



Regular Article

Bridging the gap in oral language assessment: ORAToR, a comprehensive tool for measuring oral summarization competence

Christine Wiertz^{a,*}, Thibault Coppe^b, Maël Pire^a, Benoît Galand^a, Stéphane Colognesi^a^a UCLouvain, 1 Place de l'Université, Belgique, 1348, Louvain-la-Neuve, Belgium^b Université Libre de Bruxelles, 50 avenue F.D. Roosevelt B, Belgique, 1050, Bruxelles, Belgium

ARTICLE INFO

Keywords:
Oral
Assessment
Rubrics
Summary

ABSTRACT

Oral communication is a fundamental skill in education, yet its assessment remains a significant challenge due to the lack of valid and reliable tools designed for children, particularly for research purposes. This gap limits both the understanding and study of oral skills in educational contexts. In response, we developed ORAToR (Oral Recap Assessment Tool with Reliability), a comprehensive and rigorously validated tool designed specifically for researchers to assess pupils' proficiency in orally summarizing knowledge. This paper details the 10-step process involved in constructing ORAToR, highlighting the importance of expert consultation and pilot studies to ensure robust content validity. Exploratory factor analysis revealed three factors: Informational Content, Interaction with the Interlocutor, and Clarity of the Message Conveyed. While inter-rater reliability was strong for most criteria, evaluation of the relevance of the message and paraverbal elements needs further attention. These findings underscore the need for validated tools such as ORAToR to advance research on oral communication.

1. Introduction

Oral language plays a major role in everyday and school life and is an essential element of active participation in society and a facilitator of learning (Mercer et al., 2017). Despite its importance, numerous studies have shown that oral language instruction does not yet have the place it deserves in the classroom (Wurth et al., 2022). Moreover, when speaking is addressed in class, it often involves the same tasks, such as poetry recitation or oral presentations (Colognesi & Deschepper, 2019; Oliver et al., 2005). These tasks, derived from school traditions, are not how oral language is most frequently used by pupils either in or out of school, and do not necessarily prepare them for the social tasks encountered in everyday life (Oliver et al., 2005). This situation can be partly attributed to the fact that the teachers who are responsible for instructing native speakers (both in primary education and across various secondary-school subjects) often receive little training on the nature of oral language and its specific characteristics (Dumais, 2016; Moncarey et al., 2025). As a result, they frequently do not know what to teach and assess with regard to oral communication skills (Jones, 2017).

Research on oral language therefore appears to be essential for helping teachers to better identify what to teach and assess when

working on speaking in the classroom. In this regard, didactics researchers have developed typologies of oral communication that specify its components and characteristics (e.g., Dolz et al., 1993; Dumais, 2016; Millard & Menzies, 2016). While there are parallels between these typologies (Colognesi & Hanin, 2020), there are also notable divergences. Moreover, they have mostly been grounded in theoretical conceptualizations with a lack of empirical validation.

Other studies, more rooted in the fields of psychometrics and educational measurement (e.g., Massonnié et al., 2022; O'Connell et al., 2017) have also identified constituent factors of oral language. These studies provide rigorous but standardized tasks that lack "ecological validity" (Kihlstrom, 2021), meaning they do not reflect the reality of classroom interactions (Hosson & Orange, 2019). Some of these standardized assessments (e.g., TOEFL, IELTS) do evaluate participants' oral proficiency in situations that approximate real-life communication. However, those instruments are specifically designed for adults or second-language learners (Karim & Haq, 2014; Ockey, Koyama, Setoguchi, & Sun, 2015) and rarely account for peer-to-peer interactions in pupils' first language in primary school.

In both cases, the lack of a fine-grained and comprehensive measurement of speaking in real-life situations is apparent (Mercer et al.,

* Corresponding author.

E-mail addresses: christine.wiertz@uclouvain.be (C. Wiertz), thibault.coppe@ulb.be (T. Coppe), mael.pire@uclouvain.be (M. Pire), benoit.galand@uclouvain.be (B. Galand), stephane.colognesi@uclouvain.be (S. Colognesi).

<https://doi.org/10.1016/j.ssaho.2026.103178>

Received 16 May 2025; Received in revised form 24 June 2026; Accepted 26 June 2026

Available online 15 July 2026

2590-2911/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

2017). Consequently, few empirically validated tools are available that adequately capture classroom oral practices. However, in research contexts, accurate, valid and reliable measurement is essential to obtain interpretable and generalizable results (Coupaud et al., 2019; Koo & Li, 2016).

To address this gap, this study brings together didactic approaches, which emphasize contextualized and meaningful communication, and psychometric approaches, which ensure methodological rigor, in order to develop a valid and reliable instrument for assessing pupils' oral performance in their first language of schooling (in this case, French) in primary education. This instrument is called ORAToR, which stands for 'Oral Recap Assessment Tool with Reliability'. More specifically, this paper presents the three-year process undertaken to design, refine, and validate this tool. This process encompassed the tool's initial development, a pilot study providing a first opportunity to test its functioning, and a final phase of data collection that enabled us to examine its reliability and validity.

To create this measurement tool, we chose to focus on a specific genre. Indeed, genre constitutes a didactic tool recommended in oral didactics for both teaching and assessing oral skills (Dolz & Gagnon, 2008). We can define a genre as "a particular class of speech events [written texts included] which are considered by the speech community as being of the same type, e.g., prayers, conversations, songs, speeches, poems" (Richards et al., 1993, p. 125). A genre-based approach makes it possible to anchor the teaching of oral skills in an authentic communicative context (Abbaszadeh, 2013).

To work with situations that learners are likely to encounter in both everyday and academic life, we focus specifically on the knowledge summary. This oral genre consists of a speaker informally providing an interlocutor with concise and accurate information about a given topic, drawing on their long-term memory (Wiertz et al., 2025). This type of oral production typically takes place in an informal context, involves few interlocutors, and allows for limited preparation time. For example, it may occur when a pupil summarizes to their parents what they have learned during their day at school. We selected knowledge summary on the basis of a previous study in French-speaking Belgium (Colognesi et al., 2022). In that study, primary school teachers identified oral summarization as the most important oral situation for pupils to master, as it is frequently required in classroom practice (e.g., recalling what was covered the previous day, summarizing prior knowledge before a new lesson sequence). However, the study also highlighted a paradox: While teachers considered oral summarization essential, they reported that they did not explicitly teach this genre because they did not know what its components were. Our aim, therefore, was also to help teachers identify what constitutes an oral knowledge summary, enabling us to guide teachers more effectively on what can be taught and assessed within this genre.

To achieve this, as we explained above, we first sought to develop an assessment tool capable of evaluating pupils' knowledge summaries, drawing on both didactic principles and psychometric approaches. More precisely, we created a task in which upper-primary pupils (in the French-speaking Belgian context) are invited to summarize for an age-matched peer what they know (drawing solely on their memory) about a given topic within a limited amount of time. We then developed an assessment tool, ORAToR, to evaluate pupils' performance on this task. The tool was designed with broadly applicable components and can be adapted to various linguistic and curricular contexts. To facilitate its use and support its transfer to different settings, we present it in the form of a criterion grid. With this tool, we aim to contribute to international efforts to better understand, teach, and assess oral language skills in authentic classroom situations. Ultimately, the entire process of development and validation of this criterion grid may serve as a guide for creating research-based measurement tools for the various oral genres that can be taught in class (e.g., movie review, debate, interview). Accordingly, this study has two complementary aims: first, to describe the procedure followed to develop ORAToR, a criterion-referenced tool

for assessing oral knowledge summaries; and second, to examine the empirical quality of the resulting tool by assessing its reliability and exploring its factor structure.

In the following section, we explain why the criterion grid is a promising tool for evaluating oral expression, define oral expression and its components, and introduce the notion of genre, with a focus on knowledge summarization.

2. Conceptual framework

2.1. The criterion grid as a means of assessing oral competence

Numerous studies have shown that a criterion-referenced grid is an effective tool for evaluating speaking. Indeed, the use of criteria enables better identification of what needs to be evaluated (Berthiaume et al., 2011). It provides a structured way to operationalize oral performance, which is otherwise difficult to capture due to its ephemeral nature (Dunbar et al., 2006).

However, its effectiveness depends on careful design and empirical validation (Panadero et al., 2013). For example, it is complex to translate all possible responses into a set of criteria, which can impact the validity of the tool (Bouwer et al., 2016). Furthermore, the inter-rater reliability of the grid can be threatened if raters do not interpret the criteria or their measurement scales in the same way (Balan & Jönsson, 2018). Thus, it seems essential to check the validity and reliability of a criterion-referenced grid before using it. Although many approaches have been developed in this direction (Agan & Deniz, 2020; Rochford & Borchert, 2011), many studies (particularly in the field of didactic) still use criterion-referenced grids, or more broadly, tools, for which the validity and reliability have not been empirically verified (Grugeon-Allys & Grapin, 2015). Yet, without this important step, we can neither ensure that what is being measured does indeed represent what the researchers want, nor verify the replicability of the results with another sample or another researcher (Coupaud et al., 2019; Koo & Li, 2016). To carry out this evaluation stage, it is first and foremost necessary to identify the learning objectives and what will be assessed, which in this case is oral communication. With this in mind, we will next define oral communication and its components.

2.2. Oral communication and its components

In the literature, several authors have endeavored to construct a typology of oral communication in order to better capture it (Colognesi & Hanin, 2020). Our work drew on two of these attempts. First, we adopted the three-ability typology proposed by Dolz and colleagues (1993), which is widely used in international research on oral language. This typology aligns with genre-based pedagogy, a framework increasingly recommended in recent literature for the teaching of oral language (Melissourgou & Frantzi, 2017). In Dolz et al.'s (1993) typology, there are three abilities required for oral communication. First is the ability for action. This corresponds to the adjustment of speaking to the communication context. Second is discursive ability. This refers to the production of content that follows a coherent structure. Third is linguistic-discursive ability. This includes the use of a lexicon and appropriate grammatical rules.

We also chose to rely on Dumais's (2016) typology of 10 types of speech components, which allowed a certain exhaustiveness. In Dumais's (2016) typology, there are 10 types of oral-language related aspects: paraverbal (flow, intonation, etc.), morphological (word structure), syntactic (links between words), lexical-semantic (vocabulary), non-verbal (gestures, mimicry and postures), material (media), communicative (adaptation of communication to context), discursive (discursive conduct and language operation), content (discourse information) and emotional (presence of emotional expression and attitude).

Both of these typologies (Dolz et al., 1993; Dumais, 2016) were based on French-speaking research. Stordeur (2023) identified possible

parallels of these typologies in the English-language literature, in particular, the typology developed by Millard and Menzies (2016). Those authors identified four dimensions of speaking. First is the physical dimension, which includes elements related to voice (rate, volume, etc.) and body language. Second is the cognitive dimension, which includes elements related to content, structure, clarification and critical reflection. Third, the linguistic dimension covers vocabulary, grammatical rules and rhetorical techniques. Last is the socio-emotional dimension, which encompasses interaction with the interlocutor, confidence in speaking and consideration of the audience's level of understanding.

While these typologies allow a certain characterization of oral skills, no statistical analysis has been carried out to verify the validity and reliability of the dimensions they identify. However, quantitative studies have identified factors underlying oral language, using factor analysis. For example, Massonnié et al. (2022) conducted a study of 126 4- to 5-year-olds and 124 5- to 6-year-olds in England. In that study, they assessed both speaking and listening comprehension. Measurements were taken via standardized tests of performance on several independent tasks (lexical knowledge, phonology, etc.). Factor analyses revealed a 4-factor model for 4- to 5-year-olds and a 3-factor model for 6- to 7-year-olds. These factors mainly focused on discourse content related to the production and reception of ideas, and grammatical and lexical elements, which do not correspond to all the dimensions of the typologies presented above. Furthermore, the authors noted that pupils' oral communication skills were measured via several tasks completed separately, which does not correspond to a real communication situation. The ecological validity of these factors may therefore be questioned (Kihlstrom, 2021).

In the end, it seems particularly complex to obtain a valid and reliable measure of speaking in real oral communication contexts (Gagnon et al., 2018; Mercer et al., 2017). However, the textual genre approach, which is considered in the literature to be the anchor to follow when teaching oral language (Flowerdew, 2015; Melissourgou & Frantzi, 2017), makes it possible to place pupils precisely in an authentic communicative situation. Thus, this approach seems to be a relevant avenue to explore in order to adequately measure pupils' speaking skills. More specifically, the oral knowledge summary genre, used daily by pupils, seems to be a particularly opportune genre for achieving this objective.

2.3. Oral summary as a speaking task to be investigated and evaluated

As explained in the Introduction, a genre is a set of oral or written linguistic products that share common characteristics within a given culture (Chartrand et al., 2015). Each genre is characterized by specific communicative purposes and linguistic features (Dirgeyasa, 2016), making it possible to define more precise assessment criteria. Furthermore, the genre approach anchors the teaching of oral language in authentic situations that pupils are likely to encounter in their daily and school lives (Flowerdew, 2015). Moreover, this is advocated by the action-oriented approach highlighted in the CEFR (Common European Framework of Reference for Languages; Council of Europe, 2001). The learner is then seen as a language user as well as a social actor (Germain-Rutherford, 2021).

As mentioned above, we decided to focus on the genre of oral knowledge summarization. While few studies have focused on oral summarization, many have addressed written summarization (Yu, 2009). Those studies highlighted that the aim of written summarizing is to reduce the amount of original information by selecting the most relevant information, without errors (Stewart, 2009). This is achieved by constructing a coherent, comprehensible text of a length that enables the reader to take in the information quickly, and by respecting the source's lexicon and structure (Li, 2014; Stewart, 2009). This source can be quite varied (e.g., film, book, scientific article) and influences the type of summary produced (Li, 2014). Although most studies have

focused on a written source text that is summarized in a written text (Stewart, 2009), we postulate that when a person is asked to make an oral summary of knowledge they already have, they call upon elements specific to the oral modality that are not present in a written summary, such as paraverbal (intonation, volume, etc.) and nonverbal (gestures, glances, etc.) aspects, along with direct interaction with an interlocutor (Fritz et al., 2019; Petlyuchenko & Chernyakova, 2019).

3. This study

In view of the lack of validated tools for assessing speaking skills and the difficulty of obtaining empirical knowledge based on measurements taken from pupils' speaking in authentic situations, we decided to develop a valid and reliable criterion-referenced grid for researchers to use in a real communication context, in order to overcome these shortcomings. The present paper therefore describes the development and validation of ORAToR, a criterion-referenced grid for measuring pupils' oral summarization skills. We chose oral summary as the genre to focus on, while hoping to propose a replicable approach to assessment of performance on other oral production tasks as well. This choice of oral summary was for the reasons given above: it is often used in class and on a daily basis, and teachers classify it as a speaking task to be mastered. But there is no existing inventory of pupils' oral summarizing skills, just as there is no typology to detail its various components.

4. Methods

We developed a criterion-referenced grid to assess the oral summarization of knowledge. We followed 10 steps inspired by the recommendations of Reddy (2011) and Martin et al. (2016). Three main researchers worked on this project, but four additional researchers assisted us with some of its steps (for example, to verify the reliability of the different criteria). In this section, we present these steps, organizing them into three main stages (Fig. 1).

This study was reviewed and approved by the IPSY ethics committee.

4.1. Stage 1: development of ORAToR

4.1.1. Identify learning objectives and assessment criteria (Step 1)

This step took us five months and allowed us to identify the objectives of our measurement tool.

More specifically, after consulting the literature (Rochford & Borchert, 2011), we decided that the aim of our grid would be to assess the quality of an oral summary of knowledge produced by 11- to 12-year-old pupils. We chose this group because there is still little information on pupils' competence to produce oral summaries at this age. What is more, they already have well-developed language skills, which provides sufficient content for a fine-grained analysis of their performance.

Then, to construct the first version of our criterion-referenced grid, we (1) used the literature (typologies by Dolz et al., 1993; Dumais, 2016), (2) analyzed the curriculum expectations in force in our territory at the time of the study (Fédération Wallonie-Bruxelles, 2013) and (3) analyzed already-existing grids for measuring oral skills (Colognesi et al., 2020; Gregg & Mather, 2002).

We also decided to construct a specific language production task that could be evaluated by our grid, in order to anchor its creation in a concrete situation, although our ultimate aim is for this grid to be transferable to any other oral knowledge summarization situation. We therefore began by creating a task in which pupils were asked to summarize what they knew about one of three animals - cats, dogs or rabbits - within a maximum of 3 min. This task was designed to enable pupils to talk about a subject familiar to most of them, so as to avoid assessing their level of knowledge rather than their oral summarization skills. To do such a summary, they had to select and structure the relevant elements of their knowledge, a process that aligns with the principles of

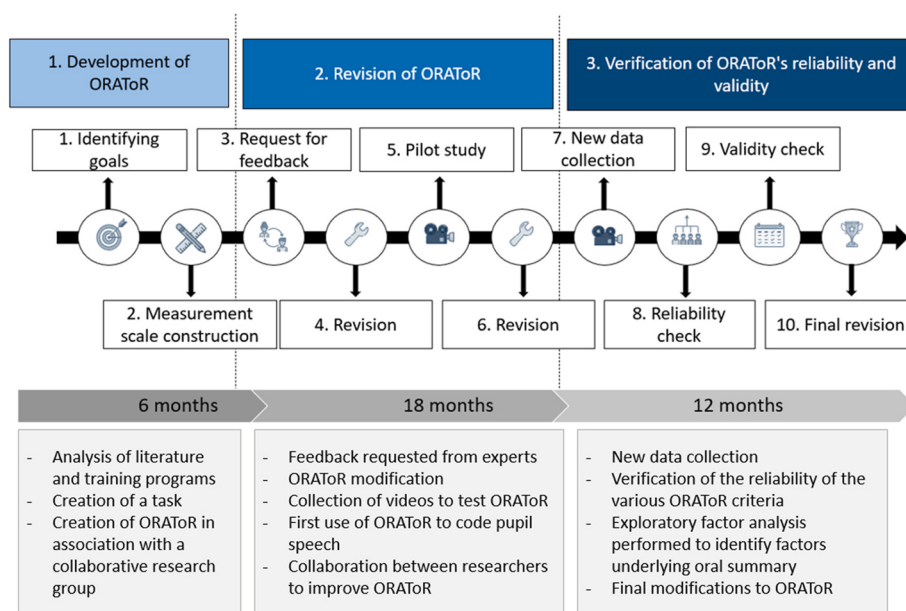


Fig. 1. The procedure followed, inspired by Reddy (2011) and Martin et al. (2016) to create ORAToR.

deletion and generalization described by Kintsch and van Dijk (1978). This task was carried out in pairs of pupils: one speaker (who spoke and performed the task) and one interlocutor (who listened to the speaker). At this point, there was only one task, and the interlocutor was instructed not to speak during it. Their role was to listen to the speaker, thereby making the situation more realistic than if the speaker were talking alone or to an instructor. Their presence also allowed us to observe how the speaker adapted to the interlocutor's reactions.

The first version of the grid and the task were presented to the collaborative oral-expression working group to which we belong (GCEO). The group includes 4 researchers, 6 teacher trainers and 5 teachers. We requested feedback from all group members. Following this feedback, we continued working exclusively with the researchers from this group, who are the potential users of the instrument, and discussed the points to bear in mind for this grid (for full details of this procedure, see Wiertz et al., 2020).

Based on these exchanges, we decided to build the measurement tool with three parts. The first part, *performance note-taking*, would transcribe the essential elements of the pupil's discourse in an organized and structured way. The second part, *criteria*, would be the grid itself. To complete it, the researcher would use the note-taking section. Finally, the third part would be a *user's guide*, designed to explain how any researcher can use the tool.

4.1.2. Develop a measurement scale for each criterion in the grid created (Step 2)

Following these decisions, the lead researcher improved the tool and built a 13-page user's guide. It includes information on the task, instructions on how to complete the note-taking section and instructions on the different levels of criterion measurement. This process was carried out in parallel with a more detailed exploration of the literature. A name was chosen for the tool: ORAToR (Oral Recap Assessment Tool with Reliability). The entire tool was initially developed in French. The various criteria have been translated into English for this paper. This stage took us two months.

4.2. Stage 2: revision of ORAToR

4.2.1. Get feedback (Step 3) and revise the tool accordingly (Step 4)

We then took two months to seek expert feedback on our measurement tool and modify it accordingly.

More specifically, to strengthen the validity of the tool (Agan & Deniz, 2020), we sought the expertise of four researchers (two national, two international) working in the field of French oral language didactics. They were asked to give their point of view and to suggest improvements. Based on their comments, four types of changes were made. The task was modified so that pupils could choose between summarizing what they knew about an animal or about an electronic device of their choice. Certain criteria were added, notably characteristics specific to spoken French. The measurement scales for the criteria were revised to make them more precise and concrete. More specifically, an explanation for each measurement level was added in the user's guide. Finally, the objectives of the measurement tool presented in the guide were made more explicit.

4.2.2. Conduct first pilot study (Step 5) and revise the tool (Step 6)

We then spent just over a year testing the tool through a pilot study to test the data collection protocol and the tool, as recommended by Rochford and Borchert (2011). The results allowed us to refine ORAToR and better align it with the realities of pupils' oral productions.

We collected 48 videos, in each of which a different 11- to 12-year-old pupil performed the speaking task. There were equal numbers of girls and boys. The pupils came from two classes from two different schools. Both schools had an average socio-economic index.

Two of the researchers coded these videos with the tool in order to identify possible difficulties when using it. Following this coding, two types of modifications were made, related to data collection and to the different parts of ORAToR. As far as data collection is concerned, we decided to modify the protocol in order to avoid bias in the data obtained. It now specifies that the researcher must position the camera so as to see both pupils, speaker and interlocutor, and that before the pupils come in, the researcher must check that the sound picked up by the camera in this position is audible. Moreover, in this pilot study, participants were given the strict instruction not to interact. However, some pupils did not follow this instruction and explicitly interacted with the speaker. We therefore decided that, in subsequent phases, when the researcher explains the task to the pair of pupils, he should no longer prevent them from interacting. Furthermore, once the researcher has explained the task, he must leave the room before the speaker begins the task, in order to avoid distracting them.

For modifications related to the tool itself, an iterative process (Rochford & Borchert, 2011) combining discussion, video viewing,

literature reading and modifications to ORAToR was repeated several times by two researchers. Finally, nine individual coding sessions, in which the two researchers watched randomly selected videos, and 18 meetings were held to modify the tool as a whole. The three parts of the tool were adapted for ease of use.

For the first part of the tool (*performance note-taking*), the evaluator was initially invited to indicate three elements: the categories of ideas addressed by the speaker (e.g., “food”), the ideas themselves (e.g., “the cat eats kibble”), and ideas that were not relevant to the task (e.g., “my mother does not like animals”). In the new version of the grid, this first part was expanded so that the evaluator could transcribe information that would be useful for completing all the criteria in part two of ORAToR. Thus, the evaluator is now invited in this first note-taking section to provide information on the connectors used by the speaker, the syntactic or lexical errors identified, and paraverbal elements, as well as information about the interactions between the speaker and the interlocutor (for example, the speaker asking the interlocutor whether they have any questions).

For the second part of the tool, which corresponds to the *criteria* themselves, some criteria were removed (e.g., the speaker's use of emotions) and many criteria were added, so that all the elements appearing in the videos could be coded in one way or another. However, in view of the poor quality of the videos made during this pilot study, the items relating to paraverbal and non-verbal components were not revised during this stage. Excluding those items, the previous version of the tool was composed of 11 criteria, whereas the revised version at this point included 17 criteria. The added criteria made it possible to assess more precisely the different types of information provided by the speaker, as well as to evaluate whether the expectations of the communication situation are met, the coherence of the speech, and what the speaker does to ensure that their message is understood by the interlocutor. Moreover, among the criteria that were kept, the wording was revised so that the speaker became the subject of the sentence and to better guide the evaluator (e.g., “*The overall speech remains centered on the topic*” was changed to “*The speaker remains focused on the summary and not on their personal feelings*”).

The final major change made in this section concerned the scoring scale. In the previous version, criteria were evaluated on scales ranging from 0 to 2 or from 0 to 3. In the revised version of ORAToR, for most criteria, the rater is simply required to indicate, based on Part 1 of the tool, the number of occurrences (for example, the number of ideas provided by the speaker or the number of connectors used). This approach makes it possible to accurately reflect what was produced by the speaker. When it is not possible to count occurrences (for instance, when assessing whether the speaker maintains sufficient eye contact with the interlocutor, or whether the speech meets the task instructions), the criteria are evaluated on a scale from 1 to 5. This scale, ranging from 1 to 5, allows for greater nuance than previously and enables a better distinction between different levels of performance.

The third part of the tool (*user's guide*) was also adapted to provide information about the added elements in Parts 1 and 2 of the revised version of ORAToR. Furthermore, to help the evaluator understand the expectations, examples based on actual pupil speeches were added for each criterion. Thus, for each measurement level for every criterion, precise indicators and examples are provided to the evaluator in order to make certain subjective terms found in the criteria more concrete, such as ‘relevant’ or ‘appropriate’.

4.3. Stage 3: checking the reliability and validity of the criterion grid

4.3.1. Collect new data (Step 7)

Following the revisions based on the pilot study carried out in the previous stage, new data were collected over three months using the revised version of ORAToR, taking into account all the modifications mentioned above. Only these data were used in this and subsequent stages.

In concrete terms, 104 videos were collected from pairs of French-speaking Belgian pupils aged 11-12, in which one pupil spoke while the other listened. Of the 104 pupils who spoke, 52% were girls, 46% boys and 2% identified themselves as non-binary. The average age of the pupils was between 11 and 12 years ($M = 11.17$; $SD = 0.47$); 73% spoke only French at home, 23% French and another language, and 4% French and two or more languages. Informed consent was obtained from each of these pupils and their parents. The pupils came from 15 classes in 12 different schools. Eight of these are considered to be schools with a high socio-economic status and seen as having a more privileged population.

The participating pupils were first given a questionnaire to complete in class, to obtain general information (age, gender, etc.). The rest of the data collection took place in an isolated room at the school. In consultation with the teacher, we were careful to ensure that the pairs created were pupils who felt comfortable with each other. Once in the room, the pupils drew lots to decide who would play the speaker and who would play the interlocutor. After the instructions had been explained, the researcher left the room to put the pupils in a more comfortable situation. The interlocutor had to turn on the camera, listen attentively to the speaker, and come and get the researcher once the performance was over. This time, the interlocutor was given no instructions on how to interact with the speaker. Observation of the videos shows that some participants remain passive during the speech, whereas others smile, nod, or interact directly with the speaker by taking the floor.

For their part, the speaker had to summarize what they knew about either an animal of their choice or an electronic device of their choice within a maximum of 3 min, and turn off the camera at the end of their speech. The speaker had no preparation time before starting to speak. If necessary, the researcher briefly entered the room after 3 min to signal to the speaker that they must finish speaking, in order to avoid too much variability in speaking time between pupils.

Once the data had been collected, two researchers reworked the criteria for paraverbal and non-verbal elements, which had not been revised in Step 4 due to poor video quality. This refinement required parallel exploration of the literature and viewing of pupil performances. As a result, eight new criteria emerged: four focused on paraverbal elements and four on non-verbal elements.

4.3.2. Check reliability (Step 8)

We then spent four months verifying the tool's inter-rater reliability.

Two raters were involved in assessing the inter-rater reliability of the new version of ORAToR: the lead researcher, who participated in the entire process, and another researcher who had not been involved in any of the previous stages. The latter was invited to read the user guide for the tool and to ask any questions if needed. The ORAToR user guide was subsequently revised to include clarifications addressing these questions.

Next, the two raters independently watched and coded five videos using ORAToR, after which they compared their results. In cases of scoring discrepancies, they discussed their understanding of the criteria. These discussions allowed the raters to refine their interpretation of the criteria and rating scales, and to further adjust the user guide accordingly.

Finally, the two researchers independently coded 15 new videos using the updated set of 25 criteria. Inter-rater reliability was assessed based on this final coding phase. For criteria corresponding to quantitative variables, an intraclass correlation coefficient (ICC) was calculated to check the inter-rater reliability of the coding (Koo & Li, 2016). The option selected was two-way random, single rater and absolute agreement (Koo & Li, 2016). To interpret the results, we used the suggestions of Koo and Li (2016), who stipulated that the level of inter-rater reliability can be considered low when the ICC is below 0.5, moderate between 0.5 and 0.74, good between 0.75 and 0.9, and excellent above 0.9. For criteria where raters were asked to indicate a score between 1 and 5, which corresponded to ordinal categorical variables, a weighted kappa was calculated with quadratic weights. We used the thresholds

proposed by Santos (2018) to interpret these kappas, where reliability is considered very low for kappas from 0 to 0.2, low from 0.21 to 0.4, average from 0.41 to 0.6, satisfactory from 0.61 to 0.8 and excellent from 0.81 to 1.

4.3.3. Check validity (Step 9) and make final revision of the tool (Step 10)

In this step 9, ORAToR's construct validity was checked to identify whether scores on the various criteria can indeed be grouped into different latent factors that underlie the task being performed (Boateng et al., 2018; Kahn, 2006).

All of the preceding steps enabled us to create a measurement tool consisting of a note-taking section, a grid with 24 criteria (Fig. 2) and a 26-page user guide. Each of these steps strengthened ORAToR's content validity (Martin et al., 2016). This version of ORAToR is available in the supplementary materials. The original version is in French and has been translated into English for this contribution.

To check the validity of this version, we performed an exploratory factor analysis (EFA). To carry out this ninth step, one researcher coded all 104 videos (since inter-rater reliability had been verified). The videos ranged in length from 8 s to 4 min 35 s. Using ORAToR in its entirety (the note-taking, criteria and guide parts), the time taken to code a video varied between 5 and 50 min. The criteria section of this version of ORAToR included nine criteria assessed on a scale from one to five, and 15 other criteria for which the evaluator is asked to count certain elements using the note-taking part (such as the number of connectors). To carry out the EFA, we converted all criteria to a 1-5 scale. Once the data from the researcher's coding had been transformed, we ran the EFA with Jamovi software version 2.3.26. We chose the "minimum residuals" option for the extraction method and performed an oblique "oblimin" rotation.

As a result of the EFA, certain criteria were removed (Step 10), based on combining reflections on the results of the EFA with theoretical and methodological reflections on the relevance of these items and how they were rated (Boateng et al., 2018; Yong & Pearce, 2013). This last step is presented later, in the Results section.

4.4. Summary of the steps

The development of ORAToR was a three-year process. During the first year, the objectives of the tool were defined, and an initial version was created, informed by a review of the literature, discussions within a collaborative research group, and input from experts in oral language didactics. A pilot study was then conducted to test the tool, during which pupils' oral performances were recorded. In the second year, the videos collected during the pilot study were analyzed. Through an iterative process led by two researchers, the tool was refined to better align with real-world performances. Finally, during the third year, data were collected from 104 pupils, which made it possible to assess the reliability and validity of ORAToR. Table 1 provides a summary of these different stages.

5. Results

In this section, we examine the findings from the three main stages described in the Methods section. We begin by highlighting the contribution of Stage 1 to the direction taken in building ORAToR. Then, we highlight the importance of Stage 2 in refining the task administration protocol and the content of the assessment tool. Finally, we present the results of the statistical analyses carried out in Stage 3 to verify the reliability and validity of the tool, as well as the final version of ORAToR.

5.1. Development of ORAToR (Stage 1)

After reading the literature, we were able to identify that criterion-referenced grids for assessing pupils' oral products are often intended

for teachers' use. However, discussions with the members of our collaborative research group (in step 1) revealed that teachers and researchers do not expect the same things from a grid. Teachers are more interested in time-saving grids, while researchers are more concerned with measurement accuracy. Based on the literature and these discussions, we decided to create a tool specifically for researchers. This choice enabled us to create a complete tool without being constrained by the time needed to complete it, which made it possible to have a three-part tool. In addition, with the members of our collaborative group, we decided to structure our evaluation tool in line with the three types of ability identified by Dolz and colleagues (1993), and to use the communicative aspects highlighted by Dumais (2016) in his typology to create the different criteria of our evaluation grid. This choice thus guided the initial content of ORAToR and served as the foundation for the construction of our assessment tool.

5.2. Revision of ORAToR (Stage 2)

The various steps followed in this stage were essential to ensure the validity and reliability of ORAToR. In fact, the experts' feedback (in step 3) made us realize that ORAToR's objectives were not easily understandable for people who had not taken part in the project. These reactions forced us to rethink ORAToR's aims and specific features. As a result, we were able to add a section in Part 3 (the user's guide), which explains the objectives and issues related to the use of ORAToR to the reader, as well as the framework within which it should be used.

The pilot study was essential for refining the protocol for administering the task, which helped to reduce measurement errors in the data collection. For example, these reflections led us to refine the pupils' task so that it could truly assess the "oral summary of knowledge" genre (e.g., we gave the speaker more freedom in choosing the subject to summarize). Moreover, the first coding exercises enabled us to compare the theoretical recommendations with the reality of the data. For example, the criterion "The speaker uses their emotions appropriately" was removed, as it was too complicated to evaluate this criterion solely on the basis of the videos obtained. Contrary to what we had initially expected, many interactions between the speaker and the interlocutor were present in the videos. In this vein, the criteria related to interactions were added. It was therefore essential to add criteria addressing this to avoid having certain elements present in pupils' oral productions that were not being analyzed by our tool. Our final refinement of the user guide was based on the coded videos of pupil performances, which we used to provide indicators and examples of what is expected for each criterion, as well as what helps distinguish the different measurement levels from one another. Once these changes had been made, we were able to carry out the necessary analyses to verify the tool's reliability and validity.

5.3. Checking the reliability and validity of the criterion grid (Stage 3)

5.3.1. The inter-rater reliability of the tool

The ICCs and weighted kappas (used for the criteria marked with an asterisk in Table 2) allowed us to determine the level of inter-rater reliability of the 25 criteria. Based on these results, we removed the "intelligibility" criterion, which was one of the two criteria judged to have poor reliability, but we kept the criterion "The elements discussed are not erroneous", in order to preserve the content validity of the grid (Moskal & Leydens, 2000). Specifically, among the criteria we retained, one showed poor inter-rater reliability, five were moderate, eight were good, and ten showed excellent reliability (Table 2), using the recommendations of Koo and Li (2016) for the ICCs and Santos (2018) for the weighted kappa.

5.3.2. Results of the EFA

Following the results of the inter-rater reliability analysis, we carried out exploratory factor analyses using the 24 remaining criteria, with the

A. Ability to act

Competence 1 – Orient your speaking to the communication situation						Comments
1a. The speaker follows the directions.	1	2	3	4	5	
1b. The speaker uses communication strategies.						
1c. The interactions are focused on content.						
1d. The speaker remains focused on the summary and not on his or her feelings.						

B. Discursive ability

Competence 2 – Develop meanings :						Comments
2a. The speaker addresses different categories of information.						
2b. The speaker addresses many different ideas.						
2c. The speaker elaborates on ideas.						
2d. The elements covered are relevant.						
2e. The elements discussed are not erroneous.						
Competence 3 - Ensure message organization and coherence						Comments
3a. The introduction is qualitative.						
3b. The different elements appear coherently one after the other.						
3c. The speech is punctuated by elements that move it forward.						
3d. The conclusion is qualitative.						
3e. The speaker uses numerous connectors.	1ous : All		Different :			

C. Language and discourse ability

Competence 4 – Use lexicon and morphosyntax						Comments
4a. The speaker uses words specifically related to the characteristics of the subject.						
4b. The Speaker constructs orally correct statements.						

Competence 5 - Use verbal means						Comments
5a. The speaker's flow is appropriate.	information integration	0	1	2		
	Discourse continuity	0	1	2		
5b. The speaker's volume is appropriate.						
5c. The speaker's intonation is appropriate.	1	2	3	4	5	
Competence 6 - Use non-verbal means						Comments
6a. The speaker doesn't use self-centered gestures.	1	2	3	4	5	
6b. The speaker uses coverbal gestures.	1	2	3	4	5	
6c. The speaker looks at interlocutor.	1	2	3	4	5	
6d. The speaker uses the draft sheet appropriately.						

 : Hatched criteria are criteria that were removed after EFA.

Fig. 2. ORAToR criteria grid after EFA.

help of the Jamovi program. A three-factor solution (shown in Table 3), with a total of 18 criteria having a minimum loading of 0.3, was identified as the best extraction solution using the parallel analysis method (Yong & Pearce, 2013).

The first factor identified by the factor analysis was made up of seven criteria. We called it Informational Content. This factor answers the question: "What information does the speaker give in their summary?" This factor includes criteria related to the amount of information given, as well as those that identify the lexicon used by the pupil. The second factor, Interaction with the Interlocutor, was made up of five criteria and answers the question "How does the speaker maintain the attention of the interlocutor?" This factor is made up of criteria relating to non-verbal, explicit interaction and intonation. However, the criterion "The interactions are focused on content" showed a loading of 0.398 with the first factor and a loading of 0.428 with this second factor. For this criterion, the speaker is assessed on the frequency with which they explicitly and relevantly include their interlocutor in their summary (by addressing the interlocutor by name, asking questions, checking comprehension). These marks of interaction were often integrated directly into the information given by the speaker, which explains why this criterion was linked with Factor 1, Informational Content. However, in the end, these various markers served more to include the interlocutor in the discourse, and thus arouse their interest, than to provide new information. That is why we decided to keep this criterion in Factor 2, Interaction with the Interlocutor, despite the cross-loading that existed with the content-related factor. The third factor is made up of six

criteria. We called it Clarity of the Message Conveyed. This factor answers the question "What does the speaker pay attention to in order to make their message comprehensible?" This factor is made up of criteria relating to paraverbal elements, speech structure, relevance of information conveyed and respect for characteristics specific to spoken French.

Bartlett's test was significant ($p < .001$) for the model obtained, indicating that the correlations between our different criteria were adequate (Yong & Pearce, 2013). According to the Kaiser-Meyer-Olkin (KMO) measure, the adequacy of the data for factor analysis was 0.813, which is considered good (Yong & Pearce, 2013). In view of these results, the criteria not retained in the EFA were removed from ORAToR to arrive at the final version of the measurement tool (Fig. 2).¹

6. Discussion and conclusion

This paper presented the process of developing and validating a tool for measuring pupils' oral communication skills. Indeed, few such tools currently exist, making it difficult to understand the specifics of oral communication and thus to clarify what is teachable and assessable in oral communication (Mercer et al., 2017; Oliver et al., 2005). To

¹ If you wish to obtain the final version of ORAToR (which is only available in French), you may contact the first author of this article at the following. christine.wiertz@uclouvain.be

Table 1

Summary of the development stages of ORAToR.

Step	Description	Approximate Duration
Identifying goals	Literature review and definition of the tool's objectives	5 months
Measurement scale construction	Creation of the initial version of the tool	2 months
Request for feedback	Collection of feedback from four experts in oral language didactics	2 months
Revision	Incorporation of expert feedback to revise the tool	
Pilot study	Collection of 48 pupil videos to test ORAToR	3 months
Revision	Analysis of the collected videos and refinement of ORAToR through discussions between two researchers; revisions were also informed by the literature	Approximately 12 months (This revision phase was spread over the year, as the two researchers involved were also engaged in other research projects.)
New data collection	Collection of 104 additional pupil videos	3 months
Reliability check	Two raters assessed inter-rater reliability by coding 15 videos	4 months
Validity check	Exploratory factor analyses (EFA) were conducted	3 months
Final revision	Final modifications to ORAToR based on the results	2 months

exemplify this construction process, we anchored our work in the development of a criterion-referenced grid for assessing pupils' competence in producing an oral summary of their knowledge about a topic.

In terms of results, discussions with various professionals in the field led us to refine our objectives and build a tool for researchers rather than teachers. In particular, teachers need a quick and easy-to-use measurement tool (Wiertz et al., 2020). Our aim, however, was to gain a better understanding of speaking. We believe that we must first measure oral language in a research environment, so that we can then adapt the tool in a way that respects the needs of teachers. The ultimate aim of this study is to serve as a guide for the creation of research-based measurement tools for the various oral genres that can be taught in the classroom. A second result highlighted in our study is the importance of following an iterative process (Blanc et al., 2025; Rochford & Borchert, 2011) to reinforce the content validity of a measurement tool. Through this iterative process, we were able to adapt our tool to the reality of pupils' performances. In particular, these steps enabled us to add criteria for assessing interactions between speaker and interlocutor. The relevance of taking interaction into account was not immediately evident to us when we read the literature. Thus, we believe that these steps enabled us to reinforce the content validity of our tool and to better anchor it in the specifics of oral communication (Martin et al., 2016). Finally, the results highlight three factors that underlie the production of an oral summary of one's knowledge.

The first factor, Informational Content, is related to the content the speaker provides to support their discourse, as well as elements linked with the oral language code (Dumais, 2016). This factor includes the criteria evaluating the amount of information given (e.g., if the speaker says that a cat has four legs and eats kibble, they are providing two different pieces of information), the elaboration of this information through various strategies (e.g., the speaker provides examples to illustrate the information they are giving), the number of different categories of information addressed (e.g., if the speaker says that a cat has four legs, two ears, and eats kibble, they are providing three different pieces of information but only two different categories of information: the animal's physical characteristics and its diet), the number and variety of the connectors used (e.g., and, because, yet), and the fact that the

Table 2

Level of reliability for each criterion.

Reliability level	Criterion	Intraclass Correlation/Weighted kappa	95% Confidence Interval	
			Lower Bound	Upper Bound
Excellent	The speaker uses words specifically related to the characteristics of the subject.	0.96	0.88	0.99
	The speaker addresses different categories of information.	0.95	0.83	0.98
	The speaker addresses many different ideas.	0.95	0.58	0.99
	The conclusion is qualitative.*	0.93	0.85	1
	The speaker elaborates on ideas.	0.91	0.75	0.97
	The speaker uses the draft sheet appropriately.*	0.88	0.72	1
	The speaker follows the directions.*	0.85	0.62	1.1
	The speaker doesn't use self-centered gestures.*	0.83	0.69	0.98
	The speaker looks at the interlocutor.*	0.82	0.66	0.99
	The introduction is qualitative.*	0.81	0.68	0.94
	The speaker uses communication strategies.	0.86	0.62	0.95
	The speaker uses numerous connectors.	0.85	0.58	0.95
Moderate	The speaker uses many different connectors.	0.82	0.51	0.94
	The different elements appear coherently one after the other.	0.8	0.5	0.93
	The speaker remains focused on the summary and not on his or her feelings.	0.75	0.55	0.88
	The speaker constructs orally correct statements.	0.75	0.54	0.87
	The speaker's volume is appropriate.	0.75	0.55	0.87
	The speaker's flow is appropriate.*	0.7	0.5	0.91
	The interactions are focused on content.	0.7	0.46	0.84
	The elements covered are relevant.	0.68	0.27	0.88
	The speaker uses coverbal gestures.*	0.57	0.36	0.78
	The speech is punctuated by elements that move it forward.	0.55	0.24	0.75
	The speaker's intonation is appropriate.*	0.48	0.05	0.99
	The elements discussed are not erroneous.	0.33	0.23	0.72

Note. For criteria with an asterisk, we have used the weighted kappa. The original items were in French and have been translated for this article.

speaker responds well to the speaking situation by respecting the characteristics of the summary genre (e.g., the speaker does not talk about their own emotions). A link can be made between this factor and the characteristics of written summarization identified in the theoretical literature. Namely, in a written summary, the writer conveys the most relevant information from a source text, while respecting its lexicon (Li, 2014).

The second factor, Interaction with the Interlocutor, relates to the

Table 3
Results of exploratory factor analysis.

	Factor		
	Content	Interaction	Clarity
The speaker addresses many different ideas.	0.839		
The speaker uses communication strategies.	0.813		
The speaker uses numerous connectors.	0.804		
The speaker uses words specifically related to the characteristics of the subject.	0.723		
The speaker uses many different connectors.	0.650	0.362	
The speaker follows the directions.	0.626		
The speaker addresses different categories of information.	0.584		
The speaker uses coverbal gestures.		0.705	
The speaker doesn't use self-centered gestures.		0.599	
The speaker's intonation is appropriate.		0.438	
The interactions are focused on content.	0.398	0.428	
The speaker looks at the interlocutor.		0.357	
The elements covered are relevant.			0.618
The speaker's flow is appropriate.			0.604
The speaker's volume is appropriate.			0.594
The speaker remains focused on the summary and not on his or her feelings.			0.513
The different elements appear coherently one after the other.			0.484
The speaker constructs orally correct statements.			0.305

Note. The 'Minimum residuals' extraction method was used in conjunction with an 'oblimin' rotation. Loadings below 0.30 are not shown on the table. The original items were in French and have been translated for this article.

interactions the speaker has with the interlocutor. More precisely, the criteria that make up this factor refer to non-verbal elements (gestures and gaze), intonation and the number of interactions occurring. It includes elements that can arouse the interlocutor's interest. This factor, which is not directly related to the characteristics of a written summary, encompasses certain specific features of the oral modality, namely, the use of non-verbal, paraverbal elements and the rules of the social game associated with interaction with the interlocutor (Casteleyn, 2019; Fritz et al., 2019).

The third factor, Clarity of the Message Conveyed, corresponds to the speech elements that ensure the clarity of the message conveyed and make the speech comprehensible to the interlocutor. These include the relevance of the information conveyed and the structure followed, as well as criteria for judging volume, flow and adherence to characteristics specific to spoken French. In this factor, we also find characteristics that have been identified as part of a written summary: the relevant selection of information as well as the ability to structure it (Li, 2014; Stewart, 2009). Added to this are paraverbal elements characteristic of the oral modality (Petlyuchenko & Chernyakova, 2019).

Thus, these three factors can easily be linked with the primary function of speaking, that is, interacting with others (Factor 2) by expressing content (Factor 1) clearly and comprehensibly for the interlocutor (Factor 3) (Millard & Menzies, 2016). However, these factors identified by our analyses do not correspond to the typologies of speaking presented in our theoretical framework (Dolz et al., 1993; Dumais, 2016; Millard & Menzies, 2016), nor to any other typology, to our knowledge. This raises important questions. The theoretical typologies presented in the literature can indeed be used to simplify the concept of orality by dividing it into several components. For instance, without conducting a factor analysis, we might have assessed pupils' performance on our task through the components proposed in those typologies. However, our findings show that this approach would have led us to evaluate factors that do not align with the latent constructs that actually emerge from pupils' verbal and non-verbal behaviors (Boateng et al., 2018; Yong & Pearce, 2013). In other words, we would have ended up with a set of criteria that do not form a coherent whole, creating a gap between what is assumed to be measured and what is actually measured, thereby compromising the validity of the reported results (Cohen et al.,

2017).

Furthermore, other factor analyses that have been carried out in an attempt to identify the factors underlying the concept of orality (e.g., Massonnié et al., 2022; O'Connell et al., 2017) have also failed to identify the same factors that we did. This is because factor analyses are influenced by the way oral competence is measured. For this reason, we attempted to measure speaking in an integrated way, using a single task that we wanted to be close to a real communication situation, in order to increase the ecological validity of the factors found in the EFA (Kihlstrom, 2021). To that end, we focused here on a specific, fairly precise genre and a specific ecologically valid task. More precisely, we ensured that the speaker interacted with a peer of the same age rather than with an adult examiner. In the same vein, and in contrast to most standardized oral assessments, we did not evaluate skills in isolation (e.g., through tasks targeting only vocabulary). Instead, all skills were assessed within a single speaking task in order to more closely reflect the demands of real-life communication situations.

The discrepancy between our factors and those found in the literature seems to confirm the importance for research of building valid and reliable measurement tools for each genre studied, in order to gain a finer and more specific perception of the skills required to perform the different tasks required of pupils. In this sense, our methodological approach enabled us to learn more about a field that is still relatively unknown and in which research faces many obstacles (Casteleyn, 2019).

More precisely, the fact that this process was carried out over a three-year period allowed us to develop a comprehensive tool that was revised multiple times and informed by expert feedback as well as both theoretical and empirical data. This entire procedure contributed to strengthening the validity and reliability of ORAToR. We therefore believe that it can serve as a guide for developing robust tools aimed at measuring other oral genres from a research-based perspective.

6.1. Limitations and perspectives

Despite our aim to create a tool to measure pupils' ability to perform an authentic oral summarization task, the ecological validity of our results can be questioned. Indeed, in everyday life, it is highly unlikely that 11- to 12-year-old pupils would actually be asked to summarize for a classmate of the same age what they know about an animal or a technological device. What is more, some pupils may not have seen the point of the task and may not have performed it to the best of their ability. Moreover, speaking in front of others can be quite anxiety-inducing for some pupils (Stordeur et al., 2022, 2025). This is even more true in a setup like ours, where pupils were filmed and had a limited amount of time to speak and no preparation time. In the current version of our tool, we have not yet found a way to account for this anxiety, either in the assessment of pupils' performance or in the way ORAToR is administered. Although we did not observe any pupils who appeared particularly anxious in the videos, such anxiety may still have influenced the results of this study. One way to partially reduce these various difficulties could be to ask pupils for feedback on their experience of taking part in an assessment with ORAToR, so that we can effectively improve our tool.

The generalizability of our factors can also be questioned. Would we have arrived at the same results with another oral knowledge summary situation? Are the factors specific to the knowledge-summary genre, or to the particular task that we used? Can these factors be generalized to other genres? It would be interesting to verify this by using ORAToR to evaluate other knowledge summarization tasks, as well as tasks representing other genres. Furthermore, even though we had more participants than in many studies on oral communication (e.g., Casteleyn, 2019), we would need to replicate our study with a larger sample, in particular to be able to carry out confirmatory factor analyses (Boateng et al., 2018). What is more, the majority of pupils in our study came from rather advantaged socio-economic backgrounds. Increasing the number of participants could also counter this potential bias. Moreover, it is

possible that how the task is performed, as well as the expected outcomes, may vary from one culture to another (particularly regarding paraverbal aspects). Since our tool was created and tested in Belgium, within a Western context, it would therefore be interesting for future research to examine whether ORAToR remains an appropriate measurement tool when used in other cultural contexts.

In addition, we chose to use a criterion-referenced grid as a measurement tool, in particular because it can address various difficulties specific to oral assessment (Dunbar et al., 2006). However, the entire process of constructing our grid was time-consuming (Martin et al., 2016; Reddy, 2011). Furthermore, it should be noted that, taken individually, the ICCs and weighted kappas for certain criteria indicated that their inter-rater reliability was not always optimal (Koo & Li, 2016; Santos, 2018). This was particularly true for the criteria assessing the relevance of information and paraverbal elements, both of which are particularly influenced by the ephemeral nature of oral communication and the subjective appreciation of the interlocutor (Mercer et al., 2017). Thus, exploring evaluation methods other than the criterion-referenced grid could be a complementary avenue to better understand how to evaluate oral communication competence effectively (Bouwer et al., 2016). For one thing, the use of a criterion-based rubric corresponds to an analytical type of assessment, in which the evaluation is the sum of scores assigned to several criteria. However, other types of assessment also exist, such as holistic assessment, where the evaluator assigns an overall score. It is also possible to design a tool with mixed procedures, in which, for example, certain factors are evaluated using specific criteria while others are assessed holistically (Harsch & Martin, 2013). Another possible strategy would be to use comparative judgment, where the evaluator compares pupils' products against one another (Lesterhuis et al., 2017). This approach appears to provide greater validity and reliability in assessment (Pollitt, 2012). In a related study, some of us compared these three assessment methods for evaluating paraverbal elements in 11- to 12-year-old pupils' oral productions. The results showed that, for this specific dimension of oral communication, the absolute holistic method reached the highest inter-rater reliability, whereas the criterion-based rubric and the comparative holistic method showed moderate reliability (Colognesi et al., 2024). Finally, to complement the evaluation, we could ask the interlocutor questions about what they think of the speaker's performance. Since one of the core functions of communication is ultimately to be understood by one's interlocutor, it seems relevant to take their opinion into account when assessing an oral product.

6.2. Practical implications

Our study addresses several needs relevant to both researchers and teachers. First, our methodological approach can inspire further research in the field of oral communication and in the construction of tools for measuring oral competence (Mercer et al., 2017). In addition, the three factors highlighted by our study provide a more detailed understanding of oral summarization of knowledge. For instance, they could be used to measure pupils' progress following an intervention or teaching sequence focused on knowledge summarization, or even to understand the progressive acquisition of these factors as a function of pupils' age, which seems relevant for this still little-known genre. In classroom settings, these factors can serve as the basis for targeted instructional activities related to oral knowledge summarization, a practice frequently used, but rarely taught explicitly (Colognesi et al., 2022). They also constitute potential guideposts for teachers in the teaching and assessment of speaking, areas that remain insufficiently supported (Gagnon & Colognesi, 2025; Jones, 2017; Mercer et al., 2017; Oliver et al., 2005). Oral skills are still seldom taught or assessed, and teachers often attribute this to uncertainty about what should be taught or evaluated (Jones, 2017). The identification of three specific factors could therefore help teachers determine what can be meaningfully addressed in class. Each factor can be taught independently within

dedicated learning sequences, and the associated criteria can support teachers in deciding what to focus on. The grid and its underlying factors thus offer a practical framework for designing instructional sequences.

This contribution also responds to another challenge frequently identified by teachers: Assessing oral skills is time-consuming due to the large number of elements to examine (Wiertz et al., 2020, 2022). Breaking down the genre into scientifically validated factors allows for a more meaningful distribution of the workload and may help reduce assessment time, since fewer criteria need to be evaluated simultaneously. For example, teachers could design a sequence focusing on the Clarity factor, and another sequence later in the year (or in a subsequent school year) focusing on the Interaction factor. Although initially developed for research purposes, our tool could therefore also help teachers better identify what can be taught in relation to oral knowledge summarization.

6.3. Conclusion

This article contributes to research on oral language assessment by presenting the development and validation of ORAToR (Oral Recap Assessment Tool with Reliability), a tool designed to assess oral knowledge summarization among primary school pupils. The study combines didactic and psychometric perspectives and proposes a 10-step process for designing and validating a criterion-referenced grid. The exploratory factor analysis identified three factors: Informational Content, Interaction with the Interlocutor, and Clarity of the Message Conveyed. By assessing pupils' oral performance in a peer-to-peer, first-language task, ORAToR offers a way to study oral competence in a situation closer to classroom communication than many standardized assessments.

CRedit authorship contribution statement

Christine Wiertz: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Thibault Coppe:** Validation, Methodology, Conceptualization. **Maël Pire:** Data curation. **Benoît Galand:** Writing – review & editing, Supervision, Conceptualization. **Stéphane Colognesi:** Writing – review & editing, Validation, Supervision, Conceptualization.

Ethics statement

This study was reviewed and approved by the Ethics Committee of the Institut de Recherche en Sciences Psychologiques (UCLouvain), with the approval number Projet2020-01, dated February 17, 2020.

Declaration of the use of AI assisted technologies

During the preparation of this work, the authors used Microsoft Copilot and ChatGPT (OpenAI) to support language editing, translation checking, and consistency checks. The authors reviewed and edited the output as needed and take full responsibility for the content of the submitted manuscript.

Funding statement

No specific funding was received for this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.ssaho.2026.103178>.

References

- Abbaszadeh, Z. (2013). Genre-based approach and second language syllabus design. *Procedia - Social and Behavioral Sciences*, 84, 1879–1884. <https://doi.org/10.1016/j.sbspro.2013.07.052>
- Agan, Ş.İ., & Deniz, S. (2020). A rubric study for assessing paragraph level written texts. *Journal of Education and Training Studies*, 8(1), 43–50. <https://doi.org/10.11114/jets.v8i1.4610>
- Balan, A., & Jönsson, A. (2018). Increased explicitness of assessment criteria: Effects on student motivation and performance. *Frontiers in Education*, 81(3). <https://doi.org/10.3389/feduc.2018.00081>
- Berthiaume, D., David, J., & David, T. (2011). Réduire la subjectivité lors de l'évaluation des apprentissages à l'aide d'une grille critériée : repères théoriques et applications à un enseignement interdisciplinaire [Reducing subjectivity in learning assessment using a criteria-based grid: theoretical benchmarks and applications to interdisciplinary teaching]. *Revue Internationale de Pédagogie de l'Enseignement Supérieur*, 27(2). <http://ripes.revues.org/524>
- Blanc, A.-C., Colognesi, S., & Gagnon, R. (2025). Évaluer la production d'anecdotes sous forme orale et écrite au primaire : construction d'un outil de mesure [Assessing the production of anecdotes in oral and written forms at the elementary level: Development of a measurement tool]. *Mesure et évaluation en éducation*, 48(3), 70–99.
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6, 149. <https://doi.org/10.3389/fpubh.2018.00149>
- Bouwer, R., Koster, M., & Van den Bergh, H. (2016). Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner. Available online: <https://dspace.library.uu.nl/bitstream/1874/338041/1/bouwerkoster>
- Casteleyn, J. (2019). Improving public speaking in secondary education—exploring the potential of an improvisational training. *L1: Educational Studies in Language and Literature*, 19(3). <https://doi.org/10.17239/L1ESL-2019.19.03.04>
- Chartrand, S.-G., Emery-Bruneau, J., & Sénéchal, K. (2015). *Caractéristiques de 50 genres pour développer les compétences langagières en français [Characteristics of 50 genres for developing language skills in French]* (2nd ed.). DIDACTICA, C.E.F.
- Cohen, L., Manion, L., & Morisson, K. (Eds.). (2017). *Research methods in education* (6th ed.). Routledge.
- Colognesi, S., Bonnaert, V., & Wiertz, C. (2024). La grille critériée : toujours une bonne méthode d'évaluation ? Comparaison de différentes méthodes d'évaluation des éléments paraverbaux dans les productions orales d'élèves de 11–12 ans. [The criterion-based rubric: Always a good assessment method? A comparison of different methods for evaluating paraverbal elements in the oral productions of 11–12-year-old students]. *Mesure et évaluation en éducation*, 47(3). <https://doi.org/10.7202/1119632ar>
- Colognesi, S., & Deschepper, C. (2019). Les pratiques déclarées de l'enseignement de l'oral au primaire. Qu'en est-il en Belgique francophone? [Declared practices of oral language teaching in primary school: What is the situation in French-speaking Belgium?]. *Language and Literacy*, 21(1), 1–18.
- Colognesi, S., & Hanin, V. (2020). Quelles pratiques efficaces pour enseigner l'oral ? Expérimentations dans huit classes du primaire et suivi de seize futurs enseignants [What effective practices are there for teaching oral language? Experiments in eight primary school classrooms and follow-up of sixteen future teachers]. *Recherches. Revue de didactique et de pédagogie du français*, 73(1), 35–54.
- Colognesi, S., Moser, V., Deschepper, C., & Hanin, V. (2022). Bonne nouvelle : les enseignants du fondamental estiment qu'il est important d'enseigner l'oral en classe et se sentent compétents pour le faire ! Mais certains ne le font quand même pas... [Good news: Primary school teachers believe it is important to teach oral language in class and feel competent to do so! But some still do not do it...]. *Veredas-Revista de Estudos Linguísticos*, 26(1), 141–169. <https://doi.org/10.34019/1982-2243.2022.v26.37930>
- Colognesi, S., Vassart, C., Blondeau, B., & Coertjens, L. (2020). Formative peer assessment to enhance primary school pupils' oral skills: Comparison of written feedback without discussion or oral feedback during a discussion. *Studies In Educational Evaluation*, 67, Article 100917. <https://doi.org/10.1016/j.stueduc.2020.100917>
- Council of Europe. (2001). *Cadre européen commun de référence pour les langues—Apprendre, Enseigner, Évaluer (CECRL)* [Common European Framework of Reference for Languages – Learning, Teaching, Assessment].
- Coupaud, M., Castéra, J., Cheneval-Armand, H., Brandt-Pomares, P., & Delserieys, A. (2019). Développer un questionnaire pour étudier les conceptions de l'évolution du vivant d'élèves de collège : entre cadres didactique et psychométrique [Developing a questionnaire to study secondary school pupils' conceptions of the evolution of life: between didactic and psychometric frameworks]. *Recherches en didactique des sciences et des technologies*, 20, 27–59. <https://doi.org/10.4000/rdst.2641>
- Dirgeyas, I. W. (2016). Genre-based approach: What and how to teach and to learn writing. *English Language Teaching*, 9(9), 45–51. <https://doi.org/10.5539/elt.v9n9p45>
- Dolz, J., & Gagnon, R. (2008). Le genre du texte, un outil didactique pour développer le langage oral et écrit. *Pratiques*, 137/138, 179–198. <https://doi.org/10.4000/pratiques.1159>
- Dolz, J., Pasquier, A., & Bronckart, J. P. (1993). L'acquisition des discours : Émergence d'une compétence ou apprentissage de capacités langagières diverses ? [Discourse acquisition: Emergence of a skill or learning of various language abilities. *Études de linguistique appliquée*, 92, 23–37.
- Dumais, C. (2016). Proposition d'une typologie des objets d'enseignement/apprentissage de l'oral [Proposal for a typology of teaching/learning objects for oral communication]. *Les dossiers des sciences de l'éducation*, 36, 37–56. <https://doi.org/10.4000/dse.1347>
- Dunbar, N. E., Brooks, C. F., & Kubicka-Miller, T. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education*, 31(2), 115–128. <https://doi.org/10.1007/s10755-006-9012-x>
- Flowerdew, J. (2015). John Swales's approach to pedagogy in genre analysis: A perspective from 25 years on. *Journal of English for Academic Purposes*, 19, 102–112. <https://doi.org/10.1016/j.jeap.2015.02.003>
- Fritz, I., Kita, S., Littlemore, J., & Krott, A. (2019). Information packaging in speech shapes information packaging in gesture: The role of speech planning units in the coordination of speech-gesture production. *Journal of Memory and Language*, 104, 56–69.
- Gagnon, R., Bourdages, R., Frossard, D., & Panchout-Dubois, M. (2018). Conter à l'oral au primaire ; mesure de quelques effets d'un outil d'aide à la structuration du récit : les cartes à raconter [Oral storytelling in primary school: measuring some of the effects of a tool to help structure narratives: story cards]. *ISHS Web of Conferences*, 46. <https://doi.org/10.1051/shsconf/20184607009>
- Gagnon, R., & Colognesi, S. (2025). Quelles recommandations pour l'évaluation de l'oral à l'école et en formation ? Enjeux de professionnalisation dans la formation à l'enseignement [What recommendations for assessing oral communication at school and in teacher education? Professionalization issues in teacher education]. *Phronesis*, 14(1), 17–26. <https://doi.org/10.7202/1116122ar>
- Germain-Rutherford, A. (2021). Action-oriented approaches: Being at the heart of the action. In T. Beaven, & F. Rosell-Aguilar (Eds.), *Innovative language pedagogy report* (pp. 91–96). Research-publishing.net. <https://doi.org/10.14705/rpnet.2021.50.1241>
- Gregg, N., & Mather, N. (2002). School is fun at recess: Informal analyses of written language for student with learning disabilities. *Journal of Learning Disabilities*, 35(1), 7–22. <https://doi.org/10.1177/002221940203500102>
- Grugeon-Allys, B., & Grapin, N. (2015). Validité d'une évaluation externe. Complémentarité des approches didactiques et psychométriques [Validity of external assessment. Complementarity of didactic and psychometric approaches]. In A.-C. Mathé, & E. Mounier (Eds.), *Actes du séminaire national de didactique des mathématiques de l'ARDM* (pp. 13–26). ARDM. <https://hal.science/hal-01317134>
- Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, 20, 281–307. <https://doi.org/10.1080/0969594X.2012.742422>
- Hosson, C., & Orange, C. (2019). Les résultats des recherches en didactique des sciences et des technologies : Quelle validité et à quelles conditions ? [The results of research in science and technology education: How valid are they, and under what conditions? *Recherches en Didactique des Sciences et des Technologies: RDST*, 20, 9–26. <https://doi.org/10.4000/rdst.2626>
- Jones, D. (2017). Talking about talk: Reviewing oracy in English primary education. *Early Child Development and Care*, 187, 498–508. <https://doi.org/10.1080/03004430.2016.1211125>
- Kahn, J. H. (2006). Factor analysis in counseling psychology research, training, and practice: Principles, advances, and applications. *The Counseling Psychologist*, 34(5), 684–718. <https://doi.org/10.1177/0011000006286347>
- Karim, S., & Haq, N. (2014). An assessment of IELTS speaking test. *International Journal of Evaluation and Research in Education*, 3(3), 152–157.
- Kihlstrom, J. F. (2021). Ecological validity and “ecological validity”. *Perspectives on Psychological Science*, 16(2), 466–471. <https://doi.org/10.1177/1745691620966791>
- Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363–394. <https://doi.org/10.1037/0033-295X.85.5.363>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Lesterhuis, M., Verhaver, S., Coertjens, L., Donche, V., & De Maeyer, S. (2017). Comparative judgement as a promising alternative to score competences. In *Innovative practices for higher education assessment and measurement* (pp. 119–138). IGI Global. <https://doi.org/10.4018/978-1-5225-0531-0.ch007>
- Li, J. (2014). Examining genre effects on test takers' summary writing performance. *Assessing Writing*, 22, 75–90. <https://doi.org/10.1016/j.asw.2014.08.003>
- Martin, E., Lefrançois, C., Guichard, A., Tapp, D., & Arsenault, L. (2016). Processus de coconstruction d'une grille critériée pour l'évaluation de productions écrites complexes à l'université [Process of co-constructing a criteria-based grid for assessing complex written work at university]. *Revue Internationale de Pédagogie de l'Enseignement Supérieur*, 32(2). Available online: <http://journals.openedition.org/ripest/1094>
- Massonnié, J., Llauro, A., Sumner, E., & Dockrell, J. (2022). Oral language at school entry: Dimensionality of speaking and listening skills. *Oxford Review of Education*, 48(6), 743–766. <https://doi.org/10.1080/03054985.2021.2013189>
- Melissourgou, M. N., & Frantzi, K. T. (2017). Genre identification based on SFL principles: The representation of text types and genres in English language teaching

- material. *Corpus Pragmatics*, 1, 373–392. <https://doi.org/10.1007/s41701-017-0013-z>
- Mercer, N., Warwick, P., & Warwick, A. (2017). An oracy assessment toolkit: Linking research and development in the assessment of students' spoken language skills at age 11–12. *Learning and Instruction*, 48, 51–60. <https://doi.org/10.1016/j.learninstruc.2016.10.005>
- Millard, W., & Menzies, L. (2016). The state of speaking in our schools. *LKMco*. <https://cfey.org/wp-content/uploads/2016/11/Oracy-Report-Final.pdf>
- Moncayey, C., Deschepper, C., Hanin, V., Van Mosnenck, S., Oliveri, S., & Colognesi, S. (2025). Les croyances des formateurs de futurs enseignants : influences sur leurs pratiques d'enseignement et d'évaluation de l'oral [Teacher educators' beliefs: Influences on their oral language teaching and assessment practices]. *Phronesis*, 14(1), 117–137. <https://doi.org/10.7202/1116127ar>
- Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research and Evaluation*, 7(1), 10. <https://doi.org/10.7275/q7rm-gg74>
- Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing*, 32(1), 39–62. <https://doi.org/10.1177/0265532214538014>
- O'Connell, J. P., Petrill, S. A., Piasta, S. B., & Nelson, R. (2017). Oral language and listening comprehension: Same or different constructs? *Journal of Speech, Language, and Hearing Research*, 60(5), 1273–1284. https://doi.org/10.1044/2017_JSLHR-L-16-0039
- Oliver, R., Haig, Y., & Rochecouste, J. (2005). Communicative competence in oral language assessment. *Language and Education*, 19(3), 212–222. <https://doi.org/10.1080/09500780508668675>
- Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies In Educational Evaluation*, 39(4), 195–203. <https://doi.org/10.1016/j.stueduc.2013.10.005>
- Petlyuchenko, N., & Chernyakova, V. (2019). Features of expressive female speech in the political discourse of Spain and Latin America. *Advanced Education*, 12, 212–219. <https://doi.org/10.20535/2410-8286.155500>
- Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22, 157–170. <https://doi.org/10.1007/s10798-011-9189-x>
- Reddy, M. Y. (2011). Design and development of rubrics to improve assessment outcomes: A pilot study in a master's level business program in India. *Quality Assurance in Education*, 19(1), 84–104. <https://doi.org/10.1108/096848811111107771>
- Richards, J. C., Platt, J., & Platt, H. (1993). *Dictionary of language teaching & applied linguistics*. Longman.
- Rochford, L., & Borchert, P. (2011). Assessing higher level learning: Developing rubrics for case analysis. *The Journal of Education for Business*, 86(5), 258–265. <https://doi.org/10.1080/08832323.2010.512319>
- Santos, F. (2018). Le kappa de Cohen: un outil de mesure de l'accord inter-juges sur des caractères qualitatifs. *Université de Bordeaux. PACEA—De la Préhistoire à l'Actuel: Culture, Environnement et Anthropologie*.
- Stewart, J. G. (2009). *Genre oriented summarization*. Doctoral dissertation. Carnegie Mellon University, Language Technologies Institute, School of Computer Science.
- Stordeur, M. F. (2023). *Les exposés oraux à l'école primaire: émotions des élèves et pratiques des enseignants* [Oral presentations in primary school: pupils' emotions and teachers' practices] (Doctoral dissertation, UCL-Université Catholique de Louvain). DIAL. <http://hdl.handle.net/2078.1/283354>
- Stordeur, M. F., Nils, F., & Colognesi, S. (2022). No, an oral presentation is not just something you prepare at home! Elementary teachers' practices supporting preparation of oral presentations. *L1: Educational Studies in Language and Literature*, 22, 1–29. <https://doi.org/10.21248/l1esll.2022.22.1.417>
- Stordeur, M.-F., Wiertz, C., & Colognesi, S. (2025). S'exposer en exposant : prendre en compte les besoins des élèves, une condition pour permettre la prise de parole [Exposing oneself while presenting: Taking pupils' needs into account as a condition for enabling speaking]. *Tréma*, 64. <https://doi.org/10.4000/14r99>
- Wallonie-Bruxelles, F. (2013). *Socles de compétences. Administration générale de l'enseignement et de la recherche scientifique, Service général du pilotage du système éducatif* [Skills bases. General Administration of Education and Scientific Research, General Department for Steering the Education System]. <http://enseignement.be/index.php?page=24737>
- Wiertz, C., Blondeau, B., Francotte, E., Galand, B., & Colognesi, S. (2022). Utiliser une grille critériée pour évaluer les explications orales de ses pairs : quels fonctionnements et quels effets ? [Using a criterion-based rubric to assess peers' oral explanations: How does it work and what are its effects?]. *e-JIREF*, 8(2), 51–88. <https://doi.org/10.48782/m3kwdh11>
- Wiertz, C., Galand, B., & Colognesi, S. (2025). « Dis-moi tout ce que tu sais sur... » Demander aux élèves du primaire de résumer oralement n'est pas si simple, même avec un appui documentaire [“Tell me everything you know about...”: Asking primary school pupils to summarize orally is not so simple, even with documentary support]. *Phronesis*, 14(1), 158–180. <https://doi.org/10.7202/1116129ar>
- Wiertz, C., Van Mosnenck, S., Galand, B., & Colognesi, S. (2020). Évaluer l'oral quand on est enseignant ou chercheur : points de discussion et prises de décision dans la coconception d'une grille critériée [Assessing oral skills as a teacher or researcher: Points for discussion and decision-making in the co-design of a criterion-based rubric]. *Mesure et évaluation en éducation*, 43(3), 1–37. <https://doi.org/10.7202/1083006ar>
- Wurth, J. G. R., Tigelaar, E. H., Hulshof, H., de Jong, J. C., & Admiraal, W. F. (2022). Teacher and student perceptions of L1-oral language lessons in Dutch secondary education. *L1: Educational Studies in Language and Literature*, 22(1). <https://doi.org/10.21248/l1esll.2022.22.1.376>
- Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79–94. <https://doi.org/10.20982/tqmp.09.2.p079>
- Yu, G. (2009). The shifting sands in the effects of source text summarizability on summary writing. *Assessing Writing*, 14(2), 116–137. <https://doi.org/10.1016/j.asw.2009.04.0>