MICHIGAN LANGUAGE ASSESSMENT

MET Go!

Development of the MET Go! Speaking Test

Technical Report

CONTACT INFORMATION

All correspondence and mailings should be addressed to:

Michigan Language Assessment
Argus 1 Building
535 West William St., Suite 310
Ann Arbor, Michigan
48103-4978 USA

T: +1 866.696.3522
T: +1 734.615.9629
F: +1 734.763.0369

info@michiganassessment.org
michiganassessment.org

11/2018

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

MET Go! is a multi-level test of English language ability designed for beginner to intermediate level learners of middle and secondary school age. Developed and produced by Michigan Language Assessment, the test covers the four language skills (listening, reading, speaking, and writing), assessing learners' ability in each area and assisting them as they progress in their learning.

The MET Go! Speaking Test is designed to assess test takers' spoken English proficiency by evaluating their ability to produce comprehensible speech in response to a range of tasks, such as describing objects, people, actions, and experience, and expressing and supporting opinions, on a variety of familiar school and everyday topics. It is conducted and assessed by a Michigan Language Assessment certified speaking examiner and graded using a fit-for-purpose rating tool. The MET Go! Speaking Test is intended to be useful in a variety of educational settings. The results can be used to monitor the progress of English as a Second Language (ESL) or English as a Foreign Language (EFL) learners, as well as for placement or diagnostic purposes to inform instructors of the strengths and weakness of the learners and areas where instruction is needed. Language programs can also use the test to certify whether or not learners have achieved the goals of a language course.

This report describes the development of the MET Go! Speaking Test. It provides information on the development of the test construct, task types, and rating tool, as well as information on score interpretation.

2. TEST CONSTRUCT

2.1. Targeted CEFR Levels

The Common European Framework of Reference (CEFR) provides a common basis for evaluating the ability level of language learners. The framework identifies six broad levels of language ability, and offers illustrative scales and can-do statements that describe "what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively" (Council of Europe 2001, p. 1).

The MET Go! Speaking Test targets speaking abilities at the A1-B1 levels of the CEFR. Both the CEFR (Council of Europe, 2001) and the CEFR companion volume (Council of Europe, 2018) were used by the MET Go! Speaking Test development team throughout the development process as references to inform the design of the test construct, task types, and rating tool.

The can-do statements from numerous CEFR scales were heavily referenced during development. These scales included the overall oral production, sustained monologue: describing experiences, sustained monologue: giving information, general linguistic range, vocabulary range, vocabulary control, grammatical accuracy, phonological control, and spoken fluency scales (Council of Europe, 2001, 2018). Table 1 summarizes the progression in overall spoken production from levels A1 to B1 for learners aged 11 – 15 (Council of Europe, 2001, 2016). As learners progress through each CEFR level they are expected to have mastered abilities described under lower levels of competence. The table shows that A1 level test-takers are able to describe people and places in isolated phrases. More proficient test-takers are able to speak on an increasing range of topics using increasingly complex language (Council of Europe, 2001).

Table 1: Overall Spoken Production (Council of Europe, 2001, 2016)

CEFR Level	Descriptor
B1	Can reasonably fluently sustain a straightforward description of one of a variety of subjects related to school life or within his/her field of interest, presenting it as a linear sequence of points.
A2	Can give a simple description or presentation of people, living or working conditions, daily routines, likes/dislikes, etc. as a short series of simple phrases and sentences linked into a list
A1	Can produce simple mainly isolated phrases about people and places.

2.2. Construct Definition

The MET Go! Speaking Test adopts the interactionalist approach to construct definition which considers performance as the result of traits, contextual features, and their interaction, and therefore, views performance as "a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts" (Chapelle, 1998, p. 43). This perspective takes into account the role of contextual factors while allowing for the generalization of test scores beyond the immediate testing instance. The test construct definition is illustrated in Figure 1 below.

