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

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

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 inter-évaluateurs. 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.

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.

Note des auteurs : la correspondance liée à cet article peut être adressée à 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 intra-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

L'enseignement de l'oral joue un rôle central dans la construction des compétences langagières des élèves, que ce soit pour interagir, pour exprimer sa pensée ou pour s'engager dans des échanges scolaires (Dobinson & Dockrell, 2021 ; Dumais, 2016 ; Kaldahl, 2019). Cependant, malgré son importance reconnue, l'oral reste souvent marginalisé au profit de l'enseignement de l'écrit, en raison, notamment, de la perception d'une moindre utilité scolaire (Delcambre, 2011 ; Lafontaine & Messier, 2009 ; Sales-Hitier & Dupont, 2025 ; Sénéchal, 2020), du manque de formation des enseignants (Gagnon et al., 2017 ; Moncarey et al., 2025 ; Wurth et al., 2022) et de ses caractéristiques qui restent encore complexes à saisir (Dumais 2016 ; Lafontaine & Préfontaine, 2007 ; Nonnon, 2016). Ce déséquilibre s'explique également, et surtout, par les nombreuses difficultés rencontrées dans l'enseignement et dans l'évaluation de l'oral, sur le plan tant des pratiques pédagogiques que des outils disponibles (Colognesi et al., 2022 ; Stordeur et al., 2022).

En effet, l'évaluation de l'oral présente des défis spécifiques. En premier lieu, la dimension éphémère et volatile de l'oral complique la fixation de critères objectifs et reproductibles (Mercer et al., 2017 ; Wiertz et al.,

2020). Contrairement à l'écrit, où la production reste tangible, les manifestations orales disparaissent dès qu'elles sont produites, ce qui rend difficile leur analyse *a posteriori*, sans s'équiper de matériel technologique pour le faire (Colognesi et al., 2023 ; Stordeur et al., 2025). De plus, l'oral sollicite simultanément plusieurs dimensions (verbale, non-verbale, paraverbale), ce qui rend son évaluation plus complexe que celle d'autres compétences scolaires (Garcia-Debanc, 1999). La subjectivité des évaluateurs est un autre facteur critique, car les jugements peuvent être influencés par des biais personnels, des effets d'ordre ou encore, des attentes implicites quant aux performances attendues (Garcia-Debanc, 1999). Enfin, il existe un manque général d'outils d'évaluation standardisés, fiables et facilement utilisables pour les enseignants (Alrabadi, 2011 ; Wiertz, 2024).

Les défis se multiplient encore lorsqu'il s'agit d'évaluer les éléments paraverbaux de l'oral, tels que l'intonation, le volume et les pauses (Boureux, 2017). Ces composantes, qui façonnent la manière dont le message est transmis, sont cruciales pour la compréhension de la parole, mais restent souvent négligées ou mal comprises (Weber, 2021). Le paraverbal est à la fois subtil et difficile à isoler, car il se mêle aux autres dimensions de la communication orale, comme le verbal et le non-verbal (Chabanne, 1999). De plus, il n'existe pas de consensus clair sur la manière de définir et de mesurer ces éléments (Pinard-Prévost, 2009). Les enseignants, même lorsqu'ils sont conscients de l'importance du paraverbal, se trouvent souvent démunis face à l'absence d'outils adaptés pour le capturer et pour le juger de manière rigoureuse (Aouchiche-Ait Yala & Zoubida, 2022).

Parmi les outils couramment utilisés pour évaluer les performances orales des élèves, la grille critériée s'impose comme une méthode privilégiée (Deschepper, 2021 ; Stordeur et al., 2021). Cependant, plusieurs limites sont associées à son utilisation. Tout d'abord, les critères ne sont pas toujours interprétés de la même manière par les évaluateurs, ce qui peut entraîner des variations dans les notations (Balan & Jönsson, 2018). Ensuite, il est souvent difficile pour une grille de capturer pleinement la complexité et les nuances d'une performance orale réelle, en raison de la rigidité des critères et des échelons de mesure (Bouwer et al., 2023). Enfin, obtenir une bonne fiabilité interjuges reste un défi majeur, notamment pour l'évaluation des éléments paraverbaux où les jugements sont particulièrement subjectifs (Wiertz et al., soumis). En effet, bien que la grille critériée permette une analyse détaillée des différents aspects d'une performance, elle peut parfois manquer de souplesse face aux nuances

paraverbales, ou encore, entraîner des divergences entre les jugements des évaluateurs. Dès lors, il est légitime de se demander si cette méthode est toujours la plus adaptée dans le cadre de l'évaluation de l'oral à l'école primaire, particulièrement en ce qui concerne les aspects paraverbaux.

Dans cette perspective, il était nécessaire de tester et de comparer plusieurs méthodes d'évaluation de l'oral pour identifier celles qui permettent le mieux de juger les aspects paraverbaux. C'est ce que nous avons fait en répliquant une recherche antérieure menée par Landrieu et al. (2022) dans laquelle ils se sont concentrés sur l'évaluation du texte argumentatif chez des élèves de cinquième secondaire en comparant trois méthodes d'évaluation. Les méthodes holistiques, qui reposent sur une appréciation globale de la performance, sont souvent privilégiées pour leur simplicité et pour leur rapidité, mais elles risquent de passer à côté de certaines dimensions de l'objet évalué (Landrieu et al., 2022 ; Ounis, 2017). À l'inverse, les méthodes analytiques, telles que les grilles critériées, permettent de détailler les différents aspects d'une production orale, mais demandent beaucoup de temps et peuvent entraîner une variabilité dans l'interprétation des critères (Berthiaume et al., 2011 ; Reddy, 2011). Enfin, les méthodes comparatives, qui consistent à classer les performances les unes par rapport aux autres, constituent une option en matière de validité et de réduction des biais, mais elles sont encore peu exploitées en contexte scolaire (Bramley, 2007 ; Pollit, 2012) et pourraient poser des problèmes éthiques de par la référence exclusivement normative.

Cet article se propose d'explorer la fiabilité et la pertinence de ces trois méthodes (holistique absolue, analytique absolue et holistique comparative) dans l'évaluation des éléments paraverbaux de l'oral, en rapportant une étude menée auprès d'élèves de 11-12 ans du primaire en Belgique francophone. En comparant ces méthodes, nous visons à déterminer laquelle est la plus adaptée pour offrir une évaluation fiable et pertinente des composantes paraverbales. Cette recherche s'inscrit dans la lignée des recherches de Landrieu et al. (2022) et Wiertz et al. (2020), et cherche à apporter des réponses aux enseignants confrontés à l'évaluation complexe de l'oral dans leur pratique quotidienne.

Cadre théorique

Dans cette section, nous définissons les principaux concepts liés à l'oral et au paraverbal, en mettant en lumière leurs interactions et leur rôle dans la communication. Nous présentons également les différentes approches d'évaluation de l'oral en milieu scolaire, en nous intéressant plus particulièrement aux méthodes utilisées pour prendre en compte la dimension paraverbale. Après une définition du paraverbal et de ses composantes, nous examinerons trois méthodes d'évaluation de l'oral (holistique absolue, analytique absolue et holistique comparative), en comparant leurs avantages et leurs limites, notamment en ce qui concerne l'évaluation des éléments paraverbaux.

