Tydskrif vir Taalonderrig - Journal for Language Teaching
- Ijenali yokuFundisa iLimi - IJenali yokuFundisa iiLwimi -
Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta Polelo
- Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo - Ijenali
Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo
- Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig -
Journal for Language Teaching - Ijenali yokuFundisa iLimi
- IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi
- Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo -
Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi
- Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa
Ririmi - Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa
iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta
Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo -
Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo
- Jenala yo Dyondzisa Ririmi
- Tydskrif vir Taalonderrig
- Journal for Language
Teaching - Ijenali
yokuFundisa iLimi -
IJenali yokuFundisa
iiLwimi - Ibhuku
Lokufundisa Ulimi
- Tšenale ya tša
Go ruta Polelo -
Buka ya Thuto
ya Puo - Jenale
ya Thuto ya Dipuo
- Ijenali Yekufundzisa
Lulwimi - Jena?a ya u
Gudisa Nyambo - Jenala yo
Dyondzisa Ririmi - Tydskrif vir Taalonderrig
- Journal for Language Teaching - Ijenali
yokuFundisa iLimi - IJenali yokuFundisa iiLwimi -
Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta
Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya
Dipuo - Ijenali Yekufundzisa Lulwimi - Jena?a ya
u Gudisa Nyambo - Jenala yo Dyondzisa Ririmi
- Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali
yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi -
Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo -
Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi
- Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa
Ririmi - Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa
iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta
Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo -
Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo
- Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig -
Journal for Language Teaching - Ijenali yokuFundisa iLimi
- IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi
- Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo -
Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi
- Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa
Ririmi - - Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa
iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta

Carina Grobler
North-West University

and

Tom FH Smits
University of Antwerp
Rhodes University

The development and refinement of a rating scale for beginner students' foreign language speaking skills

ABSTRACT

Interactional oral language proficiency is a core component of modern foreign language education with 'speaking fluently' a central learning objective (Eisenmann & Summer, 2012). In spite of its importance, speaking is a difficult skill to assess (Tajeddin et al., 2011; Yan, 2014), and there is a gap in literature regarding the reliable and valid assessment of the very basic level of speaking skills. This paper reports on the development process of an assessment instrument, which was conducted as action research and involved both quantitative and qualitative data collection and an analysis. The learning environment and its related technology-enhanced out-of-class practice environment within which the research was conducted, focus on beginner foreign language students. Activities include computer-mediated communication and face-to-face oral activities.

In order to address the need for a valid and reliable assessment instrument, a first version of the instrument was created and subsequently used to assess both computer-mediated communication and face-to-face oral activities. The reliability of the instrument was investigated during two action research cycles by means of studying the consistency, consensus estimates, and intra-rater reliability. The results from the two cycles of investigation informed changes to the instrument, and this ultimately resulted in two assessment instruments that differentiate between technology-enhanced activities and personal interaction. Similar to Gruhn and Weideman (2017), this study was of an exploratory nature and additional design principles would have to be evaluated over a longer period of time.

Keywords: Designing a rating scale; reliability of a rating scale; assessing speaking skills; foreign language assessment; computer-mediated communication

***Journal for Language Teaching*, Volume 55, Number 2, pp. 117 – 145. 2021.**
ISSN 0259-9570. https://dx.doi.org/10.4314/jlt.v55i2.5

1. Introduction

Given the tendency to upgrade the value of oral skills, and with 'speaking fluently' as a central teaching target presently (Eisenmann & Summer, 2012: 415), assessing oral competence must be done in the best and fairest way (Bachman et al., 1995; Lumley & McNamara, 1995). This could be achieved by making use of an assessment grid that is elaborate enough to provide meaningful feedback. The lack of available assessment grids for the speaking skills of beginners that could serve as a guideline in addressing the issue compounds the problem.

Studies on language testing within a South African context tend to focus on creating tests rather than on creating rating instruments for use in different situations. Emphasis is, for example, placed on different aspects of a test assessing the levels of the academic literacy of students entering university by means of a series of multiple-choice questions that include the design, development, and refinement of the test (cf. for example, Van Dyk, 2010; Rambiritch, 2013; Weideman, 2011; Weideman et al., 2016), its validation (for example, Sebolai, 2018), its stability (cf. Van der Slik & Weideman, 2009), and its appropriateness (cf. Scholtz, 2017), the translation of language tests (cf. Butler, 2017; Van Dyk et al., 2011; Van Dyk et al., 2021), and literacy in the first grades of formal education (cf. Kimathi & Bertram, 2020; Prinsloo & Harvey, 2016). There is the occasional study available on using alternative methods to assess oral proficiency (Van der Walt et al., 2008), or on the development and validation of a rating scale. In this particular case, for essay writing in English as a second language as part of the final national senior certificate examination within a South African context (Hatting & Van der Walt, 2013).

When looking at the requirements for the assessment of orals in all official languages offered for the national senior certificate (DoBE, 2020), the descriptors for assessing speaking in a second additional language include items, such as "negotiating a position", "justifying opinion", and "development of ideas and argument". It is, therefore, clear that these descriptors are at a too high level to be useful at the beginning stages of learning a foreign language. As in the study conducted by Hatting and Van der Walt (2013), it is apparent that one would have to explore alternative ways to find an appropriate rating scale for the context of the current study.

Although studies on the assessment of oral performance are conducted regularly, to our knowledge it is a rare occurrence for studies to investigate rating scales for novice foreign language learners. An extensive search for modern assessment grids for speaking skills yielded very few instruments specifically designed for beginner language learners. All but one of these instruments focus primarily on more advanced students. There is, therefore, a gap in literature where the assessment of student performance at the very basic level of foreign language acquisition is concerned.

This article aims to close this gap by presenting the first steps in creating a rating scale for spoken interaction at the beginning stages of foreign language learning. This includes determining the rating criteria to be included, and determining the reliability of the rating scale with reliability describing the level of "agreement between the results of one test with itself or with another test” (Davies et al., 1999: 168) — the consistency or reproducibility of rating outcomes to guarantee their relevance. Although high reliability does not guarantee the validity of results, acceptable reliability is a necessary prerequisite for the valid interpretation of test results, mainly because results cannot be dependable or meaningful without reliability (ALTE, 2011: 16).

2. Context and problem

Learning a new language is time consuming, and requires constant contact with and interaction in the target language. Within the tertiary context of this study, contact time with beginner French foreign language students is limited to two classes of 75 minutes per week. At the end of a 12-week semester, students should have mastered skills at pre-A1 level, meaning they can “produce short phrases about themselves, giving basic personal information (e.g., name, address, family, nationality)” (OECD, 2021: 117).

To create opportunities for students to be exposed to and to practice their speaking skills in French, a technology-enhanced out-of-class practice environment was designed (Grobler, 2020). Within this environment, students have the opportunity to participate in a series of activities that focus on the development of their oral communication skills. The first implementation of these activities starts after five weeks of learning French, and includes the following steps:

1. Listening to a model dialogue illustrating the desired outcomes of a learning cycle, and completing a short quiz on the content of the dialogue on the learning management system.

