Pertanika J. Soc. Sci. & Hum. 29 (S3): 133 - 156 (2021)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

Aligning the Language Criteria of a Group Oral Test to the CEFR: The Case of a Formal Meeting Assessment in an English for Occupational Purposes Classroom

Priscilla Shak[1]* and John Read[2]

[1]Centre for the Promotion of Knowledge and Language Learning, Universiti Malaysia Sabah, Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
[2]School of Cultures, Languages and Linguistics, The University of Auckland, 18 Symonds Street, Auckland Central, Auckland 1010, New Zealand

ABSTRACT

The Malaysian Education Blueprint (MEB) 2015-2025 has set in motion efforts from all stages of education to align programs, courses, and syllabuses to the Common European Framework of Reference (CEFR) benchmark. This exercise has brought on major revamps in all aspects of English language education in the nation. This study will present such an undertaking in a public university in Malaysia and detail how the language criteria for an oral group test of an English for Occupational Purposes course have been aligned to the stipulated CEFR level. The actual assessment task involved groups of four or five students conducting a meeting of their established company. Data for the study came from an analysis of the audio recordings of nine group meetings, along with post-assessment interviews and focus group discussions involving three EOP instructors. Based on the data analysis, this study recommends a revised set of language criteria for the assessment. Furthermore, it demonstrates how an alignment of the scoring criteria with the descriptors of the targeted CEFR scale can be achieved through a systematic comparison of the language functions (LFs) produced in the meeting task to the targeted CEFR descriptor scales. The revised language component for the meeting assessment could help ease instructors' assessment of students interactional skills and allow them to gauge

ARTICLE INFO

Article history:
Received: 16 July 2021
Accepted: 04 October 2021
Published: 30 November 2021

DOI: https://doi.org/10.47836/pjssh.29.S3.08

E-mail addresses:
pshak@ums.edu.my (Priscilla Shak)
ja.read@auckland.ac.nz (John Read)
*Corresponding author

ISSN: 0128-7702
e-ISSN: 2231-8534

better their students' attainment of the skills required in a formal meeting context.

Keywords: Assessment criteria, CEFR descriptor scales, EOP, formal meeting, group oral, language function analysis

INTRODUCTION

The English Language Education Reform prompted recent prominent transformations of Malaysia's English language education landscape due to the implementation the Malaysian Education Blueprint (MEB) 2015-2025. The MEB, launched in 2015, is a reform plan spanning all stages of education from preschool to tertiary levels, which has resulted in the unified alignment of the English curricula of these institutions to the Common European Framework of Reference (CEFR) (Council of Europe, 2001). The CEFR includes specifications of six levels of proficiency, each of which has been adopted in the MEB as the aspirational target for one level of education in Malaysia: A1for preschool, A2 for primary, B1 for secondary, B2 for post-secondary, and B2 to C1 for university (Ministry of Education Malaysia, 2016).

The CEFR originated as a project sponsored by the Council of Europe in the late 20th Century to promote language learning among adults who had completed their compulsory education. However, it has subsequently become influential at all levels of education in Europe and many other countries worldwide (Byram & Parmenter, 2012; Read, 2019). It is often seen primarily as an assessment scale, and it does serve as a point of reference for many standardized international tests, including IELTS, TOEFL, and TOEIC (Don & Abdullah, 2019; Abidin & Jamil, 2015). However, it has a much broader scope than that: there are multiple scales in the framework that "are accompanied by a detailed analysis of communicative contexts, themes, tasks and purposes" and the "CEFR is used in teacher education, the reform of foreign language curricula, the development of teaching materials and for the comparability of qualifications" (Council of Europe, 2020b).

There have been numerous critics of the CEFR, both in general terms (Fulcher, 2004; Hulstijn, 2007) and more specifically about problems in defining the B2 level for university admission in Europe and Australia (Deygers et al., 2018a; Deygers et al., 2018b). In addition, closer to home Foley (2019) has raised concerns about how the use of the CEFR as a benchmark has been implemented in various ASEAN countries, including Malaysia. Nevertheless, applied linguists have recognized the appeal of the framework to policymakers as a means of articulating language education goals according to internationally defined levels of proficiency and as a tool for accountability in education. As McNamara (2014) has pointed out, "the functionality of a universal letter/number system to code the six levels is a key feature of the CEFR, which makes it attractive to administrators and policymakers" (p. 227).

In Malaysia's case, policymakers insist that a form of standardization is required,

especially to align English graduates' language proficiency across universities and as a form of quality control. As such, it is the public higher learning institutions' role to help the Ministry achieve this target. Accordingly, this article aims to investigate how the assessment of a specific course at a Malaysian university can be aligned to the CEFR B2 benchmark.

The EOP Meeting Assessment as a Test Task

The context of the present study is a course in English for Occupational Purposes (EOP) at a Malaysian university. The students undertake a group project to establish a company, and they are assessed based on their language performance in the task of a simulated company meeting. The main objective of the EOP course is to improve the students' employability by enhancing their language skills to secure future employment and communicate effectively in future workplaces. These include interviewing, presentation, and meeting skills. Specifically, this study focuses on the formal meeting assessment of the EOP course, which is detailed in the next section.

A review of the literature reveals that the meeting test task is somewhat unconventional. For example, Shehadeh (2017) pointed out that there are relatively few studies that investigated the use of task-based language testing (TLBT) in the English for Specific Purposes (ESP) realm despite both sharing similar underlying principles, which are "goal-oriented," "has a real outcome" and "reflects real-life language use and language need" (Shehadeh, 2018, p. 1).

When learners are engaged in a task, they actively focus on meaning-making through interaction in the target language (Nunan, 1989). At the same time, tasks naturally encourage collaboration between learners (Bruton, 2002). In attempting their tasks, learners interact with one another and engage in collaborative efforts to complete the task assigned as there is a real need to do so for mutual benefits (Nakatsuhara, 2013; Shak, 2014; Shak, 2016; Taylor 1983). Therefore, tasks enable language learners to function in "extended, realistic discourse" and help them learn how to use language appropriately for real communicative purposes (Taylor, 1983, p. 70). According to Skehan (1998), managing tasks engages the "naturalistic acquisitional mechanism" that helps learners to develop language skills (p. 95).

