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[1]

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

A grade criterial: sempre um bom método de avaliação? Comparação de diferentes métodos de avaliação dos elementos paraverbais nas produções orais de alunos de 11 a 12 anos

Stéphane Colognesi
ID ORCID: 0000-0001-5763-5873
UCLouvain

Valentine Boonaert
UCLouvain

Christine Wiertz
ID ORCID: 0000-0001-5719-6612
UCLouvain

1. The French version was published in issue 47(3) 2024: https://doi.org/10.7202/1119632ar

Key words: oral assessment, comparison of methods, paraverbal aspects, criterion-based rubric, inter-rater reliability

This article compares three methods for assessing paraverbal aspects in oral productions: the holistic absolute method (overall score), the analytical absolute method (criterion-based rubric), and the holistic comparative method (Comproved software). The study addresses three key questions: 1) What is the inter-rater reliability of each evaluation method? 2) What is the correlation between these methods? 3) What differences in scoring can be observed among these methods? Each method was used to evaluate oral productions based on paraverbal criteria such as intonation, volume, and pauses. The results reveal that, contrary to expectations, the holistic absolute method demonstrated the highest inter-rater reliability. Although significant correlations were found between the methods, notable discrepancies remained in the grading of the same productions. These findings question the systematic use of criterion-based rubrics and emphasize the need to adapt evaluation methods to specific objectives, particularly for assessing paraverbal aspects of oral productions.

Mots clés : évaluation orale, comparaison des méthodes, aspects paraverbaux, fiabilité inter-évaluateurs, grille critériée

Cet article compare trois méthodes d'évaluation des aspects paraverbaux dans les productions orales : la méthode holistique absolue (note globale), la méthode analytique absolue (grille critériée) et la méthode holistique comparative (logiciel Comproved). Trois questions principales guident l'étude : 1) Quelle est la fiabilité inter-évaluateurs de chaque méthode ? 2) Quelle est la corrélation entre ces méthodes ? et 3) Quels écarts de notation observe-t-on entre elles ? Chaque méthode a été utilisée pour évaluer des productions orales sur des critères paraverbaux tels que l'intonation, le volume et les pauses. Les résultats révèlent que, contrairement aux attentes, la méthode holistique absolue présente la meilleure fiabilité interévaluateurs. Bien que des corrélations significatives existent entre les méthodes, des écarts de notation importants subsistent. Ces résultats remettent en question l'utilisation systématique des grilles critériées et montrent qu'il est crucial d'adapter les méthodes d'évaluation aux objectifs spécifiques, notamment pour les aspects paraverbaux des productions orales.

Author's note: Correspondence regarding this article can be addressed to Stephane.colognesi@uclouvain.be

Palavras-chave: aspectos paraverbais, avaliação oral, comparação de métodos, fiabilidade entre avaliadores, rubrica

Este artigo compara três métodos de avaliação dos aspetos paraverbais nas produções orais: o método holístico absoluto (nota global), o método analítico absoluto (rubrica) e o método holístico comparativo (software Comproved). Três questões principais orientam o estudo: 1) Qual é a fiabilidade inter-avaliadores de cada método? 2) Qual é a correlação entre esses métodos? 3) Que discrepâncias de pontuação se observam entre eles? Cada método foi utilizado para avaliar produções orais com base em critérios paraverbais como a entoação, o volume e as pausas. Os resultados revelam que, contrariamente às expectativas, o método holístico absoluto apresenta a melhor fiabilidade entre avaliadores. Embora existam correlações significativas entre os métodos, subsistem discrepâncias importantes nas pontuações. Estes resultados colocam em causa o uso sistemático de rubricas e mostram que é crucial adaptar os métodos de avaliação aos objetivos específicos, especialmente no que diz respeito aos aspetos paraverbais das produções orais.

Introduction

Oral language skills play a central role in students' language development, whether in interaction, in expressing their thinking, or in academic exchanges (Dobinson & Dockrell, 2021; Dumais, 2016; Kaldahl, 2019). Despite their recognized importance, oral skills are rarely given the same importance as writing skills. This is mainly because they are considered less useful at school (Delcambre, 2011; Lafontaine & Messier, 2009; Sales-Hitier & Dupont, 2025; Sénéchal, 2020), but also due to limited teacher training (Gagnon et al., 2017; Moncarey et al., 2025; Wurth et al., 2022). Oral communication is more complex to teach and assess (Dumais, 2016; Lafontaine & Préfontaine, 2007; Nonnon, 2016) due to difficulties with teaching practices and the available tools (Colognesi et al., 2022; Stordeur et al., 2022).

Indeed, assessing oral skills presents specific challenges. First, the ephemeral nature of oral communication makes it difficult to establish objective and reproducible criteria (Mercer et al., 2017; Wiertz et al., 2020). Unlike writing, where the output is tangible, oral performances disappear as soon as they are produced, making retrospective analysis difficult without appropriate recording tools (Colognesi et al., 2023; Stordeur et al., 2025). Furthermore, oral communication involves several

dimensions simultaneously (verbal, non-verbal, and paraverbal), which makes its assessment more complex than that of other academic skills (Garcia-Debanc, 1999). Assessor subjectivity is another critical factor, as judgement can be influenced by personal bias, order effects, and implicit expectations (Garcia-Debanc, 1999). Finally, there is a general lack of standardized, reliable, and user-friendly assessment tools for teachers (Alrabadi, 2011; Wiertz, 2024).

Assessing paraverbal features of speech, such as intonation, volume, and pauses (Boureux, 2017) is even more challenging. Crucial for comprehension, these components, which shape how a message is conveyed, are often overlooked or misunderstood (Weber, 2021). The paraverbal dimension is both subtle and difficult to isolate, as it is intertwined with other dimensions of oral communication, such as verbal and non-verbal communication (Chabanne, 1999). Furthermore, there is no clear consensus on how to define and measure these elements (Pinard-Prévost, 2009). Even teachers aware of the importance of the paraverbal dimension, often feel ill-equipped due to a lack of suitable tools for capturing and assessing it rigorously (Aouchiche-Ait Yala & Zoubida, 2022).

Among the tools commonly used to assess students' oral performance, criterion-based rubrics are often preferred (Deschepper, 2021; Stordeur et al., 2021). However, several limitations have been identified. First, the criteria are not always interpreted in the same way by assessors, which can lead to score variation (Balan & Jönsson, 2018). Second, a rubric rarely fully captures the complexity and nuances of an authentic oral performance, due to the rigidity of the criteria and measurement scales (Bouwer et al., 2023). Finally, achieving good inter-rater reliability remains a major challenge, particularly for assessing paraverbal elements, where judgement is particularly subjective (Wiertz et al., submitted). Criterion-based rubrics provide detailed analysis of the different aspects of a performance, but they can lack flexibility for paraverbal nuances or even lead to discrepancies between assessors' judgements. It is therefore legitimate to question whether this approach is always the most appropriate for assessing oral skills in primary school, particularly for paraverbal aspects.