The construct the MET Go! Speaking Test aims to assess is defined as test-takers' ability to perform in English across a range of spoken communicative functions that beginner to intermediate level learners of middle and secondary school age might encounter in the course of routine daily/school life. The language knowledge that the test aims to measure is specified in previous applied linguistics research on components of language ability. Specifically, the first four components of the language knowledge measured in this test, namely phonological knowledge, grammatical knowledge, textual knowledge, and sociolinguistic knowledge, are based on Bachman's and Palmer's (1996) framework of language ability and Fulcher's (2003) framework for describing the speaking construct. First, phonological knowledge refers to knowledge about pronunciation, stress, and intonation. Grammatical and lexical knowledge involves knowledge of syntax and vocabulary to produce formally accurate sentences. Third, textual knowledge refers to ability to produce explicitly marked relationships among utterances (knowledge of cohesion) and to produce organizational development in speech (knowledge of rhetorical organization). Lastly, pragmatic knowledge, including functional knowledge and sociolinguistic knowledge, allows learners to create speech appropriate to a particular language use setting, i.e., to respond appropriately when asked to describe and compare pictures, describe their personal experience, and state and support their opinion or preference. Figure 2 summarizes the types of language knowledge and fundamental processes the test aims to measure.

In the course of responding to the test takers, learners also need to use their strategic competence, defined as "a set of metacognitive components, or strategies, which can be thought of as higher order executive processes that provide cognitive activities" (Bachman & Palmer, 1996, p. 40). These strategies allow test-takers to assess the situation, decide how to respond to a question (i.e., goal setting), and decide the types of language knowledge and background knowledge to use to achieve that goal (i.e., planning).

In summary, the MET Go! Speaking Test is intended to measure the ability of test-takers, who are between 11 – 15 years old, to use phonological