L'objet et le paraverbal

L'oral, défini comme l'ensemble des actes d'écoute et de langage (Colognesi & Deschepper, 2019), intègre une dimension verbale, non-verbale, mais aussi paraverbale. Ces trois dimensions interagissent pour produire un discours complet et efficace (Dumais, 2016). Le verbal correspond au contenu du message, c'est-à-dire les mots, la syntaxe et les structures langagières utilisées. Il inclut le lexique, la grammaire, la cohérence et la clarté du discours. Le verbal est l'élément le plus facilement observable et mesurable, car il peut être transcrit sous forme écrite (Chabanne, 1999). Le non-verbal englobe tous les éléments corporels qui accompagnent l'expression orale, comme les gestes, les expressions faciales, la posture, et le regard. Il permet de soutenir ou d'accentuer le message verbal, en transmettant des émotions ou en soulignant l'importance de certaines idées (Dumais, 2016). Si ces deux dimensions sont couramment évaluées dans le cadre scolaire, le paraverbal, quant à lui, mérite une attention particulière, car il modifie et nuance le message verbal de manière subtile, bien que plus difficile à saisir et à évaluer (Pinard-Prévost, 2009).

Le paraverbal désigne ainsi les contours de la parole (Gagnon & Colognesi, 2021). Il regroupe les caractéristiques vocales du discours qui ne concernent ni le contenu (les mots) ni la gestuelle. Contrairement au verbal, les éléments paraverbaux ne peuvent pas être capturés par écrit et nécessitent un enregistrement audio pour être analysés (Chabanne, 1999). Ces éléments jouent un rôle crucial dans la compréhension et dans la perception de l'oralité, en permettant de moduler et de nuancer le message transmis (Pinard-Prévost, 2009).

Les auteurs n'ont pas encore atteint de consensus terminologique concernant les éléments constitutifs du paraverbal (Pinard-Prévost, 2009), ce qui peut constituer en soi un obstacle supplémentaire à son enseignement et à son évaluation. Certains le considèrent comme un synonyme de la prosodie (Di Cristo et al., 2004 ; Pinard-Prévost, 2009), tandis que d'autres y voient un ensemble plus vaste composé de plusieurs éléments, y compris des composants prosodiques (Boureux, 2017 ; Bourhis, 2012 ; Chabanne, 1999). C'est cette deuxième approche qui sera privilégiée ici, en considérant que le paraverbal est constitué d'éléments paraverbaux (l'accent et les pauses) et des paramètres prosodiques (la fréquence, le volume, le débit et l'intonation).

Dans les éléments paraverbaux, l'accent joue un rôle important dans l'articulation du discours. On distingue deux types d'accent en français : l'accent démarcatif, marqué sur la syllabe finale d'un mot pour délimiter les groupes syntaxiques, et l'accent d'insistance, qui met en relief certains éléments du discours, souvent liés aux émotions et à l'expressivité (Pinard-Prévost, 2009 ; Poiré, 2000).

Les pauses sont des moments d'arrêt dans le discours qui permettent de structurer et de rythmer le message (Chabanne, 1999). Elles jouent un rôle crucial dans la communication en offrant à l'auditoire un temps de réflexion ou d'assimilation. Il est possible de distinguer deux grandes catégories de pauses : les pauses silencieuses et les pauses pleines (Bourhis, 2014 ; Di Cristo, 2013). Les pauses silencieuses incluent les silences eux-mêmes, qui doivent durer au moins vingt centisecondes pour être reconnus comme tels (Bourhis, 2014 ; Pinard-Prévost, 2009), ainsi que les pauses respiratoires, où l'orateur inspire ou expire (Di Cristo, 2013). Les pauses pleines, quant à elles, se caractérisent par des sons ou des comportements particuliers comme les rires, les répétitions, les faux départs, les allongements de syllabes, ou encore, les interjections (Bourhis, 2014 ; Di Cristo, 2013). Ces pauses pleines, bien qu'elles puissent être perçues comme des hésitations, sont des éléments interactifs à part entière dans les échanges verbaux et contribuent à la gestion du discours (Chabanne, 1999).

Le paraverbal comprend également des paramètres prosodiques, souvent définis comme des éléments plus concrets, car ils peuvent être mesurés et quantifiés (Di Cristo, 2013). Ces paramètres incluent la fréquence fondamentale qui se rapporte à la hauteur de la voix, l'intensité, qui désigne le volume perçu de la voix, et la durée ou le débit, qui fait référence à la

vitesse d'articulation des sons (Aubergé, 2002 ; Bourhis, 2012 ; Chabanne, 1999 ; Di Cristo, 2013) et l'intonation (Aguert et al., 2010 ; Boureux, 2017). L'intonation est essentielle pour structurer le discours et exprimer des émotions ou des intentions (Gaudreau et al., 2011 ; Verschueren, 1999 ; Wells et al., 2004). Une bonne intonation permet de différencier les questions des affirmations, d'ajouter des nuances au discours et de rendre le message plus vivant. Une intonation monotone peut rendre le discours ennuyeux et difficile à suivre, tandis qu'une modulation excessive peut perturber la clarté du message (Dumais, 2016). La maîtrise de l'intonation est donc essentielle pour rendre un discours attractif et fluide.

La spécificité du paraverbal, avec ses nuances difficiles à saisir, soulève la question de la manière dont il peut être évalué de façon rigoureuse et fiable en milieu scolaire.

Les méthodes d'évaluation de l'oral

Évaluer la production d'un élève, qui consiste à statuer de la qualité de celle-ci (Bélec, 2017), vise deux objectifs principaux. Le premier porte sur l'apprentissage et sur le développement des élèves (Bélec, 2017). L'évaluation permet en premier lieu de vérifier si l'élève a acquis les compétences visées (Dimmitt, 2009), tout en offrant un *feedback* qui lui donne l'opportunité de se perfectionner (Bélec, 2017). Ce processus d'amélioration continue est central dans l'acquisition des compétences, particulièrement à l'oral, où la progression est souvent itérative et graduelle. Le second objectif concerne un aspect de nature sociale (Bélec, 2017). L'évaluation vise à lutter contre les inégalités (Alpes, 2012), car elle permet à l'enseignant de vérifier que ses enseignements bénéficient à tous les élèves de manière équitable et significative (Dimmitt, 2009).

Toutefois, comme mentionné *supra*, l'évaluation de l'oral présente des défis spécifiques par rapport à d'autres compétences scolaires. La volatilité des productions orales, qui complique l'établissement de critères objectifs et standards (Garcia-Debanc, 1999), et le caractère éphémère de l'oral, combiné à la variabilité interindividuelle des élèves, rendent difficile la définition de normes claires pour les enseignants (Mercer et al., 2017). Plusieurs méthodes ont été proposées pour évaluer l'oral, mais chacune présente des avantages et des limites, en particulier lorsqu'il s'agit d'évaluer les aspects paraverbaux : les méthodes holistique absolue, analytique absolue et holistique comparative.

La méthode holistique absolue

La méthode holistique absolue consiste à porter un jugement global sur la production orale d'un élève sans en détailler les différents aspects (Ounis, 2017). Cette approche est souvent rapide à mettre en œuvre et repose sur une impression générale de la qualité de la performance. Toutefois, elle présente l'inconvénient de ne pas permettre un *feedback* précis aux élèves (Aouchiche-Ait Yala & Zoubida, 2022). En outre, l'évaluation holistique peut être influencée par des biais personnels, comme l'effet de halo ou l'ordre de passation des productions (Landrieu et al., 2022 ; Schwarz et al., 2008). Cette méthode, bien que pratique, peut se révéler inadaptée pour l'évaluation d'aspects qui nécessitent une analyse fine et détaillée (Metruk, 2018).