Given this context, we decided to test and compare several oral assessment methods to identify the most effective for evaluating paraverbal aspects. We replicated the approach used by Landrieu et al. (2022) who compared three assessment methods for evaluating argumentative texts

among fifth-year secondary school pupils. Holistic methods, based on an overall assessment of performance, are often preferred for simplicity and speed, but there is the risk that certain dimensions of the construct being assessed are overlooked (Landrieu et al., 2022; Ounis, 2017). Conversely, analytical methods, such as criterion-based rubrics, provide a detailed assessment of the various aspects of oral production, but are time-consuming and there is a risk of variability in criteria interpretation (Berthiaume et al., 2011; Reddy, 2011). Finally, comparative methods, where performances are compared and ranked, offer validity and reduce bias, but are still underused in educational contexts (Bramley, 2007; Pollit, 2012) and may raise ethical concerns because they are purely norm-referenced.

This article explores the reliability and relevance of three methods of assessing paraverbal features in oral communication: absolute holistic, absolute analytical, and comparative holistic. These methods were tested in a study conducted with primary school pupils aged 11-12 years in French-speaking Belgium. We compare the three methods to determine which most effectively provides a reliable and meaningful assessment of paraverbal components. This research, which builds on the work of Landrieu et al. (2022) and Wiertz et al. (2020), aims to support teachers in the everyday assessment of oral skills.

Theoretical framework

This section defines the key concepts related to oral communication and the paraverbal dimension, and highlights how these dimensions interact and contribute to communication. It also outlines the main approaches used to assess oral communication in schools, with particular attention to how the paraverbal dimension is taken into account. After defining the paraverbal dimension and its components, we examine three approaches to assessing oral communication: absolute holistic, absolute analytical, and comparative holistic, then compare their respective strengths and limitations, particularly for assessing paraverbal elements.

Oral communication and the paraverbal dimension

Oral communication, defined as a set of listening and speaking acts (Colognesi & Deschepper, 2019), incorporates verbal, non-verbal and paraverbal dimensions. These three dimensions interact to produce coherent and effective discourse (Dumais, 2016). The verbal dimension

refers to message content: words, syntax and language structures, including vocabulary, grammar, coherence and clarity. The verbal dimension is the easiest to observe and measure, as it can be transcribed in written form (Chabanne, 1999). Non-verbal communication refers to the physical elements accompanying oral expression, such as gestures, facial expressions, posture and eye contact. It supports or emphasises the verbal message by conveying emotions or highlighting the importance of certain ideas (Dumais, 2016). These two dimensions are commonly assessed in schools. The paraverbal dimension is more difficult to identify and assess (Pinard-Prévost, 2009) but is of particular interest because it subtly shapes and nuances the verbal message.

The paraverbal dimension refers to the contours of speech (Gagnon & Colognesi, 2021). It encompasses vocal characteristics that are not related to content (words) or gestures. Unlike verbal elements, paraverbal features cannot be captured in writing and can only be analyzed from audio recordings (Chabanne, 1999). These features play a crucial role in how oral communication is understood and perceived, as they modulate and nuance the message (Pinard-Prévost, 2009).

Authors have not yet reached a consensus on the terminology used to describe the constituents of the paraverbal dimension (Pinard-Prévost, 2009), which is in itself potentially one of the obstacles to teaching and assessing it. Some authors treat it as synonymous with prosody (Di Cristo et al., 2004; Pinard-Prévost, 2009), whereas others view it as a broader set of elements that includes prosodic components (Boureux, 2017; Bourhis, 2012; Chabanne, 1999). This article adopts the latter approach, considering the paraverbal dimension to consist of paraverbal elements i.e., stress and pauses, as well as prosodic parameters, i.e., frequency, volume, rate, and intonation. Stress is an important paraverbal feature of speech articulation. In French, two types of stress are distinguished: demarcative stress, placed on the final syllable of a word to mark boundaries between syntactic groups, and emphatic stress, which highlights specific elements of speech and is often linked to emotion and expressiveness (Pinard-Prévost, 2009; Poiré, 2000).

Pauses are moments of silence that help structure speech and give it rhythm (Chabanne, 1999). They play a crucial role in communication by giving the audience time to reflect or assimilate information. Two main categories can be identified: silent pauses and filled pauses (Bourhis, 2014;

Di Cristo, 2013). Silent pauses include simple silences, which must last at least twenty centiseconds to be recognized as such (Bourhis, 2014; Pinard-Prévost, 2009), and breathing pauses, when the speaker inhales or exhales (Di Cristo, 2013). Filled pauses are characterized by sounds or behaviors such as laughter, repetitions, false starts, syllable lengthening, or interjections (Bourhis, 2014; Di Cristo, 2013). They may be perceived as hesitations, but filled pauses in verbal exchanges are interactive elements in their own right and contribute to discourse management (Chabanne, 1999).

The paraverbal dimension also includes prosodic parameters, often described as more concrete because they can be measured and quantified (Di Cristo, 2013). These parameters include fundamental frequency (F0), which relates to voice pitch; intensity, which refers to perceived loudness; and duration (or rate), which refers to articulation speed (Aubergé, 2002; Bourhis, 2012; Chabanne, 1999; Di Cristo, 2013), and intonation (Aguert et al., 2010; Boureux, 2017). Intonation is essential for structuring speech and expressing emotions or intentions (Gaudreau et al., 2011; Verschueren, 1999; Wells et al., 2004). Used effectively, it helps distinguish questions from statements, adds nuance, and makes speech more engaging. By contrast, monotonous intonation can make speech dull and difficult to follow, while excessive modulation can undermine clarity (Dumais, 2016). Mastering intonation is therefore important for producing clear, engaging, and fluent speech. Given the specificity of the paraverbal dimension, with its subtle and intangible nuances, rigorous and reliable assessment in school settings raises challenges.

Methods for assessing oral skills

Assessing a student's performance, i.e., judging its quality (Bélec, 2017), serves two main objectives. The first concerns student learning and development (Bélec, 2017). Teachers use assessment to check whether students have acquired the targeted skills (Dimmitt, 2009), and to provide feedback that helps students improve (Bélec, 2017). This ongoing improvement process is central to skill acquisition, particularly in oral communication, where progress is often iterative and gradual. The second objective is social in nature (Bélec, 2017). Assessment aims to reduce inequality (Alpes, 2012) by helping teachers to ensure that all students benefit from teaching in a fair and meaningful way (Dimmitt, 2009).

However, as noted above, assessing oral skills poses specific challenges compared with other academic skills. The volatility of oral production, which complicates the establishment of objective and standardized criteria (Garcia-Debanc, 1999), and the ephemeral nature of oral communication—combined with variability between pupils—make it difficult for teachers to define clear standards (Mercer et al., 2017). Several methods have been proposed to assess oral skills: absolute holistic, absolute analytical, and comparative holistic methods. Each method has advantages and limitations, particularly for evaluating paraverbal aspects.

