Pertanika J. Soc. Sci. & Hum. 29 (S3): 385 - 400 (2021)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

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

Relating a Sustained Monologue Speaking Production Test to CEFR: Towards Alignment

Hazita Azman[1]*, Zarina Othman[2], Chairozila Mohd. Shamsuddin[2], Wahiza Wahi[2], Mohd Sallehuddin Abd Aziz[1], Wan Nur'ashiqin Wan Mohamad[2], Shazleena Othman[2] and Mohd Hafiszudin Mohd Amin[2]

[1]Centre for Research in Language and Linguistics, Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia
[2]School of Liberal Studies (CITRA UKM), Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

ABSTRACT

This article evaluates a sustained monologue speaking production test to validate its link to the CEFR model. The monologue test is a low-stakes production test that engages the test taker in sustained monologue tasks targeted at B2-C1 of the CEFR levels. The evaluation of the test included determining the extent to which the monologue speaking tasks and the single assessment criterion-related rating scale developed for the test are valid and reliably aligned to CEFR benchmarked descriptors. The socio-cognitive framework for test evaluation was adopted, and an explanatory sequential mixed-methods research design was implemented. The evaluation revealed some contentious points of contrast between the test items and the language demand that each item prompted in production. Consequently, selected items were improved or deleted to ensure the appropriate competency levelled at B2-C1 are correctly prompted. Additionally, the findings underlined the imperative need for test developers to adhere to five inter-related sets of procedures in the justification of a claim that the monologue speaking test is aligned to the CEFR. These include familiarisation, specification, standardisation and benchmarking, standard-setting, and validation. It emerged that thorough familiarity with the CEFR by test item writers and examiners is a fundamental requirement for a test closely related to CEFR construct and levels. Thus, familiarisation training of CEFR and its illustrative

ARTICLE INFO

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

DOI: https://doi.org/10.47836/pjssh.29.S3.20

E-mail addresses:
hazita@ukm.edu.my (Hazita Azman)
zothman@ukm.edu.my (Zarina Othman)
chairozila@ukm.edu.my (Chairozila Mohd. Shamsuddin)
wawa@ukm.edu.my (Wahiza Wahi)
salleh@ukm.edu.my (Mohd Sallehuddin Abd Aziz)
wanshiqin@ukm.edu.my (Wan Nur'ashiqin Wan Mohamad)
shazleena@ukm.edu.my (Shazleena Othman)
hafiszudin@ukm.edu.my (Mohd Hafiszudin Mohd Amin)
* Corresponding author

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

descriptors is a mandatory prerequisite for ensuring test items and assessment of the elicited production correspond to the levels and ratings described in the CEFR model.

Keywords: Aligning to CEFR, assessing ESL speaking, speaking production, sustained monologue tasks

INTRODUCTION

When the Ministry of Education Malaysia (2015) decided on CEFR as the governing framework of international standards for developing English language proficiency programmes at preschool, school and tertiary levels, the need to align language curriculum, teaching and learning, and assessments to CEFR became obligatory. In doing so, the corresponding content and performance levels descriptors drawn from CEFR were made the target proficiency level for each of the education stages (Ministry of Education Malaysia, 2015): preschool at A1, primary at A2, secondary at B1, post-secondary at B2, tertiary at B2-C1, and teacher education at C1-C2 (Khan et al., 2019; Uri & Aziz, 2020). Furthermore, in line with the Ministry of Education Malaysia's (MoE) aspirations, the Ministry of Higher Education of MoHE (2018) required universities to align their English language assessments to CEFR or adopt CEFR aligned proficiency tests.

Hence, the initiative to develop and implement a sustained monologue speaking production test at a local university was motivated by three major factors. Firstly, the ability to speak and communicate proficiently in English has been commonly identified as a competency sought after by employers when recruiting new graduates. Second, the onset of globalisation has made this requirement increasingly imperative for non-native speakers of the language (Manokaran et al., 2021).

Secondly, the launch of the roadmap for English language education reform by the Ministry of Education Malaysia in 2015 provided direction for the standards of English language competencies that language curriculum from preschool to tertiary levels are expected to reach. These standards, informed by the Common European Framework of Reference for languages or CEFR (Council of Europe, 2011), stipulated students at the tertiary level to graduate with at least a minimum proficiency level equivalent to an independent user at CEFR B2-C1 levels (Ministry of Education Malaysia, 2015). Towards this end, MoHE required universities to employ CEFR aligned tests only to report their students' proficiency levels (MoHE, 2018).

However, subjecting students to CEFR aligned examinations that are readily available in the market raises the issue of affordability, especially for most students at public universities. Thus, developing an internal low stake test became the preferred option for our university. As such, this circumstance is the third impetus for developing the Sustained Monologue Speaking Production Test or SMSPT, henceforth, using the CEFR model as a referred criterion of standards.