La méthode analytique absolue

La méthode analytique absolue, quant à elle, consiste à évaluer les différentes composantes d'une production de manière séparée, souvent à l'aide d'une grille (Berthiaume et al., 2011 ; Metruk, 2018). Chaque critère fait l'objet d'une évaluation distincte, ce qui permet de fournir un retour précis aux élèves (Metruk, 2018). Bien que cette méthode soit plus précise que la précédente, elle demande un investissement en temps important de la part de l'enseignant et nécessite une formation pour assurer une interprétation cohérente des critères (Khabbazbashi & Galaczi, 2020 ; Wiertz et al., 2022). Par ailleurs, cette méthode peut parfois manquer de validité, notamment si les critères choisis ne reflètent pas pleinement les spécificités de la production (Reddy, 2011).

La méthode holistique comparative

La méthode comparative repose sur une comparaison des productions entre elles, plutôt que sur une évaluation isolée de chaque production (Ounis, 2017). Elle permet de classer les performances en fonction de leur qualité relative (Bramley, 2007). Cette méthode, inspirée de la loi du jugement comparatif de Thurstone (1927), vise à limiter certains biais individuels liés à l'évaluateur en favorisant un jugement relatif plutôt qu'absolu (Landrieu et al., 2022 ; Pollit, 2012).

Toutefois, elle n'élimine pas tous les biais potentiels. L'effet d'ordre, par exemple, peut influencer les jugements : les premières productions comparées peuvent être surévaluées ou sous-évaluées en fonction des références immédiates disponibles (Issaieva & Crahay, 2010). L'effet de contraste,

quant à lui, peut amener un évaluateur à exagérer les différences entre deux productions successives, alors qu'une évaluation indépendante aurait pu aboutir à un jugement plus nuancé (Hadji, 1992).

Par ailleurs, cette méthode repose sur une évaluation normative, où chaque production est jugée en fonction des autres et non selon un référentiel explicite de performance. Cela signifie que la qualité perçue d'une production dépend directement du niveau global du corpus évalué, ce qui peut poser des problèmes d'équité lorsque la distribution des performances est très homogène ou très hétérogène.

Enfin, la méthode comparative peut s'avérer longue et fastidieuse à mettre en œuvre, en particulier lorsqu'il s'agit de comparer un grand nombre de productions (van Daal et al., 2016).

Ces trois méthodes sont synthétisées dans le tableau 1 incluant le focus de l'évaluation, les avantages, les limitations ainsi que l'investissement en temps et la qualité du *feedback* fourni.

Tableau 1
Tableau comparatif des trois méthodes d'évaluation

Critères	Méthode holistique absolue	Méthode analytique absolue	Méthode holistique comparative
Focus de l'évaluation	Jugement global de la performance	Évaluation de chaque critère de manière séparée	Comparaison relative des performances
Avantages	Rapide à mettre en œuvre	*Feedback* détaillé, plus précis pour déceler les forces et difficultés	Réduction des biais des évaluateurs
Limitations	Pas de *feedback* détaillé, biais possibles	Nécessite du temps et une formation.	Long et fastidieux pour de nombreuses productions. Nécessite plusieurs évaluateurs.
Investissement en temps	Faible	Très élevé	Élevé
Qualité du feedback	Général, peu précis	Très précis, orienté sur chaque critère	*Feedback* relatif, mais moins spécifique que la méthode précédente

Dans les lignes suivantes, nous présentons la méthodologie mobilisée pour comparer ces trois méthodes relativement à la dimension paraverbale de l'oral.

Méthodologie

Pour comparer finement les trois méthodes d'évaluation des aspects paraverbaux de l'oral, nous avons formulé trois questions de recherche spécifiques :

- QR1 : Quelle est la fiabilité inter-évaluateurs de chaque méthode d'évaluation ?
- QR2 : Dans quelle mesure les trois méthodes produisent-elles des résultats similaires ?
- QR3 : Quels écarts de notation peut-on observer entre les trois méthodes d'évaluation ?

Nous détaillons ci-après les choix méthodologiques réalisés pour y répondre.

Les participants et la collecte de données

L'étude a été menée auprès de 29 duos d'élèves de 11-12 ans, provenant de trois classes différentes issues de deux établissements de Belgique francophone à un indice socioéconomique équivalent, relativement favorisés. Chaque duo était composé d'un orateur et d'un interlocuteur, conformément au protocole établi. Avant la passation, nous avons obtenu les autorisations parentales et établi un ordre de passage aléatoire pour déterminer l'orateur et l'interlocuteur dans chaque duo. Sur les 29 orateurs, 14 sont des filles et 15 sont des garçons. Leur moyenne d'âge est de 11 ans et 6 mois (le plus jeune a 10 ans et 6 mois et le plus âgé a 12 ans et 3 mois). Ces élèves parlent le français à la maison. Plus précisément, 17 d'entre eux parlent uniquement le français, 9 parlent le français et une autre langue et 3 parlent le français et plusieurs autres langues

Les orateurs ont été invités à réaliser un résumé d'informations à l'oral sur un sujet de leur choix : soit un animal, soit un appareil électronique (pour des détails, voir Wiertz et al., 2025). La tâche a été réalisée en dehors de la classe, dans des conditions standardisées pour tous les participants, dans un environnement calme et sans interruptions. Après avoir donné toutes les consignes, explicité aux élèves qu'ils seraient filmés et répondu

aux éventuelles questions, la chercheuse quittait la pièce pour permettre à l'orateur d'être le plus à l'aise possible. Il avait été précisé qu'il n'y avait aucune limite de temps imposée. Une fois son résumé terminé, l'orateur devait simplement se lever pour couper l'enregistrement vidéo. L'auditeur, quant à lui, avait pour rôle d'informer la chercheuse de la fin de la tâche.

L'ensemble des résumés oraux a ainsi été enregistré, permettant de capturer avec précision les nuances vocales. Les enregistrements, d'une durée allant de trois à cinq minutes, ont ensuite été anonymisés (l'image a été supprimée, laissant uniquement le son) pour garantir la confidentialité des participants et éviter tout biais de jugement basé sur la connaissance des élèves.

Les évaluations des productions

La démarche utilisée par Landrieu et al. (2022) a été répliquée pour l'ensemble des analyses. Les trois méthodes d'évaluation ont été employées pour évaluer certains des aspects paraverbaux des productions orales. Plus précisément, nous avons choisi d'évaluer les pauses, le volume et l'intonation, d'une part, car ce sont les aspects les plus facilement modifiables consciemment par les élèves, et donc, les plus aisément enseignables et, d'autre part, parce que ce sont les aspects pour lesquels il n'est pas nécessaire d'avoir un matériel spécifique pour les mesurer (Bourhis, 2014). Par exemple, la hauteur de la voix est difficilement modifiable par les élèves et il faut un matériel spécifique pour mesurer le nombre de Hertz (Candea, 2000). Or, ce type de matériel n'est *a priori* pas disponible dans un contexte de classe, c'est pourquoi nous ne nous sommes pas intéressés à ces aspects dans le cadre de cette étude.

Les trois méthodes d'évaluation ont été réalisées par deux évaluatrices, les deux chercheuses impliquées dans ce travail et dans un projet de grande envergure sur l'évaluation de l'oral (voir Wiertz, 2024).

Tout d'abord, chacune des évaluatrices a écouté les 29 productions orales et attribué une note globale sur une échelle de 1 à 10, basée sur une impression générale de la performance. Pour mener cette évaluation holistique, un guide de notation explicite a été utilisé (Aouchiche-Ait Yala & Zoubida, 2022). Il y était indiqué que « les éléments paraverbaux sont essentiels pour juger de la qualité d'une explication orale. Dans le cadre de cette prise de mesure, l'évaluateur est invité à prendre en compte le volume,

l'intonation et le débit (dont les pauses) pour donner une note sur 10 à l'élève ». Ainsi, pour cette méthode, les évaluatrices, en ayant ces aspects en tête, ont dû attribuer une note globale sur 10 à chaque production.