2. Participating in a simulated dialogue using a custom-designed software package (*Papotons!*).

3. Receiving personalised audio teacher feedback, and the results of teacher assessments.

4. Making student videos with a conversation scenario, and receiving peer feedback.

5. Participating in a summative student-teacher F2F oral test (cf. Grobler & Smits, 2016).

In Step 2, students record their individual contributions of a basic question-and-answer session or “simulated conversation” (Council of Europe, 2001: 178). With the help of the software package, students listen to a pre-recorded question as many times as they wish. They then record an answer to this question, listen to their reply, and either choose to save it and continue to the next question, or to re-record the answer. After completion, the teacher listens to the production and assesses it.

In Step 5, students take part in a F2F oral with the teacher to give students another opportunity to interact and communicate, which is regarded as being of particular importance (Smith & Schulze, 2013). The oral is recorded and assessed afterwards. As with the simulated conversation, the questions asked are aligned with the outcomes that should have been reached at that point in the learning process.

Threats to the validity of these assessment activities are (a) that the simulated conversation is assessed on a vaguely defined sliding scale, with descriptors of the lowest and the highest level for the different categories (Listening*: Understood none of the questions – Understood all of the questions;* Replies to questions*: Gave one-word answers – Always answered in full sentences;* Formulation of sentences*: Did not formulate the answers correctly – Formulated all the answers correctly;* Pronunciation*: Speaks with a very heavy accent – Sounds almost French*), and (b) that the F2F oral is assessed question by question on a 3-point scale (0 = *Cannot answer*, 3 = *Answers correctly in a full sentence*). The assessment is, therefore, not based on the students' overall performance that include aspects such as overall task achievement and quality of replies.

Both teachers and students need to be conscious of the objectives to be achieved. The quality criterion of transparency requires students to be informed of the goals that will be evaluated, and also the criteria used (Van Petegem & Vanhoof, 2002b). The need for an assessment grid, and one that is available to students beforehand, is again stressed.

3. Assessing speaking skills

The process of constructing rating scales and the origin of criteria are rarely reported on, and very little guidance on how rating scales should be created is available (Hatting & Van der Walt, 2013: 74; Knoch, 2011: 84). Rating scales are often created by means of a group of teachers' intuition or through a process of adapting existing scales (Fulcher, 2003; North & Schneider, 1998).

Literature differentiates between holistic and analytic scales. Holistic instruments assess task performance globally by making use of comprehensive descriptions of performance levels, whereas analytic assessment scales contain a number of categories or rating criteria as performance indicators, scaled (and rated) separately. An analytical assessment grid consisting of rating scales (Van Petegem & Vanhoof, 2002a: 51) also provides input for feedback, because the grid contains well-defined performance descriptions doubling as stages in the development of a proficient speaker (Geyskens et al., 2010: 22–24).

Inter-reliability and intra-rater reliability should be investigated, due to research on written and spoken language performance assessments that has indicated several rater effects (cf. Caban, 2003; Eckes, 2005; Knoch et al., 2007). Despite the qualities of rating scales, such as a high intra-rater reliability (i.e., the scoring stability in repeated assessments of the same performance by the same rater), they do not guarantee inter-rater reliability (i.e., the likeliness of different raters to score student productions identically) without necessary rater training (cf. Yan, 2014: 504), the goal of which is to reduce differences in scores from different raters (Davis, 2016: 119). Rater training is a well-established and widely accepted practice that is used in different contexts, and plays an important role in the process of establishing the validity of rating scales (McNamara, 2000: 58). The purpose of this training is to lead raters to an understanding and application of the scoring criteria that accurately reflects the language abilities the test is intended to measure (Davis, 2016: 119).

4. Design and methodology

The purpose of this empirical study was to develop a scoring instrument for beginner foreign language students' oral production, and to establish its level of reliability. This was executed as a design experiment (as defined by Cohen et al., 2018: 413) by adopting a pragmatic research paradigm, which is aimed at finding the most appropriate solution to a specific real-world problem (Creswell & Poth, 2018: 27; Newby, 2014: 48). A design-based approach is interventionist and iterative. Iterative cycles include an analysis, a design, development, implementation, testing, and redesign (Amiel & Reeves, 2008: 35) that lead to a more refined product.

To strengthen the design process, the project was embedded in an action research (AR) framework (as defined by Burns, 2005: 57; and Piggot-Irvine et al., 2015: 548). The iterative process included the development, trialling, and refinement of the scoring instrument. The AR approach seemed to best fit the aims of this study: addressing the issue of a reliable assessment of oral skills to improve practice (Liu, 2013: 102).

The two research questions the design experiment seeked to answer were:

> Research question 1. What rating criteria (sub-constructs) should be used for assessing foreign language students' oral communication skills at an elementary level?
>
> Research question 2. To what extent will using a generic rating scale to grade two different oral activities on a beginner's level influence the raters' consistency and consensus estimates, and the level of intra-rater reliability?

These questions underpinned the investigation of the design of a rating scale for elementary speaking skills in a F2F and CMC foreign language learning environments.

An action research design is cyclical with each cycle informing the next, and is based in practice to allow researchers to learn in and from practice. The different steps that constitute a cycle are described differently by various researchers (cf. Burns, 2005; Tripp, 2004; Kemmis et al., 2014; Riel, 2010). In this study, the following proposition was retained: (1) Reconnaissance (only at the beginning of the first cycle), (2) Plan, (3) Act and Observe, and (4) Reflect (Kemmis et al., 2014: 89). The final stage in the process leads to the planning of the next cycle, and the reflections of the second cycle lead to the planning of the third cycle, and so on. Each cycle has its own question to answer and problem to solve.

Two cycles were run within the scope of the present study. The individual phases are summarised below. The two distinct cycles are presented in separate sub-sections containing information about the question that inspired each of the cycles, and on the different steps taken during the phases of each cycle (Kemmis et al., 2014), followed by the results of each cycle. The cycles consisted of the following phases:

Cycle 1

1. Reconnaissance: developing (the first version of) the rating scale.

2. Plan: preparing rater training for inter-rater consistency.

3. Act and observe: training raters and implementing the scale for the first round of assessments; collecting the scores given for the two oral activities (CMC and F2F), and recording the feedback of raters on the use of the grid; analysing the consistency and consensus estimates for the two raters; analysing the internal consistency of each rater and noting the comments made by raters.

4. Reflect: reflecting on changes to be made to the grid in light of (a) the results of the statistical estimates, and (b) the comments of raters on different aspects of the grid (i.e., ease of use, specific difficulties they encountered, adequacy of one grid for two types of oral activities). Their comments will inform the changes to be made to the assessment grid, opting to fine-tune the rating scales on a continual basis (cf. Yan, 2014: 506).

Cycle 2

2. Plan: preparing the modified version(s) of the grid.

3. Act and observe: assessing a new round of student performances using the new grid(s); collecting the scores given for the two different oral activities and recording the eedback of raters on the use of the grid; analysing the consistency and consensus estimates for the two raters; analysing the internal consistency of each rater and noting the comments made by raters.