The absolute holistic method

In the absolute holistic method, raters make an overall judgement of a student's oral production without detailing its various aspects (Ounis, 2017). Often quick to implement, this approach is based on a general impression of the quality of the performance. It does not, however, allow teachers to provide precise feedback to students (Aouchiche-Ait Yala & Zoubida, 2022). Holistic assessment can also be influenced by personal biases, such as the halo effect, or by the order in which performances are assessed (Landrieu et al., 2022; Schwarz et al., 2008). It is a practical method, but unsuitable for assessing aspects requiring detailed analysis (Metruk, 2018).

The absolute analytical method

The absolute analytical method, by contrast, involves evaluating the different components of a performance separately, often using a rubric (Berthiaume et al., 2011; Metruk, 2018). Each criterion is assessed independently, enabling teachers to provide more precise feedback to students (Metruk, 2018). Although this method is more detailed than the absolute holistic method described above, it is more time-consuming for teachers and requires training to ensure consistent interpretation of the criteria (Khabbazbashi & Galaczi, 2020; Wiertz et al., 2022). Furthermore, this method can lack validity, particularly if the selected criteria do not fully capture the specific features of the performance (Reddy, 2011).

The holistic comparative method

The comparative method is based on comparing outputs, rather than assessing each output in isolation (Ounis, 2017). Performances are ranked according to their relative quality (Bramley, 2007). This method, inspired by Thurstone's law of comparative judgement (1927), aims to limit certain individual rater biases by favoring relative rather than absolute judgement (Landrieu et al., 2022; Pollitt, 2012).

This method does not, however, eliminate all potential biases. The order effect, for example, can influence judgements: the first productions compared may be over- or under-rated depending on the immediate references available (Issaieva & Crahay, 2010). On the other hand, the contrast effect can cause a rater to exaggerate the differences between two successive productions, whereas an independent evaluation might have resulted in a more nuanced judgement (Hadji, 1992).

Furthermore, this method is based on normative assessment, where each piece of work is judged in relation to the others and not against an explicit performance benchmark. As a result, perceived quality depends directly on the overall level of the corpus being assessed, raising issues of fairness when the distribution of performance is either very homogeneous or very heterogeneous.

Finally, the comparative method can be time-consuming and tedious to implement, particularly when comparing a large number of productions (van Daal et al., 2016).

These three methods are summarized in Table 1, which shows the focus of the assessment, the advantages, the limitations, the time investment, and the quality of the feedback provided.

Table 1
Comparative table of the three assessment methods

Criteria	Absolute holistic method	Absolute analytical method	Comparative holistic method
Focus of the assessment	Overall performance assessment	Separate assessment of each criterion	Relative comparison of performance
Advantages	Quick to implement	Detailed feedback, more accurate for identifying strengths and weaknesses	Reduction in evaluator bias
Limitations	No detailed feedback, potential for bias	Requires time and training.	Long and tedious for many productions. Requires multiple evaluators.
Time investment	Low	Very high	High
Quality of feedback	General, not very specific	Very specific, focused on each criterion	Relative feedback, but less specific than the previous method

In the following section, we present the methodology used to compare these three methods for the paraverbal dimension of oral communication.

Methodology

We formulated three specific research questions to precisely compare how the three methods assess the paraverbal aspects of oral communication:

- RQ1: How reliable is each assessment method across assessors?
- RQ2: To what extent do the three methods produce similar results?
- RQ3: What score differences are observed between the three assessment methods?

The methodological choices made to address these questions are described below.

Participants and data collection

The study was conducted with 29 pairs of pupils aged 11-12 years, from three classes in two schools in French-speaking Belgium, with comparable and relatively advantaged socio-economic status. Each pair of pupils consisted of a speaker and a listener, in line with the established protocol. Before the test, parental consent was obtained and the roles of speaker and listener in each pair were assigned at random. Of the 29 speakers, 14 were girls and 15 were boys. Their mean age was 11 years and 6 months (range: 10 years and 6 months to 12 years and 3 months). All pupils spoke French at home: 17 spoke only French, 9 spoke French and one other language, and 3 spoke French and several other languages.

Speakers were asked to orally summarize a topic of their choice, either an animal or an electronic device (for details, see Wiertz et al., 2025). The task was carried out outside the classroom under standardized conditions, in a quiet environment without interruptions. After providing instructions, explaining that pupils would be filmed, and answering any questions, the researcher left the room so the speaker would feel as comfortable as possible. Pupils were told there was no time limit. The speaker gave their oral summary, then simply stood up to stop the video recording. The listener was to inform the researcher when the task was complete.

All oral summaries were recorded for fine-grained capture of vocal nuances. The recordings of three to five minutes were then anonymized (by removing the video, leaving only the audio) to ensure participant confidentiality, and to avoid judgement bias linked to a pupil's identity.

Assessment of the work produced

We used the approach of Landrieu et al. (2022) for all analyses. The three assessment methods were used to evaluate specific paraverbal aspects of oral productions. We focused on pauses, volume and intonation for two reasons: first, these aspects can be most readily and consciously modified and are therefore the easiest to teach; second, no specialized equipment is required (Bourhis, 2014). For example, voice pitch is difficult for students to modify, and measuring the frequency in Hertz requires specific equipment (Candea, 2000). Such equipment is not typically available in classroom settings, which is why these aspects were not examined in this study.

The three assessment methods were implemented by two raters, namely the two researchers involved in this study and in a broader project on oral assessment (see Wiertz, 2024). First, for the holistic method, each rater listened to the 29 oral productions and assigned an overall score on a scale of 1 to 10 based on their general impression of the performance. The explicit scoring guide applied (Aouchiche-Ait Yala & Zoubida, 2022) stated that "… paraverbal elements are essential for judging the quality of an oral explanation. In this assessment, the rater is asked to consider volume, intonation and delivery (including pauses) in order to assign the student a score out of 10."

Three weeks later, the same two raters conducted an absolute analytical assessment of the same 29 productions using a criterion-based rubric. The rubric was developed specifically to assess paraverbal elements for this study, drawing on recommendations from the oral assessment literature (Berthiaume et al., 2011; Metruk, 2018). It comprised three main criteria: (a) volume, i.e., control of vocal loudness during oral production; (b) pauses, i.e., appropriate use of pauses, including silent and filled pauses; and (c) intonation, i.e., the ability to vary intonation to convey nuance. Each criterion was scored on a scale of 1 to 5, with descriptors for each performance level. For example, a score of 1 for intonation corresponded to monotonous delivery with little or no variation, whereas a score of 5 reflected dynamic, expressive intonation that could capture the audience's attention. The full rubric is provided in Appendix A.