Ensuite, trois semaines plus tard, ces deux mêmes évaluatrices ont réalisé l'évaluation analytique absolue pour ces mêmes 29 productions, mais cette fois, avec une grille d'évaluation critériée. Cette grille a été développée spécifiquement pour l'évaluation des éléments paraverbaux dans le cadre de cette étude, en tenant compte des recommandations de la littérature sur l'évaluation orale (Berthiaume et al., 2011 ; Metruk, 2018). Elle comprenait trois critères principaux : 1) le volume, soit le contrôle du volume de la voix au cours de la production orale ; 2) les pauses, soit l'utilisation appropriée des pauses, incluant les pauses silencieuses et non silencieuses et 3) l'intonation, soit la capacité à varier l'intonation pour exprimer des nuances dans le discours. Chaque critère a été noté sur une échelle de 1 à 5, avec des descriptions précises pour chaque niveau de performance. Par exemple, une note de 1 pour l'intonation correspondait à une intonation monotone et sans variation, tandis qu'une note de 5 représentait une intonation dynamique et expressive, capable de capter l'attention de l'auditoire. La grille complète est disponible en annexe A.

Enfin, l'évaluation comparative a été réalisée à l'aide du logiciel *Comproved*, qui permet une comparaison par paires des productions orales. Ce logiciel demande aux évaluateurs de comparer deux productions à la fois et de choisir celle qu'ils considèrent comme la meilleure, en fonction des éléments paraverbaux. Ce processus est répété jusqu'à ce que chaque production soit comparée plusieurs fois, générant ainsi un classement global des performances. Pour cela, 34 autres évaluateurs, en plus des deux évaluatrices de départ, ont été mobilisés. Il s'agit d'enseignants-chercheurs qui suivaient un cours de didactique de l'oral et sensibilisés à la question de son évaluation. Avant de commencer leur évaluation, ils ont reçu des explications et des précisions sur ce que sont l'oral et les éléments paraverbaux (volume, pauses et intonation) ainsi qu'une introduction au logiciel *Comproved*. Chaque évaluateur a ensuite été invité à effectuer au moins cinq comparaisons, conformément aux recommandations du logiciel. En tout, 192 comparaisons ont été réalisées, permettant ainsi de classer les 29 productions selon leur qualité perçue.

L'analyse des données

Une fois ces différentes méthodes d'évaluation réalisées, des analyses ont été menées en cohérence avec nos trois questions de recherche.

Pour évaluer la fiabilité des jugements entre les différents évaluateurs (QR1) pour la méthode holistique absolue et analytique, nous avons utilisé le coefficient de corrélation intraclasse (ICC), un indicateur couramment utilisé pour mesurer la cohérence des jugements entre plusieurs évaluateurs (Koo & Li, 2016). L'ICC permet d'évaluer à quel point les jugements sont similaires pour une même production orale, indépendamment de la méthode d'évaluation employée. Dans cette étude, l'ICC a été calculé pour chaque méthode d'évaluation (holistique absolue, holistique comparative et analytique absolue) afin de déterminer laquelle offre la meilleure cohérence entre les évaluateurs. Les critères d'interprétation du coefficient ICC étaient les suivants : ICC > 0,75 : bonne fiabilité inter-évaluateurs ; ICC entre 0,60 et 0,75 : fiabilité modérée ; ICC < 0,60 : faible fiabilité (Cicchetti, 1994). Les calculs ont été effectués séparément pour chaque méthode d'évaluation afin de comparer les résultats obtenus et d'identifier les méthodes les plus fiables pour l'évaluation des aspects paraverbaux. Pour la méthode comparative, le logiciel *Comproved* calcule un indice de fiabilité, le SSR (*Scale Separation Reliability*), qui est considéré comme bon à partir de 0,7.

Afin de déterminer dans quelle mesure les méthodes holistique absolue, analytique absolue et holistique comparative produisaient des résultats similaires (QR2) et donc, de voir si les résultats obtenus par ces différentes méthodes d'évaluation étaient cohérents entre eux, nous avons calculé les coefficients de corrélation de Pearson. La corrélation de Pearson est un indicateur statistique qui mesure la force et la direction de la relation linéaire entre deux variables (Cohen, 1988). Les coefficients de corrélation ont été interprétés selon les seuils suivants : r > 0,70 : forte corrélation ; r entre 0,50 et 0,70 : corrélation modérée ; r < 0,50 : corrélation faible. Cette analyse a permis d'identifier si certaines méthodes étaient plus alignées entre elles, ce qui pourrait indiquer leur interchangeabilité ou leur complémentarité dans le cadre de l'évaluation des aspects paraverbaux.

Pour examiner les écarts de notation entre les trois méthodes d'évaluation (QR3), nous avons analysé à la fois la fréquence et l'ampleur des écarts pour chaque élève. Cette étape a permis de déterminer si certaines méthodes tendaient à être plus strictes ou plus indulgentes que d'autres. Pour visualiser les écarts entre les différentes méthodes d'évaluation, nous

avons généré un graphique alluvial. Ce type de représentation graphique permet de suivre l'évolution des notations entre les différentes méthodes et de visualiser comment les jugements se distribuent et varient selon la méthode employée (Tsai et al., 2022). Le graphique alluvial met en évidence les correspondances ou les divergences entre les méthodes, permettant ainsi de mieux comprendre les relations entre elles.

Résultats

Nous présentons les résultats relativement à nos trois questions de recherche.

QR1 : La fiabilité inter-évaluateurs n'est pas identique pour chaque méthode d'évaluation.

L'analyse de la fiabilité inter-évaluateurs à l'aide du coefficient de corrélation intraclasse (ICC) a permis d'évaluer la cohérence des jugements entre les évaluateurs pour chacune des trois méthodes d'évaluation des aspects paraverbaux. Pour rappel, la fiabilité est considérée comme bonne à partir de 0,75.

Pour la méthode holistique absolue (note globale), la fiabilité inter-évaluateurs a montré une bonne cohérence (ICC = 0,78). Bien que cette méthode soit perçue comme rapide et globale, elle a permis une évaluation relativement fiable des productions orales pour les aspects paraverbaux. Cette cohérence des jugements peut s'expliquer par le fait que les évaluatrices ont utilisé un guide de notation commun et partagé des critères implicites relatifs aux éléments paraverbaux, malgré l'absence d'une grille détaillée. Ainsi, bien que basés sur des impressions globales, les jugements ont une certaine stabilité entre les évaluatrices.

La méthode analytique absolue (grille critériée) a montré une fiabilité modérée (ICC = 0,68). Malgré l'utilisation d'une grille d'évaluation critériée, les différences d'interprétation des critères par les évaluateurs ont contribué à une variabilité plus importante. Cela suggère que, bien que cette méthode soit plus détaillée, elle ne garantit pas nécessairement une meilleure cohérence des jugements entre évaluateurs.

Pour la méthode holistique comparative (*Comproved*), une fiabilité modérée est observée (SRR = 0,69). La comparaison par paires a aidé à réduire l'impact des biais individuels, mais la cohérence entre évaluateurs était inférieure à celle de la méthode holistique absolue. Bien que la

méthode comparative aide à structurer le jugement en comparant systématiquement les productions entre elles, des variations entre évaluateurs demeurent.

En conclusion, la méthode holistique absolue (note globale) a montré la meilleure fiabilité inter-évaluateurs, tandis que les méthodes holistique comparative et analytique absolue ont présenté une fiabilité modérée.

QR2 : Les trois méthodes produisent des résultats cohérents, mais avec des variations.

Les résultats des analyses de corrélation entre les différentes méthodes d'évaluation sont présentés dans le tableau 2.