4. Reflect: reflecting on the impact of the improved grid(s) on the inter reliability and intra-rater reliability and adapting the grid(s) – if necessary – based on the results from cycles 1 and 2.

The action research cycles were conducted during the first semester of the academic year. Both quantitative and qualitative data were collected, and analysed to inform the validation process. The quantitative data consist of 362 rating scores awarded by the two raters to the performance of students in the two cycles of two oral activities (150 rating scores for the simulated conversation, and 212 for the face-to-face oral, respectively). The qualitative data consist of interviews with the two raters following each round of student assessment performances. The collection and analysis of data are discussed in more detail in the next section.

5. Action research cycles of investigation: data collection and results

The conception of the initial grid was based on the answer to research question 1 (cf. 4 — Design and methodology). This allowed for an investigation to determine the reliability of the newly conceived grid, therefore answering research question 2 (cf. 4 — Design and methodology). The results of the first action research cycle informed the question to be addressed in the second action research cycle.

Cycle 1

Reconnaissance

The question asked by Tajeddin et al. (2011) that was not extensively explored in literature on speaking assessment, served as the starting point for the conception of an assessment grid for students with elementary proficiency in (spoken) French: *What criteria do teachers use for rating speaking?*

As part of the literature review for this study and in order to answer research question 1, the assessment criteria and level descriptors of ten assessment grids used for the assessment of oral skills in second and foreign language learners in different countries (including Japan, America, Austria, Iran, Britain and Taiwan) and for different target groups (learners from elementary school to university level, and independent learners wanting to validate their competences) were analysed in an attempt to find common ground (CIEP, 2016; Eisenmann & Summer, 2012; Hwang et al., 2016; IELTS, 2016; Mewald et al., 2013; Nakamura, 2009; Nakatsuhara, 2007; Plough et al., 2010; Tajeddin et al., 2011; Van Moere, 2006).

The commonalities between the grids were identified, and it was decided to use three categories contained in all the proposed assessment grids: “Grammar”, “Vocabulary”, and “Pronunciation”, as the backbone of the new grid. The sub-constructs of grammar and vocabulary (especially used in combination) have been found to be the main factors distinguishing speaking competence levels, as opposed to accent or fluency. At this low proficiency level, pronunciation too contributes significantly to the rated ability of test-takers (Iwashita et al., 2008). As all but one of the grids mentioned above focus primarily on more advanced students, it was decided to use the grid proposed by the CIEP (Centre International d'Études Pédagogiques) for the A1 (beginner) level of the international certification examinations of language abilities for non-native speakers of French (DELF) as a strong guideline. The category "Task achievement" from CIEP was, therefore, added as a fourth criterion. Performance descriptors were formulated for a context-specific criterion-referenced analytic scale ranging from 0 to 5.

To ensure that the rating scale was not simply "a patchwork quilt created by bundling descriptors from other scales together based on scaled teacher judgments" (Fulcher, 2015: 199), the first draft of the grid was reviewed by a panel of eight colleagues involved in the teaching of different languages to obtain their input and comments on (a) the clarity of descriptors, (b) the appropriateness of the descriptions of each level of competency for the different categories, and (c) the level of cohesion between the same level of competency across the different categories. Following their comments, changes were made to correct the use of terminology and language and, as a last step, the range of the scale was diminished from 0–5 to 0–4, as some found it difficult to distinguish between the middle levels (level 2 and 3). These two levels were consequently collapsed into a single level (cf. Figure 1 for the complete grid).

	Task achievement	Vocabulary	Grammar	Pronunciation
4	Could answer all of the questions by providing the information required	Used the appropriate words in all the answers	Every reply was structured correctly. Full sentences were used throughout	Pronounced the words in a clear and understandable way
3	Provided the correct information to the questions with one or two exceptions	Vocabulary adequate with one or two errors	Occasional errors which didn't cause misunderstanding. Mostly formulated full sentences	Occasional mispronunciations which didn't interfere with understanding
2	Provided the required information to most of the questions	Wrong choice of words in some instances	Errors hampered understanding in some instances. One-word responses were frequent	Mispronunciation led to instances of misunderstanding

	Task achievement	Vocabulary	Grammar	Pronunciation
1	Provided the correct information in only a few instances	Vocabulary largely inadequate for this task	Grammar mostly inaccurate	Pronunciation frequently unintelligible
0	Couldn't answer of the questions	Couldn't find the appropriate words to answer the questions	Couldn't structure a correct reply	Couldn't pronounce the words in an understandable way

Figure 1: Assessment grid A – the initial version

Task achievement Vocabulary Grammar PronunciationThe next stage in the action research cycle was the plan stage, which involved the planning of rater training.

Plan

The training was planned according to the format of a F2F moderation meeting, as described by McNamara (2000: 44). The goal of the rater training process was twofold: (1) to minimise differential rating perceptions (Tadjeddin et al., 2011: 126) and to bring raters into alignment on the use of the rating rubric (Yan, 2014: 506) through training and discussions, and (2) to make raters internally consistent (Tadjeddin et al., 2011: 130). This was done in an attempt to ensure equal and fair assessment. Rater scores were subsequently used to calculate consistency and consensus estimates, which served as indicators of the reliability of the assessment instrument.

Act and observe

The rater training was conducted by the teacher-researcher at the North-West University (NWU), South Africa (Potchefstroom Campus). The raters involved in the training were two native speakers of French – one from Cameroon (rater 1) and the other from France (rater 2). They did not have any previous experience in assessing oral tasks done by foreign language students, and they did not teach or know any of the students, therefore ensuring objectivity. However, as teachers of French they had been involved in assessing foreign language students in general. The assessment subjects were 70 students at this particular university who were novice learners of French. The recordings of all the student productions for two oral practice cycles were used. Ethical clearance for this project was obtained from the NWU, and all the participants signed an informed consent form before taking part in the study.

The training started with an introduction to the assessment grid (cf. Figure 1) that included the categories and descriptors for each level, to ensure that the instrument was understood in the same way by the different raters. Uncertainties about the language used to describe the different levels were clarified. The training was done using the recordings of four student productions from previous cycles to reduce rater variability (Tadjeddin et al., 2011: 130). The raters were asked to independently listen to two of the four recordings of the simulated dialogue (CMC activity), and to score the performances using the grid provided. The results were then discussed with each rater, allowing them to provide the details of the assessments and the reasons why they gave the specific marks. This was done to establish to what degree the raters assessed in the same manner, and to discuss the grid to determine if there were elements that needed to be refined or re-defined. The student recordings were played again, and the trainer and raters discussed the specific mistakes that played a role in the scores given. The following observations were made after the first assessment:

(1) The raters found it difficult at first to distinguish between the different categories. For example, rater 1 tended to penalise students in other categories for not pronouncing words correctly. This could be explained by what Van Moere (2006: 424) refers to as “trade-off behaviour between categories”. Rater 1 tended to be stricter that rater 2.

