LINKING THE TOEPAS WITH THE CEFR: TECHNICAL REPORT

SLOBODANKA DIMOVA
UNIVERSITY OF COPENHAGEN

OCTOBER 9-11, 2018

This document contains the results from a three-day linking event with participants from the TAEC project and the Centre for Internationalisation and Parallel Language Use (CIP).

The project is co-funded by the Erasmus+ programme of the European Union.

Erasmus+

TAEC
Transnational Alignment
of English Competences
for University Lecturers

1. Introduction

The purpose of this ERASMUS+ funded project (Strategic Partnerships) was to link the Test of Oral English Proficiency for Academic Staff (TOEPAS) to the Common European Framework of Reference (CEFR) with the goal of increasing the TOEPAS result transparency for cross-institutional and transnational use. Using the methodologies and activities outlined in the manual *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* (Council of Europe, 2009), the linking procedure consisted of four distinct stages: familiarization, specification, standardization, and validation. This procedure helped to provide empirical evidence for the link between the TOEPAS and the CEFR.

1.1. The Common European Framework of Reference (CEFR)

The Common European Framework of Reference (CEFR) was designed to provide a common basis for development of language instruction programs, curricula, syllabi, and assessments in order to improve the communication among language professionals and the comparability of modern language teaching and learning across different European contexts. However, usage of the CEFR has grown beyond the European borders, and a number of tests and curricula in Asia and the United States have been linked to the CEFR. The framework has three broad levels of reference [Basic User (A1 & A2), Independent User (B1 & B2), and Proficient User (C1 & C2)], each subdivided into two sublevels.

Basic User	Independent User	Proficient User
A1: Breakthrough	B1: Threshold	C1: Effective Operational Proficiency
A2: Waystage	B2: Vantage	C2: Mastery

Unlike traditional language proficiency descriptions based on the four skills (listening, speaking, reading, and writing), the framework comprises a number of scales related to different modes of communication (reception, production, interaction, and mediation). These scales comprise descriptors presented as “can do” statements that characterize each proficiency level. Therefore, instead of linear progression, the newest representation of the reference levels shows the embeddedness of each lower level into a higher one (Council of Europe, 2018).

As proposed by the manual for relating scales to CEFR, the linking project was completed in four stages:

1. **Familiarization**--ensuring that participants in the linking process have the necessary knowledge of the CEFR communicative models, proficiency levels, and illustrative descriptors;

2. **Specification**--describing test content and tasks types in relation to the categories presented in CEFR in order to provide test quality evidence;

3. **Standardization**--*training* the participants to interpret and use the CEFR descriptors when rating performances, and *benchmarking* local performance samples;

4. **Validation**--monitoring of the quality of the linking activities and the reliability of the obtained results.

1.2. Test of Oral English Proficiency for Academic Staff (TOEPAS)

The Test of Oral English Proficiency for Academic Staff (TOEPAS) is an oral English proficiency test used for university lecturer certification for teaching English medium instruction (EMI) courses at the University of Copenhagen (UCPH) (Kling & Stæhr, 2012). The TOEPAS is a performance-based test that uses a simulated lecture as a sample of the lecturers' oral English performance. The TOEPAS results are reported holistically on a five-point scale in increments of 10 (20-60) (see Appendix 1). The scale has been empirically developed to reflect lecturers' use of English in the target language use domain of university instruction. Alongside a holistic score report, test-takers receive detailed written and oral feedback and a video-recording of their performance.

The main purpose of linking the TOEPAS to the CEFR is result transferability and transparency. Previous research on the uses and consequences of the TOEPAS suggests that the lecturers certified through the TOEPAS for teaching EMI courses have the need for international recognition of their proficiency level (Dimova, 2017). These lecturers express concern that stakeholders may be unable to recognize their proficiency levels based solely on the TOEPAS scale, so they need the reported results to be trans-institutionally and trans-nationally recognizable.

2. The linking procedure design

As proposed by the manual *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* (Council of Europe, 2009), the linking procedure consisted of four stages: familiarization, specification, benchmarking, and validation. The alignment took place over a three-day standardization event held from October 9-11, 2018 at the University of Copenhagen (UCPH). While Day 1 focused on familiarization with the CEFR scales, Day 2 was dedicated to training, and Day 3 to benchmarking. The specification stage was completed before the three-day event in collaboration with the test designers and coordinators, as they are most familiar with the different aspects of the test content and tasks. This stage was guided by the forms provided by the Manual (A1-8), available in Appendix 2. The TOEPAS specifications were presented to the panel judges on Day 2 before benchmarking local samples. Procedural validation data were collected during the event, but the analysis was performed after the event ended.

2.1. Participants

The panel participants were recruited based on the following criteria.

1. Experience with academic English language instruction or research

2. Familiarity with EMI contexts, defined as university contexts where content course instruction (e.g., medicine, political science, math) is conducted in English, a language which for most of the students and the lecturers is not their mother tongue.

3. MA or PhD in Applied Linguistics, Teaching English to Speakers of Other Languages (TESOL), or a related field.

It was also important that the panel consisted of a mix of participants from both UCPH and abroad, some already familiar with the TOEPAS procedures and scales and others who were not.