SMSPT was designed to elicit long turn speaking samples that can be assessed to gauge the ability to speak directly on a selected topic in a sustained monologic communication style. The test is conducted face-to-face with an interlocutor who prompts the test taker to respond to a selected speaking topic. The topics are thematically linked to social and workplace domains. The candidates are given a few minutes to understand the question before responding. Then, they are allowed to enquire for clarification from the interlocutor if necessary. Finally, they are given a maximum of three minutes to respond. The test performance is recorded and rated remotely by two trained examiners.

The developers of SMSPT were informed by several CEFR resources, which included the *Manual for relating language examinations to the CEFR* (Council of Europe, 2009), the *Structured overview of all CEFR scales* (Council of Europe, 2011), the *CEFR Companion Volume* (CoE, 2018), and the updated series of the *CEFR manual 2020* (Council of Europe, 2020). These documents helped familiarize the test developers with constructs of targeted language competencies and specified tasks that elicit language production for the targeted proficiency levels.

This article describes the evaluation conducted on SMSPT towards validating its alignment to CEFR. The evaluation is informed by the Council of Europe (CoE) manual published in 2009 and 2020, which systematically delineates "procedures in a cumulative process to situate examinations in relation to the CEFR" (Council of Europe, 2009, p. 9). The article proceeds to describe the content analysis of SMSPT to determine its cognitive and context validity concerning CEFR, guided by Weir's (2005) socio-cognitive framework for language test validation. Finally, the article illuminates contrasts found between SMSPT and the CEFR model while highlighting implications for changes to SMSPT and the sets of procedures essential towards aligning the test to CEFR.

METHODOLOGY: TOWARDS ALIGNING SMSPT TO CEFR

The first step towards aligning a test to CEFR requires test developers to show how their tests can be related to CEFR in terms of "test content and assessment criteria, and how performance on the language test is interpreted" (Council of Europe, 2011, p. 7). According to the CoE, relating an examination or a test to CEFR "entails implementing five inter-related sets of procedures" (Council of Europe, 2009, p. 9), as depicted in Figure 1. It includes familiarisation, specification, standardisation and benchmarking, standard-setting, and validation processes.

The subsequent section describes the extent to which SMSPT adhered to these five inter-related sets of procedures. The evaluation of this adherence was conducted by an external group of CEFR experts in relating the extent to which SMSPT is aligned to the criteria features of CEFR.

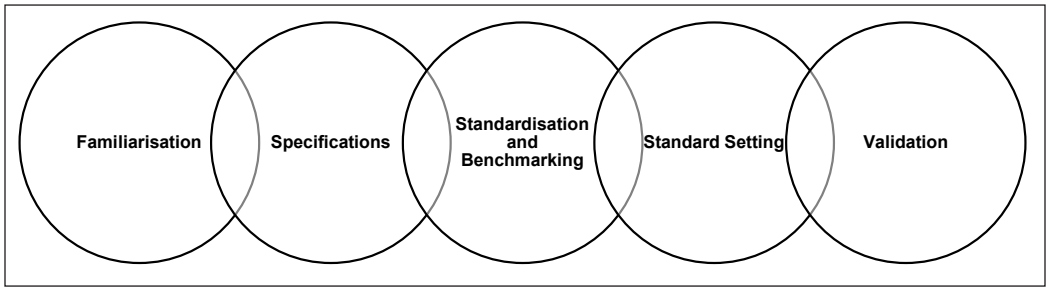

Figure 1. Five inter-related sets of procedures for relating an examination to the CEFR
Source. Relating Language Examination to the CEFR: A Manual (Council of Europe, 2009)

Linking SMSPT to the Five Inter-related Sets of Procedures

It is helpful to begin this section with a brief description of how CEFR views speaking competency. First, the CEFR model makes a distinction between spoken production and spoken interaction that is the ability to speak individually and to interact with two or more people, respectively, on a variety of topics, from familiar to less familiar, situated in domains ranging from social, educational, and occupational, and extended degrees of formality (Council of Europe, 2018).

In CEFR, the spoken production encompasses the ability to produce sustained monologue in the form of "short and simple" directional speech to detailed descriptions and "presentations on complex subjects" in long turn forms (Council of Europe, 2018, p. 68). These monologic tasks may include describing experiences, giving opinions or information, putting a case forward or addressing familiar to complex topics situated in various contexts and domains.