For an assessment task to be authentic, it should "parallel those in the real world" (Messick, 1996, p. 3). It means that a task should simulate the target context as closely as possible. Ellis (2003) also highlighted the need for task-based assessment to represent "real-world" behavior and activities (p. 285). In an earlier study undertaken by the first author to investigate the learners' perception of a task-based group project work related to the current study, it was found that the participants viewed the tasks assigned as comparable to a real-world task (Shak, 2014). In addition, for a test task to be useful, it should be informed by the

real-world language use domain (Bachman & Palmer, 1996). Finally, these authors discussed the notion of 'interactiveness,' which refers to the match between the abilities engaged by the test task and those that learners require in the target language use (TLU) context. Following Bachman and Palmer's visual representation, the TLU domains and tasks for this study are presented in Figure 1.

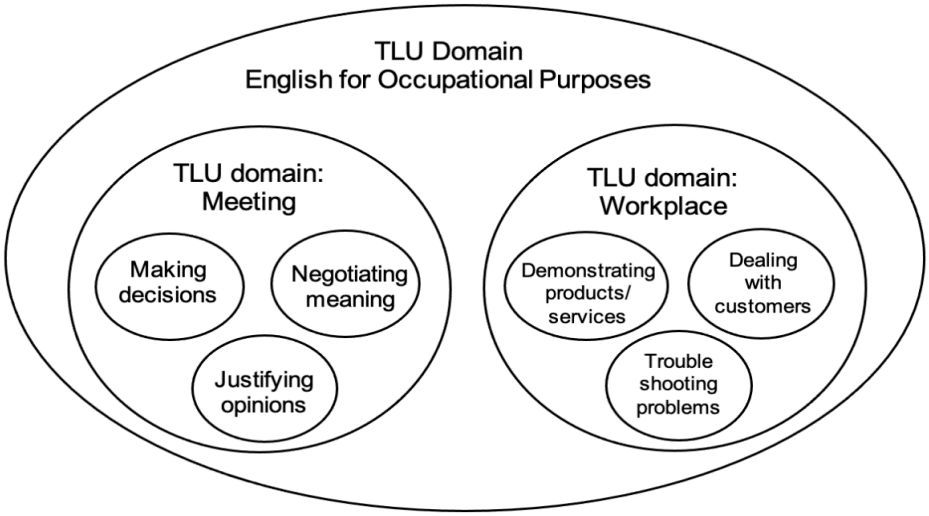

Figure 1. English for Occupational Purposes TLU domain and TLU tasks

As illustrated in Figure 1, the tasks in the TLU domain that apply to the EOP meeting require the test takers to make decisions, negotiate meaning and justify opinions. These functions are among those that are necessary for the successful completion of the meeting assessment task.

Previous studies have highlighted the central role of discourse analysis in offering insights into the nature of interactions in various testing contexts (McNamara et al., 2002; Nakatsuhara, 2013; van Batenburg et al., 2018; Woodward-Kron & Elder 2015). In addition, researchers studying institutional talk have identified formal meeting talk as a genre distinct from other institutional discourse and ordinary conversation (Angouri & Marra, 2010; Asmuß, 2013; Asmuß & Svennevig, 2009; Drew & Heritage, 1992; Svennevig, 2012a; Svennevig, 2012b). Therefore, assessments focusing on this genre should concentrate on its distinctive characteristics and the acquisition of relevant skills to perform the meeting tasks. The appropriate tool for this purpose is Language Function Analysis, which is discussed further in the Data Analysis section below.

The Present Study

The main objective of this study, which is part of a larger-scale project, is to recommend a revised marking scheme for the meeting assessment of the EOP course offered by a Language Centre in a public university in Malaysia. The paper focuses on the alignment of the assessment criteria to the stipulated CEFR B2 level. As such, the paper addresses the following two research questions:

1. What problems did the EOP instructors face when using the existing marking scheme to assess their students' interactional competence?

2. How can the existing marking scheme be revised to align with the CEFR B2 level?

Two sets of qualitative data were obtained from the EOP instructors to address the first research question: individual interviews after the assessment and a Focus Group Discussion (FGD). The synthesized data provided specific details regarding the problems faced by the instructors when assigning marks to their students and their thoughts on the alignment to the CEFR level. For the second research question, results from a Language Function Analysis (LFA) performed on audio recordings of the meeting assessment task were compared to the benchmarked CEFR B2 level descriptor scales for formal discussion (meetings), and recommendations were made based on the findings. The result is a recommended revised version for the language component of the meeting assessment marking scheme.

The EOP Meeting Assessment

The main purpose of the EOP meeting assessment was to evaluate whether the students had acquired the language skills needed to communicate successfully in a meeting setting. In addition, students were tested on their abilities to use language in a formal context and handle such workplace demands in the future. Based on their group project and the roles or positions, each of the students participated in a meeting assessment following a pre-agreed agenda for their group's meeting. The students' main task was to resolve their agenda items to their meeting objective(s). While performing the different roles assigned to them for the meeting test task, students were expected to utilize various language functions such as agreeing, clarifying, suggesting, justifying, negotiating, reciprocating, and interrupting to resolve their agenda items.

The assessment of the meeting task was guided by a marking scheme that contained a list of 16 Likert-type scale items. In accordance with the task-based nature of the EOP group project, the marking criteria focused on the abilities of the students to undertake the meeting task. The evaluation form covered three main components: content and organization (30 marks), presence (20 marks), and delivery, language, and grammar (30 marks). Table 1 lists the items for each of the components. Each item was graded according to a scale of one (very poor) to five (excellent), and each student was assigned individual marks.

