the medium of instruction. *Hermes–Journal of Language and Communication Studies, 45*, 77-96.

Björkman, B. (2011). Pragmatic strategies in English as an academic lingua franca:Ways of achieving communicative effectiveness? *Journal of Pragmatics, 43*, 950-964.

Brown, K., Fishman, P., & Jones, N. (1990). *Legal and policy issues in the language proficiency assessment of international teaching assistants*. Houston, TX: Institute for Higher Education Law and Governance, University of Houston Law Center.

Bucholtz, M. (2000). The politics of transcription. *Journal of pragmatics, 32*(10), 1439-1465.

Byrd, P., & Constantinides, J. C. (1988). FTA training programs: Searching for appropriate teaching styles. *English for Specific Purposes, 7*(2), 123-129.

Byrd, P., & Constantinides, J. C. (1992). The language of teaching mathematics: Implications for

training ITAs. *TESOL Quarterly, 26*(1), 163-167.

Byrd, P., Constantinides, J. C., & Pennington, M. C. (1989). *The foreign teaching assistant's manual*. Collier Macmillan.

Campagna, S., & Pulcini, V. (2014). English as a medium of instruction in Italian universities: Linguistic policies, pedagogical implications. *Textus, 1*, 173-190.

Chalupa, C., & Lair, A. (2000). Meeting the Needs of International TAs in the Foreign Language Classroom: A Model for Extended Training. I*n Mentoring foreign language teaching assistants, lecturers, and adjunct faculty. Issues in language program direction: A series in annual volumes* (pp. 119-142). Retreived from ERIC database. [481005]

Coxhead, A. (2000) A New Academic Word List. *TESOL Quarterl*y, 34(2), 213-238.

Creswell, J. W., & Clark, V. L. P. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Dafouz, E., & Camacho-Miñano, M. M. (2016). Exploring the impact of English-medium instruction on university student academic achievement: The case of accounting. *English for Specific Purposes, 44,* 57-67.

De Ayala, R. (2013). The IRT tradition and its applications. *The Oxford Handbook of Quantitative Methods: Foundations, 1*, 144-169.

Deuter, M. (2008). *Oxford collocations dictionary: for students of English*: Oxford University Press.

Dimova, S. (2017). Life after oral English certification: The consequences of the Test of Oral English Proficiency for Academic Staff for EMI lecturers. *English for Specific Purposes, 46*, 45-58.

Dimova, S., & Kling, J. (2015). Lecturers' English Proficiency and University Language Polices for Quality Assurance. In R. Wilkinson, & M. L. Walsh (eds.), *Integrating Content and Language in Higher Education: From Theory to Practice Selected Papers from the 2013 ICLHE Conference*, (pp. 50-65). Frankfurt: Peter Lang.

Douglas, D. (2000). *Assessing languages for specific purposes*. New York, NY: Cambridge University Press.

Douglas, D., & Selinker, L. (1989). Markedness in discourse domains: Native and non-native teaching assistants. *Papers in Applied Linguistics, 1*(1), 69-81.

Embretson, S., & Reise, S. (2000). *Item response theory for psychologists*. New York, NY: Psychology Press.

Engelhard, G., Jr. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–287). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Engelhard, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement english literature and composition program with a many-faceted rasch model. *ETS Research Report Series, 2003*(1).Farnsworth, T. L. (2004). *The effect of teaching skills on holistic ratings of language ability in performance tests for international teaching assistant selection.* (PhD), University of California Los Angeles.

Farnsworth, T. L. (2013). An investigation into the validity of the TOEFL iBT speaking test for international teaching assistant certification. *Language Assessment Quarterly, 10*(3), 274-291.

Freibourg. (2016). English medium instruction - Qualitätssicherung. Retrieved from https://www.sli.unifreiburg.de/englisch/emi/quality?set_language=en

Gokcora, D. (1992). The SPEAK Test: International Teaching Assistants' and Instructors' Affective Reactions. Retreived from ERIC database. [351731].

Hoekje, B., & Linnell, K. (1994). " Authenticity" in Language Testing: Evaluating Spoken Language Tests for International Teaching Assistants. *TESOL Quarterly*, 103-126.

Hoekje, B., & Williams, J. (1992). Communicative competence and the dilemma of international teaching assistant education. *TESOL Quarterly, 26*(2), 243-269.

House, J., & Lévy-Tödter, M. (2010). Linguistic competence and professional identity in English medium instruction. In B. Meyer & B. Apfelbaum (Eds.), *Multilingualism at work: from policies to practices in public, medical and business settings* (Vol. 9, p. 13). Hamburg: Hamburg Studies on Multlingualism (HSM).

Isaacs, T. (2008). Towards defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *Canadian Modern Language Review, 64*(4), 555-580.

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly, 9*(3), 249-269.

Kang, O., Rubin, D., & Lindemann, S. (2015). Mitigating US undergraduates' attitudes toward international teaching assistants. *TESOL Quarterly, 49*(4), 681-706.

Klaassen, R. G., & Bos, M. (2010). English language screening for scientific staff at Delft University of Technology. *Hermes–Journal of Language and Communication Studies, 45*, 61-75.

Kling Soren, J. (2013). *Teacher identity in English-medium instruction: Teacher cognitions from a Danish tertiary education context*. (PhD Thesis). University of Copenhagen. Retrieved from http://halshs.archives-ouvertes.fr/docs/00/86/38/16/PDF/Ph.d._2013_Kling_Soren.pdf

Kling, J. (2017). English medium instruction and the international classroom. In A. M. Snow & D. M. Brinton (Eds.), *The content-based classroom: Perspectives on integrating language (2ed.)*. Ann Arbor, MI: University Michigan Press.