On the other hand, spoken interaction illustrates the ability to interact with verbal exchanges in pairs and groups. The speaker demonstrates the competency in turn-taking skills to initiate, maintain and end the interaction and intervene in ongoing exchanges when appropriate (Council of Europe, 2018). These interaction tasks may include conversations, dialogues, interviews, and group discussions that elicit short turns and joint constructions of discourse to manage and sustain turn-taking in the pair or group interactions.

While CEFR descriptors specify what language learners can do at different proficiency stages (Council of Europe, 2011), it does not clarify or illustrate what materials or tasks should be designed to elicit these abilities for assessment. Furthermore, it does not explain, as it was never intended to do in the first place, how learners can develop their knowledge of spoken English to get to the next CEFR level (Don, 2020). Herein lies the gap between the CEFR specifications and how to operationalize them in translating them into language curriculum, teaching the targeted level and assessing the targeted proficiency. This section addresses the aspect concerning assessment in this lacuna. It describes how we attempted to interpret the specifications and translate them into test items towards

relating SMPST to CEFR within the frames of the five inter-related sets of procedures.

Familiarisation

Familiarisation is a procedure where "the language test developer must demonstrate an in-depth knowledge and understanding of the CEFR descriptors that illustrate the salient features of the language proficiency in different skills at the different levels" (Council of Europe, 2009, p. 17).

The SMSPT test developers comprised ESL experts, who received a one-week familiarisation training conducted by the Cambridge Assessment English (CAE) experts. Based on their shared understanding of CEFR obtained through related documents and training, the test developers derived task specifications from the B2 CEFR descriptors. Based on these specifications, 80 monologue task items were developed, and only 50 were selected for SMSPT after a pilot test analysis. Test items were randomly selected from this selection by an interlocutor during the speaking test.

Specifications

Specification procedure requires "detailed descriptions of the test, profiling its test specifications for content analysis and verification of the abilities that are tested can be related to the relevant CEFR descriptors, categories and levels" (Council of Europe, 2009, p. 29). The specifications specify (1) the speaking production abilities that can be assessed at the targeted levels of proficiency, (2) the types of real-world speaking purposes that the targeted abilities and level of proficiency will fulfil, and (3) the rating descriptors that distinguish one level of proficiency from another to rate the performance of these competencies as concisely and comprehensibly as possible.

The CEFR model (Council of Europe, 2011) identifies five spoken production tasks—*addressing audiences, public announcements, describing an experience, giving information,* and *putting a case*. The SMSPT test developers *described the experience* and *put a case* as the two categories of production tasks that are assessed. It is mainly because these speaking tasks are commonly practised in English language proficiency courses at the university.

Standardisation and Benchmarking. Standardisation training is an extended part of the familiarisation cycle where test examiners or raters work with exemplar performances and test tasks to achieve an adequate understanding of CEFR levels and develop an ability to relate the local test tasks and performances to those levels (Council of Europe, 2009). For SMSPT, both standardisation and benchmarking were conducted in the same session, following the procedures explicated in the CEFR manual (Council of Europe, 2009, p. 40-53). In addition, rater standardisation documents containing selected exemplar performance from validated pilot sessions, sample tasks, rating scales, and sample marks were compiled for the one-day examiner training session.

The benchmarking session progresses with sample tasks from the actual test, where test examiners practice rating the production videos individually and in small groups. Finally, a plenary group discussion is conducted to reach a consensus regarding assigning a particular performance to a CEFR level. A single assessment criterion that referenced CEFR with bands corresponding to A2-C1 levels was used to rate the test performance. To confirm inter and intra rater reliability, training of interlocutors and test examiners was conducted to ensure standardisation in the rating of test performances across examiners.

Standard-Setting. Standard-setting procedures (Council of Europe, 2009) is related to establishing the overall validity and reliability of the test concerning its alignment to CEFR standards, categories and levels. Concerning SMSPT, the performance level standards are drawn mainly from the "Can do" statements in CEFR for monologue spoken production descriptors (Council of Europe, 2018, pp. 68-73). The assessment criteria used for SMSPT covers levels A2 to C1, and the standardisation training provided shows cased exemplars of test performances that were gauged at the said levels. Of course, the concern with the standard-setting results applied for SMSPT is whether the CEFR level allocated to the student performances is trustworthy.

We now turn to the discussion about the validation process and procedures, the fifth and final phase in the process of linking a test to CEFR. However, this discussion after that will be restricted to the examination of the SMSPT test items, mainly to highlight salient aspects of the monologue tasks, in terms of cognitive, context, scoring and criterion-related validity (Weir, 2005), and ways in which the test can more clearly be linked to CEFR.

Validation. The validation procedure conducted on SMSPT involved a content analysis of its test items to determine the extent to cognitive validity, context validity, scoring validity, and criterion-related validity can be linked to the CEFR descriptors and established standards. Hereafter, the scope of discussion related to validation is limited to highlighting the points of similarity and contrast between SMSPT test items and CEFR descriptors. To this point, the framework analysis of the evaluation conducted on SMSPT to justify its link to CEFR is explicated.