Finally, the comparative assessment was conducted using Comproved software, which supports pairwise comparisons of oral productions. The software asks raters to compare two productions at a time and select the better one based on paraverbal elements. This procedure is repeated until each production has been compared several times, producing an overall ranking of performance. For this purpose, we recruited 34 additional raters, professors undertaking research as part of their master's, enrolled in a course on oral language teaching and familiar with issues in oral assessment. Before starting, we explained oral communication and the targeted paraverbal elements (volume, pauses and intonation), and introduced them to Comproved software. Each of them was then asked to complete at least five comparisons, in line with the software's recommendations. In total, 192 comparisons were completed, and 29 productions ranked according to perceived quality.

Data analysis

Once the three assessment procedures were completed, we analyzed the results against our three research questions.

To examine inter-rater reliability for the absolute holistic and absolute analytical methods (RQ1), we computed intraclass correlation coefficients (ICC), a widely used index of agreement among raters (Koo & Li, 2016). The ICC measures the similarity between raters' scores for the same oral production. In this study, ICCs were calculated for each assessment method (absolute holistic, comparative holistic and absolute analytical) to determine which yielded the most consistent ratings. We interpreted ICC values using the following criteria: ICC > 0.75 = good reliability; ICC between 0.60 and 0.75 = moderate reliability; ICC < 0.60 = low reliability (Cicchetti, 1994). ICCs were computed separately for each method to compare results and identify the most reliable approach for assessing paraverbal aspects. For the comparative method, Comproved provides a reliability index, the SSR (Scale Separation Reliability), which is considered satisfactory when it is at least 0.7.

In order to determine the extent to which the absolute holistic, absolute analytical and comparative holistic methods produced similar results (RQ2), and thus whether the results obtained by these different evaluation methods were consistent, we calculated Pearson correlation coefficients. Pearson's correlation measures the strength and direction of the linear relationship between two variables (Cohen, 1988). The correlation coefficients

were interpreted according to the following thresholds: $r > 0.70$: strong correlation; r between 0.50 and 0.70: moderate correlation; $r < 0.50$: weak correlation. This analysis identified whether certain methods were more aligned with each other, hence indicating potential interchangeability or complementarity for assessing paraverbal aspects.

To examine scoring differences across the three assessment methods (RQ3), we analyzed both the frequency and the magnitude of differences for each student. This served to compare whether the methods were more stringent or more lenient than others. To visualize these differences, we produced an alluvial plot. This type of visualization makes it possible to trace how ratings shift across methods and to show how judgements are distributed and vary depending on the method used (Tsai et al., 2022). The alluvial plot highlights convergences and divergences between methods, providing a clearer view of their relationships.

Results

The results are presented below for each of the three research questions.

RQ1: Inter-rater reliability is not identical for each assessment method

The intraclass correlation coefficient (ICC) was used to examine inter-rater reliability, that is, the consistency of raters' judgements across the three methods used to assess paraverbal aspects. As a reminder, reliability is considered good when ICC ≥ 0.75.

For the absolute holistic method (overall score), inter-rater reliability was good (ICC = 0.78). Although this method is often seen as quick and broad, it yielded a relatively reliable assessment of paraverbal aspects in oral productions. This consistency may be explained by the use of a shared scoring guide and common implicit criteria for paraverbal features, despite no detailed rubric. In other words, even though the ratings were based on overall impressions, the judgements were fairly stable across assessors.

The absolute analytical method (criterion-based rubric) showed moderate reliability (ICC = 0.68). Despite the use of a criterion-based rubric, assessors' interpretation of the criteria led to greater variability. This suggests that, although this method provides more detail, it does not necessarily lead to greater consistency in judgements across assessors.

For the comparative holistic method using Comproved, reliability was also moderate (SSR = 0.69). Pairwise comparisons may have reduced the impact of individual biases, but consistency between assessors was lower than for the absolute holistic method. Although the comparative approach structures judgement by systematically comparing productions, there was variability between assessors.

Overall, the absolute holistic method (overall score) showed the highest inter-rater reliability, whereas the comparative holistic and absolute analytical methods showed moderate reliability.

RQ2: The three methods produce consistent results, but with variations

The results of the correlation analyses between the different assessment methods are presented in Table 2.

Table 2
Results of correlation analyses between the different assessment methods

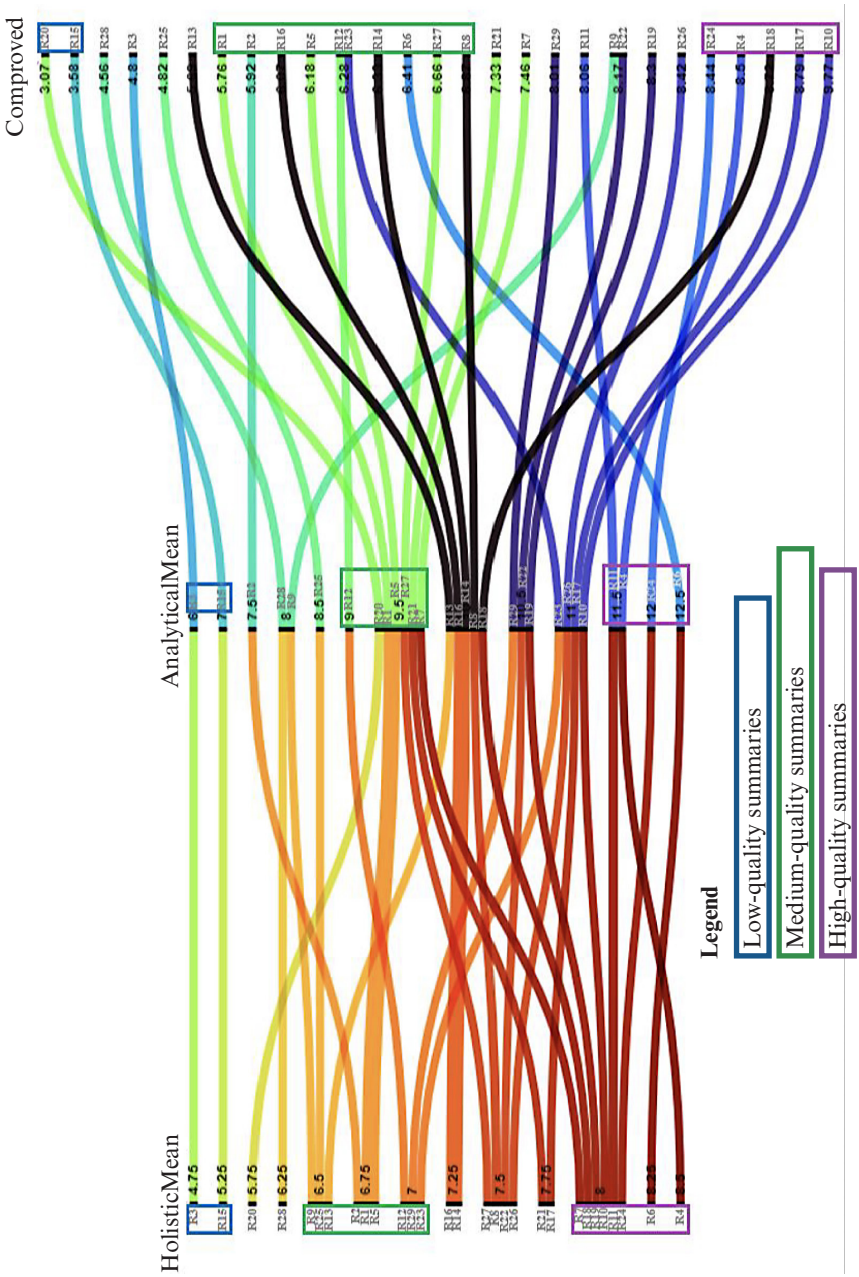
Methods compared	**Correlation (r)**	**Significance (p)**
Absolute holistic method vs. absolute analytical method	0.819	$p < 0.001$
Comparative holistic method vs absolute analytical method	0.61	$p < 0.001$
absolute holistic method vs comparative holistic method	0.585	$p < 0.001$