While the study was being conducted, the center reviewed all of its English courses

Table 1

EOP meeting assessment's marking criteria

Content and organisation (30%)	Quality of ideas or contents presented in the meeting Sufficient support for ideas Active contribution in the discussion Organized and clear presentation of ideas Perform role assigned effectively Adhere to correct meeting procedures
Presence (20%)	Physical appearance, neatness, and grooming Posture, gestures, mannerism, and movement Eye contact and rapport with group members Listens attentively and shows respect when others are speaking
Delivery, language and grammar (30%)	Enthusiasm and vocal variation (freedom from monotone) Preparation and knowledge of materials (confident and organized) Vocabulary and use of appropriate words (meeting terminologies) Freedom from distracting "uh"s and "like"s Pronunciation, enunciation, audibility, and clarity Grammar

to align them to the Common European Framework of Reference (CEFR) to implement the nationwide English Language Education Roadmap standardization process under the Malaysian Education Blueprint (MEB). As mentioned in the Introduction, part of the MEB requirements is for all English courses in public universities across Malaysia to be aligned to the CEFR's B2 or C1 levels. Given this, the English Language Unit of the Centre determined that the EOP course would be aligned to the CEFR B2 level. This alignment meant that the EOP course would need to produce language learners capable of demonstrating a B2 level of proficiency. As such, it is important that the course assessments could determine whether the learners can perform at this level. Due to this, the assessment criteria of the course would need to be revised according to this benchmark so that an accurate assessment of the learners' proficiency can be correctly mapped to the targeted level.

MATERIALS AND METHODS

The formal meeting assessment involved groups of four or five students. Based on a meeting agenda prepared by the students in advance, each group member was assigned an agenda item based on their role in the project. It provided an information gap as each student had information not available to the others. Following formal meeting conventions, a chairperson was appointed for each group to lead the meeting. Each group was given between 20 to 25 minutes to complete the task. In total, nine meeting groups were audio-recorded.

Each test-taker was awarded individual marks based on the three main rating criteria: a). content and organizations, b). presence, and c). delivery, language, and grammar (Table 1). This paper will focus on the third criterion, the delivery, language, and grammar component.

Participants

In total, 42 second-year undergraduates taking the EOP course and three full-time EOP instructors participated in the study. The student participants had scored Band 1 or 2 in the Malaysian University English Test (MUET), which is a prerequisite for university entrants. The instructor participants recruited the student participants (30 females and 12 males) from their respective classes. Each instructor recruited three groups from their classes. All the instructors were experienced in teaching the EOP course.

Procedures

Each meeting assessment session was attended by the instructor (as evaluator), one group of students (as test-takers), and the first author (as non-participant observer). All the assessment sessions were audio-recorded, as it is less intrusive than video recording for data collection during an assessment event. All the audio files were downloaded into the NVivo 12 software and transcribed orthographically using the transcribe feature of the software. In total, nine transcripts were obtained and analyzed.

All the instructors' post-assessment interview sessions were conducted the week after the meeting assessments. For the post-assessment interviews, a set of semi-structured questions was utilized (Appendix A). Questions relevant to this part of the study included the instructors' feedback regarding their students' performance and their difficulties assigning marks. In total, 136 minutes of recorded data were obtained. In addition, all instructor participants attended a focus group discussion (FGD) as a follow-up to their post-assessment interviews. The FGD was conducted to obtain collective input from the instructors to identify similar issues faced in assigning marks and discuss possible solutions to the problems faced. The FGD lasted for approximately 1 hr 48 min. Appendix B shows the FGD questions.

Data Analysis

The Language Function Analysis (LFA) procedures reported here are situated within a larger project focusing on using group oral assessments in the EOP classroom. For the LFA, both the audio recording and verbatim transcriptions were used concurrently. Therefore, it was necessary to identify the language functions (LFs) that required extensive re-listening and re-reading, and contextual information was essential. The O'Sullivan et al. (2002) Observation Checklist was utilized as an initial operational coding guide (Table 2) to ensure systematic coding of the LFs. Although developed for "real time" use in the Cambridge Main Suite examination paired

speaking test, the successful application of O'Sullivan et al. (2002) checklist was also reported in other studies of oral group tests (Brooks, 2003; Nakatsuhara, 2013).

To ensure that the LFs were coded reliably, the first author and a second coder specializing in English language testing coded all nine transcripts. In instances where there was coding disagreement, specifically those associated with codes where the kappa values were below 0.4, indicating less to a fair agreement (Fleiss et al., 2003; Landis & Koch, 1977; Sim & Wright, 2005; Vierra & Garrett, 2005), the items were further examined and discussed. Upon reaching a final consensus, the kappa values for these items were recalculated. The overall Cohen's kappa value for all of the codes for all the sources is 0.94. Thus, it indicates a high level of inter-coder reliability. In addition, for all codes, average kappa values between 0.71 to 1.0 were obtained.

For the instructors' post-assessment interviews and the focus group discussion (FGD), the audio files were transcribed verbatim orthographically in Word document file format (.docx). The transcripts were then uploaded to NVivo and prepared for coding. Several rounds of close and repeated reading were done before the data were segmented and subjected to thematic analysis coding, allowing researchers to focus on the content highlighted by the participants (Zacharias, 2012). Fereday and Muir-Cochrane (2006) refer to this as "a form of pattern recognition within the data" (p. 82), thus enabling the authors to focus on the specific theme of interest. After the initial coding, the codes and categories were further refined for final data coding before the data was reported.

For the instructors' post-assessment interviews, the themes were coded under two main categories. The first category coded was the challenges in group discussion assessment, which was further sub-coded into i) the scripted discussion; ii) quantity versus quality; iii) role assignment; iv) personality and v) proficiency. The second category coded focused on the challenges posed by the marking criteria. Similarly, for the FGD, the two main categories identified in the post-assessment interviews were used in the NVivo coding. The sub-themes coded under the theme of the challenges in group discussion assessment were i) the scripted discussion, ii) role assignment, iii) monopoly of talk, and iv) proficiency.