Framework Analysis of the Evaluation Conducted on SMSPT

Content Analysis of SMSPT Test Items. As part of the validation, the procedure to link SMSPT to CEFR, a content analysis of the test was conducted with three primary purposes in mind: 1) To examine the test items for evidence of cognitive validity and context validity (Cambridge Assessment English, 2019) in order to validate the extent to which the monologue tasks elicit the competencies described in CEFR for level B2 specifically. 2) To investigate the scoring validity of the test to determine the reliability of the judgment by the test

examiners in rating the test performances to a CEFR level. 3) To reference the test, for evidence of criterion-related validity, to external validations. In the case of SMSPT, this entailed comparing the judgments of external experts trained with CEFR knowledge with the scores allocated by SMSPT test examiners.

The mediating theoretical framework employed for the validation process is Weir's (2005) socio-cognitive framework for language test validation. It is in line with the use of language for social purposes as defined in CEFR.

The framework adopts an interactionist position in defining language ability construct where "ability is defined both in terms of cognitive abilities and mental processing of individual learners as well as the interaction of these abilities with the surrounding social and contextual factors" (Cambridge Assessment English, 2019, p. 8). Weir (2005) identifies five critical components of test validity as indicated by the darkened boxes in Figure 2. Only four components, namely cognitive, context, scoring and criterion-related validity, will be predominantly discussed in the findings. Consequential validity would require an extensive study which is beyond the scope of this article.

An explanatory sequential mixed-methods approach (Creswell & Clark, 2011) was adopted for the validation process of the test item analysis as "it allows for qualitative methods to establish a rich explanation of the quantitative results from the participants' perspectives" (Zeiglar & Kang, 2016, p. 56). Thus, the first phase of the data collection was quantitative, and which was then fed into a qualitative focused stage before both the quantitative and qualitative data were combined to provide an integrated interpretation of the findings, as shown in Table 1.

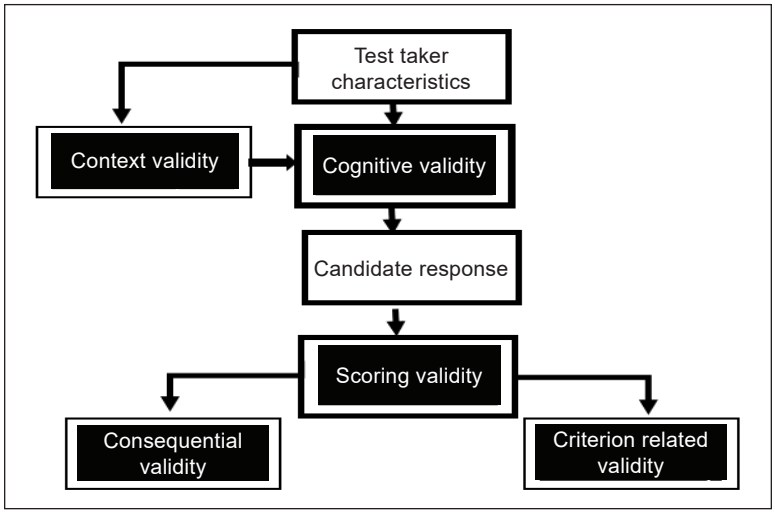

Figure 2. A socio-cognitive framework
Source. Weir, C. J. (2005), pp. 44-47

Table 1
Sequential explanatory mixed-method research design

Focus of analysis	Phase 1: Quantitative ⇨	Phase 2: Qualitative ⇨	Integration & Interpretation
Test appraisal of monologue task items by evaluators	• 50 Test items were randomly selected and examined by a panel of 10 experts. • The panel comprised CEFR trained experts with more than ten years of experience in assessing speaking. • The panel were asked to judge and assign each item a CEFR level. It was done separately and individually.	1. Data gathered from the quantitative phase was processed and reviewed. 2. The quantitative data were referred to in a focus group discussion with the panel of experts. 3. Annotations from this phase are transcribed and inform the interpretation and integration of the data. 4. The discussion was framed in line with aspects of the socio-cognitive framework.	Quantitative and qualitative data are combined to provide a rich review of the content analysis of the test items and the extent to which they can be linked to the CEFR.
Test performance appraisal: video and scores allocated	• Video recordings of Test performances taken during SMSPT were viewed and judged by the panel of 10 experts. • The panel were asked to rate and allocate a score to each performance based on a single-criterion scale drawn from the CEFR 2017 descriptor tables for spoken language. • These judgements were compared with the scores already allocated by examiners of the SMSPT.		