Tableau 2
Résultats des analyses de corrélation entre les différentes méthodes d'évaluation

Méthodes comparées	**Corrélation (r)**	**Significativité (p)**
méthode holistique absolue vs méthode analytique absolue	0,819	$p < 0,001$
méthode holistique comparative vs méthode analytique absolue	0,61	$p < 0,001$
méthode holistique absolue vs méthode holistique comparative	0,585	$p < 0,001$

Les résultats montrent que la corrélation la plus forte ($r = 0,819$) est observée entre la méthode holistique absolue (note globale) et la méthode analytique absolue (grille critériée). La corrélation observée entre la méthode holistique comparative (*Comproved*) et la méthode analytique absolue (note globale) est modérée ($r = 0,610$), mais elle révèle tout de même une certaine cohérence entre ces deux méthodes. C'est aussi le cas pour la corrélation entre la méthode holistique absolue et la méthode holistique comparative ($r = 0,585$).

En somme, les résultats montrent que, bien que des variations existent dans le degré de cohérence entre les méthodes, celles-ci présentent une certaine concordance globale dans l'évaluation des aspects paraverbaux des productions orales.

QR3 : Les écarts de notation révèlent des divergences significatives entre les méthodes.

L'analyse des écarts de notation entre les trois méthodes d'évaluation a révélé des différences significatives en matière tant de fréquence que d'ampleur. Ces écarts mettent en évidence les divergences entre les approches d'évaluation utilisées pour juger les productions orales.

Afin de visualiser ces écarts, nous avons utilisé un graphique alluvial (figure 1) qui permet de suivre les correspondances et les divergences entre les méthodes d'évaluation. Chaque courbe colorée représente une production orale spécifique, identifiée par un numéro, et les résumés sont classés de manière cohérente du plus faible (en haut) au plus élevé (en bas). Le graphique présente les résultats de la méthode holistique absolue (note globale) à gauche, ceux de la méthode analytique absolue (grille critériée) au centre et ceux de la méthode holistique comparative (*Comproved*) à droite.

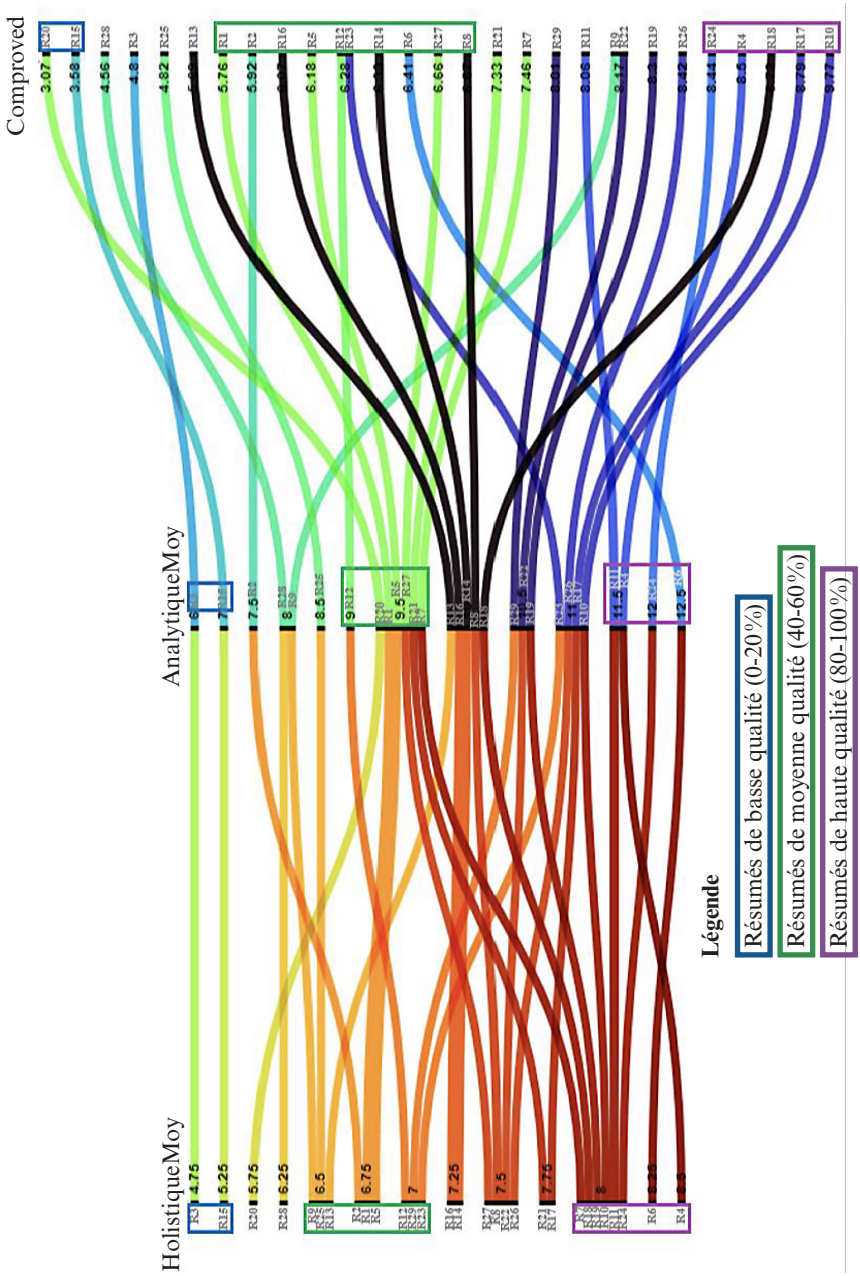
Pour faciliter l'interprétation, nous avons divisé les résumés en cinq intervalles de 20 %, en suivant la méthode proposée par Tsai et al. (2022). Ainsi, les productions situées entre 0 et 20 % sont classées comme de basse qualité (encadré bleu), celles comprises entre 40 et 60 % comme de qualité moyenne (encadré vert) et celles situées entre 80 et 100 % comme de haute qualité (encadré mauve).

Les productions sont visualisées selon ces intervalles, ce qui permet d'identifier les écarts de notation entre les méthodes. Le tableau 3 résume cette distribution, en mettant en gras les productions classées de manière cohérente dans la même catégorie par les trois méthodes.

Tableau 3
Distribution des productions selon les méthodes d'évaluation

	Méthode holistique absolue (note globale)	**Méthode analytique absolue (grille critériée)**	**Méthode comparative (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**

Figure 1
Cycle de construction et de gestion qualité des évaluations en formation



Certaines similitudes apparaissent. Par exemple, le résumé 15 (R15) est jugé de faible qualité dans les trois méthodes, tandis que les résumés 4 (R4) et 24 (R24) sont systématiquement classés comme de haute qualité. Néanmoins, certaines productions obtiennent des scores élevés avec une méthode et des scores nettement inférieurs avec une autre. Par exemple, le résumé 6 (R6) est classé dans l'intervalle 80-100 % pour les méthodes holistique absolue et analytique absolue, mais dans l'intervalle 40-60 % pour la méthode holistique comparative. De la même manière, le résumé 20 (R20) est situé dans l'intervalle 40-60 % pour la méthode analytique absolue, mais dans l'intervalle 0-20 % pour la méthode holistique comparative. De plus, certaines productions ont été jugées comme réussies selon une méthode et insuffisantes selon une autre.

Discussion

L'objectif de cette étude était de comparer la fiabilité et la validité de trois méthodes d'évaluation des aspects paraverbaux dans les productions orales d'élèves de 11-12 ans : la méthode holistique absolue, la méthode analytique absolue et la méthode holistique comparative. Plus spécifiquement, nous avons cherché à connaître la fiabilité inter-évaluateurs de chaque méthode d'évaluation et à savoir si les trois méthodes produisaient des résultats similaires et si des écarts de notation allaient apparaitre entre les trois méthodes d'évaluation.