Meanwhile, the sub-themes coded under the theme of the challenges in group discussion assessment were i) generic language component, ii) group collaboration, and iii) interpretation of the assessment items. For this study, codes related to the language component of the marking criteria were highlighted in the results section. Data obtained from the post-assessment interviews and the FGD were instrumental in providing the writers with the directions in which the revised assessment criteria should take; most importantly, they need to move towards a more CEFR-aligned format.

RESULTS

Table 2 presents the range of language functions and corresponding percentage of

test-takers use. Additional LFs not found in the original checklist (O'Sullivan et al., 2002) are shown in bold italic typeface. For example, eight additional LFs under *Interactional* functions were identified, while four additional functions under the *Managing interaction* functions were found.

Table 2

The percentage of test-takers for each of the language functions used

Informational functions	%	Interactional functions	%	Managing interaction	%
Expressing opinions	90.5	Asking for opinions	61.9	Reciprocating	42.91
Providing information	83.3	***Asking for confirmation***	59.5	***Nominating***	33.3
Elaborating	76.2	***Confirming***	59.5	***Concluding***	26.2
Justifying opinions	71.4	***Commenting***	54.8	Changing	23.8
Suggesting	66.7	Agreeing	54.8	***Interrupting***	21.4
Describing	31.0	Negotiating meaning	52.4	Deciding	19.0
Staging	14.3	Asking for information	50.0	***Prompting***	4.8
Speculating	14.3	***Acknowledging***	47.6	Initiating	4.3
Summarizing	14.3	***Instructing***	33.3		
Comparing	7.1	***Assisting***	33.3		
Expressing preferences	4.8	***Assuming responsibility***	26.2		
		Modifying	16.6		
		Disagreeing	9.5		
		Granting permission	9.5		

*Additional LFs in ***bold italics*** typeface

As can be seen in Table 2, the meeting assessment elicited the highest number of *Interactional* functions (14 LFs), followed by *Informational* functions (11 LFs) and *Managing Interaction* Functions (8 LFs). It demonstrated the propensity of the meeting test task to elicit the desired functions, which in turn indicated the overall effectiveness of the group oral in prompting interaction among the meeting participants. Thus, it

can be regarded as validating the use of the task to assess the test-takers interactional competence.

Apart from that, the additional LFs identified under the *Interactional* and *Managing Interaction* functions were also unique to the test task, which exemplifies how a specific-purpose assessment task could elicit LFs distinct from other types of group interaction. As presented in this section, identifying the LFs elicited from the test task is crucial in recommending a revised language component for the meeting assessment. It will be addressed further in the Discussion section.

The Instructors' Perspectives

This section presents the data collected from the three EOP instructors' post-assessment interview and focus group discussion (FGD) sessions. It primarily discusses the instructors' concerns regarding their difficulties in evaluating their students' interactional skills and assigning student marks. The instructors' post-assessment interviews were necessary to gain their feedback based on their assessed groups and their personal opinions regarding the assessment task. Meanwhile, the FGD was utilized to obtain collective input regarding what the instructors recognized were the main assessment issues regarding the use of the meeting test task. It was especially useful to gauge their views on what needed to be done to improve the meeting assessment further. The results in this section are based on the synthesized findings.

As the meeting discussion was individually assessed, Instructor 2 expressed that some students did not "care about other people" but focused only on speaking during their turns. As such, interaction and input to each other's topics were minimal, and the desired scaffolding did not occur. These test-takers, it seemed, focused only on presenting their ideas, and, as soon as they had voiced their opinions, they ceased to contribute. "When they're not speaking, you know that they're not in the meeting already… Only doing their part, and that's it", said Instructor 2. Although she observed such behavior, Instructor 2 could not penalize her students as such criteria were not stipulated in the marking scheme. Nevertheless, it was an issue for Instructor 2 as she could not adequately assess her students' interactional skills.

Since the meeting assessment was meant to gauge the test-takers abilities to engage in group interaction, they needed to be involved in the co-construction of the interaction rather than merely presenting their ideas. Therefore, the existing marking criteria that focus on language and grammar components are not particularly relevant for assessing the test-takers interactional abilities. For example, one component focused on vocabulary use, specifically meeting terminologies and useful meeting expressions, but that did not cover the test-takers abilities to use such expressions to co-construct the discussion by continuing, elaborating, negotiating and sustaining the topics being considered.

Both Instructor 1 and Instructor 2 agreed that aligning the existing marking scheme to the CEFR would help improve the validity of the marking scheme in assessing the test-takers interactional skills more effectively and fairly. Instructor 1 believed that the test-takers language abilities could be better gauged if they were assessed based on more specific criteria and "not just by performing [the meeting task]." It implies that the test-takers performance should not be judged solely based on their language abilities to complete their own assigned role but also the means through which they collaborated with the others to accomplish the joint task.

Instructor 2 stressed the need to assess both language and meeting management skills as "they are inter-related. Because if you are able to conduct the meeting, definitely, you have a certain degree of language ability in order to carry out all the procedures, convey ideas clearly and understand others." Hence, in her opinion, the assessment criteria should take these aspects into account. As East (2016) has argued, although to a certain extent, task completion is dependent on linguistic abilities, it may not be a sufficient criterion to assess proficiency in this specific context, where proficiency also involves the ability to engage and interact with each other's thoughts and opinions in order to reach a consensus.

For Instructor 3, the existing marking scheme did not pose any problems for her. She typically adhered to it fairly strictly and would award marks based on the criteria stipulated. Hence, she did not assess components absent from the marking scheme. Interestingly, this was an aspect that she did not realize and only became aware of when attending the FGD. It illustrates how relevant interactional skills might have been neglected in these oral assessments as the focus was just on the linguistic aspects of the test-takers abilities. Nevertheless, Instructor 3 agreed that alignment to the CEFR would entail some revisions to the existing language criteria and believed this move would be more positive.