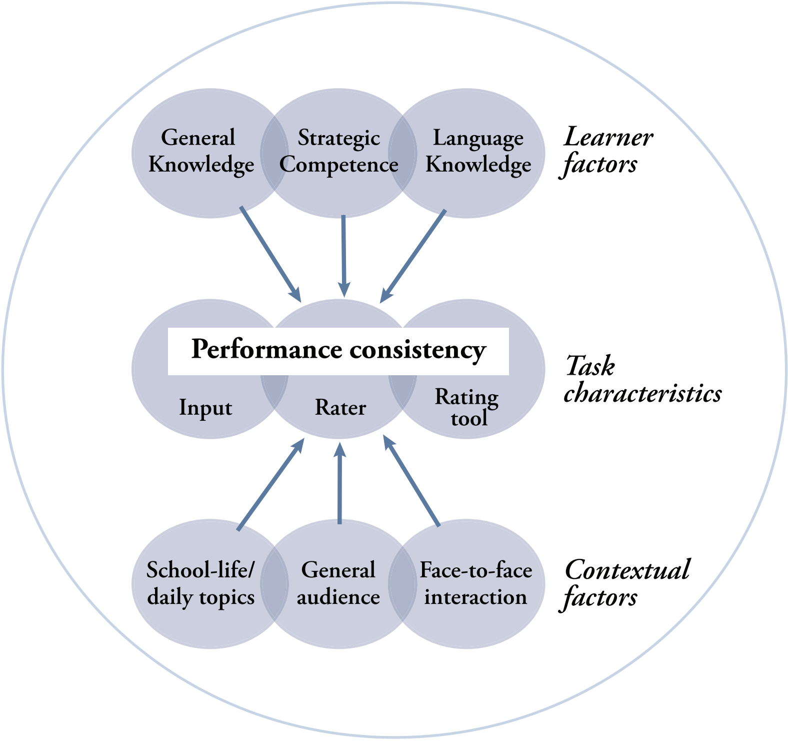

Figure 1: Construct definition of the MET Go! Speaking Test

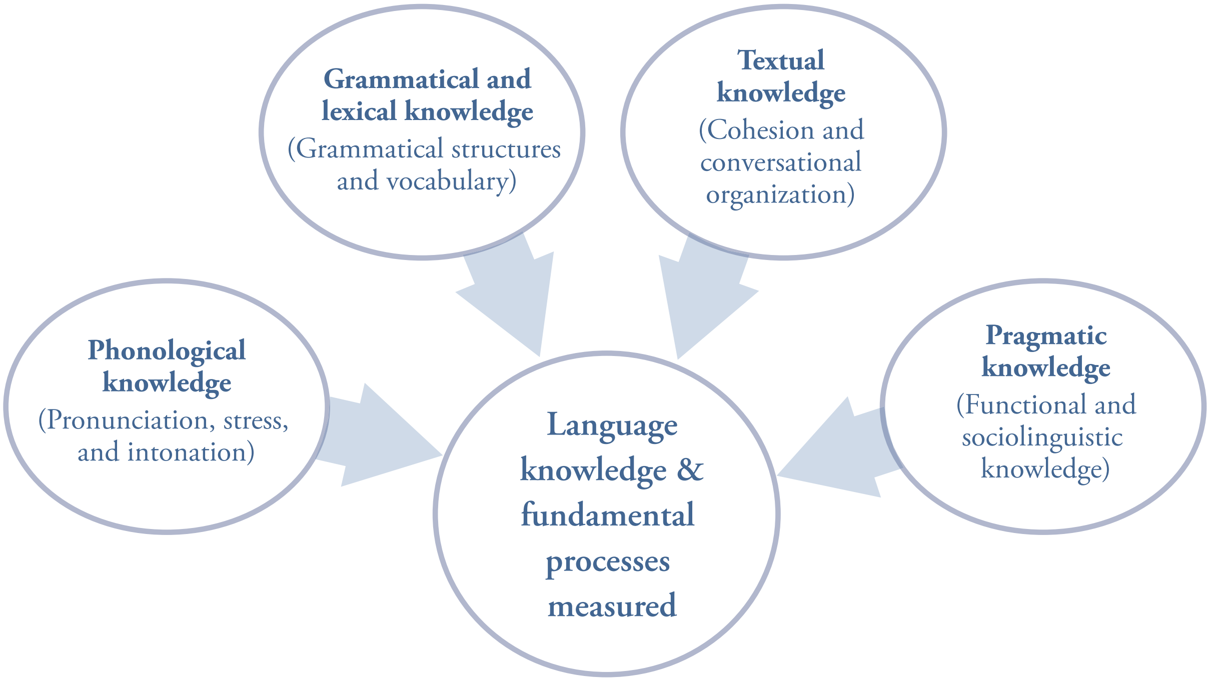
Grammatical and lexical knowledge
(Grammatical structures and vocabulary)
Textual knowledge
(Cohesion and conversational organization)
Phonological knowledge
(Pronunciation, stress, and intonation)
Language knowledge & fundamental processes measured
Pragmatic knowledge
(Functional and sociolinguistic knowledge)

Figure 2: Components of the language knowledge and fundamental processes measured by the MET Go! Speaking Test

knowledge and a growing range of grammatical structures and vocabulary accurately in their oral language to produce responses that are connected, intelligible, and fluent and perform a variety of functions in school-life and daily situations, including describing objects, people, actions, personal experiences, and stating and supporting opinions or preferences on familiar topics.

3. TASK DEVELOPMENT

3.1. Task Theory

The speaking test's tasks are designed to elicit spoken language representing a range of ability levels from beginner to low intermediate (CEFR levels A1-B1). Descriptors of these levels determined the linguistic functions that would be elicited in the test. Specifically, the test targets linguistic functions that distinguish one level from another. For example, while an A2 level learner can state a preference or an opinion, this learner cannot yet elaborate their response and provide relevant supporting details to support this preference or opinion. Production of this latter function is expected to appear in the language of candidates at the B1 level (Council of Europe, 2001).

Examiner variability, as documented in the literature (Brown, 2003; Galaczi, 2008; O'Sullivan, 2002), could introduce construct-irrelevant variance and present a potential threat to the validity of the test score interpretation. In fact, examiner behavior can have an effect both on the amount and type of language that a candidate produces as well as on the score awarded to the candidate (Brown, 2005). Asymmetry between the examiner and the test-taker has the potential to limit the language functions elicited in a speaking test, which can ultimately have severe consequences for the test-taker's final score (Plough, MacMillan, & O'Connell, 2010). In addition, each examiner has "distinct and individual styles which they tend to employ across interviews" (Brown, 2003, p. 2). Therefore, the MET Go! Speaking Test examiner uses a predetermined script to deliver instructions and task prompts. This decision is intended to control for interviewer reliability and standardize test interactions between the examiner and test-takers, providing them with equal opportunities to demonstrate the extent of their proficiency, while also allowing for variability via supplied follow up questions to deal with communication breakdowns.

3.2. Task Design

The MET Go! Speaking Test is a face-to-face test of spoken production, administered by one examiner to one test-taker, and scored in real time. The test consists of four parts accessible to both lower- and higher-level test-takers. Table 2 describes the purpose of each test task, the CEFR level targeted, and the corresponding linguistic functions.