Given the recruitment criteria, the international panel consisted of six UCPH participants who were familiar with the TOEPAS and six participants from different European countries (Italy, Croatia, Spain, and the Netherlands) who were unfamiliar with the TOEPAS. The age of the participants ranged from 29 to 60 (median 42), and all, except for one, had assessment experience of 8 years more (median 10). All participants had experience as ESP or EAP teachers (ranging from 6 to over 25 years), and all, but three, had experience training EMI teachers. Only two participants from the panel had not used the CEFR scales before the linking event.

participants		age	education	assessment experience (yrs)	EMI trainer experience (yrs)	EAP/ESP teaching experience (yrs)	CEFR familiarity
Non-UCPH	P1	29	MA	10	7	10	Yes
	P2	39	PhD	15	3	10	Yes
	P3	33	MA	9	0	9	Yes
	P4	40	PhD	10	0	10	No
	P5	42	PhD	8	2	6	Yes
	P6	41	MA	9	0	9	Yes
UCPH	P7	54	PhD	25	15	25+	Yes
	P8	53	PhD	28	28	28	Yes
	P9	40	PhD	0.5	1	7	No
	P10	53	MA	12	9	15	Yes
	P11	60	MA	9	9	9	Yes
	P12	49	MA	20	10	20	Yes

Table 1. Participant background

2.2. Procedures

As recommended by the manual for relating examinations to the CEFR, the linking procedure consisted of four stages:

1. **Familiarization**—the participants took part in four CEFR familiarization activities to ensure their familiarity with the CEFR scales. They were trained to use the relevant CEFR speaking scales using standardized speaking samples.

2. **Specification**—the linking event coordinator collected information about the TOEPAS content and task in discussion with test developers and coordinators.

3. **Standardization**—the participants benchmarked local test performance samples and standardized the levels.

4. **Validation**—the event coordinator monitored the linking activities and collected validity evidence during the three-day event.

3. Familiarization

The familiarization stage consisted of three parts: pre-event training material, scale descriptor familiarization activities, and training.

3.1. Pre-event familiarization training material

Three weeks before the event, the participants received a packet with pre-training familiarization materials consisting of the schedule for the event, the CEFR levels, the CEFR communicative model for overall language proficiency, the representations of production and activities and strategies, the scales for speaking and mediation, and links to standardized samples. To ensure effective completion of the linking event, the participants were instructed to do the following:

1. Review the CEFR communication models

2. Read the descriptors in each of the scales

3. Make sure they could categorize the descriptors according the level and activity/skill

4. Take note of the key words at each overall level that distinguish the particular level from the rest

5. Practice the use of the scales by listening to some performances online.

3.2. Familiarization activities

The first day of the event was dedicated to familiarization with CEFR scales. It started with an introduction of CEFR, followed by different scale familiarization activities and training through the use of standardized speech samples at different CEFR levels.

After a round of participant introductions, the coordinator gave a PowerPoint presentation to provide background information about the CEFR communicative models and scales, as well as their uses. Then the coordinator discussed the purposes for aligning the TOEPAS with the CEFR. She then explained the different stages of aligning tests to the CEFR and the activities in which the participants were expected to take part. During this presentation, the participants had the opportunity to ask questions and seek clarification regarding the proposed event activities.

Four familiarization activities were then undertaken. In Familiarization Activity 1 (FA1), the participants were asked to identify and discuss in groups of three the specific words/phrases that distinguish the specific level from the rest on the overall spoken production scale, the overall spoken interaction scale, and the overall mediation scale. Then, in Familiarization Activity 2 (FA2), each participant was required to complete an online activity where they needed to select the appropriate level (B2-C2) for each of the 34 CEFR descriptors presented on the screen. The participants were also instructed to type in the key word(s) in the descriptor that helped them decide on the level. Table 2 shows that the range of correctly identified levels of the provided descriptors ranged from 21%-47%.

		mediation	interaction	production	total	percent
participants	P10	2	3	2	7	0.21
	P8	1	4	3	8	0.24
	P12	2	3	3	8	0.24
	P1	5	5	1	11	0.32
	P9	5	5	2	12	0.35
	P11	2	6	5	13	0.38
	P6	5	4	5	14	0.41
	P2	6	5	4	15	0.44
	P3	4	6	5	15	0.44
	P4	5	6	5	16	0.47
	P7	5	6	5	16	0.47
	P5	5	7	4	16	0.47

Table 2. Number/percentage of responses with correct descriptor level

The participants were asked to work again in groups of three for the Familiarization Activity 3 (FA3), where they were required to reconstruct the *Qualitative features of spoken production* scale. In the online Familiarization Activity 4 (FA4), the participants were supposed to match the category and the level for each descriptor they were given (N=15). Table 3 shows the results from FA4, where the participants had a high percentage of correct responses, ranging from 83% to 100%).

participant	category	level	total
P1	1	1	1
P11	1	1	1
P3	1	1	1
P7	1	1	1
P4	1	0.94	0.97
P2	1	0.87	0.93
P6	1	0.87	0.93
P8	1	0.87	0.93
P9	1	0.87	0.93
P5	1	0.8	0.90
P12	1	0.73	0.87
P10	1	0.67	0.83

Table 3. Percentage of correct responses by category and level for the *Qualitative features of spoken production* scale

In the discussion following the familiarization activities, the participants noted that it was difficult for them to pinpoint the exact level of the descriptor in the overall scales because they were quite general. On the other hand, the participants thought that the descriptors in the *Qualitative features of the spoken production* scale were more detailed and precise, so they could associate them more easily with the relevant category and CEFR level.