(2) Rater 2 was influenced by the less than perfect intonation of the students, and penalised them for this in the “Pronunciation” category. During the discussion it was agreed that intonation does not play a crucial role at this level, and that if words are pronounced in an understandable – yet not perfect – manner, students should rather be awarded a good mark.

The scores of the two raters for student 1 and 2 were quite different ― the final score and the way in which the marks were awarded (e.g., the pattern) varied. The most marked differences in the way the raters attributed marks pertained to the “Grammar” and “Pronunciation” categories. After the discussions, the raters negotiated and agreed on a final mark for student 1 and 2. The raters were then asked to assess two other recordings to see if the extent of agreement between the marks given by the two raters had improved. The scores for student 3 differed by 4 marks (11 and 15 out of a possible 16), but the pattern of the scoring was identical, indicating that the raters had the same appreciation of the different skills, albeit with varying degrees of strictness. The overall scores for student 4 were the same even though the way in which the raters scored were not the same. These tendencies of agreement were somewhat encouraging for future assessments. It is important, nevertheless, to bear in mind that in spite of wide acceptance of the necessity of rater training to ensure the reliability of scores produced in language performance tests, rater training is no guarantee for equal scoring. In an attempt to compensate for inter-rater unreliability, it was decided that in the “Act and observe stage”, students would receive the average of the two scores given by the raters.

After the training, the raters were provided with 97 recordings of student performances (40 CMC simulated dialogues and 57 F2F oral interviews). They were asked to make notes about the grid (Grid A), while assessing the recordings of oral activities by considering the following questions:

1. Is the rating scale easy to use?

2. Are there specific difficulties in using it?

3. Is the single grid adequate for assessing the two types of activities (CMC and F2F)?

These reflections were used for possible changes to be made to the grid after the first round of assessments, therefore adhering to the requirement of “triangulating different sources of data on rater performance using a mixed-methods approach, especially in local testing contexts” (Yan, 2014: 501).

The 194 scores produced by the raters were submitted for a data analysis concerning consistency and consensus estimates (i.e., if the raters shared an understanding of the rating scale and were they in agreement of rating scores) (Yan 2014: 504). Moreover, the statistical analysis investigated the internal consistency of each rater (intra-rater consistency). The results of this analysis are discussed in the following sub-sections.

Consistency and consensus estimates – Cycle 1

The consistency estimates took the form of a Spearman rank correlation coefficient between the scores for each of the categories in the assessment grid assigned by the two raters. The results after the first round of assessments are provided in Table 1. The higher the value, the more they scored students in the same way.

Table 1: Consistency of scores (Spearman's rho[1]) from the two raters

Consistency (correlations)	**Grid A**
Simulated Conversation _ Task Achievement	0.7
Simulated Conversation _ Vocabulary	0.4
Simulated Conversation _ Grammar	0.3
Simulated Conversation _ Pronunciation	0.3
	(N=40)
Oral _ Task Achievement	0.5
Oral _ Vocabulary	0.4
Oral _ Grammar	0.6
Oral _ Pronunciation	0.5
	(N=57)

The results show a medium to large effect size with four of the eight categories showing a large effect size (~0.5). The F2F oral activity showed higher effect sizes than the simulated conversation activity for "Grammar" and "Pronunciation", whereas the CMC activity showed a higher effect size for "Task achievement". The results for "Vocabulary" were the same.

When looking at the instances in the consensus estimates[2] where the raters gave the exact same score (cf. Table 2), it is clear from the percentage-agreement figures that the rating scores for the simulated conversation activity (column on the left) had a higher level of agreement in three of the four categories than the scores for the F2F oral (column on the right): "Task achievement" (66,7% vs. 59,6%), "Vocabulary" (56,4% vs. 42,1%), and "Grammar" (41% vs. 29,8%).

[1] Spearman's rho indicates practical significance of relationship or effect sizes. Guideline values: ~0.1, small, no practical significant relationship; ~0.3, medium, practical visible relationship; ~0.5, large, practical significant relationship.

Table 2: Consensus for simulated conversation vs. consensus for F2F oral

Consensus[2]	**Grid A %**	**Consensus**[2]	**Grid A %**
Simulated Conversation Task Achievement		**Oral Task Achievement**	
Exact	66.7	Exact	59.6
Adjacent	33.3	Adjacent	36.8
Discrepant	0.0	Discrepant	3.5
Conflicting	0.0	Conflicting	0.0
Simulated Conversation Vocabulary		**Oral Vocabulary**	
Exact	56.4	Exact	42.1
Adjacent	41.0	Adjacent	43.9
Discrepant	2.6	Discrepant	14.0
Conflicting	0.0	Conflicting	0.0
Simulated Conversation Grammar		**Oral Grammar**	
Exact	41.0	Exact	29.8
Adjacent	46.2	Adjacent	64.9
Discrepant	12.8	Discrepant	3.5
Conflicting	0.0	Conflicting	1.8
Simulated Conversation Pronunciation		**Oral Pronunciation**	
Exact	48.7	Exact	71.9
Adjacent	43.6	Adjacent	28.1
Discrepant	7.7	Discrepant	0.0
Conflicting	0.0	Conflicting	0.0

[2] Exact = same mark; Adjacent = point difference; Discrepant = 2 points difference; Conflicting = more than 2 points difference.

The consensus score for “Pronunciation” was higher for the F2F oral (71,9% vs. 48,7%). This tendency changed slightly when the combined percentage of the exact and the adjacent scores is considered– where the result of “Grammar” for the F2F oral was slightly higher than for the simulated conversation. The discrepant results for “Grammar” in the simulated conversation (12,8%) and for “Vocabulary” in the F2F oral (14%) were rather high.

Internal consistency of raters – Cycle 1

The results of the level of internal consistency of the raters (cf. Table 3) permits a look at the broad spectrum of information available. These results might give some further direction for future action.

Table 3: Intra-rater reliability

Reliability	Grid A
Simulated Conversation _ Rater 1	0.63[3]
Simulated Conversation _ Rater 2	0.80
	Grid A
Oral _ Rater 1	0.73
Oral _ Rater 2	0.73

The overall score for rater 2 was slightly higher than that of rater 1, but all of the results were well above the required 0.5, and three out of the four results were above 0.7, which is satisfactory.

Feedback from the raters – Cycle 1

During a semi-structured interview with the raters after the first round of assessments, several issues emerged. The first aspect under discussion was the wording for level 2 of “Vocabulary” (cf. Figure 1). Both of the raters felt that the descriptor caused the need for another level in this category – something between level 2 and 3 – which would allow them to accurately evaluate student performance. The comments made by the raters also pertained to the use of full sentences / one-word responses that was part of the

[3] A Cronbach’s Alpha coefficient of 0.7 and above is generally accepted as an indication of reliability, but in the early stages of research like in this case values of 0.5 or above will also be sufficient (Field, 2014: 708 – 709).

descriptors for the category “Grammar”. They felt that it made grading more complex, due to a certain tension between the descriptors of this category. One of the raters expressed a need for a category assessing “enthusiasm” – equivalent to the categories “Attitude” or “Affective variables” in the studies of Nakamura (2009) and Tajeddin et al. (2011) – in order to reward students who “invest themselves” in the activity.