Kling, J., & Dimova, S. (2015). The Test of Oral English for Academic Staff (TOEPAS): Validation of standards and scoring procedures. In A. Knapp , & K. Aguado (eds.), *Fremdsprachen in Studium und Lehre - Chancen und Herausforderungen für den Wissenserwer*b (pp. 247-268). Frankfurt/Main: Peter Lang. Theorie und Vermittlung der Sprache, Bind. 57

Kling, J., & Hjulmand, L. L. (2008). PLATE–Project in language assessment for teaching in English. In R. Wilkinson & V. Zegars (Eds.), *Realizing content and language integration in higher education*, (p. 191-200). Maastricht: Universitaire Pers Maastricht.

Kling, J., & Stæhr, L. S. (2012). *The development of the Test of Oral English Proficiency for Academic Staff (TOEPAS)* (Technical Report). Centre for Internationalisation and Parallel Language Use: University Of Copenhagen. Retrieved from http://cip.ku.dk/forskning/cip_publikationer/CIP_TOPEPAS_Technical_Report.pdf/

Linacre, J. M. (2010). Predicting responses from Rasch measures. *Journal of applied measurement, 11*(1), 1.

Lumley, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language profficiency. *English for Specific Purposes, 17*(4), 347-367.

Madden, C. G., & Myers, C. L. (1994). *Discourse and Performance of International Teaching*

Assistants. Alexandria, VA: TESOL

Mauranen, A., Hynninen, N., & Ranta, E. (2010). English as an academic lingua franca: The ELFA project. *English for Specific Purposes, 29*(3), 183-190.

McNamara, T. (1996). *Measuring second language performance*. Harlow, Essex, UK: Addison Wesley Longman.

McNamara, T. (1997). Problematising content validity: the Occupational English Test (OET) as a measure of medical communication. *Melbourne Papers in Language Testing, 6*(1), 19-43.

Oppenheim, N. (1998). Undergraduates' Assessment of International Teaching Assistants' Communicative Competence. Retrieved from ERIC database(423783).

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models*. Thousand Oaks, CA: Sage.

Pecorari, D., Shaw, P., Irvine, A., & Malmström, H. (2011). English for academic purposes at Swedish universities: Teachers' objectives and practices. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE)*(22), 55-78.

Pytlik, B. P., & Liggett, S. (2002). *Preparing college teachers of writing: histories, theories, programs, practices*. New York, NY: Oxford University Press on Demand.

Rounds, P. L. (1987). Characterizing successful classroom discourse for NNS teaching assistant training. *TESOL Quarterly*, 643-671.

RUC. (2015). Certification for teaching English as an international language. Retrieved from http://www.ruc.dk/fileadmin/assets/paes/LICS/Certificering/Certification_folder.pdf

Saif, S. (2002). A needs-based approach to the evaluation of the spoken language ability of

international teaching assistants. *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquée, 5*(1), 145-167.

Saldaña, J. (2015). *The coding manual for qualitative researchers*. Thousand Oaks, CA: Sage.

Smith, J. (1989). Topic and variation in ITA oral proficiency: SPEAK and field-specific tests. *English for Specific Purposes, 8*(2), 155-167.

Spencer, D., & Warfel, T. (2004). Card sorting: a definitive guide. . Retreived from http://boxesandarrows.com/card-sorting-a-definitive-guide/.

Tange, H. (2010). Caught in the Tower of Babel: University lecturers' experiences with internationalisation. *Language and Intercultural Communication, 10,* 137-149.

Thomas, C. F., & Monoson, P. K. (1993). Oral English language proficiency of ITAs: Policy, implementation, and contributing factors. *Innovative Higher Education, 17*(3), 195-209.

Thøgersen, J., & Airey, J. (2011). Lecturing undergraduate science in Danish and in English: A comparison of speaking rate and rhetorical style. *English for Specific Purposes*, 30, 209-221.

Vinke, A. (1995). *English as the medium of instruction in Dutch engineering education* (doctoral dissertation). TU Delft: Delft University of Technology, the Netherlands. Delft University Press. Retrieved from http://repository.tudelft.nl/view/ir/uuid:491b55f9-fbf9-4650-a44d-acb9af8412a8/.

Wagner, E. (2016). A Study of the Use of the TOEFL iBT® Test Speaking and Listening Scores for International Teaching Assistant Screening. *ETS Research Report Series, 2016*(1), 1-48.

Westbrook, P. N., & Henriksen, B. (2011). Bridging the linguistic and affective gaps. In R. Cancino,

L. Dam, & K. Jæger (Eds.), *Policies, principles, practices: New directions in foreign language education in the era of educational globalization* (pp. 188-212). Newcastle upon Tyne, UK: Cambridge Scholars Press.

Williams, J., Inscoe, R., & Tasker, T. (1997). Communication strategies in an interactional context: The mutual achievement of comprehension. In G. Kasper & E. Kellerman (Eds.), *Communication strategies: Psycholinguistic and sociolinguistic perspectives*, (pp.304-322). London: Longman.

Wilkinson, R. (2005). *The impact of language on teaching content: Views from the content teacher.* Paper presented at the Bi and Multilingual Universities– Challenges and Future Prospects Conference. Retrieved from http://www.palmenia.helsinki.fi/congress/bilingual2005/presentations/wilkinson.pdf

Xi, X. (2007). Validating TOEFL® iBT Speaking and setting score requirements for ITA screening. *Language Assessment Quarterly, 4*(4), 318-351.