3.3. Training

Training panelists in the standard-setting method that will be applied through illustration and practice is essential to ensure the procedural validity (Cizek & Bunch, 2007; Raymond & Reid, 2001). Therefore, in the last part of the familiarization stage, the participants were trained how to use the scale descriptors in relation to standardized illustrative examples of the speaking levels B2 to C2. The training consisted of two segments: illustration and benchmarking practice. In the first segment, the participants watched standardized videos with speaking performances ranging from B2 to C2 level. While listening to each performance, the participants were required to identify the characteristics specific for the particular level. Then, they discussed the performances first in small groups and then in plenum.

In the second segment, the participants watched 6 videos with speaking performances, which they ranked from lowest to highest and rated using CEFR levels. Then, they discussed the rankings in groups of three and used CEFR descriptors to justify the CEFR levels they had assigned. The final ranking and CEFR levels was performed by reaching a consensus in a plenary discussion, the purpose of which was shared understanding of the speaking qualities associated with each CEFR level, as well as selection of performances to serve as anchors.

4. Specification

According to the Manual, the purpose of the specification phase is to describe the content of the test to be related to the CEFR. Specification, also known as construct congruence (Tannenbaum & Cho, 2014), is among the first stages of standard-setting events because it provides evidence that the test content and tasks are congruent with the framework of interest. It is usually conducted by filling out a number of Forms in the Manual (Ch. 4). The General Description Forms (A1–A8) were filled out by the TOEPAS developers and coordinators because they were deemed most knowledgeable about the development and the content of the test, as well as the estimated proficiency levels in relation to the CEFR. Apart from the Manual Forms, further documentation used during the Specification meeting included the technical manual and various TOEPAS research studies and reports. The completed forms are available in Appendix 2.

5. Benchmarking/Standardization

During the familiarization stage, the participants were trained to use the CEFR speaking scale descriptors with illustrative speech samples. Once they felt comfortable using the scales for rating speech performances, it was time for the benchmarking/standardization stage. This stage consisted of three main parts:

1. Boundary discussions
2. Relating TOEPAS descriptors to CEFR levels
3. Judgement rounds

5.1. Boundary discussions

In this part of the Benchmarking/Standardization stage, the participants discussed boundary points that distinguished one proficiency level from another. Given the specifications, the TOEPAS developers estimated the TOEPAS performance range to be between B2 and C2 CEFR levels. Therefore, only these levels were used in the boundary discussions. Once operational definitions of the levels were reached, the participants watched 5 performances across the TOEPAS scale. After watching, they ranked the performances from lowest to highest, deciding which performances were above or below a CEFR level boundary. The discussions lasted until a consensus was reached and the boundaries were internalized.

5.2. Relating TOEPAS descriptors to CEFR levels

To ensure complete linking between the TOEPAS scale and the CEFR, it was decided that the process should also include relating the existing TOEPAS scale descriptors with the CEFR levels to provide further evidence about the validity of the linking procedure. For this purpose, the panel members participated in two activities. In the first activity, the panel members worked in small groups to physically match the descriptors with the relevant scale categories and proficiency levels. Then, they participated individually in an online activity where they were presented with 42 descriptors, one at a time, for which they were asked to assign a CEFR level. The data from the online activity were exported in an Excel worksheet and analyzed. The analysis included correlation between the CEFR levels (B1 and below=1, B2=2, C1=3, C2=4) assigned by each panel member to the descriptors and their means. As Table 4 indicates, the correlations were high, which suggests strong agreement among the panel members.

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
0.95	0.96	0.97	0.96	0.93	0.9	0.91	0.95	0.97	0.96	0.94	0.95

Table 4. Correlations: panel member's descriptor scores and descriptor mean

5.3. Judgement rounds

Once the participants reached consensus regarding the minimally competent candidate at each boundary, they were asked to perform Round 1 of judging 10 pre-selected speaking performances using the CEFR speaking scale (see Appendix 3). The performances were selected to represent different TOEPAS (and CEFR) levels. The participants were asked to use the CEFR descriptors to justify their rating decisions.

The participants entered their judgements (B2, B2+/C1-, C1, C2) in a pre-prepared online form on the online learning platform (Canvas). The numbers 1-4 were used for the CEFR levels (B2=1 to C2=4).The judgements were exported in an Excel workbook, and individual and group mean CEFR scale levels were automatically estimated. The judgements from Round 1 are shown in Table 5. The level of agreement among the participants was quite high except for a few performances.