In Part 1, which is not scored, test-takers are asked general questions to help familiarize them with the examiner, understand the test format, and reduce test anxiety. Part 2, aimed at beginner and low-intermediate speakers (A1 and A2 on the CEFR), requires test-takers to give a description of the differences between concrete, familiar topics (such as actions, places, objects, people) based on information given in two picture prompts. Parts 3 and 4 are aimed at more proficient speakers (A2 and B1 on the CEFR). These tasks, which are thematically linked, include a picture prompt and three separate tasks where test-takers are required to give simple descriptions and tell a story, relate a personal experience, and state and support a preference. Specifically, Part 3 targets multiple proficiency levels. For the A2 range it aims to generate a description of places and situations familiar test takers, while for the B1 range it aims to generate narration of actions and/or events. Part 4 – Task 1 aims at A2 level test-takers who are asked to provide a description of a personal experience on a topic related to the picture prompt in Part 2, while Part 4 – Task 2 is targeted at B1 level speakers and requires them to state and support their opinion on a topic tangentially related to the picture prompt in Part 2. These speaking tasks are all presented to the test-taker both orally by the examiner and textually on a prompt sheet to maximize the likelihood that the test-taker will clearly understand the task.

4. RATING TOOL DEVELOPMENT

4.1. Rating Scale Theory and Target Language Features

Following Luoma's (2004) suggestion, evaluation criteria for the MET Go! Speaking Test were developed concurrently with the test construct and tasks. The rating tool was initially developed using theoretical and intuitive methods (Fulcher, 2003; Knoch, 2009). A committee of experts determined the wording of the descriptors and levels in the rating tool. The level descriptors for each criterion were designed to be brief, clear, concrete, and detailed enough (with the absence of field-specific jargon) to sufficiently guide raters from varying backgrounds to rate speaking performances consistently, and also allow them to make quick scoring decisions. Word count and length of each level's performance descriptors were also considered, as the descriptors have to be concrete yet practical to be useful for raters (Luoma, 2004, p. 81).

The CEFR and other relevant assessment literature (Bachman & Palmer, 1996; Fulcher, 2003) were consulted to identify the criteria to be applied. The CEFR presents five qualitative aspects of spoken language use—range, accuracy, fluency, interaction and coherence (Council of Europe, 2001, pp. 28–29)—that is, what a test-taker at each CEFR level "can do" when speaking. Also, as rating criteria should reflect the construct the test aims to measure (Fulcher,

Table 2: MET Go! Speaking Test Parts, CEFR Levels Targeted, and Linguistic Functions

Test part	Description	Level(s) targeted	Linguistic functions
Part 1	Warm-up	N/A	N/A
Part 2	Picture comparison	A1 – A2	Describe differences between concrete, familiar topics, such as objects, people, places, and actions
Part 3	Picture description	A2 – B1	Give simple descriptions and tell a story on concrete, familiar topics
Part 4-Task 1	Personal experience	A2	Give short, basic descriptions of events and activities
Part 4-Task 2	Expressing opinion	B1	State opinions on a general topic and briefly give reasons to support them

1996), the test construct, which is defined based on frameworks of language and speaking ability (Bachman & Palmer, 1996; Fulcher, 2003), were also taken into account. Three evaluation criteria for the rating tool emerged: Task Completion, Linguistic Resources, and Intelligibility.

Task Completion refers to the degree to which the test-taker addresses the task presented in the prompt, that is, the relevance of the response to the task. This criterion also focuses on the quantity of speech and the level of richness of the response.

Linguistic Resources refers to how test-takers use their lexical and syntactic resources to convey meaning. With respect to lexis, previous research has found a relationship between proficiency level and the number of words produced and range of words (Iwashita, Brown, McNamara, & O'Hagan, 2008). Lexical diversity has also been shown to be substantially and significantly correlated with oral proficiency (Yu, 2009). The indicators of syntactic complexity adopted include the use of dependent clauses and verb phrases (Iwashita et al., 2008). This is operationalized in the scale as the use of simple versus complex structures. Lower-level test-takers are expected to have difficulty forming sentences and fragments accurately; verbs marked with tense and aspect and embedding are expected in the performances of more advanced-level speakers (Upshur & Turner, 1995, p. 9). Less proficient students could produce fewer subordinate clauses in their speech (Neary-Sundquist, 2017). Additionally, less able learners are expected to be less accurate grammatically, especially in the use of articles, tense marking, third person singular, and preposition (Iwashita et al., 2008).