FINDINGS AND DISCUSSION

Cognitive Validity

The construct of cognitive validity establishes the types of cognitive processing or cognitive load that is activated by the test question and the extent to which the cognitive processes required to respond to the question are appropriate for the target language level, candidates, and purpose of the test (Taylor, 2011). Field (2011) proffers a cognitive processing model for speaking, depicting six stages of how a speaker processes information in preparation for speech production: conceptual, syntactic, lexical, phonological, phonetic, and articulatory stages. The model is depicted in Figure 3. The following section summarises the findings concerning the cognitive processing triggered by test items of SMSPT.

Cognitive Processing Triggered by Test Items in SMSPT

According to Taylor (2011), the design of a speaking task must be mindful of the cognitive demands that the given task may have on the test taker. Thus, the test taker's performance is highly dependent on whether the speaking task required of them is familiar and is pitched at a suitable level in terms of ideas or topics and linguistic complexity. Table 2 illustrates descriptors in the CEFR scale for overall spoken production that offers ideas for speaking tasks. At the B1 level, for instance, the focus of the speaking tasks should be on personal and everyday information, 'within his/her field of interest, presenting it as linear sequence of points.' In contrast, at the B2 level, the items should prompt developed descriptions on a wide range of subjects, expanded with supporting ideas and relevant examples on familiar and less familiar topics of interest.

Regarding CEFR, topic familiarity and any other reliance on content knowledge that can facilitate rather than inhibit performance are important features to carefully consider when selecting ideas for topics of spoken production assessment (Alderson, 2000; Galaczi & French, 2011). For example, with

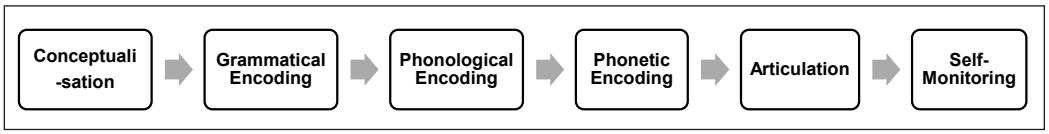

Figure 3. A model of cognitive processing for speaking assessment
Source. Cognitive validity (Field, J., 2011)

Table 2
Overall spoken production CEFR scale for A2-C1 levels (Council of Europe, 2018, p. 75)

2	Can give clear, systematically developed descriptions and presentations, with appropriate highlighting significant points and relevant supporting details. Can give clear, detailed descriptions and presentations on a wide range of subjects related to their field of interest, expanding and supporting ideas with subsidiary points and relevant examples.
B1	Can reasonably fluently sustain a straightforward description of one of a variety of subjects within their field of interest, presenting it as a linear sequence of events.

SMSPT, it was found that conceptualisation of the topics and themes for the monologue speaking tasks were relatable to candidates' living experiences and are therefore suitable.

However, as the test is a one-level criterion-referenced test targeted at B2 spoken production level, all of the 50 test items should be comparable within the test. An evaluation of each of the test items examined by the evaluators found potential issues in this regard. The evaluation revealed that 20% or only 10 test items are estimated to target at B2, whereas the majority of the items or n=24 (48%) is found to be estimated between B2 and C1 levels. Meanwhile, 8 or 16% of the items are estimated to be between B1–B2 levels, and four items or 8% were found to be ranged between C1-C2 levels. However, another four items were found to be unsuitable to the CEFR category of topics. This level of parity in terms of idea provision is a pressing concern as the overall data shows that a total of 38 (76%) of the 50 test items appear to skew towards high B2-C2 levels. Figure 4 illustrates the findings by evaluators in their comparative estimation of three SMSPT items, from the theme about learning a foreign language to the CEFR level.

As Field (2011) observes, the distinction in cognitive load between a B1 and B2 task item is often most noted in the wording of the test items themselves. The way the test question is posed entails grammatical encoding to be applied by the candidates as they attempt to comprehend the purpose of the speaking task and trigger the related linguistic patterns required to perform the task successfully.

Compare, for example, the sample from SMSPT as illustrated in Figure 4. Although the three items displayed were originally designed to be comparable at the B2 level, the analysis revealed that the way the questions were worded could potentially raise the cognitive and linguistic demand of the task, resulting in the disparity. For instance, the way a candidate may respond to "What are some benefits of learning a foreign language?" and to "Students should be required to learn a foreign language. Do you agree or disagree?" would elicit oral competency of differing levels. While the likely response to the former question is factual and may elicit a simple listing of positive factors drawn from personal experiences or opinion, the latter question, by comparison, is somewhat evaluative, inviting an appraisal of personal, public, and national policy perspectives. Thus, triggering B1 and B2 levels of competencies, respectively. Similarly, the third question in

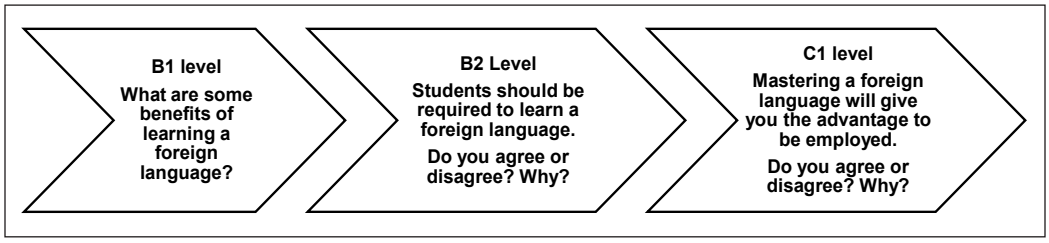

Figure 4. Evaluation of 3 items from theme on Learning a Foreign language and their CEFR level estimates

Figure 4 above is more evaluative than it is factual. It is because, as commented by an expert panel: "… the item invites nuanced views and the use of complex structures to express and defend an opinion(s)," estimated at the C1 level.

Another recurring theme that was found lacking in SMSPT is in its holistic assessment criteria scale. The current descriptions in the SMSPT scale do not reference specific spoken competencies sufficiently to provide an accurate measure of the 'cognitive processes which would prevail in a natural context' (Field, 2011, p. 66). Some of these include criterial features of phonological encoding, articulation, and self-monitoring. CEFR specifically references these features as they are also viewed as indicators of proficiency in spoken production. Table 3 shows findings related to evaluating SMSPT holistic assessment criteria regarding these criterial features that needed improvement.

Context Validity

Weir's socio-cognitive framework identifies specific aspects to context validity for speaking. Salient aspects of these contextual factors are addressed below regarding the SMSPT task items, highlighting the extent to which the evaluation of the characteristics of the test items and their administration are appropriate to the target candidates, levels, and test purpose.

The evaluation found that the long-turn monologic task format in SMSPT is a semi-controlled response format that tends to 'elicit predominantly informational functions', typical of an examiner-candidate format. In appraising the test items, the

Table 3
Comparable findings between CEFR and SMSPT for spoken competency criterial features

Cognitive Processing features indicating spoken competence	CEFR descriptors for fluency at B2 level (CoE, 2020)	SMSPT assessment descriptors
Phonological encoding: Use of pre-assembled chunk, length of run, duration of planning pauses, frequency of hesitation and pauses	Can produce stretches of language with a fairly even tempo; although the speaker can be hesitant as he or she searches for patterns and expressions, there are noticeably long pauses.	It does not refer to a tempo, hesitations, searches for patterns and expressions and pauses.
Articulation: Use of appropriate intonation, stress, sound articulation, L1 interference, intelligibility rather than accuracy	Can generally use appropriate intonation, place stress correctly, and articulate individual sounds clearly; accent tends to be influenced by speaker's L1 but has little or no effect on intelligibility.	It does not explicitly reference the quality of articulation in speech. For example, while the descriptors refer to candidates giving 'clear information' with 'few language slips', there is no direct reference to pronunciation of what is produced.
Self-monitoring: Use of self-initiated production strategy to self-repair	Can often retrospectively self-correct occasional slips and errors in sentence production that the speaker becomes conscious of.	It refers to 'correction of slips,' but it is not clear how the ability to self-monitor is assessed.

evaluators pointed out the need for each task developed to reflect real-life skills the test taker may use or need. As Shaw and Weir (2007, p. 71) point out, 'appropriateness of task purpose enhances the authenticity of the assessment because it is imbued with a real-world purpose which goes beyond the ritual display of knowledge for assessment'. Furthermore, it underlines the main point that different purposes require different cognitive processes, which impact the difficulty of a task. Furthermore, the evaluation found that while the current SMSPT items are mainly informational functions, the 'types of talk' (Galaczi & French, 2011, p.163) can be grouped into a range of functions as listed in Table 4.

Table 4
Types of informational functions found in SMSPT

Informational functions in the CEFR	SMSPT Test items
• Expressing opinion	10
• Justifying opinions	10
• Describing	7
• Expressing preferences	7
• Providing personal information	5
• Suggesting	5
• Comparing	5
• Speculating	5

SMSPT's prevalent types of items emerges as "expressing opinions" and "justifying opinions". However, Galaczi and French (2011) note that while functions are present across various levels, some such as 'comparing' and 'speculating' are only tested at higher levels. Hence items types that ask for test-takers to "compare" and "speculate" (5 items each) need to be revised.