The raters expressed the need to further adapt the grid to incorporate an aspect unique to the F2F oral: human interaction. Both felt that the interaction between the student and interviewer should be reflected on, because this setting allowed for verbal cues or prompts from the interviewer in case a student had difficulty to communicate. This interlocutor effect results from the way in which speakers co-construct meaning while communicating – in this case, the interviewer collaborates with and supports the interviewee in order to make sense of the interaction as a whole. It is of obvious importance, and should be taken into consideration.

Reflect

The answer to research question 2 (cf. 4 —Design and methodology) was not very encouraging. The statistical analyses indicated that the understanding of the raters concerning the rating scale was not satisfactory with half of the results of the consistency estimates not complying with the minimum requirement. This showed that changes were necessary for both the simulated conversation activity and the F2F oral activity to attempt to increase the number of exact scores given by the different raters. It was, therefore, important to investigate the assessment grid in a subsequent cycle. The most important implication of incorporating the qualitative data obtained from the feedback of the raters discussed above was creating a separate grid for the two activities (CMC and F2F).

Cycle 2

Wanting to investigate the impact of the changes that would be made to the initial grid following the last two stages of cycle 1, led to a new question for the next cycle of research: “To what extent will inter-rater reliability (consistency and consensus estimates) and intra-rater reliability change if raters use a differentiated grid to grade the different oral activities at a beginner’s level?”. The first step in this new cycle was to plan the differentiated grids based on the reflections obtained from cycle 1.

Plan

During the discussions after the first cycle that evaluated the two oral communication activities, it became clear that the raters were not completely satisfied with certain aspects of the grid. In a study conducted by Tajeddin and his colleagues, it was shown that there was “a sharp decline in the significance given to [...] affective variables” after raters underwent training (2011: 125). Their study was also one of only two of the ten studies used as a point of departure for the creation of an assessment grid that includes affective factors as part of the rating criteria. It was, therefore, decided not to include the affective category "Enthusiasm" in the assessment grid for beginner students.

Several changes resulted from cycle 1. The first change was that the descriptor for "Vocabulary" level 2 was changed from "*Wrong choice words in some instances*" to "*Vocabulary only just sufficient for this task*" (compare Table 1: Assessment grid A and Table 5: Assessment grid B). Secondly, it was decided to move references to the use of full sentences in the descriptors for scales 2, 3 and 4 from the category "Grammar" to the category "Vocabulary" (cf. Figure 2).

	Task achievement	**Vocabulary**	**Grammar**	**Pronunciation**
4	Could answer all of the questions by providing the information required	Used the appropriate words in all the answers. *(Full sentences were used throughout)*	Every reply was structured correctly.	Pronounced the words in a clear and understandable way
3	Provided the correct information to the questions with one or two exceptions.	Vocabulary adequate with one or two exceptions. (*Mostly formulated full sentences*)	Occasional errors which didn't cause misunderstanding.	Occasional mispronunciations which didn't interfere with understanding
2	Provided the required information to most of the questions	**Vocabulary only just sufficient for this task.** *(One-word responses were frequent*)	Errors hampered understanding in some instances.	Mispronunciation led to instances of misunderstanding
1	Provided the correct information in only a few instances	Vocabulary largely inadequate for this task	Grammar mostly inaccurate	Pronunciation frequently unintelligible
0	Couldn't answer any of the questions	Couldn't find the appropriate words to answer the questions	Couldn't structure a correct reply	Couldn't pronounce the words in an understandable way

***Figure 2: Assessment grid B – adapted version for simulated conversation activity*[4]**

[4] Descriptors regarding the amount of help given by the interviewer are given in curly brackets.

A different grid was created to assess the F2F oral to reflect the aspect of human interaction that forms an integral part of the activity. Elaborating on the descriptors in this way also reflects the level of understanding displayed by students without having to create a separate category (cf. Figure 3).

	Task achievement	**Vocabulary**	**Grammar**	**Pronunciation**
4	Could answer all of the questions by providing the information required **(without any help)**	Used the appropriate words in all the answers *(Full sentences were used throughout)*	Every reply was structured correctly	Pronounced the words in a clear and understandable way
3	Provided the correct information to the questions **(with help from the other person) in one or two instances**	Vocabulary adequate with one or two exceptions. (*Mostly formulated full sentences*)	Occasional errors which didn't cause misunderstanding.	Occasional mispronunciations which didn't interfere with understanding
2	Provided the required information to most of the questions **(with regular help from the other person)**	**Vocabulary only just sufficient for this task.** *(One-word responses were frequent*)	Errors hampered understanding in some instances.	Mispronunciation led to instances of misunderstanding
1	Provided the correct information in only a few instances **(despite a lot of help from the other person)**	Vocabulary largely inadequate for this task	Grammar mostly inaccurate	Pronunciation frequently unintelligible
0	Couldn't answer any of the questions	Couldn't find the appropriate words to answer the questions	Couldn't structure a correct reply	Couldn't pronounce the words in an understandable way

***Figure 3: Assessment grid C – adapted for F2F oral*[5]**

[5] Descriptors regarding the amount of help given by the interviewer are given in curly brackets.

The result of the changes was the creation of two grids – one for the CMC simulated dialogue and one for the F2F interview (cf. Figures 2 and 3) – that were subsequently used in a second cycle of assessment.

Act and observe

During the second round of assessment of student productions, the two scales described above were used to assess 84 student performances (35 for the CMC simulated dialogue and 49 for the F2F oral interview). These scores were used to perform the statistical analysis discussed below.

Consistency and consensus estimates – Cycle 2

The correlations between the scores given by the raters over the two rounds of assessments are provided in Table 4. The higher the value, the more the raters scored the students in the same way. The column on the left (Grid A) refers to the original grid used during the first round of assessment, and the column on the right (Grid B & C) refers to the differentiated grids used for the two types of activities during the second round of assessment.

Table 4: Consistency of scores from the two raters over two rounds

Consistency (correlations)	**Grid A**	**Grid B**
Simulated Conversation _ Task Achievement*	**0.7**[1]	**0.2**
Simulated Conversation _ Vocabulary	**0.4**	**0.5**
Simulated Conversation _ Grammar	**0.3**	**0.3**
Simulated Conversation _ Pronunciation*	**0.3**	**0.3**
	(N=40)	**(N=35)**
	Grid A	**Grid C**
Oral _ Task Achievement	**0.50**	**0.7**
Oral _ Vocabulary	**0.4**	**0.6**
Oral _ Grammar	**0.6**	**0.4**
Oral _ Pronunciation*	**0.5**	**0.5**
	(N=57)	**(N=49)**

(*= exact same descriptors in the two versions of the grid)

For the simulated conversation (CMC), the result for "Task achievement" decreased significantly even though the descriptors were exactly the same. The result for "Vocabulary" increased, and the two other results remained the same with "Grammar" and "Pronunciation" still unsatisfactory. For the F2F oral interview, there was an increase of 0.2 for "Task achievement" and "Vocabulary", while "Grammar" showed a 0.2 decrease. No changes were recorded in "Pronunciation".