Intelligibility refers to the clarity and delivery of the test-taker's response. This criterion is indicated by (1) pronunciation, stress, intonation, and rhythm, which are typically categorized as "phonology" (Brown et al., 2005; Iwashita et al., 2008), (2) the frequency and length of filled and unfilled pauses, (3) number of attempts at repair (trying to self-correct language) or repetitions, and (4) speech rate. It is expected that higher-level learners have more English-like sound patterns (pronunciation, stress, intonation, and rhythm) while lower-level learners have more non-English-like intonation (Iwashita et al, 2008). For instance, target-like syllables at both the word and sub-word level show more noticeable differences across levels. Regarding the fluency of test-takers' speech, three features have been found to discriminate among different English-language proficiency levels: speech rate, unfilled pauses, and total pause time (broadly categorized as hesitations). At lower levels of fluency, overly fast or slow speech rate have been found to cause problems for the listener; at higher levels of fluency, speech rate is typically consistent and appropriate (Brown et al., 2005, p. 38). Also, unfilled pauses have been found to characterize low-level learners while filled pauses and other types of hesitations (such as searching for content words or ideas) have been shown to be markers of higher-level learners (Iwashita et al., 2008). In addition, instances of repair have been shown to contribute to fluency judgments.

4.2. Rating Tool Design

For the MET Go! Speaking Test, the rating tool consists of two components: a checklist and a

Table 3: Evaluation Criteria for the MET Go! Speaking Test

Criteria	Rating Tool Component	Descriptions of features
Task Completion	Checklist	Relevance of response to task • Quantity of speech • Richness of the response (i.e. elaboration, supporting details)
Linguistic Resources	Rating Scale	Use of appropriate vocabulary and grammar to add meaning • Vocabulary: range and accuracy • Grammar: range and accuracy
Intelligibility	Rating Scale	Clarity of message in terms of sound patterns and fluency • Intelligibility: pronunciation of words and phrases, intonation, rhythm of speech, and stress placement on syllables in words and phrases • Fluency: hesitations or pauses; repetitions, occurrences of repair; speech rate

rating scale. While the test developers chose to use a traditional analytic rating scale to assess Linguistic Resources and Intelligibility, they chose to employ a checklist, consisting of multiple items that each referred to a different criterion, to assess Task Completion. Table 3 provides a summary of the rating tool and these evaluation criteria.

In L2 assessment, checklists are commonly used by teachers (and possibly classmates) for continuous assessment of class performances, pieces of work, and projects throughout the course. They can also be used for summative assessment at the end of the course (Banerjee & Wall, 2006; Council of Europe, 2001). Additionally, checklists are very popular for students' portfolio assessment, such as the CEFR "can-do" self-assessment checklists or those developed by the American Council on the Teaching of Foreign Languages.

For the MET Go! Speaking Test, the checklist is meant to provide a reliable assessment of Task Completion and reduce the speaking examiners' cognitive load during rating. Additionally, the checklist is intended to result in positive washback for test users by enabling the test developers to provide more detailed, personalized feedback on the test-takers' performance. This is desirable since it could help test takers to identify specific areas that need improvement, and it could help teachers, school administrators, and other test users to monitor students' progress and plan their curriculum or syllabus accordingly.

Determining the number of scale levels can be more a matter of practicality than of theoretical validity (McNamara, 2000). In deciding on the number of scale levels in the MET Go! Speaking Test, it was important to consider the number of distinctions raters could reasonably be expected to make consistently and the meaningfulness of the number of scale levels in terms of the degree to which they would correspond to the ability levels targeted. In this respect, the number of levels on the checklist was influenced by the number of CEFR levels that were targeted (i.e., A1–B1) and raters' ability to provide reliable ratings for multiple checklist items. As a result, the checklist items are rated on the three categories, namely "yes", "somewhat", and "no", each representing a CEFR level targeted, although the specific level represented by each value varies by checklist item (i.e. "no" can occasionally represent a level lower than A1 and a "yes" may represent a level other than B1). Similarly, the number of levels on the rating scale was also influenced by the number of ability CEFR levels targeted, although another level was added to the top of the rating scale account for test-takers with speaking proficiency higher than B1.

4.3. Piloting the Rating Tool

A pilot study was conducted to ensure that the MET Go! Speaking Test rating tool functioned as expected in terms of its usability, meaningfulness, and ability to distinguish appropriately between test-takers at different levels. Both quantitative and qualitative data were collected for the study. Five raters with a background in linguistics/TESOL scored spoken responses from 100 test-takers from various countries, such as Argentina (16), Bolivia (4), Colombia (23), Greece (3), South Korea (14), Mexico (24), Peru (9), and Uruguay (7). Each test performance was scored by two to four raters using the rating tool. In addition, 15 speaking examiners from Colombia (7), Argentina (2), Peru (2), Bolivia (1), Greece (1), South Korea (1), and Uruguay (1) voluntarily participated in an online survey. The survey consisted of six 4-point Likert-scale items eliciting speaking examiners comments on the checklist and the rating scale in terms of scale understandability, applicability, clarity, item/descriptor distinguishability, appropriateness, and confidence when rating. Qualitative data were speaking examiners' responses to the open-ended questions in the survey.