It was also argued that a greater range of interaction types that provide adequate coverage of open-ended formats with the interaction between peers should be included in the SMSPT test response format. The evaluation raised the issue of whether the single-question task provides adequate scaffolding for both weaker and stronger candidates. In line with the CEFR descriptors, it was suggested as well that visual and text prompts be provided as scaffolding support to facilitate the cognitive demand of the abstract questions.

Scoring Validity

Taylor and Galaczi (2011) suggest that cognitive, contextual, and scoring validity form the core of the socio-cognitive framework. As pointed out, by focusing on these three core dimensions, test developers can better develop a collection of theoretical, logical, and empirical evidence to support validity claims and arguments about the quality and usefulness of the test (p. 172). In the case of SMSPT, the evaluators made the following observations about the scoring criteria used:

- The wording of the scales is often very negative in tone. CEFR descriptors focus on what the students can do.
- There is a mismatch between some of the tasks and the descriptors in the scale. The tasks need to be revised to match the descriptors measured.
- The link between one scale level and the subsequent need to be

made more evident and specific to show gradation incompetence as illustrated in the CEFR scales.

- The contextual parameters in the current scale should include specific features such as discourse management, grammatical control, phonological encoding, articulation, self-monitoring and mediation following the updated 2020 CEFR scales.

Criterion Related Validity

Evidence of criterion-related validity can be obtained from relating a test to an external standard such as the CEFR model (Khalifa & Salamoura, 2011). For example, in the case of SMSPT, the evaluation found evidence that the test selected production functions focused on the CEFR, specifically concerning the types of talks that elicit "describing experiences and giving information" as well as "putting a case".

As noted earlier, there are aspects that SMSPT needs to further emphasise in its assessment criteria, such as include discourse management strategies, which is viewed as an indicator of fluency and competency in CEFR. For example, production strategy such as self-monitoring, i.e. "Can correct mix-ups with tenses…" at B1 level, can help distinguish competency from B2 level, where a speaker "Can correct slips and errors if he/she becomes conscious of them…." Likewise, a C1 competency who "Can back track when he/she encounters a difficulty…" can be compared with a C2 level speaker who "Can back tract and restructure around a difficulty smoothly…" (Council of Europe, 2018, p.78). Thus the SMSPT scale should be revised to include such aspects of performance for better criterion-related validity towards CEFR alignment.

DISCUSSION AND CONCLUSION

Having explored various aspects of SMSPT, this section discusses implications of the findings and lessons learned by way of conclusion. This discussion is focused on four salient issues about the aim of relating SMSPT to CEFR.

Firstly, in terms of cognitive validity, SMSPT could be enhanced by including a variety of interaction patterns-from controlled to semi-controlled and open communication. It would broaden the construct being assessed while enabling a broader spoken production to assess, ensuring a more valid assessment of the oral performance and its criterial features. Additionally, to achieve more significant cognitive validity, SMSPT should consider including prompts, visual and textual, to lessen the cognitive demand on candidates in tackling the speaking task. It will assist in balancing support for weaker candidates while allowing stronger candidates to show the full range of their speaking ability.

Secondly, the evaluation found a notable disparity in the cognitive demands of individual questions in terms of discourse mode, nature of information, lexical and functional resources required, and topics selected. It has been noted elsewhere that greater control is needed concerning the

relative demand of one question prompt versus another if they are to be genuinely comparable and show a clearer adherence to the assumptions of the CEFR model. Thus, item analysis of test tasks will be conducted to review each speaking task's construction carefully. Vocabulary analytical tools such as Text Inspector will be applied to gauge the CEFR level of each of the test structures or rubric.

The length of familiarisation training provided for the SMSPT test developers and test examiners was inadequate. There is a need to provide prolonged training to ensure a satisfactory level of familiarisation is reached before specifications of the tasks and standardisation of judgements in rating performances can be aligned to the CEFR standards. Therefore, in reviewing SMSPT, retraining the test developers is imperative and revised test validation is a vital criterion. Better rater training would also improve the delivery of the test and encourage more consistent standardised rating.

Thirdly, the current rating scale used for SMSPT, holistic for ease of use, given many candidates to be evaluated, needs to be revised. The rating scale needs more specific descriptions related to production strategies and management discourse subskills, which further distinguishes the competent speaker from the less competent according to CEFR. In addition, it calls for a more nuanced rating system measuring aspects of production strategies such as pauses, compensating and self-correcting. Furthermore, it was pointed out that an analytic scale based on CEFR's multiple illustrative scales for communicative activities, communication strategies, communicative language competence, and plurilingual and pluricultural competences (Council of Europe, 2020) is more valid than a holistic assessment scale. However, regarding SMSPT's test purpose and aims, aspects of plurilingual and pluricultural competences remain unnecessary for inclusion in the revised scale.

Finally, the fourth aspect for SMSPT to consider in its revision is that a speaking performance within the CEFR framework must reflect the underlying assumption that production, reception and interaction, as well as mediation, should be viewed as co-occurring facets of language use rather than activities which happen in isolation (Taylor, 2011). It suggests that there is a need for SMSPT to include a variety of interaction patterns and tasks to be presented to the candidate to ensure better test validity and enhance its coverage of the CEFR. Thus, a monologue speaking production test on its own is limited in its capacity to be linked to the CEFR completely. Considering these revelations, while remaining a sustained monologue test, SMSPT will instead expand its test tasks to address familiar to complex topics situated in various contexts and domains, using visual and text prompts.

In conclusion, the evaluation of SMSPT has revealed some contentious points of contrast between the test items and the language demand that each item prompted in production. Consequently, selected items will be improved or deleted to ensure the appropriate competency levelled at B2-

C1 are correctly prompted. The findings also underlined the imperative need for adherence to procedures for justifying the test aligned to CEFR. Finally, it emerged that familiarisation training of CEFR and its illustrative descriptors is a fundamental prerequisite for attaining this alignment.

ACKNOWLEDGEMENT

This evaluation of SMSPT is made possible by the research grant awarded by Universiti Kebangsaan Malaysia (KRA-2017-004) in collaboration with Cambridge Assessment English, United Kingdom.

REFERENCES

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732935

Cambridge Assessment English. (2019). *Review of a monologue spoken production test*. [Unpublished research technical report]. University of Cambridge.

Council of Europe. (2009). *Relating language examination to the CEFR: A manual.* Council of Europe Publishing.

Council of Europe. (2011). *Common European Framework of Reference for Languages: Learning, teaching, assessment: A structured overview of all CEFR scales*. Council of Europe Publishing.

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volume with new descriptors.* Council of Europe Publishing.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volume*. Council of Europe Publishing.

Creswell, J. W., & Clark, V. L. P. (2011). *Designing and conducting mixed methods research* (2nd ed.). Sage Publications.

Don, Z. M. (2020). The CEFR and the production of spoken English: A challenge for teachers. *The English Teacher, 49*(3), 77-88.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.) *Examining speaking: Research and practice in assessing second language speaking, studies in language testing* (Vol. 30, pp. 65-111). Cambridge University Press.

Galaczi, E., & French, A. (2011). Context validity. In L. Taylor (Ed.) *Examining speaking: Research and practice in assessing second language speaking, studies in language testing* (Vol. 30, pp. 112-170). Cambridge University Press.

Khalifa, H., & Salamoura, A. (2011). Criterion-related validity. In L. Taylor (Ed.) *Examining speaking: Research and practice in assessing second language speaking, studies in language testing* (Vol. 30, pp. 259-292). Cambridge University Press.

Khan, A. M. A., Aziz, M. S. A. & Stapa, S. H. (2019). Examining the factors mediating the intended washback of the English language school-based assessment: Pre-service ESL teachers' accounts. *Pertanika Journal of Social Science & Humanities, 27*(1), 51-68.

Manokaran, J., Soh, O. K., & Azman, H. (2021). Lecturers' perceptions towards students' English language proficiency: A preliminary study. *International Journal of Academic Research in Progressive Education and Development, 10*(2), 957-963. https://doi.org/10.6007/IJARPED/v10-i2/10461

Ministry of Higher Education Malaysia. (2015). *English language education reform in Malaysia: The roadmap 2015-2025.* Ministry of Education.

Ministry of Higher Education. (2018). *the ecosystem for English language learning and assessment in higher education*. Ministry of Education.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.

Taylor, L. (Ed.) (2011). *Examining speaking: Research and practice in assessing second language speaking, studies in language testing* (Vol. 30). Cambridge University Press.

Taylor, L., & Galaczi, E. (2011). Scoring validity. In L. Taylor (Ed.) *Examining speaking: Research and practice in assessing second language speaking, studies in language testing* (Vol. 30, pp. 234-258). Cambridge University Press.

Uri, N. F. M., & Aziz, M. S. A. (2020). The appropriacy and applicability of English assessment against CEFR global scale: Teachers' judgement. *3L The Southeast Asian Journal of English Language Studies, 26*(3), 53-65. https://doi.org/10.17576/3L-2020-2603-05

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

Zeigler, N., & Kang, L. (2016). Mixed methods design - Chapter 4. In A. J. Moeller, J. W. Creswell & N. Saville (Eds.), *Studies in language testing 43: Second language assessment and mixed methods research* (pp. 51-83). Cambridge University Press.