Even though the results from the consistency estimates were not all that encouraging, the consensus estimates have improved from grid A to grid B and C in several instances (cf. Tables 5 and 6). In Table 5, the results for grid B are compared with the results from grid A for the assessment of the simulated conversation. Table 6 provides a comparison between the results from grid A and grid C for assessing the F2F oral (cf. Table 2 for the results when grid A was used for both activities).

Table 5 – Consensus: grid A & B

Consensus	Grid A%	Grid B%
Simulated Conversation_ Task Achievement*		
Exact[2]	66.7	54.3
Adjacent	33.3	45.7
Discrepant	0.0	0.0
Conflicting	0.0	0.0
Simulated Conversation Vocabularly		
Exact	56.4	62.9
Adjacent	41.0	37.1
Discrepant	2.6	0.0
Conflicting	0.0	0.0
Simulated Conversation _ Grammar		
Exact	41.0	54.3
Adjacent	46.2	42.9
Discrepant	12.8	0.0
Conflicting	0.0	2.9

Table 6 – Consensus: grid A & C

Consensus	Grid A%	Grid B%
Oral _ Task Achievement		
Exact	59.6	57.1
Adjacent	36.8	42.9
Discrepant	3.5	0.0
Conflicting	0.0	0.0
Oral _ Vocabularly		
Exact	42.1	44.9
Adjacent	43.9	40.8
Discrepant	14.0	14.3
Conflicting	0.0	0.0
Oral _ Grammar		
Exact	29.8	51.0
Adjacent	64.9	40.8
Discrepant	3.5	6.1
Conflicting	1.8	2.0

Simulated Conversation Pronunciation*		
Exact	48.7	48.6
Adjacent	43.6	45.7
Discrepant	7.7	5.7
Conflicting	0.0	0.0

Oral _ Pronunciation*		
Exact	71.9	71.4
Adjacent	28.1	26.5
Discrepant	0.0	2.0
Conflicting	0.0	0.0

(*= exact same descriptors in the two versions of the grid)

The consensus results concerning the simulated conversation between grid A and B presented above were more alike than the results of the consistency estimates (cf. Table 4). This similarity clearly reflects the fact that the descriptors for “Task achievement” and “Pronunciation” are the same in the two grids. The percentage of exact scores for “Vocabulary” and “Grammar” has gone up in grid B, and no more discrepant scores for these categories were found. The scores for “Pronunciation” were mainly the same. This was to be expected, because the descriptors were the same in the two grids.

Where the F2F oral was concerned, the consensus estimates for “Task achievement” were more or less the same for the two grids, although there were no more discrepant scores for grid C. “Vocabulary” showed no significant changes from one grid to the next, whereas in “Grammar” the occurrence of exact scores increased markedly. As with the simulated conversation, the results for “Pronunciation” remained more or less the same.

Internal consistency of the raters – cycle 2

The results shown in Table 7 pertain to the internal consistency of the raters (Cronbach’s alpha) while using the grids. All of the results were above the required 0.5, but there are marked differences between the results of the simulated conversation for the two different grids.

Table 7: Intra-rater reliability over two rounds of assessments

Reliability	**Grid A**	**Grid B**
Simulated Conversation _ Rater 1*	0.63[3]	0.55
Simulated Conversation _ Rater 2	0.80	0.64
	Grid A	**Grid C**
Oral _ Rater 1	0.73	0.78
Oral _ Rater 2	0.73	0.77

The decline in the results of rater 2 for the CMC activity could be explained by the low corrected item-total correlation for “Task achievement.” This phenomenon is difficult to explain, because the descriptors for this category were exactly the same in the two grids. It is not possible to explain the decline in results of rater 1 from the data, because the corrected item-total correlation for all four categories were low.

The results of the grid used for the F2F oral performance of students were all above 0.7, and improved slightly for both of the raters from grid A to grid C.

Feedback from raters – cycle 2

After this second round of evaluations using the two adapted grids, the raters did not have any further suggestions for changes to either of the grids.

Reflect

To answer the question for cycle 2 (cf. 5.2), the results of the consistency and consensus estimates were calculated. The results varied and the implication of this is presented in the following section. The results for the two pairs of rating scales (A—B; A—C) used for the two activities are discussed separately.

Simulated conversation (CMC)

Overall, the consistency estimates dropped from grid A to grid B. It is important to note, though, that the decrease was caused by a decline in a single category – that of “Task achievement”. As stated before, this is hard to explain, because no changes were made to grid B. It is encouraging that “Vocabulary” showed an increase that might have been caused by the changes made to the descriptors in grid B.

The improvement in the level of consensus between the score of the raters for “Vocabulary” and “Grammar” is an indication that the changes made to these categories had a positive result. These changes should, therefore, be retained in a future version of the grid for the simulated conversation.

F2F oral

The overall consistency of grid C was higher than that of grid A. As with the simulated conversation, the change was caused by the results from “Task achievement”, but this time the change was positive. This could be a result of the changes made to the descriptors to include the “interaction” element. The consistency estimates for “Grammar” and “Vocabulary” were inversed from grid A to grid C, which indicates that the changes made to the descriptors for these two categories had a positive effect on “Vocabulary”, but a negative effect on “Grammar”.

The consensus estimates for “Task achievement”, “Vocabulary”, and “Grammar” did not change significantly, but the results of the exact scores of “Grammar” improved by more than 20%. Despite the negative impact of the changes in this category shown by the consistency estimates, the changes had a positive impact on the agreement of rating scores.

The next step was to identify the best version of the grids for the two types of activities by making use of information obtained from an analysis about the correlations between the marks of the raters and the degree of consensus between the marks obtained by using the different grids.

Assessment grid for simulated conversation

No further discussion was necessary, because the descriptors used in the categories “Task Achievement” and “Pronunciation” stayed exactly the same in the two versions of the grid. It is interesting to note that the consensus for “Pronunciation” between the two cycles of assessment was very nearly identical, but that of “Task Achievement” did not reflect the same high level of similarity.

In the categories “Vocabulary” and “Grammar”, the consensus between the raters indicates that the descriptors used in grid B yielded the most consistent results. This was indicated by the fact that the percentages for “Exact” in grid B were higher, and the results for “Adjacent” and “Discrepant” were lower. When one looks at the changes that have been made to grid A, it became clear that it was the move of the descriptors pertaining to the use of full sentences/one-word responses from the category “Grammar” to the category “Vocabulary” that brought about the higher consensus.

Assessment grid for oral

The descriptors for “Pronunciation” are the same for grid A and C, and were therefore used in the final version of the grid for the oral.