The strongest correlation (r = 0.819) was found between the absolute holistic method (overall score) and the absolute analytical method (criterion-based rubric). The correlation between the comparative holistic method (Comproved) and the absolute analytical method was moderate (r = 0.610), indicating a degree of consistency between these two methods. A similar moderate correlation was observed between the absolute holistic method and the comparative holistic method (r = 0.585).

Overall, although the strength of the relationships varies across methods, the findings suggest general agreement in the assessment of paraverbal aspects of oral productions.

Figure 1

Alluvial plot showing differences in production quality ratings across assessment methods.



RQ3: Scoring differences reveal significant divergences between methods

Analysis of scoring differences across the three assessment methods revealed marked variations in both their frequency and magnitude. These differences highlight divergences in how the methods judge oral productions.

We used an alluvial plot (Figure 1) to show these differences, tracking similarities and discrepancies across methods. Each colored flow represents one oral production, identified by a number. The summaries are ranked consistently from lowest (top) to highest (bottom). The figure shows the results of the absolute holistic method (overall score) on the left, the absolute analytical method (criterion-based rubric) in the center, and the comparative holistic method (Comproved) on the right.

To facilitate interpretation, we divided the summaries into five 20% intervals, following the approach proposed by Tsai et al. (2022). Productions falling between 0% and 20% were classified as low quality (blue box); between 40% and 60% as medium quality (green box); and between 80% and 100% as high quality (purple box).

Productions are visualized in these intervals, showing differences in ratings across methods. Table 3 summarizes this distribution. Productions consistently classified in the same category by all three methods are shown in bold.

Table 3
Distribution of productions across assessment methods

	Absolute holistic method (overall score)	**Absolute analytical method (criterion-based rubric)**	**Comparative method (Comproved)**
0-20%	3, **15**	3, **15**	**15**, 20
40-60%	**1**, 2, **5**, 9, **12**, 13, 23, 25, 29	1, **5**, 7, **12**, 20, 21, 27	**1**, 2, **5**, 6, 8, **12**, 14, 16, 23, 27
80-100%	**4**, 6, 7, 10, 11, 18, 19, **24**	**4**, 6, 11, **24**	**4**, 10, 17, 18, **24**

Certain similarities emerge. For example, summary 15 (R15) is rated as low quality across all three methods, while summaries 4 (R4) and 24 (R24) are consistently rated as high quality. Nevertheless, some productions receive high scores with one method and substantially lower scores with another. For instance, summary 6 (R6) falls in the 80–100% range for the absolute holistic and absolute analytical methods, but in the 40–60% range for the comparative holistic method. Similarly, summary 20 (R20) falls in the 40–60% range for the absolute analytical method, but in the 0–20% range for the comparative holistic method. More generally, some productions are judged successful by one method and insufficient by another.

Discussion

The aim of this study was to compare the reliability and validity of three methods for assessing paraverbal aspects in the oral productions of pupils aged 11-12 years: the absolute holistic method, the absolute analytical method, and the comparative holistic method. More specifically, we examined the inter-rater reliability of each method and whether the three methods yielded similar results, as well as any differences in scoring across methods.

Firstly, our results indicate that the absolute holistic method achieved the highest inter-rater reliability (ICC = 0.78). This finding is surprising, as the literature often emphasizes that this method is more susceptible to subjective bias (Bouwer & Koster, 2016; Metruk, 2018; Pollitt, 2012). However, the use of a shared scoring guide in our study probably reduced variability in assessor judgement, resulting in more consistent ratings. As Berthiaume et al. (2011) have shown, providing explicit guidelines can clarify performance expectations and reduce the influence of subjective bias.

In contrast, the other two methods showed moderate reliability. For the absolute analytical method, this finding is at odds with previous research (Landrieu et al., 2022). Difficulty assessing paraverbal elements using criterion-based rubrics is a possible explanation. As Garcia-Debanc (1999) notes, it is difficult to objectively assess paraverbal aspects of oral communication. This may explain why there was greater variability in assessor judgement for more subjective features, such as intonation (Deschepper, 2021).

Although the holistic comparative method (Comproved) is expected to be more reliable according to Thurstone's law of comparative judgement (1927), the result may reflect the inherent normative reference effect in this

approach. A limited number of evaluators reduced the number of possible comparisons, which may also have affected judgement stability. Moreover, the perceived quality of a production depends on the other productions being compared, potentially causing variations in the rankings. It would be pertinent for future studies to examine whether increasing the number of pairwise comparisons improves reliability.

Secondly, regarding consistency of results across the three methods, our findings show significant positive correlations, suggesting that productions rated as stronger or weaker are broadly identified as such regardless of the method used. Such overall consistency is reassuring, as it indicates that high-quality performances tend to stand out regardless of the assessment tool. It also supports the idea that assessment can fulfil its educational function by reliably distinguishing between students who are struggling and those who are not (Dimmitt, 2009).

The strongest correlation was observed between the two absolute methods ($r = 0.819$). This may be explained by a shared assessment logic based on absolute scores, whereby productions are judged individually rather than through direct comparison (Landrieu et al., 2022). However, this consistency may also reflect common sources of bias, such as order and rater effects (Barkaoui, 2010; Metruk, 2018). The more moderate correlations observed between the holistic comparative method and the other two methods ($r = 0.610$ and $r = 0.585$) suggest that there are discrepancies in how productions are judged and may point to a normative reference effect. Nevertheless, a degree of consistency is still evident, suggesting that the comparative approach also reliably identifies the least and most successful productions.

Thirdly, despite the significant correlations observed, variations emerged in the perceived quality of certain oral productions depending on the assessment method used. Some productions rated positively using one method were considered weaker with another method, which potentially undermines key functions of assessment, particularly giving consistent and constructive feedback to students (Bélec, 2017; Dimmitt, 2009).