Overall, although all the instructors agreed that the existing marking scheme allowed them to gauge the competencies required to perform the meeting task and could provide information regarding the test-takers abilities to participate in the discussions, the criteria lacked focus on the use of specific language functions, especially those associated with the group interaction in a meeting. This aspect could be improved with alignment to the relevant CEFR scale.

As the study was being undertaken when the alignment of the EOP course to the CEFR had been proposed in line with the Ministry's standardization exercise, there was increased awareness on the instructors of the need to comply with this requirement. As a result, both Instructor 1 and Instructor 2 could pinpoint the specific table for the Formal discussion (Meetings) scale in the CEFR. Table 3 shows the illustrative descriptors for spoken interaction in that context.

Table 3

CEFR's formal discussion (meetings) illustrative descriptors scale (Council of Europe, 2020a, p.78)

	Formal discussion (Meetings)
C2	Can hold their own in a formal discussion of complex issues, putting an articulate and persuasive argument at no disadvantage to other participants. Can advise on/handle complex, delicate, or contentious issues, provided they have the necessary specialized knowledge. Can deal with hostile questioning confidently, hold on to the turn and diplomatically rebut counter-arguments.
C1	Can easily keep up with the debate, even on abstract, complex, unfamiliar topics. Can argue a formal position convincingly, responding to questions and comments and answering complex lines of counter-argument fluently, spontaneously, and appropriately. Can restate, evaluate and challenge contributions from other participants about matters within their academic or professional competence. Can make critical remarks or express disagreement diplomatically. Can follow up questions by probing for more detail and can reformulate questions if these are misunderstood.
B2	Can keep up with an animated discussion, accurately identifying arguments supporting and opposing points of view. Can use appropriate technical terminology when discussing their area of specialization with other specialists. Can express their ideas and opinions with precision and present and respond to complex lines of argument convincingly. Can participate actively in routine and non-routine formal discussion. Can follow the discussion on matters related to their field, understand in detail the points given prominence. Can contribute, account for, and sustain their opinion, evaluate alternative proposals and make and respond to hypotheses.
B1	Can follow much of what is said related to their field, provided interlocutors avoid very idiomatic usage and articulate clearly. Can put over a point of view clearly, but has difficulty engaging in debate. Can take part in a routine formal discussion of familiar subjects clearly articulated in the standard form of the language, or a familiar variety that involves exchanging factual information, receiving instructions, or discussing solutions to practical problems. Can follow argumentation and discussion on a familiar or predictable topic, provided the points are made in relatively simple language and/or repeated, and opportunity is given for clarification.

Table 3 *(Continued)*

	Formal discussion (Meetings)
A2	Can generally follow changes of a topic in formal discussion related to their field, which is conducted slowly and clearly. Can exchange relevant information and give their opinion on practical problems when asked directly, provided they receive some help with formulation and can ask for repetition of key points if necessary. Can express what they think when addressed directly in a formal meeting, provided they can ask for repetition of key points if necessary.
A1	No descriptors available
Pre-A1	No descriptors available

DISCUSSION

As the authors were made aware of the need for the EOP course to align to the CEFR B2 benchmark, careful consideration was given to meeting this requirement. Hence, in making recommendations for improvement, the authors decided to incorporate the relevant CEFR scale for formal discussion and meetings into the assessment scheme to illustrate what the test-takers should do at the B2 level. However, it has to be pointed at this juncture that a higher number of the LFs produced by the test-takers corresponded more closely to the descriptors below the dividing line after the second statement in the B2 level descriptors. It indicated that the test-takers were likely to be at the lower range of B2 performance, which was to be expected as it represented a more realistic target for Malaysian students with MUET Band 1 and 2 scores. Nevertheless, there were also instances where the more proficient test-takers could produce LFs that reflected higher-level descriptors. Therefore, it indicated that the meeting assessment task was able to elicit LFs beyond B2 level performance. However, as the EOP course has been benchmarked at the B2 level, the revisions were made based on comparison to this level of descriptors.

In order to incorporate elements of the CEFR descriptors into revised language criteria for the meeting test, the authors examined the LFs generated from the meeting assessment, specifically those that yielded higher percentages of test-taker use (ranging from 50% to 90.5%) and compared these to the CEFR descriptors. Table 4 illustrates this comparison.-

After examining the corresponding LFs to the CEFR descriptors, the recommended revisions for the language and delivery components were put forth and presented in Table 5 to replace the existing delivery, language, and grammar components of the meeting assessment (Table 1).

Table 4

CEFR B2 descriptors scale for formal discussion and meeting and the corresponding language functions

Level	Descriptors scale for formal discussion and meetings	Corresponding Language Functions
B2	Can keep up with animated discussion, accurately identifying arguments supporting and opposing points of view. Can express his/her ideas and opinion with precision, present and respond to complex lines of arguments convincingly. Can participate actively in routine and non-routine formal discussion. Can follow the discussion on matters related to his/her field, understand in detain the points given prominence by the speaker. Can contribute, account for, and sustain his/her opinion, evaluate alternative proposals and make and respond to a hypothesis.	(Dis)agreeing Supporting Negotiating meaning Expressing/Asking for opinions Justifying opinions Suggesting Asking for confirmation/ Confirming Elaborating Commenting Asking for/Providing information

Table 5

Recommended revisions for the language and delivery components

Language and Delivery
Can present with confidence and enthusiasm (vocal variation, e.g., freedom from monotone).
Can use accurate vocabulary and grammar (appropriate meeting terminologies and sentence structure).
Can speak with correct pronunciation (enunciation, audibility, and clarity).
Can speak fluently (free from lengthy/frequent pauses and distracting fillers, independent of notes).
Can contribute ideas and suggest alternatives.
Can respond to ideas by (dis)agreeing, commenting, confirming, and negotiating meaning.
Can sustain discussion by elaborating, supporting, and justifying opinions and/or arguments.

As presented in Table 5, the recommended version incorporates 'can do' statements, characteristic of the CEFR. These statements correspond to the B2 level of the CEFR's formal discussions and meetings scale. In this revised version, four of the descriptors from the original CEFR list are integrated. Where broader behavioral features are indicated in the CEFR, they are represented more explicitly in the revised version of the marking scheme. For example, at the CEFR B2 level, students 'can keep up with animated discussion, accurately identifying arguments supporting and opposing points of view' (Table 4). These skills are represented in the revised version's abilities to 'present with confidence and enthusiasm' and sustain the discussion by 'elaborating, supporting, and justifying opinions and/or arguments.' It is also worth pointing out that the recommended version does not emphasize accuracy in grammar and pronunciation. Not because these are not important but mainly because these features could be better tested through the other types of assessment that the test-takers have to perform in the EOP course, such as the test, presentation, proposal, and portfolio tasks. As such, the assessment of the meeting task should concentrate more on the abilities of the test-takers to perform interactional functions in such a setting. As Galaczi and Taylor (2018) have recommended, CEFR descriptors should be further refined to meet stakeholder needs. In the case of this study, one of the considerations for the revision of the assessment criteria is the concept of test localization, which "stipulates that for a test to be valid, its design and development must take into consideration the population, context, and the domain in which the test is used" (Abidin & Jamil, 2015, p. 1).

This study has utilized the qualitative bottom-up approach to gain insights into the language produced by the test takers to substantiate the recommendations for a revised marking scheme. At the same time, the post-assessment interviews and FGD with the instructors revealed concerns about the marking scheme and the need to align it with the benchmarked CEFR level, which has illuminated aspects that required improvement.

One of the main aims of language proficiency testing in ESP is to assess test-takers performance based on a simulated setting to predict their capacities to tackle such real-world demands in the future (Basturkmen & Elder, 2004; Douglas, 2000; Woodward-Kron & Elder, 2015). The results of the LFA indicated that, in addition to the LFs found in the assessment of dyads, the group format could generate a wider range of LFs, which lends support to its use for assessing the interactional competence of language learners. Most importantly, the group meeting task could generate language functions that reflect those in natural workplace settings. It is an important aspect of the EOP course as students are exposed to realistic and meaningful interaction. When "the language learners are functioning in the target language in situations similar to the ones they experience every day, they may start internalising English and their motivation may increase" (İlin, 2014, p. 2).

As illustrated in this study, identifying LFs in a meeting setting is instrumental in informing the design of revised marking criteria for the language component of the meeting evaluation form. The recommended language descriptors make it easier for the instructors to evaluate a student's performance. However, as the stakeholders require, they align with the CEFR's formal discussion and meeting descriptors. Despite skeptics' claims, the CEFR can serve as a rich resource for rating scale development and adapted to various testing conditions (Deygers & Van Gorp, 2015; North, 2014; Weir, 2005a; Weir, 2005b; Abidin & Jamil, 2015).

CONCLUSION

This study has illustrated how the language criteria of an EOP meeting assessment can be aligned to the CEFR by demonstrating in detail the steps involved in the alignment process. Qualitative data obtained from the EOP instructors' post-assessment interviews and FGD were utilized to identify the specific issues they faced while assigning students marks to help determine areas requiring revision. In addition, the LFA provided empirical evidence of the LFs elicited by the task. It enabled them to be compared to the CEFR descriptors, which led to the recommended revised criteria.

The methodological implication of the study is that data from the corpus of students' meeting assessment events are a rich and viable resource for the alignment of assessment criteria with the objective and learning outcome of a course. By examining in-depth what was produced by the test-takers in an actual assessment event and comparing this to the targeted performance descriptors, CEFR-compliant assessment criteria could be devised to ensure that the assessment method correlates with the desired level of performance. In this case, the LFA was useful to help gauge the effectiveness of the meeting test task to elicit the desired language output and served as an effective method to map the elicited output to the CEFR's B2 level descriptors for formal meetings and discussions. The result was the recommended CEFR-aligned marking criteria for the language component as presented earlier.

The limitation of this study is that data were collected from just a small number of instructors. Despite this, feedback from these experienced instructors indicated that they were aware of the shortcomings of the assessment scheme utilized then. Another shortcoming is that the trial of the revised assessment has yet to be undertaken. Nevertheless, the proposed revised criteria presentation to the three instructor participants and preliminary discussions indicated that the recommended version would likely ease the challenges of grading the students. In addition, the resulting assessment marks would better reflect the students' interactional abilities. Another limitation concerns the focus of the recommended revisions based on the B2 level descriptors. It has to be acknowledged that it is possible for other lower (B1 below) or higher levels (C1 and C2) LFs can manifest during the formal

meeting assessment. Nevertheless, as highlighted earlier, since the Centre has determined the EOP course to be aligned to the B2 level, the main focus of the revisions in this study was placed on this level's descriptors. Nonetheless, similar processes may be adopted for the other CEFR level descriptors in other contexts based on the steps undertaken in aligning the marking criteria detailed in this study.

An area worth exploring in the future is the trialing and implementing this revised marking scheme to gauge its effectiveness and a further detailed examination of other assessment criteria to enhance further the overall assessment of the students' interactional abilities.

ACKNOWLEDGEMENTS

This paper is based on a completed doctoral study with financial support from the Ministry of Higher Education Malaysia and Universiti Malaysia Sabah.

REFERENCES

Abidin, S. A. Z., & Jamil, A. (2015). Toward an English proficiency test for postgraduates in Malaysia. *SAGE Open, 5*(3), 1-10. https://doi.org/10.1177/2158244015597725

Angouri, J., & Marra, M. (2010). Corporate meetings as genre: A study of the role of chair in corporate meeting talk. *Text and Talk, 30*(6), 615-363. http://doi.org/10.1515/TEXT.2010.030

Asmuß, B. (2013). Conversation analysis and meetings. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 2006-2008). http://doi.org/10.1002/978140598431.wbeal0210

Asmuß, B., & Svennevig, J. (2009). Meeting talk. *Journal of Business Communication, 46*(1), 3-22. http://doi.org/10.1177/0021943608326761

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

Basturkmen, H., & Elder, C. (2004). The practice of LSP. In A. Davies & C. Elder (Eds.), *Handbook of applied linguistics* (pp. 672-694*)*. Blackwell. https://doi.org/10.1002/9780470757000.ch27

Brooks, L. (2003). Converting an observation checklist for use with the IELTS speaking test. *Cambridge ESOL Research Notes, 11*, 20-21.

Bruton, A. (2002). From tasking purposes to purposing task. *ELT Journal, 56*(3), 280-288. https://doi.org/10.1093/elt/56.3.280

Byram, M., & Parmenter, L. (Eds.) (2012). *The Common European Framework of reference: The globalisation of language education policy*. Multilingual Matters. https://doi.org/10.21832/9781847697318

Council of Europe (2001). *Common European Framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2020a) *Common European Framework of reference for languages: Learning, teaching, assessment - Companion volume.* Council of Europe Publishing.

Council of Europe (2020b). *The Common European Framework of reference for languages*. Council of Europe Publishing.

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing, 32*(4), 521-540. https://doi.org/10.1177/0265532215575626

Deygers, B., Carlsen, C. H., Saville, N., & Van Gorp, K. (2018a). The use of the CEFR in higher education: A brief introduction to this special

issue. *Language Assessment Quarterly, 15*(1), 1-2. https://doi.org/10.1080/15434303.2017.1421957

Deygers, B., Carlsen, C. H., Saville, N., & Van Gorp, K. (Eds.) (2018b). Special issue: Language tests for academic enrolment and the CEFR. *Language Assessment Quarterly, 15*(1), 1-108. https://doi.org/10.1080/15434303.2017.1421957

Don, Z. M., & Abdullah, M. H. (2019, May 17). What the CEFR is and isn't. *Free Malaysia Today*. https://www.freemalaysiatoday.com/category/opinion/2019/05/27/what-the-cefr-is-and-isnt/

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

Drew, P., & Heritage, J. (1992). Analyzing talk at work: An introduction. In P. Drew, & J. Heritage (Eds.), *Talk at work: Interaction in institutional settings* (pp. 3-65). Cambridge University Press. https://doi.org/10.1075/fol.1.2.08ade

East, M. (2016). *Assessing foreign language students' spoken proficiency: Stakeholder perspectives on assessment innovation*. Springer.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods, 5*(1), 80-92. https://doi.org/10.1177/160940690600500107

Fleiss, J. L., Levin, B. A., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Wiley-Interscience.

Foley, J. (2019). Issues on assessment using CEFR in the Region. *LEARN Journal: Language Education and Acquisition Research Network, 12*(2), 28-28.

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly, 1,* 253-266. https://doi.org/10.1207/s15434311laq0104_4

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 218-236. https://doi.org/10.1080/15434303.2018.1453816

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *Modern Language Journal, 91*, 663-667. https://doi.org/10.1111/j.1540-4781.2007.00627_5.x

İlin, G. (2014). Student-teacher judgements on Common European Framework: Efficacy, feasibility and reality. *Journal of Language and Literature Education, 9*, 8-19. https://doi.org/10.12973/jlle.11.221

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

McNamara, T. (2014). 30 years on–evolution or revolution? *Language Assessment Quarterly, 11*, 226-232. https://doi.org/10.1080/15434303.2014.895830

McNamara, T., Hill, K., & May L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221-242. https://doi.org/10.1017/s0267190502000120

Messick, S. (1996). Validity and washback in language testing. *ETS Research Report series, 1996*(1), i-18. https://doi.org/10.1002/j.2333-8504.1996.tb01695.x

Ministry of Education Malaysia. (2016, April 27). Executive summary: *Malaysia education blueprint 2015-2025 (Higher Education)*. Kementerian Pendidikan Malaysia.

Nakatsuhara, F. (2013). *The co-construction of conversation in group oral test*. *Language Testing and Evaluation, 30*. Peter Lang.

North, B. (2014). *The CEFR in practice*. *English Profile Studies, 4*. Cambridge University Press.

Nunan, D. (1989). *Designing tasks for the communicative classroom*. Cambridge University Press.

O'Sullivan, B., Weir, C. J. & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing*, *19*(1), 33-56. http://doi.org/10.1191/0265532202lt219oa

Read, J. (2019). The influence of the Common European Framework of Reference (CEFR) in the Asia-Pacific region. *LEARN Journal, 12*(1), 12-18.

Shak, P. (2014). Incorporating task-based group project work in English for Occupational Purposes Course: The instructors' perspectives. *MANU Journal, 21*, 77-97.

Shak, P. (2016). Taken for a ride: Students' coping strategies for free-riding in group work. *Pertanika Journal of Social Science & Humanities, 24*(1), 401-414.

Shak, P. (2019). *Towards a framework for effective group oral assessment in the ESP classroom* [Unpublished Doctoral dissertation]. University of Auckland.

Shehadeh, A. (2017). Foreword: New frontiers in task-based language teaching. In M. Ahmadian & M. Mayo (Eds.). *Recent perspectives on task-based language learning and teaching* (pp. vii-xxi). De Gruyter Mouton. http://doi.org/10.1515/9781501503399-015

Shehadeh, A. (2018). Task-based language assessment. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English Language Teaching*. (pp. 1-6). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118784235.eelt0379

Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257-268. https://doi.org/10.1093/ptj/85.3.257

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Svennevig, J. (2012a). Interaction in workplace meetings. *Discourse Studies*, *14*(1), 3-10. http://doi.org/10.1177/1461445611427203

Svennevig, J. (2012b). The agenda as a resource for topic introduction in workplace meetings. *Discourse Studies*, *14*(1), 53-66. http://doi.org/10.1177/1461445611427204

Taylor, B. P. (1983). Teaching ESL: Incorporating a communicative, student-centered component. *TESOL Quarterly, 17*(1), 69-88. https://doi.org/10.2307/3586425

van Batenburg, E. S. L., Oostdam, R. J., van Gelderen, A. J. S., & de Jong, N. H. (2018). Measuring L2 speakers' interactional ability using interactive speech tasks. *Language Testing, 35*(1), 75-100. https://doi.org/10.1177/0265532216679452

Vierra, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The Kappa statistics. *Family Medicine, 37*(5), 360-363.

Weir, C. (2005a). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3) 281-300. https://doi.org/10.1191/0265532205lt309oa

Weir, C. J. (2005b). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Woodward-Kron, R., & Elder, C. (2015). A comparative discourse study of simulated clinical roleplays in two assessment contexts: Validating a specific-purpose language test. *Language Testing, 33*(2), 251-270. https://doi.org/10.1177/0265532215607399

Zacharias, N. T. (2012). *Qualitative research methods for second language education: A coursebook*. Cambridge Scholars Publishing.

APPENDICES

Appendix A

Post-assessment interview questions (adapted from Shak, 2019)

1. What do you think of your students' overall performance for the meeting assessment?
   Potential prompts:
   a) Are you happy with the performance of the groups?
   b) Are you happy with the students' performance?

2. For the formal meeting assessment, were there any successful group discussions that stood out?
   Potential prompts:
   a) Why was/were the discussion(s) successful?
   b) What did the students do to make the discussion successful?

3. Did any of the students perform well beyond your expectation of him/her?
   a) Why was the student's/students' performance successful?
   b) How did this affect your marking?

4. During the meeting assessment, were there any students who performed badly?
   a) Why were the students' performance less successful?
   b) What did the students do/fail to do?

5. Do you think the group discussion assessment format is suitable for assessing your students' language skills?

   Follow-ups if YES:
   a) Why?
   b) How?

   Follow-ups if NO:
   a) Why?
   b) What method(s)/format(s) would you suggest instead?

6. In your opinion, is the use of the group discussion assessment fair for the students?

   Follow-ups if YES:
   a) Why?
   b) Please elaborate on why you feel that it is a fair assessment.
   c) What do you do to ensure that the students are assessed fairly in the group assessment?

   Follow-ups if NO:
   a) Why?
   b) Please elaborate on why you feel that it is not a fair assessment.
   c) What do you think can be done to improve the fairness of the group discussion assessment?

7. During their group assessment, the students were assigned different roles. Do you think this will favor some students (i.e., the chairperson of the meeting) while placing the others at a disadvantage?

Follow-ups if YES:

a) Why?
b) How do you think this can be prevented?

Follow-ups if NO:

a) Why?

8. For the group assessment, is there a specific marking scheme that you adhere to? (Refer to marking scheme)
   a) Did you follow the marking scheme strictly when assessing your students? Why? If not, how did you do it?
   b) How did you use the marks sheets? Do you go according to the list of items in the score sheet?
   c) Do you think the marking scheme reflects the objectives of the meeting discussion assessment? How so? If not, how do you think this can be done?
   d) Do you think the marking criteria allow for effective assessment of the specific language skills required to perform the group discussion task?
   e) Do you think that the marking criteria are suitable for assessing the individual language abilities of the students?
   f) Do you think that the marking criteria are fair for all students?
   g) Do you agree with all the items in the marking scheme? Explain.
   h) Did you face any problems while using the marking scheme? Please explain.
   i) Did you have any difficulty assessing all the students within the duration of their group assessment?
   j) How did you ensure that the assessment was done within the timeframe for each of the students?
   k) In your opinion, how can the marking scheme be improved?

9. The course outline specified groups of four students for the group project. In groups where there were more/extra member(s),
   a) How had the extra student affected the assessment process?
   b) How did you manage the assignment of marks in bigger groups?

10. What did you pay attention to when assigning marks to your students? (eg. Language/performance/cooperation)

11. How did/would you assess students who were quiet during the assessment?
    a) Those who are naturally quiet
    b) Those who are weak in the language
    c) Those who cannot get a word in because of other members who manipulate discussion
    d) Those who chose not to contribute when given a chance (the free-rider?)

12. How did/would you assess students who manipulated most of the talk time during the assessment to get a higher score?

13. How did you use your knowledge of your students to help you in assigning their marks?
14. How did you ensure that everyone gets the marks they deserved and that you have marked them fairly?
15. Were your marks set by the end of the assessment? Did you review your marks? How did you do this?
16. What are your suggestions to make the group assessment process more effective?
17. Do you have anything to add?

Appendix B

Focus group discussion questions (adapted from Shak, 2019)

1. What do you think about the topic that has brought us here today (meeting assessment)?

2. I understand that in this Centre, the course chairperson makes most of the decisions about the course design. What are the roles of the other instructors of the course in the decision-making process?
   Items covered:
   - Course design
   - Course assessment
   - Course content

3. In your opinion, what are the major problems in implementing the group discussion assessment format?
   Items covered:
   - Time constraints
   - Numbers of students in a group
   - Students who free-ride (or do not contribute much to the discussion).
   - Students who monopolize the discussion
   - The different personalities
   - The marking scheme
   - The allocation of marks (individual versus group marks)
   - Whether the marks reflect the individual student's language abilities
   - Whether the marks given are generalizable to other settings. (i.e., whether being able to perform well in the group discussion assessment means being able to perform in other oral tasks competitively as well)

4. What do you think can be done to overcome the problems you (the instructors) face?

5. Could you provide any suggestions on how the group discussion assessment process can be improved?
   Items covered:
   - Planning
   - Strategies to ensure fair evaluation of the students
   - Marking scheme/criteria
     - Task versus construct considerations
     - How to ensure that the student's skills can be captured and are reflected in their scores

- How to ensure that the marking sheet is practical for use for the group discussion assessment

6. Do you have anything to add?