The results pertaining to the consensus reached for “Task Achievement” were not conclusive. Even though the score for “Exact” was slightly higher for grid A, the score for “Discrepant” was 3,5% that nullified the higher “Exact” grid A score. The descriptors used in grid C were retained, because the descriptors regarding interaction between the two people involved in the oral were added on request of the raters. This category could be re-examined in future to establish whether the element of interaction has an impact on the consensus between the scores awarded by raters.

The results of the categories “Vocabulary” and “Grammar” should be interpreted together, because these two categories were linked when the changes made to grid A were considered after the first cycle of assessments. Even though the initial results obtained from the correlations between the marks of raters indicated a higher score for grid A in the “Grammar” category (0.6 in contrast with 0.4 for grid C), the results regarding the consensus between the marks contradicted this finding. As in the case of the grid for the simulated conversation (grid B), the move of the descriptors pertaining to the use of full sentences/one-word responses from the category “Grammar” to the category “Vocabulary” strengthened the consensus for both these categories with an increase of 21% for “Grammar”. This confirms that the changes made to grid A contributed to the quality of the differentiated grid (grid C) for the oral.

Following the considerations discussed above, the versions to be used in future assessments for the type of activities described in this article would be grid B and C.

6. Discussion and conclusion

Proposing and implementing a rating scale for oral skills at novice level is the first step in an attempt to address the issue of reliable assessment instruments, particularly at a very basic level of oral competence. Procedures for establishing the reliability of an assessment instrument were outlined and applied.

The results of this study showed that it is not sufficient to create a generic grid for oral competence, but that within the rating criteria there should be a marked differentiation between CMC and F2F contexts. It is, therefore, recommended that the rating scale B should be studied in a manner similar to the one described in this article to verify the degree of correlation and consensus with a new group of students and different raters. This would constitute the beginning of a third action research cycle in the ongoing process of creating the best possible grid for the assessment of beginner foreign language students to see how inter-rater consistency could be improved.

The descriptor for level 2 of “Task achievement” in the rating scale B (simulated conversation) should also be adapted, because providing the correct information to “most of the questions” would surely merit a mark above 50%. It might be wise to change the descriptors pertaining to the use of full-sentences in future – especially on level 4 of “Vocabulary” and in particular for the F2F oral – for even though students know that they are required to use full sentences as often as possible, it limits the authenticity during person-to-person interaction, and will ensure that the language is not forced and unnatural even at a very basic level.

A subsequent research phase is planned to refine the rating instrument in order to “improve the meaningfulness and fairness of the conclusions reached about individual candidates” (McNamara, 2000: 56).

References

ALTE. 2011. *Manual for language test development and examining*. Retrieved from Strasbourg: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2b

Amiel, T. & Reeves, T. C. 2008. Design-based research and educational technology: rethinking technology and the research agenda. *Educational technology & society, 11*(4), 29-40.

Bachman, L. F., Brian, K. L., & Mason, M.1995. Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing 12*:238-257.

Burns, A. 2005. Action research: an evolving paradigm? *Language Teaching 38*:57-74.

Butler, G. 2017. Translating the Test of Academic Literacy Levels into Sesotho. *Journal for Language Teaching* 51(1):11–43. DOI: 10.4314/jlt.v5i1.1.

Caban, H. 2003. Rater group bias in the speaking assessment of four L1 Japanese ESL students. Second Language Studies, 21(2):1-44.

CIEP. 2016. *Grille d'évaluation de la production orale A1*. Available: http://www.delfdalf.fr/_media/a1-grille-po.pdf. [Accessed: 11 June 2019].

Cohen, L., Manion, L., & Morrison, K. 2018. *Research methods in education*. London, New York: Routledge.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Available: https://rm.coe.int/1680459f97. [Accessed: 13 February 2015].

Creswell, J. W., & Poth, C. N. 2018. *Qualitative inquiry and research design*: SAGE.

Davies, A., Brown A., Elder, C. Hill, K., Lumley, T. & McNamara, T. 1999. *Dictionary of Language Testing*. Cambridge: CUP.

Davis, L. 2016. The influence of training and experience on rater performance in scoring spoken language. *Language Testing* 33(1):117–135.

DoBE, Department of Basic Education 2020. *The requirements for the assessment of orals in all official languages offered for the national senior certificate*. Pretoria Available: https://wcedonline.westerncape.gov.za/circulars/minutes20/CMminutes/DCF/dcf0006-2020-ANNEXURE.pdf

Eckes, T. 2005. Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis. *Language Assessment Quarterly* 2(3):197–221. DOI:10.1207/s15434311laq0203_2.

Eisenmann, M. & Summer, T. 2012. Oral Exams: Preparing and Testing Students. In: Eisenmann, M. & Summer, T. Eds. 2012. *Basic Issues in EFL Teaching and Learning.* Heidelberg: Winter. pp. 415–428.

Fulcher, G. 2003. *Speaking ability*. New York: OP.

Fulcher, G. 2015. Assessing second language speaking. *Language Testing,* 48:198-216. DOI:10.1017/S0261444814000391

Geyskens, J., Donche, V. & Van Petegem, P. 2010. Effective feedback als hefboom voor begeleid zelfstandig leren. *Begeleid Zelfstandig Leren* 25:15–38.

Grobler, C. & Smits, T. F. H. 2016. Designing a digital pedagogical pattern for improving foreign language learners' oral proficiency. *Literator*, 37(2):a1273. DOI: http://dx.doi.org/10.4102/lit.v37i2.1273.

Grobler, C. 2020. Designing a model for a technology-enhanced environment developing the oral interactional competence of beginner language learners. (PhD), University of Antwerp, Antwerp, Belgium.

Gruhn, S. & Weideman, A. 2017. The initial validation of a test of emergent literacy. *Per Linguam,* 33(1):25-53.

Hatting, K. & Van der Walt, J. L. 2013. The development and validation of a rating scale for ESL essay writing. *Journal for Language Teaching,* 47(1):73-105. doi:http://dx.doi.org/10.4314/jlt.v47i1.4

Hwang, W., Shadiev, R., Hsu, J., Huang, Y., Hsu, G., & Lin, Y. 2016. Effects of storytelling to facilitate EFL speaking using web-based multimedia system. *Computer Assisted Language Learning,* 29(2):215–241. DOI: 10.1080/09588221.2014.927367

IELTS (Producer). 2016. Speaking assessment criteria (band descriptors - public version). Available: https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en. [Accessed: 21 February 2016].

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.

Kemmis, S., McTaggart, R., & Nixon, R. 2014. *The Action Research Planner*. Singapore: Springer.

Kimathi, F. K., & Bertram, C. 2020. Oral language teaching in English as first additional language at the foundation phase: A case study of changing practice. *Reading & Writing, 11*(1), a236. doi:10.4102/rw.v11i1.236

Knoch, A. 2011. Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing,* 16:81-96. doi:10.1016/j.asw.2011.02.003

Knoch, U., Read, J. & von Randow, J. 2007. Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing* 12(1):26–43.