These discrepancies highlight the importance of selecting a pertinent assessment method for the purpose. However, our results tend to support the use of a holistic overall assessment for the paraverbal dimension. As Barkaoui (2010) notes, this approach is particularly useful when time is limited. Thus, if the objective is to provide a quick overall judgement, the holistic method

may be appropriate. It is especially relevant for intermediate formative assessments, where brief feedback is sufficient to guide students' progress. However, if the objective is to provide students with detailed, more nuanced and precise feedback on each aspect of the paraverbal dimension (Aouchiche-Ait Yala & Zoubida, 2022), an analytical assessment using a criterion-based rubric is needed. This approach is better suited to summative assessments or to situations where the focus is on the fine-grained development of skills. That said, appropriate implementation for oral assessment requires more time and expertise. Alternating between these methods is therefore to be considered: a holistic assessment at the beginning of the sequence to position students, followed by a more detailed analytical assessment when paraverbal skills are a central focus of oral work.

On the other hand, the use of the holistic comparative method via the Comproved tool seems less relevant for paraverbal assessment. Although this approach may yield more reliable results for certain types of assessment, it does not appear to add significant value for evaluating oral productions with varying degrees of paraverbal complexity, especially due to the challenges of its implementation in school settings (Landrieu et al., 2022).

Practical implications

Our findings promote flexibility when choosing assessment tools and recommend against the systematic use of a single method. Based on our results, the message is clear: Teachers should select different assessment methods depending on their objectives. If the aim is to obtain a quick, overall evaluation of an oral production, the absolute holistic method is recommended for assessing the paraverbal dimension, given its simplicity, speed of implementation, and relatively good reliability. This approach is particularly useful when time is limited, and a broad formative assessment is sufficient. However, criterion-based rubrics are not to be rejected. Their targeted and purposeful use is recommended depending on assessment needs. In fact, relying on more extensive comparison mechanisms does not appear to be a relevant option for paraverbal assessment, as it is based on relative judgement. This makes a directly individualized assessment of each student's skills less likely, as judgements are constructed through comparison with other productions rather than against fixed criteria.

Secondly, although the absolute analytical method provides more detail, it may show inter-rater variability because criteria can be interpreted subjectively. This clearly points to an area of focus for teacher training. Training

is required in the use of such tools and student support for peer assessment—which often raises questions (Vassart et al., 2022)—but also, and above all, about how to communicate results. Indeed, while criterion-based rubrics help teachers identify a student's specific strengths and weaknesses (Berthiaume et al., 2011; Deschepper, 2021), difficulties have been reported in using and communicating rubric-based information for effective feedback (Colognesi et al., 2024) that supports progress in oral learning. It therefore seems important to strengthen teacher training in this area, and to encourage reflection on the alignment between the assessment method and its purpose, be it formative or summative. The practical conditions of implementation must also be taken into account: For example, with or without student involvement; interactive or individual; immediate or delayed feedback, etc. It appears essential that more support is required for these issues of assessment in teacher training (Barbier & Colognesi, 2024).

Limitations of the study

Like any study, this research has several limitations worth noting. First, the assessments were based on a sample of 29 audio recordings, which is relatively small. Second, a higher number of evaluators using Comproved would have generated more pairwise comparisons, strengthening the robustness of the resulting rankings. Greater diversity among evaluators might also have reduced the influence of individual biases and improved judgement stability. Third, the difficulty of focusing exclusively on paraverbal aspects—at the expense of the content of the summaries—may have introduced bias into the evaluations. Fourth, potential bias due to the researchers' involvement should be considered, since the same two individuals conducted both the absolute holistic and the absolute analytical assessments. Although there was a three-week interval between the two sessions, the initial holistic ratings may have influenced subsequent analytical judgements. Finally, the study was not conducted in an authentic classroom setting. Although the findings are intended to inform oral assessment in schools, they were obtained under controlled experimental conditions. Further research should examine how these methods function in real-world contexts, particularly given the time and organizational constraints faced by teachers and pupils.

Conclusion

Our study challenges the systematic reliance on a single assessment method such as criterion-based rubrics, which are widely recommended in the literature (Metruk, 2018). Our findings show that the absolute holistic method (an overall score) is more reliable, while also being less time-consuming and less demanding to implement than criterion-based rubrics (Barkaoui, 2010). This method does not, however, provide students with precise feedback (Aouchiche-Ait Yala & Zoubida, 2022). This highlights the need to select assessment methods aligned with the purpose: a global approach for rapid assessment, and a more detailed approach when specific feedback is required.

Proofreading: Caroline Lefour

Formatting: Emmanuel Gagnon

Portuguese abstract: Eusebio Andre Machado

LIST OF REFERENCES

Aguert, M., Laval, V. & Bernicot, J. (2010). Understanding the speaker's communicative intent: a study of the role of intonation and context in children aged 5 to 9. *L'Année psychologique*, 110, 49-70. https://doi.org/10.3917/anpsy.101.0049

Alpes, Y. (2012). About PISA: why and for whom should we assess and compare students' skills? *Questions Vives. Educational Research*, *6*(16), 11-14. https://doi.org/10.4000/questionsvives.892

Alrabadi, E. (2011). What method should be adopted for teaching/learning oral communication? *Didáctica. Lengua y Literatura*, *23*, 15-34. https://doi.org/10.5209/rev_DIDA.2011.v23.36308

Aouchiche-Ait Yala, O. & Zoubida, B. (2022). The challenges of assessing oral communication. *Pratiques & didactique*, *1*(1), 72-86. https://www.asjp.cerist.dz/en/downArticle/764/1/1/177627

Aubergé, V. (2002). Prosody and emotion. *Proceedings of the second national conference of the GdR*, *I3*. https://www.researchgate.net/publication/228760016_Prosodie_et_emotion

Balan, A., & Jönsson, A. (2018). Increased explicitness of assessment criteria: Effects on student motivation and performance. Frontiers in Education, 3, 81. https://doi.org/10.3389/feduc.2018.00081

Barbier, É. & Colognesi, S. (2024). Do the practices recommended in teacher training influence the lesson plans of future French teachers? *Canadian Journal of Education*, *47*(1), 113-148.

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modelling approach. *Language Testing*, *27*(4), 515-535. https://doi.org/10.1177/0265532210368717

Bélec, C. (2017). Why assess? *College teaching*, *30*(4), 10–16. https://eduq.info/xmlui/bitstream/handle/11515/35711/belec-30-4-2017.pdf?sequence=2&isAllowed=y

Berthiaume, D., David, J., & David, T. (2011). Reducing subjectivity in learning assessment using a criterion-based rubric: theoretical benchmarks and applications to interdisciplinary teaching. *Revue internationale de pédagogie de l'enseignement supérieur*, *27*(2). http://ripes.revues.org/524

Boureux, M. (2017). Better perception for better pronunciation. Which corrective phonetics to help Italian learners speak French well. *Rivista Interculturale, Università di Lecce*, 43-68. http://magali.boureux.com/IMG/pdf/2017-03-11_actesrome2016boureux.pdf

Bourhis, V. (2012). Reading situations in nursery school: the role of paraverbal communication. *Le Français aujourd'hui*, (4), 85-97. https://doi.org/10.3917/lfa.179.0085

Bourhis, V. (2014). Teacher's voice, student's voice: interlocutory dialogism. Éla. Études de linguistique appliquée, 173, 73-85. https://doi.org/10.3917/ela.173.0073.