Results from exploratory factor analysis showed that the checklist items measured the same underlying construct, but only correlated moderately, indicating that they were measuring distinct aspects of the ability being measured. Multi-facet Rasch measurement analyses indicated that the checklist and the rating scale generally functioned as intended. All of the Task Completion criteria included in the final rating tool functioned well, although a small number of changes were made as a result of the pilot results. Specifically, a criterion about describing similarities between two pictures was removed due to its poor performance, and two criteria pertaining to topic relevance were combined with two other items because of their high correlations with these items and the lack of a corresponding ability level in the CEFR for these criteria. The speaking examiners generally reported positive attitudes toward the rating tool on the survey, generally agreeing that the rating tool was easy to understand and apply, and that the checklist and rating scale descriptors were clear, distinguishable from one another, and appropriate for rating test-takers' responses. They also reported high confidence in their scores when using the rating tool.

5. INTERPRETING SPEAKING TEST SCORES

MET Go! Speaking Test scores are intended to reflect test-takers' ability to communicate successfully using phonological knowledge and a range of grammatical structures and vocabulary accurately in their oral language to produce responses that are connected, intelligible, and fluent and perform a variety of functions in school-life and daily situations on familiar topics. Test takers who complete the speaking test will receive a score report that includes a scaled score (0-52) and CEFR level (Below A1-B1) based on their overall performance on the speaking test, as well as personalized feedback in the form of a performance descriptor statement and a recommended learning activity based on their performance on the different parts of the speaking test. For test takers, these results can help them to recognize their strengths and weaknesses and decide on strategies for improving their English. For ESL/EFL instructors, these results can help them place students into appropriate classes, monitor the progress of students in a class, and provide diagnostic information to identify areas where instruction is needed.

6. REFERENCES

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

Banerjee, J., & Wall, D. (2006). Assessing and reporting performances on pre-sessional EAP courses: Developing a final assessment checklist and investigating its validity. *Journal of English for Academic Purposes, 5*(1), 50-69.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1–25. doi: 10.1191/0265532203lt242oa

Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt, Germany: Peter Lang.

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Report Series*. Retrieved on July 18, 2018 from http://www.ets.org/Media/Research/pdf/ RR-05-05.pdf

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (eds.), *Interfaces between second language acquisition and language testing research* (pp. 32-70). New York: Cambridge University Press.

Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Council of Europe. (2016). Collated representative samples of descriptors of language competences developed for young learners aged 11-15 years. Retrieved from https://rm.coe.int/1680697fc9 on June 7, 2018.

Council of Europe. (2018). Common European Framework of Reference for Languages: learning, teaching, assessment. Companion Volume with New Descriptors. Retrieved from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208–238. doi: 10.1177/026553229601300205

Fulcher, G. (2003). *Testing second language speaking*. Harlow, UK: Pearson Longman.

Galaczi, E. D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*(2), 89–119.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (Ed.) *Proceedings of the Computer-based Assessment (CBA) of Foreign Language Speaking Skills* (pp. 29-51). Brussels, Belgium: European Union.

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24–49. doi.10.1093/applin/amm017

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale*. New York: Peter Lang.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

Neary-Sundquist, C. A. (2017) Syntactic complexity at multiple proficiency levels of L2 German speech. *International Journal of Applied Linguistics, 27*, 242–262. doi: 10.1111/ijal.12128.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency pair-task performance. *Language Testing, 19*(3), 277–295.

Plough, I., MacMillan, F. M., & O'Connell, S. P. (2011). Changing tasks…changing evidence: A comparative study of two speaking proficiency tests. In G. Granena, J. Koeth, S. Lee-Ellis, A. Lukyanchenko, G. P. Botana, and E. Rhoades (Eds). *Proceedings of the 2010 Second Language Research Forum* (pp. 91–104). College Park, MD: Cascadilla Press.

Upshur, J. & Turner, C. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3–12. doi: 10.1093/elt/49.1.3

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*(2), 236–259. https://doi.org/10.1093/applin/amp024