		Panel members														Avg. level
		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	mean	SD	
examinees	E8	4	4	4	4	4	4	4	4	4	4	4	4	4	0	C2
	E7	4	4	4	4	4	4	4	4	4	3	4	4	3.92	0.28	C2
	E3	4	3	4	4	4	4	4	4	4	4	4	4	3.92	0.28	C2
	E1	3	3	3	4	3	3	4	3	4	3	4	4	3.42	0.49	C1
	E4	3	2	2	4	2	4	3	3	3	3	3	3	2.92	0.64	C1
	E5	2	2	2	3	3	4	3	4	1	4	4	3	2.92	0.95	C1
	E2	3	2	2	2	2	3	3	4	3	2	3	4	2.75	0.72	C1
	E10	2	1	2	1	2	1	3	2	2	1	2	1	1.67	0.62	B2+/C-1
	E9	1	2	2	2	2	2	2	2	1	1	1	1	1.58	0.49	B2
	E6	1	1	1	2	1	1	2	2	1	1	1	2	1.33	0.47	B2

Table 5. Round 1 of judgements

Discussions followed the Round 1 of judgements to further clarify the use of descriptors and the decision-making process. When the panel members felt comfortable with the decisions made during the discussions, they were asked to watch and rate the same performances once more. During Round 2, some participants made changes to the ratings they had assigned during Round 1. According to the data from Round 2 of judgements in Table 6, the variation among the scores assigned by the participants was lower in this round (0-.62) than in Round 1 (0-.95).

		Panel members														Ave level
		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	mean	SD	
examinees	E8	4	4	4	4	4	4	4	4	4	4	4	4	4	0	C2
	E7	4	4	4	4	4	4	4	4	4	4	4	4	4	0	C2
	E3	4	4	4	3	4	3	4	4	4	4	4	4	3.83	0.37	C2
	E1	3	3	3	3	3	3	4	3	3	3	3	3	3.08	0.28	C1
	E4	3	2	3	2	3	3	3	3	3	3	3	3	2.83	0.37	C1
	E5	2	2	2	3	3	3	3	3	3	3	3	3	2.75	0.43	C1
	E2	2	2	2	2	3	3	3	4	3	2	3	3	2.67	0.62	C1
	E10	2	1	1	2	2	2	2	1	1	1	2	1	1.5	0.5	B2+/C1-
	E9	1	1	1	1	1	1	1	2	1	1	1	2	1.17	0.38	B2
	E6	1	1	1	1	2	1	1	1	1	1	1	1	1.08	0.28	B2

Table 6: Round 2 of judgements

Based on a short discussion after the results from Round 2 were presented, the panel members expressed satisfaction with the final decision and lack of necessity for further calibrations. Then, the TOEPAS scores of the 10 performances were revealed in relation to the calibrated CEFR levels.

6. Validation

Two types of validity evidence were gathered during the linking event: procedural validity and internal validity. The procedural validity evidence was gathered through questionnaires which elicited information about the degree to which the panel members understood the linking activities, felt comfortable using the CEFR scales, and were able to follow the procedures. The internal validity evidence was obtained through analysis of panel members' level of agreement when rating the speaking performances.

6.1. Procedural validity

At the end of each day of the linking event, the panel members filled out a questionnaire related to clarity of instructions, usefulness of activities, and ability to complete the required procedures. The purpose of these questionnaires was to evaluate the procedures and to identify any difficulties or problems before moving to the following stage.

The first questionnaire was administered at the end of the familiarization stage. It contained 9 Likert-type questions and 2 open-ended questions. The responses were generally positive as the participants agreed that they could follow the instructions and complete the activities. The responses are presented in Table 7.

questions	strongly agree	agree	don't know	disagree	strongly disagree
I understand the purpose of relating tests to the CEFR.	8	4	-	-	-
The coordinator clearly explained the process of relating of the TOEPAS to the CEFR.	7	4	1	-	-
The familiarization activities were helpful for understanding the CEFR levels.	10	2	-	-	-
The panel discussions after the familiarization activities were helpful for understanding the CEFR levels.	9	2	1	-	-
The group discussions during the familiarization activities were helpful for understanding the CEFR levels.	8	4	-	-	-
The illustrative videos were helpful for understanding the CEFR levels.	8	4	-	-	-
I was able to complete the practice judgement process as required.	7	5	-	-	-
I was able to follow the instructions and complete the activities.	8	4	-	-	-
There was enough time to complete the familiarization activities.	9	3	-	-	-

Table 7. Day 1: Questionnaire after the familiarization stage

The first open-ended question on this questionnaire asked the panel members to write what they liked best during the familiarization stage. While several participants liked the familiarization activity where they were supposed to reconstruct the speaking scale by matching descriptors with category and level, most found the group and the panel discussions useful; these discussions helped the participants compare their understanding of the level descriptors with the other team members and improve their understanding of the variation within the CEFR bands. One participant stated, she liked…

> …seeing how different members of the team perceive different speakers. Although we know in practice how to apply the CEFR, individual judgement is still present. As a group we seem to arrive at the right answer, but even then personal opinion is visible. This was very good because we all can see where our bias lies and try to adjust it accordingly.

The second open-ended question was related to what aspects the participants did not like. While most participants did not state anything here, three participants needed more time to work on the descriptors and the illustrative samples.

The second questionnaire was administered at the end of the benchmarking stage. It contained 8 Likert-type questions and 2 open-ended questions. The responses were generally positive as the participants agreed that they understood the purpose and the process of benchmarking, and they could complete the benchmarking activities. The responses are presented in Table 8.

questions	strongly agree	agree	don't know	disagree	strongly disagree
The TOEPAS specifications were informative.	8	4	-	-	-
The benchmarking process was clear to me.	8	4	-	-	-
I understand the purpose of benchmarking.	10	2	-	-	-
I am comfortable using the CEFR descriptors for TOEPAS performances.	4	8	-	-	-
I can relate the TOEPAS descriptors to the CEFR levels.	2	9	1	-	-
I feel familiar with the CEFR descriptors.	3	9	-	-	-
The discussions were helpful.	10	2	-	-	-
I had enough time to complete the benchmarking activities.	8	4	-	-	-

Table 8. Day 2: Questionnaire after the benchmarking stage

When asked about what they liked best, most panel members indicated again that the discussions were very useful. Many also added that it was interesting to use the CEFR levels and descriptors for the TOEPAS performances. One participants claimed, “Many elements of the assessment process became clearer to me.” In response to the open-ended question regarding what they liked least, most participants did not have anything to add, though a couple of them found that listening to the performances required a great deal of concentration and was therefore exhausting.

The final questionnaire was administered at the end of the three-day event. It contained 5 Likert-type questions and one open-ended question. The participants gained an improved judging confidence during the second round as the discussions helped them with decision-making. The responses are presented in Table 10.

questions	strongly agree	agree	don’t know	disagree	strongly disagree
I felt confident judging the speaking performances in the first round.	5	7	-	-	-
The discussion after the first round of judgements was helpful.	10	2	-	-	-
Submitting the judgements online was easy.	11	1	-	-	-
I felt confident judging in the second round.	9	3	-	-	-
The three-day event improved my understanding of the CEFR levels.	12	-	-	-	-

Table 10. Day 3: Final questionnaire after the standardization stage

Most additional comments in the open-ended question were positive. The participants found the linking event useful, well organized, informative, and enjoyable, though it was quite intensive and required a great deal of attention.

6.2. Internal validity

In addition to the validity evidence related to panel members’ understanding and participation in the linking event, it was important to analyze the degree to which the members agreed in their rating, or the interrater consistency. Three measures were used for that purpose: Chronbach’s alpha, inter class consistency (ICC), and Pearson correlation between each panel members’ rating and the mean scores. Values greater than .70 are typically acceptable for consistency estimates of interrater reliability (Barrett, 2001).

When more than two raters are used, Chronbach’s alpha coefficient and interclass consistency (ICC) are appropriate consistency estimates of interrater reliability, i.e. how consistently the raters use the rating scale. Cronbach's alpha helps examine the degree to which the ratings from a group of judges are similar when measuring a common dimension (Stemler & Tsai, 2008). Interclass correlation coefficient (ICC), on the other hand, is a more conservative estimate of interrater reliability because it confounds two ways in which raters differ: consensus (mean differences) and consistency (association). When the Chronbach’s alpha coefficient and ICC are closer to 1, it means that the raters’ agreement is very high (Stemler & Tsai, 2008). As shown in Table 11, a high level of interrater consistency was found both among the non-UCPH and the UCPH participants. Although both the ICC and the Chronbach’s alpha increased in Round 2, this increase was more evident in ICC.

	Round 1		**Round 2**	
	ICC	Chronbach’s alpha	ICC	Chronbach’s alpha
Non-UCPH	.79	.96	.92	.99
UCPH	.73	.94	.9	.98

Table 11. Interrater consistency in Rounds 1 and 2 of judgements

Pearson correlation coefficients are usually used to estimate interrater reliability between two raters. Strong correlation coefficients mean that raters are follow consistent rating patterns (not necessarily exact agreement in scores). Instead of computing the correlation coefficients between all rater pairs, correlations were computed between each rater's scores and the all raters' mean. Strong correlations between individual participant's scores and the means were found both in Round 1 and Round 2, though these correlations were stronger after the second round. It is expected that there will be some variation among the panel members, but since all coefficients were above the minimum of .70, they are acceptable. The correlations from each round are presented in Table 12.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**
Round 1	0.94	0.9	0.89	0.87	0.91	0.89	0.92	0.86	0.86	0.9	0.94	0.89
Round 2	0.95	0.96	0.97	0.92	0.96	0.95	0.96	0.9	0.99	0.98	0.98	0.95

Table 12. Panel members' score correlations with means

7. Conclusions

This technical report provided information about the alignment of the TOEPAS scores with the CEFR levels. A panel of 12 participants participated in the linking event. Given that the TOEPAS is a performance-based speaking test, the Benchmark method was applied as most suitable. During the event, data were collected online via the learning platform Canvas. This platform eased the data collection process and allowed for immediate feedback.

Based on the evidence collected and the data analysis, the alignment of the TOEPAS scale and the CEFR levels are presented in Figure 1.

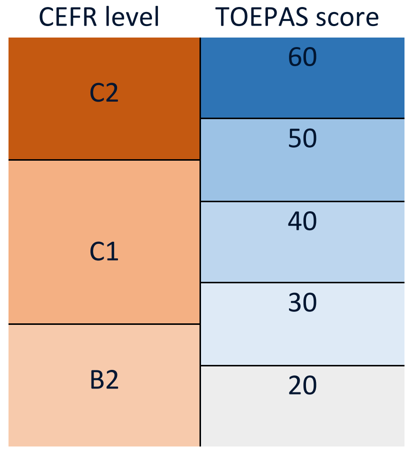


Fig. 1. TOEPAS alignment with CEFR

References

Barrett, P. (2001). Assessing the reliability of rating data. Retrieved from http://www.pbarrett.net/presentations/rater.pdf

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A Manual.* Strasbourg: Language Policy Division.

Council of Europe. (2018). *COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES: LEARNING, TEACHING, ASSESSMENT COMPANION VOLUME WITH NEW DESCRIPTORS.* Strasbourg: Language Policy Division.

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. DOI: 10.1016/j.esp.2016.12.004

Kling, J. M. & Stæhr, L. S. (2012). *The development of the Test of Oral English Proficiency for Academic Staff (TOEPAS).* Centre for Internationalisation and Parallel Language Use.

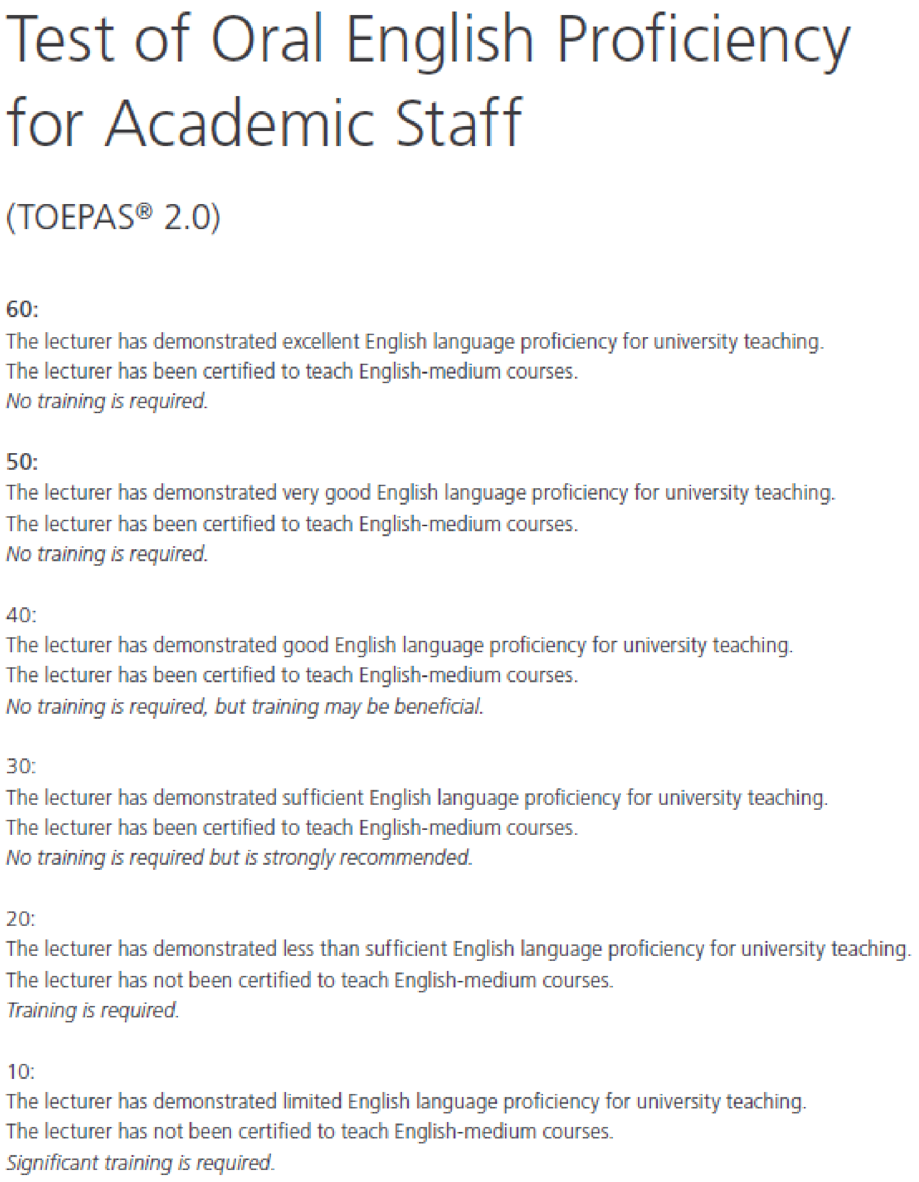
Raymond, M. R., & Reid, J. R. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–157). Mahwah, NJ: Erlbaum

Stemler. S.E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J.W. Osborne (Ed.), *Best Practices in Quantitative Methods* (pp. 29 – 49). California: Sage Publications, Inc.

Tannenbaum, R.J. & Cho, Y. (2014). Critical Factors to Consider in Evaluating Standard-Setting Studies to Map Language Test Scores to Frameworks of Language Proficiency. *Language Assessment Quarterly, 11,* 233-249. DOI: 10.1080/15434303.2013.869815

Appendices

Appendix 1: TOEPAS scale



Test of Oral English Proficiency for Academic Staff

(TOEPAS® 2.0)

60:
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.

50:
The lecturer has demonstrated very good English language proficiency for university teaching.
The lecturer has been certified to teach English-medium courses.
No training is required.

40:
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.

30:
The lecturer has demonstrated sufficient English language proficiency for university teaching.
The lecturer has been certified to teach English-medium courses.
No training is required but is strongly recommended.

20:
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.

10:
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.

Appendix 2
Form A1: General Examination Description (part)

GENERAL EXAMINATION DESCRIPTION	
1. General Information Name of examination	Test of Oral English Proficiency for Academic Staff (TOEPAS)
Language tested	English
Examining institution	Centre for Internationalisation and Parallel Language Use (CIP), University of Copenhagen
Versions analysed (date)	TOEPAS 2.0
Type of examination	□ International □ National □ Regional ● Institutional
Purpose	Oral English certification for teaching in English medium instruction (EMI) programs in higher education
Target population	□ Lower Sec □ Upper Sec □ Uni/College Students ● Adult
No. of test takers per year	50-100
2. What is the overall aim? To allow the test user (university leadership) to draw inferences, based on test performance, on lecturers' oral English proficiency level for teaching purposes.	
3. What are the more specific objectives? If available describe the needs of the intended users on which this examination is based. TOEPAS is targeted specifically at universities who require some evidence of the language ability of their teaching staff. TOEPAS provides university teaching staff with formative feedback regarding their oral English language strengths and weaknesses for teaching EMI courses.	
4. What is/are principal domain(s)?	X Public □ Personal X Occupational X Educational

5. Which communicat-ive activities are tested?	□ **1** Listening comprehension □ **2** Reading comprehension **X 3** Spoken interaction □ **4** Written interaction **X 5** Spoken production □ **6** Written production □ **7** Integrated skills **X 8** Spoken mediation of text □ **9** Written mediation of text □ **10** Language usage □ **11** Other: (specify): ____________	**Name of Subtest(s)** _warm-up_______ _lecture + assignment_ _Q&A__________ __________________ __________________ __________________ __________________ __________________ __________________ __________________	**Duration** _5-10 min_______ _20 min_________ _7-10 min________ ________________ ________________ ________________ ________________ ________________ ________________ ________________
6. What is the weighting of the different subtests in the global result?	Warm-up is not rated Lecture and Q&A are rated holistically		
7. Describe briefly the structure of each subtest	Warm-up: some discussion about teaching at the university Lecture: test-takers give a lecture they normally teach or will teach; they are also supposed to give an assignment (e.g. homework) Q&A: other two participating lectures play the role of students and ask questions		
8. What type(s) of responses are required?	□ Multiple-choice □ True/False □ Matching □ Ordering □ Gap fill sentence □ Sentence completion □ Gapped text / cloze, selected response □ Open gapped text / cloze □ Short answer to open question(s) X Monologue □ Interaction with examiner X Interaction with peers □ Other	Subtests used in (Write numbers above) □	

9. What information is published for candidates and teachers?	X Overall aim X Principal domain(s) □ Test subtests □ Test tasks □ Sample test papers □ Video of format of oral	□ Sample answer papers □ Marking schemes X Grading schemes □ Standardised performance samples showing pass level □ Sample certificate
10. Where is this accessible?	X On the website □ From bookshops □ In test centres □ On request from the institution X Other	______________________________
11. What is reported?	X Global grade and feedback □ Grade per subtest	□ Global grade plus graphic profile □ Profile per subtest

Form A2: Test Development

Test development	**Short description and/or references**
1. What organisation decided that the examination was required?	X Own organisation/school □ A cultural institute □ Ministry of Education □ Ministry of Justice □ Other: specify: ___________________
2. If an external organisation is involved, what influence do they have on design and development?	□ Determine the overall aims □ Determine level of language proficiency □ Determine examination domain or content □ Determine exam format and type of test tasks □ Other: specify: ___________________
3. If no external organisation was involved, what other factors determined design and development of examination?	X A needs analysis X Internal description of examination aims X Internal description of language level □ A syllabus or curriculum X Profile of candidates
4. In producing test tasks are specific features of candidates taken into account?	□ Linguistic background (L1) □ Language learning background □ Age X Educational level □ Socio-economic background X Social-cultural factors □ Ethnic background □ Gender
5. Who writes the items or develops the test tasks?	The test is based on simulated lecture—no specific items are designed on regular basis
6. Have test writers guidance to ensure quality?	□ Training □ Guidelines □ Checklists □ Examples of valid, reliable, appropriate tasks: □ Calibrated to CEFR level description □ Calibrated to other level description: ___________________________________
7. Is training for test writers provided?	□ Yes □ No
8. Are test tasks discussed before use?	□ Yes □ No
9. If yes, by whom?	□ Individual colleagues □ Internal group discussion □ External examination committee □ Internal stakeholders

	☐ External stakeholders
10. Are test tasks pretested?	☐ Yes ☐ No
11. If yes, how?	
12. If no, why not?	
13. Is the reliability of the test estimated?	X Yes ☐ No
14. If yes, how?	X Data collection and psychometric procedures ☐ Other: specify: ___________________
15. Are different aspects of validity estimated?	X Face validity X Content validity ☐ Concurrent validity ☐ Predictive validity ☐ Construct validity
16. If yes, describe how.	

Form A3: Marking

Marking: Subtest	**Complete a copy of this form for each subtest. Short description and/or reference**
1. How are the test tasks marked?	For receptive test tasks: □ Optical mark reader □ Clerical marking For productive or integrated test tasks: X **Trained examiners** □ Teachers
2. Where are the test tasks marked?	□ Centrally ⊘ **Locally:** X **By local teams** X **By individual examiners**
3. What criteria are used to select markers?	
4. How is accuracy of marking promoted?	X **Regular checks by co-ordinator** X **Training of markers/raters** X **Moderating sessions to standardise judgments** ⊘ **Using standardised examples of test tasks:** □ Calibrated to CEFR X **Calibrated to another level description** □ Not calibrated to CEFR or other description
5. Describe the specifications of the rating criteria of productive and/or integrative test tasks.	□ One holistic score for each task □ Marks for different aspects for each task X **Rating scale for overall performance in test** X **Rating Grid for aspects of test performance** □ Rating scale for each task □ Rating Grid for aspects of each task X **Rating scale bands are defined, but not to CEFR** □ Rating scale bands are defined in relation to CEFR
6. Are productive or integrated test tasks single or double rated?	□ Single rater X **Two simultaneous raters** X **Double marking of scripts / recordings** □ Other: specify:_________________
7. If double rated, what procedures are used when differences between raters occur?	□ Use of third rater and that score holds X **Use of third marker and two closest marks used** □ Average of two marks X **Two markers discuss and reach agreement** □ Other: specify:_________________

8. Is inter-rater agreement calculated?	X **Yes** □ No
9. Is intra-rater agreement calculated?	X **Yes** □ No

Form A4: Grading

Grading: Subtest ________________________	**Complete a copy of this form for each Subtest.** **Short description and/or reference**
1. Are pass marks and/or grades given?	□ Pass marks X Scores based on TOEPAS scale
2. Describe the procedures used to establish pass marks and/or grades and cut scores	Borderline method was used for setting the certified/not certified cutoff
3. If only pass/fail is reported, how are the cut-off scores for pass/fail set?	
4. If grades are given, how are the grade boundaries decided?	The Body of Work method was used to decide on the grade boundaries
5. How is consistency in these standards maintained?	Routine Rasch analysis of consistency of levels and rater behavior

Form A5: Reporting Results

Results	Short description and/or reference
1. What results are reported to candidates?	X Global grade or pass/fail □ Grade or pass/fail per subtest □ Global grade plus profile across subtests □ Profile of aspects of performance per subtest
2. In what form are results reported?	□ Raw scores □ Undefined grades (e.g. "C") X Level on a defined scale X Diagnostic profiles
3. On what document are results reported?	X Letter or email □ Report card X Certificate / Diploma X On-line
4. Is information provided to help candidates to interpret results? Give details.	The candidates receive detailed feedback report where the performance is described in detail. The candidates participate in a follow-up meeting with the examiners to discuss the results.
5. Do candidates have the right to see the corrected and scored examination papers?	The candidates access the video recording.
6. Do candidates have the right to ask for remarking?	Yes

Form A6: Data Analysis

Data analysis	Short description and/or reference
1. Is feedback gathered on the examinations?	X Yes □ No
2. If yes, by whom?	X Internal experts (colleagues) □ External experts □ Local examination institutes X Test administrators □ Teachers □ Candidates
3. Is the feedback incorporated in revised versions of the examinations?	X Yes □ No
4. Is data collected to do analysis on the tests?	X On all tests □ On a sample of test takers: How large?: ________. How often?:_________ □ No
5. If yes, indicate how data are collected?	□ During pretesting X During live examinations X After live examinations
6. For which features is analysis on the data gathered carried out?	□ Difficulty □ Discrimination X Reliability X Validity
7. State which analytic methods have been used (e.g. in terms of psychometric procedures).	Pearson correlation for intr-a and inter-rater reliability MFRM for scale, rater, and bias analysis.
8. Are performances of candidates from different groups analysed? If so, describe how.	MFRM is used for analysis of bias.
9. Describe the procedures to protect the confidentiality of data.	Data are stored on a protected server. Different users (raters, admins, candidates) have access only to the data relevant to them. Database access is password protected. Feedback report and video recordings are only distributed to candidates.
10. Are relevant measurement concepts explained for test users? If so, describe how.	Measurement and linguistic concepts are explained using general descriptions in written and oral feedback.

Form A7: Rationale for Decisions

Rationale for decisions (and revisions)	Short description and/or reference
Give the rationale for the decisions that have been made in relation to the examination or the test tasks in question.	The test format (simulated lecture), the scale, and the score/feedback report have been developed based on extensive needs analyses, which included classroom observations, interviews with lecturers, and discussions with university leaders (e.g., chairs, deans, program managers, union representatives).
Is there a review cycle for the examination? (How often? Who by? Procedures for revising decisions)	No fixed review cycle is in place. However, the scale was revised in 2013 (four years after the launch in 2009), and the feedback procedure was revised in 2015 (six years after the launch). The revisions are performed by the TOEPAS team at the Centre for Internationalisation and Parallel Language Use (CIP) at the University of Copenhagen based on continuous analysis of test data and discussion with candidates and score users.

Form A8: Initial Estimation of Overall Examination Level

Initial Estimation of Overall CEFR Level
Short rationale, reference to documentation TOEPAS is not designed to offer a measure of ability at a single level, instead it measures across levels B2 to C2 – no attempt is made to distinguish levels below B2.

Appendix 3 CEFR Table 3: Qualitative features of spoken language

	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turntaking, referencing, allusion making etc.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skilfully to those of other speakers.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.
B2+					
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.	Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.

The TAEC project has been co-funded by the Erasmus+ programme of the European Union.

"The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein."

Co-funded by the
Erasmus+ Programme
of the European Union

The TAEC project is a collaboration between the following partners:

UNIVERSITY OF COPENHAGEN

Universitat de Lleida

Maastricht University

UNIVERSITÀ DEGLI STUDI DI TORINO