Bouwer, R. & Koster, M. (2016). *Bringing writing research into the classroom. The effectiveness of Tekster, a newly developed writing programme for primary school pupils*. [Doctoral thesis, Utrecht University]. Utrecht University Repository https://dspace.library.uu.nl/handle/1874/338041

Bouwer, R., Koster, M. & van den Bergh, H. (2023). Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner. *Assessment in Education: Principles, Policy & Practice*, *30*(3–4), 302–319. https://doi.org/10.1080/0969594X.2023.2241656

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), T*echniques for monitoring the comparability of examination standards* (pp. 246–300). QCA.

Candea, M. (2000). *Contribution to the study of silent pauses and phenomena known as "hesitations" in spontaneous spoken French. Study based on a corpus of narratives in French classes* [Doctoral thesis, Université de la Sorbonne nouvelle-Paris III]. HAL. https://theses.hal.science/tel-00290143v1

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardised assessment instruments in psychology. Psychological Assessment, 6(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Chabanne, J. C. (1999). Verbal, paraverbal, and nonverbal aspects of humorous verbal interaction. In J. M. Defays & L. Rosier (Eds.), *Approaches to comic discourse* (pp. 35–53). Mardaga. https://hal.science/hal-00921934

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Colognesi, S., Coppe, T., Leroux, L., & Wiertz, C. (2024). Does pedagogical metamorphosis exist? Exploring the practices of primary school teachers at different stages of their careers. *British Educational Research Journal*, *50*, 2062–2090. https://doi.org/10.1002/berj.4014

Colognesi, S., Coppe, T. & Lucchini, S. (2023). Improving the oral language skills of primary school pupils through video-recorded performances. *Teaching and Teacher Education*, *128*, 104141.

Colognesi, S. & Deschepper, C. (2019). Declared practices of oral language teaching in primary schools. What is the situation in French-speaking Belgium? *Language and Literacy*, *21*(1), 1-18.

Colognesi, S., Moser, V., Deschepper, C. & Hanin, V. (2022). Good news: primary school teachers believe that it is important to teach oral skills in the classroom and feel competent to do so! But some still don't do it... *Veredas-Revista de Estudos Linguísticos*, *26*(1), 141-169

Delcambre, I. (2011). How to think about oral/written relationships in a school setting. *Recherches*, *54*(1), 7-15.

Deschepper, C. (2021). How and why should we question oral assessment rubrics? Description of an initial training programme and prospects for research. *Évaluer. International Journal of Research in Education and Training*, *7*(2), 61-78. https://doi.org/10.48782/e-jiref-7-2-61

Di Cristo, A. (2013). *The prosody of speech*. De Boeck Supérieur.

Di Cristo, A., Auran, C., Bertrand, R., Chanet, C., Portes, C. & Régnier, A. (2004). Prosodic tools and discourse analysis. *Cahiers de l'Institut de Linguistique de Louvain*, *30*, 27-84. https://hal.science/hal-00349856

Dimmitt, C. (2009). Why evaluation matters: Determining effective school counselling practices. *Professional School Counselling*, *12*(6), 395-399. https://journals.sagepub.com/doi/pdf/10.1177/2156759X0901200605

Dobinson, K. L. & Dockrell, J. E. (2021). Universal strategies for the improvement of expressive language skills in the primary classroom: A systematic review. *First Language*, *41*(5), 527–554. https://doi.org/10.1177/0142723721989471

Dumais, C. (2016). Proposal for a typology of oral teaching/learning objects. *Les dossiers des sciences de l'éducation*, *36*, 37-56. https://doi.org/10.4000/dse.1347

Gagnon, R. & Colognesi, S. (2021). Editorial: Assessing oral performance without distorting it? *Évaluer. International Journal of Research in Education and Training*, *7*(2), 1-5. https://doi.org/10.48782/e-jiref-7-2-1

Gagnon, R., de Pietro, J.-F. & Fisher, C. (2017). Introduction. In J.-F. de Pietro, C. Fisher and R. Gagnon (Eds.), *Oral communication today: didactic perspectives* (pp. 11-40). Namur University Press.

Garcia-Debanc, C. (1999). Evaluating oral skills. *Pratiques*, *103-104*, 193-212.

Gaudreau, G., Hudon, C., & Monetta, L. (2011). Psycholinguistic and neuroanatomical bases of irony comprehension in adults. *Revue de neuropsychologie*, *3*, 148–154. https://doi.org/10.3917/rne.033.0148

Hadji, C. (1992). The evaluation of educational actions. Presses Universitaires de France. https://doi.org/10.3917/puf.hadji.1992.01.

Issaieva, É. & Crahay, M. (2010). Conceptions of school assessment by pupils and teachers: validation of scales and study of their relationships. Measurement and Evaluation in Education, 33(1), 31–61. https://doi.org/10.7202/1024925ar

Kaldahl, A.-G. (2019). Assessing oracy: Chasing the teachers' unspoken oracy construct across disciplines in the landscape between policy and freedom. *L1-Educational Studies in Language and Literature*, *19*, 1-24. https://doi.org/10.17239/L1ESLL-2019.19.03.02

Khabbazbashi, N., & Galaczi, E. D. (2020). A comparison of holistic, analytic, and part marking models in speaking assessment. *Language testing*, *37*(3), 333-360. https://doi.org/10.1177/0265532219898635

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

Lafontaine, L. & Préfontaine, C. (2007). Descriptive teaching model for oral production in secondary school French as a first language classes. *Revue des sciences de l'éducation*, *33*(1), 47–66. https://doi.org/10.7202/016188ar

Lafontaine, L. and Messier, G. (2009). Representations of oral teaching and assessment among secondary school teachers and students of French as a language of instruction. Revue du Nouvel-Ontario, 34, 119-144.

Landrieu, Y., De Smedt, F., Van Keer, H. & De Wever, B. (2022). Assessing the quality of argumentative texts: Examining the general agreement between different rating procedures and exploring inferences of (dis) agreement cases. *Frontiers in Education*, *7*. https://doi.org/10.3389/feduc.2022.784261

Lavoie, C. & Bouchard, É. (2017). University training in oral assessment: a look at the self-assessment skills of future teachers. In J.-F., De Pietro, C. Fisher & R. Gagnon (Eds.), *Oral communication today: teaching perspectives* (pp. 259-274). Namur University Press.

Mercer, N., Warwick, P. & Ahmed, 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*(1), 51-60. https://doi.org/10.1016/j.learninstruc.2016.10.005

Metruk, R. (2018). Comparing holistic and analytic ways of scoring in the assessment of speaking skills. *Journal of Teaching English for Specific and Academic Purposes*, *6*(1), 179–189. https://doi.org/10.22190/JTESAP1801179M

Moncarey, C., Deschepper, C., Hanin, V., Van Mosnenck, S., Oliveri, S. & Colognesi, S. (2025). The beliefs of future teacher trainers: influences on their teaching practices and oral assessment. *Phronesis*, *14*(1), 117–137. https://doi.org/10.7202/1116127ar

Nonnon, É. (2016). Forty years of discourse on oral teaching: didactics confronts its questions. *Pratiques*, 169-170. https://doi.org/10.4000/pratiques.3115

Ounis, M. (2017). A comparison between holistic and analytic assessment of speaking. *Journal of Language Teaching and Research*, *8*(4), 679. http://dx.doi.org/10.17507/jltr.0804.06

Pinard-Prévost, G. (2009). A terminological consensus in prosody? *Proceedings of the XXIIIrdLinguistics Days (JDL)*, 5-6. 77

Poiré, F. (2000). The focal accent and the stress accent in the description of French intonation. *Canadian Journal of Linguistics/Revue Canadienne De Linguistique*, *45*(3-4), 275-302. https://doi.org/10.1017/S0008413100017710

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. Assessment in *Education: Principles, Policy & Practice*, *19*, 281–300. http://dx.doi.org/10.1080/0969594X.2012.665354

Reddy, M. Y. (2011). Design and development of rubrics to improve assessment outcomes: A pilot study in a Master's level business programme in India. *Quality assurance in education*, *19*(1), 84-104. DOI 10.1108/09684881111107771

Sales-Hitier, D. & Dupont, P. (2025). Assessment to support oral teaching and learning: the SEMO system. *Phronesis*, *14*(1), 71–94. https://doi.org/10.7202/1116125ar

Schwarz, N., Knäuper, B., Oyserman, D. & Stich, C. (2008). The psychology of asking questions. *International handbook of survey methodology*, 18-34.

Sénéchal, K. (2020). Rethinking the teaching sequence model for teaching oral skills in primary school: results from the first year of research. *Recherches*, *73*, 75-92.

Stordeur, M. F., Nils, F. & Colognesi, S. (2021). Seven dilemmas encountered by primary school teachers when assessing pupils' oral presentations. *e-JIREF*, *7*(2), 7-37.

Stordeur, M. F., Nils, F. & Colognesi, S. (2022). No, an oral presentation is not just something you prepare at home! Primary school teachers' practices supporting the preparation of oral presentations. *L1-Educational Studies in Language and Literature*, 22, 1-29.

Stordeur, M.-F., Nils, F., Francotte, È. & Colognesi, S. (2025). The challenge of using self-confrontation to support primary school pupils in producing oral presentations. *Phronesis*, *14*(1), 215–241. https://doi.org/10.7202/1116132ar

Tsai, Y-C., Chien, T-W., Wu, J-W., & Lin, C-H. (2022). Using the Alluvial plot to visualise the network characteristics of 100 top-cited articles on attention-deficit/hyperactivity disorder (ADHD) since 2011: bibliometric analysis. *Medicine*, *101*(37), 1–11. http://dx.doi.org/10.1097/MD.0000000000030545

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological review*, *34*(4), 273. https://doi.org/10.1037/h0070288

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V. & De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in Education: Principles, Policy & Practice*, *26*(1), 59–74. https://doi.org/10.1080/0969594X.2016.1253542

Vassart, C., Blondeau, B. & Colognesi, S. (2022). Behind the scenes of peer assessment of oral skills in primary school. *Education and Francophonie*, *50*(1).

Verschueren, J. (1999). *Understanding pragmatics*. Edward Arnold.

Weber, C. (2021). Oral, assessment and reflexivity. Towards an integrative approach to oral skills. *Évaluer. International Journal of Research in Education and Training*, *7*(2), 79-94. https://doi.org/10.48782/e-jiref-7-2-79

Wells, B., Peppé, S. & Goulandris, N. (2004). Intonation development from five to thirteen. *Journal of Child Language*, *31*(4), 749-778. https://doi.org/10.1017/S030500090400652X

Wiertz, C. (2024). *Oral information summarisation: an empirical approach to its characterisation through the development of a measurement tool* [Unpublished doctoral thesis, Catholic University of Louvain].

Wiertz, C., Coppe, T., Galand, B. & Colognesi, S. (submitted). Bridging the Gap in Oral Language Assessment: ORAToR, a Comprehensive Tool for Measuring Oral Summarisation Competence.

Wiertz, C., Blondeau, B., Francotte, E., Galand, B. & Colognesi, S. (2022). Using a criterion-based rubric to evaluate 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). 'Tell me everything you know about...': asking primary school pupils to summarise 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). Assessing oral skills as a teacher or researcher: points for discussion and decision-making in the co-design of a criterion-based rubric. *Measurement and Evaluation in Education*, *43*(3), 1-37.

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*, *20*, 1–27. https://doi.org/10.21248/l1esll.2022.22.1.376

Appendix A

Criteria	1	2	3	4	5
Volume	Most of the speech is inaudible: more than 4 words/phrases are inaudible **AND** the volume of a large part of the explanation is very low.	Much of the speech is inaudible: 2, 3 or 4 words/phrases are inaudible **AND** the volume of much of the explanation is very low.	Part of the speech is inaudible: more than 2 words/expressions are inaudible **OR** the volume of a large part of the explanation is very low.	One or two words/ expressions are inaudible.	All of the speech is audible and the volume is adequate.
Pauses	No pauses are made, preventing the listener from processing the information as it is delivered, **AND** pauses are too long, too frequent, or occur at inappropriate moments, creating breaks in the flow of speech and hindering understanding.	No pauses are made, preventing the listener from processing the information as it is delivered, **OR** pauses are too long, too frequent, or poorly placed, creating breaks in the flow of speech and hindering understanding.	The speaker does pause, but pauses are too few and/ or too short for the information to be fully processed, **AND** speech is generally continuous, but a few pauses that are too long, too frequent, or poorly placed create discontinuity in certain parts of the explanation.	The speaker makes some pauses, but they are too few and/or too short for optimal processing of the information, **OR** speech is generally continuous, but a few pauses that are too long, too frequent, or poorly placed create discontinuity in certain passages.	Pauses are made, allowing the listener to process the information conveyed by the speaker, AND pauses do not create any breaks in the flow of speech.

Appendix A (suite)

Criteria	1	2	3	4	5
Intonation	The student hardly modulates their voice, which makes their speech monotonous. It is very difficult to maintain attention throughout the speech and to identify the important elements of the speech.	Between 1 and 3.	The student is neither monotonous nor particularly eloquent. They modulate their voice moderately.	Between 3 and 5.	The student modulates their voice a great deal and is eloquent. It is very easy to maintain attention throughout the speech and identify the important elements of the speech.