Premièrement, concernant la fiabilité inter-évaluateurs, nos résultats montrent que la méthode holistique absolue a présenté le meilleur indice de fiabilité inter-évaluateurs (ICC = 0,78). Ce résultat est surprenant, car la littérature existante souligne souvent que cette méthode est plus susceptible d'être influencée par des biais subjectifs (Bouwer & Koster, 2016 ; Metruk, 2018 ; Pollitt, 2012). Toutefois, dans notre étude, l'utilisation d'un guide de notation commun a probablement contribué à réduire la variabilité des jugements entre évaluateurs, permettant ainsi d'obtenir des résultats plus cohérents. Comme l'ont montré Berthiaume et al. (2011), fournir des directives explicites peut clarifier les attentes liées à la performance et diminuer l'influence des biais subjectifs.

En revanche, les deux autres méthodes ont montré des fiabilités modérées. Pour la méthode analytique absolue, ce résultat est en décalage avec les travaux existants (Landrieu et al., 2022). Cette contradiction pourrait être due à la difficulté d'évaluer des éléments paraverbaux à travers des

grilles critériées. Comme le souligne Garcia-Debanc (1999), l'évaluation des aspects paraverbaux de l'oral est difficilement objectivable. Cela pourrait expliquer pourquoi certains aspects plus subjectifs, comme l'intonation, ont entrainé des jugements variables chez les évaluateurs (Deschepper, 2021). Pour la méthode holistique comparative (*Comproved*), bien que supposée plus fiable selon la loi du jugement comparatif de Thurstone (1927), ce résultat pourrait s'expliquer par l'effet de la référence normative inhérent à cette méthode. Un nombre limité d'évaluateurs a réduit le nombre de comparaisons possibles, ce qui a pu affecter aussi la stabilité des jugements. De plus, la qualité perçue d'une production dépend des autres productions mises en comparaison, ce qui peut expliquer des variations dans les classements. Il serait pertinent d'explorer dans des études futures si un nombre accru de comparaisons par paires améliorerait cette fiabilité.

Deuxièmement, concernant la cohérence des résultats entre les trois méthodes, nos résultats montrent des corrélations significatives et positives, ce qui indique que les productions jugées bonnes ou moins bonnes le sont de manière relativement cohérente à travers les méthodes. Cette cohérence globale est rassurante, car elle suggère que, quel que soit l'outil d'évaluation utilisé, les élèves qui produisent des performances de qualité se démarqueront. Cela renforce l'idée que l'évaluation peut remplir sa fonction pédagogique en identifiant de manière fiable les élèves en difficulté ou ceux qui ne le sont pas (Dimmitt, 2009).

La corrélation la plus forte a été observée entre les deux méthodes absolues ($r = 0{,}819$), ce qui peut s'expliquer par une logique commune d'évaluation basée sur des scores absolus, où les productions sont jugées individuellement sans comparaison directe avec d'autres productions (Landrieu et al., 2022). Cependant, cette cohérence peut aussi être influencée par des biais communs tels que l'effet d'ordre ou l'influence des évaluateurs (Barkaoui, 2010 ; Metruk, 2018). Les corrélations plus modérées observées entre la méthode holistique comparative et les deux autres méthodes ($r = 0{,}610$ et $r = 0{,}585$) indiquent l'existence d'écarts entre la manière dont les productions sont jugées et, potentiellement, l'existence d'un effet de la référence normative. Toutefois, une certaine cohérence demeure entre les méthodes, suggérant que l'approche comparative permet également d'identifier de manière fiable les productions les plus et les moins réussies.

Troisièmement, malgré les corrélations significatives identifiées, nous avons observé des variations dans la qualité perçue pour certaines productions orales en fonction de la méthode d'évaluation utilisée. Ainsi, certaines productions jugées de manière positive par une méthode ont été considérées moins bonnes par une autre, ce qui fragilise les fonctions de l'évaluation, en particulier lorsqu'il s'agit de fournir des *feedbacks* cohérents et constructifs aux élèves (Bélec, 2017 ; Dimmitt, 2009).

Ces écarts soulignent donc l'importance de choisir la méthode d'évaluation adaptée à l'objectif visé. Mais, pour le paraverbal, nos résultats vont plutôt en faveur de l'utilisation d'une évaluation holistique globale. Comme le souligne Barkaoui (2010), cette méthode est particulièrement utile lorsque le temps est limité. Ainsi, si l'objectif est de fournir une appréciation globale rapidement, cette méthode peut convenir. Elle est particulièrement pertinente pour des évaluations formatives intermédiaires, où un retour synthétique est suffisant pour guider l'élève dans sa progression. Néanmoins, si l'objectif est de fournir un retour détaillé à l'élève, plus nuancé et plus précis sur chaque aspect du paraverbal (Aouchiche-Ait Yala & Zoubida, 2022), il faudra néanmoins avoir recours à une évaluation analytique, en utilisant une grille critériée. Cette méthode est plus adaptée aux évaluations sommatives ou aux moments où l'accent est mis sur le développement détaillé des compétences. Mais elle demande plus de temps et d'expertise pour être mise en œuvre correctement dans le cadre de l'évaluation de l'oral. Une alternance entre ces méthodes pourrait être envisagée : une évaluation holistique en début de séquence pour situer les élèves, suivie d'une évaluation analytique plus détaillée lorsque les compétences paraverbales deviennent un enjeu central du travail oral.

En revanche, l'utilisation de la méthode holistique comparative via l'outil *Comproved* semble moins intéressante dans le contexte de l'évaluation du paraverbal. Bien que cette approche puisse offrir des résultats plus fiables pour certaines évaluations, elle ne semble pas apporter une valeur ajoutée significative lorsqu'il s'agit d'évaluer des productions orales avec des degrés de complexité paraverbale, d'autant plus que son application en contexte scolaire reste difficile (Landrieu et al., 2022).

Implications pratiques

L'une des premières implications pratiques est de pouvoir considérer la souplesse des outils d'évaluation et d'éviter l'utilisation systématique d'une seule méthode. En effet, sur la base de nos résultats, le message est le

suivant : les enseignants peuvent choisir entre différentes méthodes d'évaluation en fonction des objectifs fixés. Si l'objectif est d'obtenir une appréciation rapide et générale des productions orales, la méthode holistique absolue est recommandée pour le paraverbal, en raison de sa simplicité, de sa rapidité d'exécution et de sa fiabilité relative. Cela est particulièrement utile dans des contextes où le temps est limité et où une évaluation formative globale est suffisante. Toutefois, ce choix ne signifie pas un rejet des grilles critériées, mais plutôt une invitation à les utiliser de manière ciblée et pertinente en fonction des besoins d'évaluation. Par ailleurs, mettre en place des dispositifs de comparaison de productions plus conséquents ne semble pas une voie pertinente dans le cadre de l'évaluation du paraverbal, car cette méthode repose sur une évaluation relative, ce qui peut limiter la possibilité d'attribuer une appréciation directement individualisée des compétences de chaque élève, dans la mesure où le jugement se construit par comparaison avec d'autres productions et non sur des critères fixes.

Ensuite, la méthode analytique absolue, bien que plus détaillée, peut souffrir de variabilité inter-évaluateurs en raison de l'interprétation subjective des critères. C'est sans doute un élément qui pourrait faire l'objet de plus d'attention dans la formation des enseignants : comment utiliser les outils, comment amener les élèves à les utiliser dans le cadre de l'évaluation par les pairs, ce qui pose souvent des questions (Vassart et al., 2022) mais aussi, et surtout, comment en communiquer les résultats. En effet, si les grilles critériées permettent aux enseignants de cibler les forces et les faiblesses spécifiques des élèves (Berthiaume et al., 2011 ; Deschepper, 2021), il n'en demeure pas moins que des difficultés ont été pointées quant à l'utilisation et à la communication des informations de la grille pour donner une rétroaction efficace (Colognesi et al., 2024) permettant des améliorations dans l'apprentissage de l'oral. Accentuer la formation des enseignants sur ce plan semble également pertinent, tout comme la nécessité de réfléchir au lien entre la méthode d'évaluation et sa fonction (formative ou sommative) ainsi qu'aux modalités spécifiques de mise en œuvre (avec ou sans implication des élèves, en interaction ou en autonomie, avec un *feedback* immédiat ou différé, etc.). Il semble effectivement important de soutenir ces aspects liés à l'évaluation en formation des enseignants (Barbier & Colognesi, 2024).

Limites de l'étude

Comme toute recherche, cette étude présente des limites. Pointons-en quelques-unes. Premièrement, l'évaluation a été réalisée sur un échantillon de 29 fichiers audios, un nombre relativement restreint. Deuxièmement, le nombre d'évaluateurs pour *Comproved* aurait pu être plus important. Un effectif plus élevé aurait permis d'obtenir un plus grand nombre de comparaisons par paires, ce qui aurait renforcé la robustesse des classements obtenus. De plus, une diversité accrue des évaluateurs aurait pu réduire l'impact des biais individuels et offrir une meilleure stabilité des jugements. Troisièmement, la difficulté à se concentrer uniquement sur les aspects paraverbaux, au détriment du contenu des résumés, a également pu biaiser les évaluations. Quatrièmement, un biais potentiel lié à l'implication des chercheuses dans cette étude est à considérer, puisque ce sont les deux mêmes personnes qui ont réalisé à la fois l'évaluation holistique absolue et l'évaluation analytique absolue. Bien qu'un délai ait été respecté entre les deux évaluations, il est possible que l'évaluation holistique absolue, effectuée en premier, ait influencé les jugements lors de l'évaluation analytique absolue. Enfin, une autre limite possible concerne l'absence de mise en œuvre de cette étude en contexte réel de classe. Bien que les résultats visent à avoir des implications pratiques pour l'évaluation de l'oral en milieu scolaire, ils ont été obtenus dans un cadre expérimental contrôlé. Il serait pertinent d'examiner comment ces méthodes fonctionnent en situation réelle, notamment en tenant compte des contraintes temporelles et organisationnelles propres aux enseignants et aux élèves.

Conclusion

Notre étude remet en question l'usage systématique d'une méthode unique d'évaluation comme les grilles critériées, largement recommandées dans la littérature (Metruk, 2018). En effet, nos résultats montrent que la méthode holistique absolue (une note globale) est plus fiable, tout en étant moins chronophage et énergivore que les grilles critériées (Barkaoui, 2010). Toutefois, cette méthode ne permet pas de donner un *feedback* précis aux élèves (Aouchiche-Ait Yala & Zoubida, 2022). Cela soulève la question de l'utilisation de méthodes différentes selon les objectifs : une méthode globale pour des évaluations rapides et une méthode détaillée pour un retour précis.

Révision linguistique : Marie-Claire Legaré
Mise en page : Emmanuel Gagnon
Résumé en portugais : Eusébio André Machad

Réception : 23 septembre 2024
Version finale : 27 février 2025
Acceptation : 17 mars 2025

LISTE DE RÉFÉRENCES

Aguert, M., Laval, V., & Bernicot, J. (2010). Comprendre l'intention communicative du locuteur : une étude du rôle de l'intonation et du contexte chez des enfants de 5 à 9 ans. *L'Année psychologique*, 110, 49-70. https://doi.org/10.3917/anpsy.101.0049

Alpes, Y. (2012). À propos de PISA : pourquoi, pour qui, évaluer et comparer les compétences des élèves ? *Questions Vives. Recherches en* éducation, *6*(16), 11-14. https://doi.org/10.4000/questionsvives.892

Alrabadi, E. (2011). ¿Qué método se debe adoptar para la enseñanza/aprendizaje de la comunicación oral ? [Quelle méthodologie faut-il adopter pour l'enseignement/apprentissage de l'oral ?] *Didáctica. Lengua y Literatura*, *23*, 15-34. https://doi.org/10.5209/rev_DIDA.2011.v23.36308

Aouchiche-Ait Yala, O., & Zoubida, B. (2022). Les défis de l'évaluation de l'oral. *Pratiques & didactique, 1*(1), 72-86. https://www.asjp.cerist.dz/en/downArticle/764/1/1/177627

Aubergé, V. (2002). Prosodie et émotion. *Actes des deuxièmes assises nationales du 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). Les pratiques préconisées en formation pour faire la classe interviennent-elles dans les planifications des futurs enseignants de français ? *Revue Canadienne de l'Éducation*, *47*(1), 113-148.

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

Bélec, C. (2017). Pourquoi évaluer ? *Pédagogie collégiale, 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). Réduire la subjectivité lors de l'évaluation des apprentissages à l'aide d'une grille critériée : repères théoriques et applications à un enseignement interdisciplinaire. *Revue internationale de pédagogie de l'enseignement supérieur, 27*(2). http://ripes.revues.org/524

Boureux, M. (2017). Mieux percevoir pour mieux prononcer. Quelle phonétique corrective pour aider les apprenants italiens à bien parler français. *Rivista Interculturale, Università di Lecce,* 43-68. http://magali.boureux.com/IMG/pdf/2017-03-11_actesrome2016boureux.pdf

Bourhis, V. (2012). Situation de lecture en toute petite section : le rôle du paraverbal. *Le Français aujourd'hui,* (4), 85-97. https://doi.org/10.3917/lfa.179.0085

Bourhis, V. (2014). Voix du maître, voix de l'élève : un dialogisme interlocutif. 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 program for elementary students. [Thèse de doctorat, 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. Dans P. Newton, J. A. Baird, H. Goldsteing, H. Patrick, & P. Tymms (dir.), *Techniques for monitoring the comparability of examination standards* (p. 246–300). QCA.

Candea, M. (2000). Contribution à l'étude des pauses silencieuses et des phénomènes dits « d'hésitation » en français oral spontané. Étude sur un corpus de récits en classe de français [Thèse de doctorat, Université de la Sorbonne nouvelle-Paris III]. HAL. https://theses.hal.science/tel-00290143v1

Chabanne, J. C. (1999). Verbal, paraverbal et non-verbal dans l'interaction verbale humoristique. Dans J. M. Defays, & L. Rosier (dir.), *Approches du discours comique* (p. 35-53). Mardaga. https://hal.science/hal-00921934

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2e éd.). 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 elementary school students through video-recorded performances. *Teaching and Teacher Education*, *128*, 104141.

Colognesi, S., & Deschepper, C. (2019). Les pratiques déclarées de l'enseignement de l'oral au primaire. Qu'en est-il en Belgique francophone ? *Language and Literacy*, *21*(1), 1-18.

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…. *Veredas-Revista de Estudos Linguísticos*, *26*(1), 141-169

Delcambre, I. (2011). Comment penser les relations oral/écrit dans un cadre scolaire. *Recherches*, *54*(1), 7-15.

Deschepper, C. (2021). Comment et pourquoi questionner les grilles d'évaluation de l'oral ? Description d'un dispositif de formation initiale et perspectives pour la recherche. *Évaluer. Journal international de recherche en éducation et formation, 7*(2), 61-78. https://doi.org/10.48782/e-jiref-7-2-61

Di Cristo, A. (2013). *La prosodie de la parole*. De Boeck Supérieur.

Di Cristo, A., Auran, C., Bertrand, R., Chanet, C., Portes, C., & Régnier, A. (2004). Outils prosodiques et analyse du discours. *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 counseling practices. *Professional School Counseling*, *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). Proposition d'une typologie des objets d'enseignement/apprentissage de l'oral. *Les dossiers des sciences de l'éducation*, *36*, 37-56. https://doi.org/10.4000/dse.1347

Gagnon, R., & Colognesi, S. (2021). Éditorial : Évaluer les performances orales sans les dénaturer ? *Évaluer. Journal international de Recherche en Education et Formation, 7*(2), 1-5. https://doi.org/10.48782/e-jiref-7-2-1

Gagnon, R., de Pietro, J.-F., & Fisher, C. (2017). Introduction. Dans J. -F. de Pietro, C. Fisher et R. Gagnon (dir.), *L'oral aujourd'hui : perspectives didactiques* (p. 11-40). Presses universitaires de Namur.

Garcia-Debanc, C. (1999). Évaluer l'oral. *Pratiques*, *103-104*, 193-212.

Gaudreau, G., Hudon, C., & Monetta, L. (2011). Bases psycholinguistiques et neuroanatomiques de la compréhension de l'ironie chez l'adulte. *Revue de neuropsychologie*, 3, 148-154. https://doi.org/10.3917/rne.033.0148

Hadji, C. (1992). L'évaluation des actions éducatives. Presses Universitaires de France. https://doi.org/10.3917/puf.hadji.1992.01.

Issaieva, É., & Crahay, M. (2010). Conceptions de l'évaluation scolaire des élèves et des enseignants : validation d'échelles et étude de leurs relations. Mesure et évaluation en éducation, *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., & Messier, G. (2009). Les représentations de l'enseignement et de l'évaluation de l'oral chez des enseignants et des élèves du secondaire en français langue d'enseignement. *Revue du Nouvel-Ontario*, 34, 119-144.

Lafontaine, L., & Préfontaine, C. (2007). Modèle didactique descriptif de la production orale en classe de français langue première au secondaire. *Revue des sciences de l'éducation*, *33*(1), 47–66. https://doi.org/10.7202/016188ar

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). Formation universitaire à l'évaluation de l'oral : regard sur la capacité d'autoévaluation de futurs enseignants. Dans J. -F., De Pietro, C. Fisher, & R. Gagnon (dir.), *L'oral aujourd'hui : perspectives didactiques* (p. 259-274.). Presses universitaires de Namur.

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). Les croyances des formateurs de futurs enseignants : influences sur leurs pratiques d'enseignement et d'évaluation de l'oral. *Phronesis*, *14*(1), 117–137. https://doi.org/10.7202/1116127ar

Nonnon, É. (2016). 40 ans de discours sur l'enseignement de l'oral : la didactique face à ses 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). Un consensus terminologique en prosodie ? *Actes des XXIII^e journées de la linguistique (JDL)*, 5-6. 77

Poiré, F. (2000). L'accent focal et l'accent d'emphase dans la description de l'intonation du français. *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 program in India. *Quality assurance in education*, *19*(1), 84-104. DOI 10.1108/09684881111107771

Sales-Hitier, D., & Dupont, P. (2025). Une évaluation pour soutenir l'enseignement et les apprentissages de l'oral : le dispositif SEMO. *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). Repenser le modèle de la séquence didactique pour enseigner l'oral au primaire : résultats d'une première année de recherche. *Recherches*, *73*, 7592.

Stordeur, M. F., Nils, F., & Colognesi, S. (2021). Sept dilemmes rencontrés par les enseignants du primaire quand ils évaluent les exposés oraux des élèves. *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! Elementary teachers' practices supporting preparation of oral presentations. *L1-Educational Studies in Language and Literature*, 22, 1-29.

Stordeur, M.-F., Nils, F., Francotte, È., & Colognesi, S. (2025). Le pari de l'utilisation des auto-confrontations pour accompagner les élèves du primaire dans la production d'exposés oraux. *Phronesis*, *14*(1), 215–241. https://doi.org/10.7202/1116132ar

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

Tsai, Y-C., Chien, T-W., Wu, J-W., & Lin, C-H. (2022). Using the Alluvial plot to visualize 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

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). Dans les coulisses de l'évaluation de l'oral par les pairs au primaire. *Éducation et francophonie*, *50*(1).

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

Weber, C. (2021). Oral, évaluation et réflexivité. Vers un mode intégratif des traits d'oralité. *Évaluer. Journal international de recherche en éducation et formation*, *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). *Le résumé d'informations à l'oral : démarche empirique de sa caractérisation via l'élaboration d'un outil de mesure* [Thèse de doctorat inédite, Université catholique de Louvain].

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 ? *e-JIREF*, *8*(2), 51-88. https://doi.org/10.48782/m3kwdh11

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

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. *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. *Mesure et Évaluation en Éducation*, *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

Anmexe A

Critères	1	2	3	4	5
Volume	La majorité du discours est inaudible : plus de 4 mots/ expressions sont inaudibles **ET** le volume d'une grande partie de l'explication est très bas.	Une grande partie du discours est inaudible : 2, 3 ou 4 mots/expressions sont inaudibles **ET** le volume d'une grande partie de l'explication est très bas.	Une partie du discours est inaudible : plus de 2 mots/expressions sont inaudibles **OU** le volume d'une grande partie de l'explication est très bas.	Un ou deux mots/ expressions sont inaudibles.	L'ensemble du discours est audible et le volume sonore est adéquat.
Débit	Les pauses sont inexistantes, ce qui ne permet pas à l'interlocuteur d'intégrer les informations transmises au fur et à mesure de l'explication **ET** les pauses sont trop longues, trop nombreuses ou réalisées à des endroits inopportuns, ce qui engendre une discontinuité dans le discours et nuit à la compréhension de l'explication.	Les pauses sont inexistantes, ce qui ne permet pas à l'interlocuteur d'intégrer les informations transmises au fur et à mesure de l'explication **OU** les pauses sont trop longues, trop nombreuses ou mal placées, ce qui engendre une discontinuité dans le discours et nuit à la compréhension de l'explication.	Certaines pauses sont réalisées par l'orateur, mais elles sont trop peu nombreuses et/ou trop courtes pour une intégration optimale des informations **ET** une certaine continuité du discours est assurée, mais quelques pauses trop longues, trop fréquentes ou mal placées engendrent une discontinuité dans certains passages de l'explication.	Certaines pauses sont réalisées par l'orateur, mais elles sont trop peu nombreuses et/ou trop courtes pour une intégration optimale des informations **OU** une certaine continuité du discours est assurée, mais quelques pauses trop longues, trop fréquentes ou mal placées engendrent une discontinuité dans certains passages de l'explication.	Des pauses sont réalisées et permettent à l'interlocuteur d'intégrer les informations transmises par l'orateur **ET** les pauses n'engendrent aucune discontinuité dans le discours.

Anmexe A (suite)

Critères	1	2	3	4	5
Intonation	L'élève ne module presque pas sa voix, ce qui rend le discours monotone. Il est très difficile de maintenir son attention tout au long du discours et d'identifier les éléments importants du discours.	Entre 1 et 3.	L'élève n'est ni monotone ni particulièrement éloquent. Il module sa voix moyennement.	Entre 3 et 5.	L'élève module énormément sa voix, il fait preuve d'éloquence. Il est très facile de maintenir son attention tout au long du discours et d'identifier les éléments importants du discours.