Liu, X. 2013. Action Research on the Effects of an Innovative Use of CALL (Computer Assisted Language Learning) on the Listening and Speaking Abilities of Chinese University Intermediate Level English Students. PhD thesis. University of Exeter. (Unpublished).

Lumley, T., & McNamara, T. F. 1995. Rater characteristics and rater bias: Implications for training. *Language Testing,* 12:54-71.

McNamara, T. 2000. *Language Testing.* Oxford University Press.

Mewald, C., Gassner, O., Brock, R., Lackenbauer, F., & Siller, K. 2013. *Testing speaking for the E8 standards*. Available: https://www.bifie.at/wp-content/uploads/2017/05/TR_Speaking_130805.pdf. [Accessed: 22 February 2016].

Nakamura, Y. 2009. Rating criteria for the three speaking test format: monologue, dialogue and multilogue. *Educational Studies* 51:133–141.

Nakatsuhara, F. 2007. Developing a rating scale to assess English speaking skills of Japanese upper-secondary students. *Essex Graduate Student Papers in Language and Linguistics,* 9:83103.

Newby, P. 2014. *Research methods for education*. London & New York: Routledge.

North, B. & Schneider, G. 1998. Scaling descriptors for language proficiency scales. *Language testing*, 15(2):217-263.

OECD. 2021. *PISA 2025 Foreign language assessment framework*. Paris: OECD Publishing.

Rambiritch, A. 2013. Validating the Test of Academic Literacy for Postgraduate Students (TALPS). *Journal for Language Teaching*, 47(1):175-193.

Plough, I. C. 2010. A multi-method analysis of evaluation criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing* XX(X):1–26.

Piggot-Irvine, E., Rowe, W. & Ferkins, L. 2015. Conceptualizing indicator-domains for evaluating action research. *Educational Action Research* 23(4):545–66.

Prinsloo, C. H., & Harvey, J. C. 2016. The viability of individual oral assessments for learners: Insights gained from two intervention evaluations. *Perspectives in Education, 34*(4):1-14.

Riel, M. 2010. Understanding collaborative action research. Available: http://cadres.pepperdine.edu/ccar/define.html. [Accessed: 3 July 2019].

Scholtz, D. 2017. The appropriateness of standardised tests in academic literacy for diploma programmes of study. *Language Matters, 48*(1):27-47. doi:10.1080/10228195.2016.1271350

Sebolai, K. 2018. Revisiting the meaning of validity for language testing: The case of two tests of English language ability. *Journal for Language Teaching, 52*(1):152-168. doi: https://dx.doi.org/10.4314/jlt.v52i1.8

Smith, B., & Schulze, M. 2013. Thirty years of the Calico journal – Replicate, replicate, replicate. *CALICO Journal,* 30(1):i-iv.

Tajeddin, Z., Alemi, M. & Pashmforoosh, R. 2011. Non-native teachers' rating criteria for L2 speaking: Does a rater training program make a difference? *TELL* 5(1):125–153.

Tripp, D. 2004. Available: https://www.researchgate.net/profile/David_Tripp/ publication/299861325_Action_Reaearch_Introduction_to_your_project_design/data/570673ac08aea3d28021141a/AR-Intro-EdProject.ppt. [Accessed: 2 July 2019].

Van der Slik, F., & Weideman, A. 2009. Revisiting test stability: further evidence relating to the measurement of difference in performance on a test of academic literacy. *Southern African Linguistics and Applied Language Studies (Special issue: Assessing and developing academic literacy),* 27(3):253-263.

Van der Walt, C., De Wet, F., & Niesler, T. 2008. Oral proficiency assessment: The use of automatic speech recognition systems. *Southern African Linguistics and Applied Language Studies, 26*(1):135-146.

Van Dyk, T. 2010. *Konstitutiewe voorwaardes vir die ontwerp van 'n toets van akademiese geletterdheid.* PhD thesis. University of the Free State. Available: http://scholar.ufs.ac.za:8080/xmlui/handle/11660/1918. [Accessed: 1 March 2022].

Van Dyk, T., Murre, P., & Kotzé, H. 2021. Does one size fit all? Some considerations for test translation. In A. Weideman, J. Read, & T. Du Plessis (Eds.), *Assessing academic literacy in a multilingual society: Transition and transformation* (pp. 52-74). doi:10.21832/9781788926218

Van Dyk, T., Van Rensburg, A., & Marais, F. 2011. Levelling the playing field: An investigation into the translation of tests. *Journal for Language Teaching,* 45(1):153-169.

Van Moere, A. 2006. Validity evidence in a university group oral test. *Language Testing,* 23(4):411–440.

Van Petegem, P. & Vanhoof, J. 2002a. *Een alternatieve kijk op evaluatie. In Begeleid zelfstandig leren*. Antwerpen, Mechelen: Wolters-Plantyn.

Van Petegem, P. & Vanhoof, J. 2002b. *Evaluatie op de testbank. Een handboek voor het ontwikkelen van alternatieve evaluatievormen*. Antwerpen, Mechelen: Wolters-Plantyn.

Weideman, A. 2011. Academic literacy tests: Design, development, piloting and refinement. *Journal for language teaching, 45*(2):100-113.

Weideman, A., Patterson, R., & Pot, A. (2016). Construct refinement in tests of academic literacy. In J. Read (Ed.), *Post-admission language assessment of university students*. Cham: Springer.

Yan, X. 2014. An examination of rater performance on a local oral English proficiency test: mixed-methods approach. *Language Testing,* 31(4):501–527.

ABOUT THE AUTHORS

Carina Grobler

North-West University
ORCID number: 0000-0001-7451-7115

E-mail: Carina.Grobler@nwu.ac.za

Carina Grobler is a senior lecturer in French at the North-West University, South Africa. She obtained her PhD in Education Science at the University of Antwerp, Belgium. The study focused on the design of a model for a technology-enhanced practice environment that aims to develop the oral interactional competence of beginner foreign language learners. Her research interests include instructional design, computer-assisted language learning, the assessment of foreign language competency at beginner level, teacher training and foreign language didactics.

Tom FH Smits

University of Antwerp
Rhodes University
ORCID number: 0000-0001-6615-4019

E-mail: tom.smits@uantwerpen.be

Tom Smits obtained his PhD on sub-standard language variation in German and Dutch, and its influence on attitudes to regional language use in media and education at Antwerp University's Department of Linguistics, where he teaches German grammar and variational linguistics. Since 2008, as an English, German and CLIL teacher educator at the Antwerp School of Education, he has developed expertise in differentiation, multiperspectivity and urban education, awarded with the 2019 UAntwerp Teaching Award. His research activities cover language variation, foreign language education, including CALL, and intercultural competences with students in Flanders (Belgium), Turkey, Indonesia, South Africa and DR Congo.

Tydskrif vir Taalonderrig - Journal for Language Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig - Journal for Language Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig - Journal for Language Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig - Journal for Language Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig - Journal for Language Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig - Journal for Language Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig - Journal for Language Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa Ririmi - - Tydskrif vir Taalonderrig - Journal for Language Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta