L'évaluation de la production orale en allemand L2 : quels outils de correction ? Comparaison d'une grille holistique et d'une grille analytique

Verónica Sánchez Abchi, Institut de recherche et de documentation pédagogique
Sophie Sieber Meylan, Haute École Pédagogique Vaud
Alina Matei, Institut de recherche et de documentation pédagogique

Cette contribution a pour but d'explorer la pertinence et la complémentarité des outils de correction et d'évaluation de la production orale en langue étrangère. En particulier, nous nous intéressons à la comparaison de deux échelles ou grilles d'analyse – une analytique et l'autre holistique – visant à évaluer la production de l'oral en allemand langue étrangère chez des élèves de fin de primaire (niveau A1), scolarisés dans le contexte francophone de la Suisse romande. Les productions orales réalisées par 212 élèves de toute la Suisse romande, à partir de deux tâches spécifiques, ont été évaluées par 4 juges, à l'aide de deux types de grilles différentes. Les résultats mettent en évidence les apports et les limites des deux outils.

1. Introduction

Cette contribution, qui s'inscrit dans le domaine de l'évaluation des langues étrangères, a pour but d'explorer la pertinence et la complémentarité des outils de correction et d'évaluation de la production orale en langue étrangère. En particulier, nous nous intéressons à la comparaison de deux échelles ou grilles d'analyse – une analytique et l'autre holistique – visant à évaluer la production de l'oral en allemand langue étrangère chez des élèves de fin de primaire (niveau A1), scolarisés dans le contexte francophone de la Suisse romande.

Le travail fait partie du projet Eprocom développé par l'Institut de recherche et de documentation pédagogique de la Conférence intercantonale de l'instruction publique et de la culture de la Suisse romande et du Tessin. Ce projet vise à sélectionner et valider des tâches permettant d'évaluer les objectifs du Plan d'études romand (Conférence intercantonale de l'instruction publique et de la culture de la Suisse romande et du Tessin [CIIP], 2024a, dorénavant, PER), afin de les mettre à disposition des enseignant·e·s des cantons de Suisse romande via une plateforme en ligne (PistEval)[1]. Dans les étapes précédentes, le projet s'est concentré sur la mise à disposition de tâches d'évaluation en mathématiques et en français L1 (Roth et al., 2021 ; Sánchez Abchi et al., 2016 ; Sánchez Abchi et al., 2022 ; Roth et Ruf, 2023). Plus récemment, la nécessité d'intégrer des tâches d'évaluation de l'allemand langue étrangère a été soulevée, en donnant la priorité à l'évaluation des compétences orales (réception et production). Ainsi, des tâches pour évaluer la production de l'oral (PO) pour le niveau A1 (niveau correspondant à la 8e année de scolarité obligatoire) ont été créées et validées (voir Sánchez Abchi et al., 2024).

Ces tâches ont fait l'objet d'une passation auprès de 212 élèves de 8e HarmoS en Suisse Romande. Les productions de ces élèves ont été évaluées à l'aide d'une grille analytique et d'une grille holistique ; conçues spécialement pour l'appréciation des tâches proposées.

Les deux grilles ont été discutées avec des expert·e·s et des enseignant·e·s afin d'examiner leur pertinence et de les adapter au contexte romand. Il faut rappeler, cependant, que tous les outils d'évaluation ont des limites et qu'il est essentiel de s'assurer, via l'expérimentation, si les grilles mesurent de manière satisfaisante la compétence à évaluer.

Cet article vise ainsi à analyser le fonctionnement, la pertinence et la complémentarité de ces deux grilles – analytique et holistique – d'évaluation de la PO en allemand langue étrangère (niveau A1), dans le

1 Les *Pistes pour l'évaluation (PistEval)* est un site Internet d'accès interne, via le site du PER (CIIP, 2024a), sur lequel sont mises à disposition des enseignant·e·s romand·e·s des ressources évaluatives pour la 8e HarmoS (élèves de 11-12 ans) en français et en mathématiques. Elles visent à les soutenir dans l'élaboration de leurs évaluations, qu'elles servent à des fins diagnostiques, formatives ou certificatives (Roth et Ruf, 2023).

DOI 10.24452/sjer.47.3.6 ISSN 2624-8492

but de mettre à disposition des enseignant·e·s non seulement des tâches validées, mais également des outils de correction pertinents.

1.1 Les échelles analytiques et holistiques

L'interprétation de résultats d'un test construit sur la base de plusieurs tâches constitue un aspect crucial de l'évaluation, car c'est le construit même de ce qu'on évalue qui est en jeu. La notation ou l'attribution d'un score est liée à la validité d'un test (Savard et Sévigny, 2007) et, en même temps, des scores non valides peuvent entrainer des interprétations discutables ou trompeuses. Dans le cas de la pondération des productions orales ou écrites, de nombreux facteurs doivent être pris en compte afin d'éviter une interprétation erronée des résultats, pouvant affecter la validité de l'ensemble d'une tâche.

Un premier aspect est la subjectivité des évaluateurs et des évaluatrices, qui peut entrainer une variabilité dans la notation des productions et avoir une influence sur les inférences faites sur les compétences linguistiques des candidat·e·s (Kim, 2015).

Un autre point d'attention concerne les aspects à observer dans le jugement de la qualité et la notation de productions. Quel est le poids de différentes composantes de la production ? Quels aspects faut-il prendre en considération ? Ce point a suscité des discussions variées (North, 2005), concernant, par exemple, le rôle des aspects grammaticaux ; l'influence de la prononciation et de l'intonation (Park, 2020) ; l'importance de la fluidité du discours face à l'intelligibilité (Isaacs, 2016) ; ou le poids de la précision – *accuracy* – de la langue dans la tâche (Chavez, 2007), parmi d'autres.

Un troisième point, directement lié aux aspects observés, concerne les échelles utilisées pour noter les productions. En effet, les scores de production – orale ou écrite – s'obtiennent, le plus généralement, par des échelles ou grilles. Ces échelles permettent de décrire la compétence linguistique consistant en une série de niveaux, par rapport auxquels la performance des apprenant·e·s est jugée. Les niveaux ou bandes sont généralement caractérisés en termes de ce que les sujets peuvent faire avec la langue cible et leur maitrise de ressources linguistiques (voir, Isaacs, 2016, pour une synthèse).

Les échelles utilisées pour l'évaluation de la production sont, le plus souvent, de type holistique, de type analytique (Fulcher, 2015) ou une combinaison de deux ; chaque procédure ayant toujours leurs avantages et leurs inconvénients (voir Cushing Weigle, 2002 ; Savard et al., 2008 ; Yetiş, 2017, entre autres, pour une synthèse).

Les échelles ou grilles holistiques donnent une impression générale de la capacité d'un·e élève sur la base de plusieurs aspects considérés simultanément et non selon des composantes individuelles (syntaxe, structure, orthographe, etc.) Les échelles holistiques sont pratiques pour la prise de décision parce qu'elles permettent d'évaluer rapidement : les personnes qui évaluent ont moins d'aspects à considérer que dans une grille complexe, comportant de nombreux critères (Luoma, 2004).

Toutefois, malgré leur praticité, certaines limites ont été signalées. Par exemple, les échelles holistiques ne permettent pas de spécifier les faiblesses et les forces des personnes qui passent les tests. Il a également été soulevé que les évaluatrices et évaluateurs peuvent être désorienté·e·s lorsque plusieurs aspects sont évalués simultanément (Isaacs, 2016). Ces grilles laissent plus de place à l'interprétation de juges (Savard et Sévigny, 2007), car elles mobilisent davantage des descripteurs qui laissent une certaine marge à la subjectivité (Luoma, 2004), ce qui pourrait avoir un impact sur la fiabilité de tels outils (Alderson et Banerjee, 2002).

Les échelles analytiques, de leur côté, présentent séparément les différentes dimensions de l'évaluation, avec des descripteurs pour les différents niveaux, ce qui permet d'attribuer un score pour chacune de ces dimensions (ou « critères »). Puis une note composite est générée sur la base de ces différentes dimensions. Les dimensions évaluées émergent d'une conception ou d'une hypothèse sur ce qu'est l'expression orale. Généralement, on trouve des dimensions telles que l'organisation du texte, le contenu, la grammaire, le vocabulaire, la prononciation, la fluidité, etc. Il faut pourtant veiller à que les échelles analytiques présentent un nombre adéquat de critères, idéalement entre 3 et 5, afin de ne pas surcharger les juges (Luoma, 2004). Un avantage des échelles analytiques est notamment la richesse des informations fournies en lien avec ces différentes dimensions, ainsi que les conseils détaillés qui peuvent être formulés de manière précise sur les forces et les faiblesses spécifiques des performances des élèves (Luoma, 2004).

L'élaboration des échelles ainsi que leur utilisation présentent donc des défis importants. D'un côté, ce n'est pas simple de saisir la complexité de la capacité d'expression avec un bref descripteur pour les différents critères (Isaacs, 2016) ; d'un autre côté, avec les échelles analytiques en particulier, il existe un risque d'analyser la production de manière très compartimentée, sans pouvoir prendre en compte le lien et l'articulation des différents aspects, et sans considérer la production comme un tout.

C'est la raison pour laquelle certain·e·s auteur·e·s proposent l'utilisation combinée d'une grille holistique et d'une grille analytique pour évaluer les productions. Bachman et Savignon (1986) ont suggéré que deux notes -une note holistique, ainsi qu'une note analytique- devraient être attribuées pour fournir un profil précis de la capacité d'expression orale de la personne examinée.

Pourtant, les études comparant les scores de deux types de grilles ont donné des résultats parfois bien différents. Chuang (2009) a étudié les notations, pour évaluer la PO en anglais, en utilisant des grilles holistiques et analytiques. Dans cette étude, aucune différence significative n'a été constatée entre les scores des deux grilles. Ounis (2017), de son côté, s'est intéressé à l'évaluation des compétences orales des apprenant·e·s d'anglais langue étrangère, en Tunisie, en comparant les scores des grilles analytiques et holistiques. Malgré le fait que les deux méthodes aient donné de faibles taux de fiabilité, l'échelle holistique était plus utile, plus fiable et plus cohérente pour le contexte respectif. Metruk (2018), quant à lui, en comparant les grilles pour évaluer la PO en anglais langue étrangère dans le contexte universitaire en Slovaquie, a trouvé de différences significatives entre les scores, mais il a conclu que les deux méthodes d'évaluation pouvaient être considérées comme complémentaires. Witzigmann et Sachse (2020), de leur côté, en s'intéressant aux compétences de PO en français d'élèves de 10-11 ans, ont analysé la manière dont les critères d'une évaluation analytique étaient pris en compte dans une évaluation holistique. Les résultats ont montré que les caractéristiques linguistiques de la grille analytique ont été évaluées de manière fiable et intégrées par les évaluatrices et les évaluateurs dans un jugement global. Chen et al. (2022) se sont intéressés à la manière dont les différentes grilles affectent la fiabilité des personnes qui évaluent. Dans l'ensemble, bien que les deux types de grilles classent les performances de manière similaire, la notation holistique a conduit à une fiabilité des notes relativement plus élevée.

Ce bref aperçu des études utilisant et comparant les deux types de grilles montre que les résultats ne peuvent pas être généralisés, mais doivent être considérés dans chaque contexte particulier, en tenant compte des différentes variables qui entrent en jeu. En effet, l'identification d'une échelle d'évaluation appropriée dépend des objectifs de l'évaluation et de la disponibilité des instruments existants.

Afin de savoir quel outil de correction est le mieux adapté à notre contexte et aux tâches utilisées, cet article vise à analyser deux échelles d'évaluation de la PO. Plus précisément, il a pour objectif d'examiner la pertinence, l'utilité et la complémentarité d'une grille d'évaluation holistique et d'une grille analytique (basée sur trois critères), conçues pour évaluer la production orale en allemand, langue étrangère, d'apprenant·e·s débutant·e·s (niveau A1). La passation auprès de 212 élèves de 8e en Suisse Romande a permis d'associer les productions aux scores de deux grilles.

En ce sens, la présente étude pose les questions de recherche suivantes pour l'évaluation de la production orale en allemand L2 (niveau débutant) :

1. Quel est le lien entre le score à chaque critère de l'échelle analytique et le score de l'échelle holistique ?
2. Est-ce que le score de l'échelle holistique peut remplacer l'ensemble des scores de chaque critère de l'échelle analytique ?

Répondre à ces deux questions devrait nous permettre de savoir si les deux grilles présentent une cohérence entre elles et si elles nous permettent de mesurer de manière satisfaisante ce que l'on souhaite évaluer : la compétence de PO pour le niveau concerné. À partir de l'analyse des productions orales des élèves, obtenues lors de la passation de deux tâches, et de l'application des deux grilles, nous tenterons de donner des réponses à nos interrogations.

2. Méthodologie

2.1 Caractéristiques de tâches

Deux tâches, validées du point de vue du contenu préalablement par un groupe de didacticien·ne·s (Sánchez Abchi et al., 2024), ont été utilisées dans la présente étude. La première tâche demande à l'élève de se présenter et de présenter son école, dans le but de participer à un concours. Pour la deuxième tâche, l'élève doit se mettre dans la peau de quelqu'un qui s'est perdu dans un parc d'attractions et qui doit donner des informations personnelles et décrire les personnes avec lesquelles elle ou il a fait l'excursion.

Les deux tâches permettent d'évaluer l'objectif du PER suivant *:* « Présentation de soi, de sa famille ou d'une tierce personne (nom, âge, provenance, domicile, école, emploi du temps, hobbys) » (CIIP, 2024b, para 4 ; objectifs L2, 24), ce qui correspond au descripteur du CECR (A1.2) « [l'apprenant·e] peut produire des expressions simples isolées sur les gens. » (Conseil de l'Europe, 2001, p. 49)

Les tâches étaient présentées sur tablette tactile. Les élèves pouvaient ainsi gérer de manière indépendante l'écoute des consignes, autant de fois que nécessaire, ce qui permettait d'enlever la difficulté potentielle de la lecture. En plus, les élèves pouvaient décider le moment pour lancer l'enregistrement. Les productions pouvaient être ensuite écoutées et les élèves pouvaient décider de valider les enregistrements ou de continuer les essais. D'après les instructions, les élèves pouvaient réaliser jusqu'à trois enregistrements. Les deux tâches étaient accompagnées de deux grilles de correction que nous explicitons dans le point suivant.

2.2 Les échelles d'évaluation

Élaboration des échelles

Les grilles ont été construites compte tenu la littérature sur l'évaluation de l'oral, ainsi que le construit spécifique à évaluer.

La première étape a consisté à l'identification des objectifs d'apprentissage et à la définition des critères à retenir. Pour ce faire, nous avons principalement tenu compte des objectifs du plan d'études romand pour le niveau concerné. Nous avons, également, pris en considération le descripteur du CECR pour le niveau A1 pour la production orale : « Peut produire des expressions simples isolées sur les gens et les choses » (Conseil de l'Europe, 2001, p. 49). Nous avons ensuite formulé des critères pour les différents niveaux des grilles, qui ont été opérationnalisés sous la forme de descripteurs.

Dans un deuxième temps, les grilles ont été soumises à des groupes d'expert·e·s différents, dans le but de les préciser et les améliorer : un groupe de 3 formatrices et formateurs spécialistes dans la didactique de l'allemand langue étrangère et un groupe constitué par 5 enseignant·e·s chevronné·e·s qui enseignaient à des élèves de 8e HarmoS, soit dans le contexte pour lequel ces grilles étaient censées être utilisées. Les deux groupes se sont prononcés sur l'adéquation de ces grilles pour estimer le niveau de performance des élèves, ainsi que sur la clarté des critères et des descripteurs mobilisés. Les deux grilles ont été revues et améliorées en fonction de la rétroaction obtenue auprès des deux groupes.

Caractéristiques des échelles

Chaque production d'élève était notée à l'aide d'une grille analytique et d'une grille holistique, que nous présentons ci-dessous.

Grille ou échelle holistique

Le score holistique assigné par les juges à chaque production permettait de la classer dans un des trois niveaux de performance : 0 (niveau plus bas) ; 1 (niveau intermédiaire) et 2 (niveau plus élevé) ; chaque niveau prenant en considération un ensemble de critères. Le tableau 1 présente l'échelle holistique.

Tableau 1

Échelle holistique

Score	Description
2	La production est claire, bien structurée et développe tous les points de la consigne. Malgré des erreurs mineures au niveau lexical et grammatical, la production est très compréhensible et présente une certaine variété lexicale et de structures.
1	La production correspond à la situation de communication, bien que certains éléments de la consigne puissent manquer. La production reste claire et compréhensible, malgré quelques erreurs grammaticales et lexicales importantes.
0	La production ne correspond pas à la situation et n'atteint pas l'objectif de communication en raison de son développement insuffisant, de sa désorganisation ou de la présence de nombreuses erreurs (lexicales et/ou grammaticales) qui rendent difficile la compréhension du message.

Grille ou échelle analytique

La grille analytique intégrait trois critères liés qui permettaient de décrire une production :

a) *Le contenu :* Ce critère prenait en considération si les éléments de contenu et les informations sollicitées dans la consigne étaient présents dans la production. Le message devait correspondre à la situation de communication et intégrer une majorité des éléments demandés.

b) *L'étendue du vocabulaire et la correction grammaticale :* Ce critère consistait à vérifier si le vocabulaire disponible était suffisant pour effectuer la tâche demandée, si les erreurs entravaient ou non la communication et si les structures utilisées étaient adéquates. Nous avons choisi de fusionner ces deux critères pour plusieurs raisons. Tout d'abord, parce que les études précédentes montrent que ces deux composantes combinées permettent de mieux évaluer les performances des candidat·e·s (Ma, 2022). Ensuite, parce que, dans le contexte qui nous occupe, les moyens d'enseignement employés privilégient une approche centrée sur les *chunks*, c'est à dire des unités lexico-grammaticales, composées de plusieurs mots (Lenz et Barras, 2017), qui associent des structures grammaticales à des expressions directement utilisables par les élèves dans des situations de communication concrètes. Et enfin, pour une raison pragmatique : pour garder un nombre de critères aussi restreint que possible (Metruk, 2018, pour une synthèse). Ainsi, compte tenu de l'âge et du niveau des élèves (A1), ces deux critères sont fréquemment évalués ensemble comme faisant partie de la compétence lexicale, qui relève à la fois d'éléments lexicaux et grammaticaux et de la capacité à les utiliser (Conseil de l'Europe, 2001).
c) *La fluidité et la prononciation :* Ce critère évaluait si la prononciation et la fluidité – cette dernière opérationnalisée par la longueur et la fréquence des pauses- affectaient la qualité de la communication. Aux niveaux initiaux, la fluidité n'est pas prise en compte dans les descripteurs du PER considérés comme référence (CIIP, 2024). Pourtant, nous avons décidé d'associer la fluidité à la prononciation, afin d'évaluer si ces aspects pris ensemble, liés tous les deux à la nature « physique » de la parole, pouvaient avoir un impact sur la communication.

Les correcteurs et correctrices devaient attribuer un score pour chacun de critères qui constituaient la grille. Sur l'axe horizontal, la grille présentait trois bandes correspondantes à différents niveaux de performance : niveau plus bas, score de 0 ; niveau intermédiaire, score de 1 et niveau plus élevé, score de 2. L'échelle analytique est présentée dans le tableau 2.

Tableau 2

Échelle analytique

Critère/ Score	0	1	2
Contenu (tâche)	La production correspond partiellement à la consigne (plusieurs informations (plus de trois) manquent) ou le message ne répond pas à la situation de communication (informations incohérentes).	La production correspond à la consigne. Il peut manquer quelques informations (une à trois). Message cohérent en lien avec la situation de communication, malgré quelques maladresses.	La production correspond à la consigne. L'élève donne tous les éléments demandés, voire plus. Message cohérent en lien avec la situation de communication
Étendue du vocabulaire et correction grammaticale*	Ne dispose pas d'un vocabulaire suffisant pour effectuer correctement la tâche et recourt à d'autres langues. Erreurs fréquentes sur les structures simples.	Peut mobiliser le lexique nécessaire à la réalisation de la tâche ou à une partie de la tâche. Certains mots peuvent manquer. Quelques erreurs de structure subsistent sans gêner la compréhension.	Maitrise bien les structures courantes, voire fait un effort pour mobiliser un lexique plus varié, ou plus précis, sans répétition fréquente de structures identiques.
Utilisation correcte de la prononciation et de l'intonation. **Fluidité**	Souvent incorrecte, gênant fréquemment la compréhension. Pauses longues et fréquentes.	Prononciation correcte, malgré quelques erreurs. Pauses longues et fréquentes.	Prononciation correcte malgré quelques erreurs, Effort pour adopter une intonation authentique. Rythme fluide malgré quelques hésitations

*Étant donné le niveau A1, les deux critères ne sont pas distincts.

2.3 Passation

Les deux tâches retenues ont été passées auprès d'un échantillon de 212 élèves de 8e HarmoS (11-12 ans), provenant de 56 classes de toute la Suisse romande (environ huit classes par canton). Les conditions de passation ont été identiques pour chaque élève. La passation de tâches de PO se faisait à la suite d'une série de tâches de

compréhension orale en allemand sur tablette, qui étaient passées par la totalité d'une classe. Avant le début de la passation, les élèves ont réalisé de tâches de type « prise en main » pour se familiariser avec le dispositif, ainsi que les fonctionnalités, notamment écouter, enregistrer. En ce qui concerne l'appropriation de la fonction « enregistrement », presque la totalité des élèves a réalisé les opérations sans problème et très rapidement. Seulement deux élèves ont eu besoin de plus de temps pour se familiariser avec cette fonction (plus de deux minutes).

Après avoir finalisé la partie de compréhension orale, quatre élèves de chaque classe, sélectionné·e·s au hasard, recevaient une notification sur leurs écrans pour réaliser une des deux tâches de PO. En tout, 104 élèves ont réalisé la première tâche et 108 élèves ont réalisé la deuxième tâche.

Les élèves concerné·e·s sortaient de la salle commune pour réaliser une tâche dans une salle séparée et plus calme. Les élèves pouvaient réaliser les tâches de PO de manière indépendante. Un·e adulte était toutefois présent·e dans la salle pour répondre à d'éventuelles questions techniques ou simplement pour rassurer aux élèves. Les élèves ne recevaient ni note ni rétroaction pour la réalisation de ce travail. Les productions ont été notées à postériori à l'aide des deux grilles d'évaluation.

2.4 Correction de tâches

Toutes les productions ont été analysées, avec les deux échelles, par quatre correctrices. Elles se sont familiarisées avec les tâches et les grilles de correction et se sont entrainées avec des productions modèles.

Ensuite, les correctrices ont travaillé de manière indépendante, deux utilisant la grille analytique et deux la grille holistique. Les deux personnes de chaque sous-équipe discutaient les résultats en cas de désaccord, dans le but d'avoir seulement une valeur holistique et une valeur analytique pour chaque production. La focale de l'étude étant sur les différences des grilles et pas dans le comportement des évaluatrices ou évaluateurs, le consensus entre les correctrices était important pour pouvoir comparer les grilles.

3. Résultats

Les résultats des analyses nous ont permis de chercher des éléments de réponse pour nos questions de recherche. Tout d'abord, les scores de grilles analytiques et ceux de la grille holistique ont été associés aux productions des élèves. Ces scores prennent les valeurs 0, 1 ou 2, ce dernier constituant le meilleur résultat possible (voir les tableaux 1 et 2).

Ensuite, en vue de mesurer la cohérence interne de la grille analytique, le coefficient de Cronbach (Cronbach, 1951) a été calculé sur les scores des élèves aux trois critères conformant l'échelle analytique, sans distinction entre les deux tâches. La valeur obtenue étant assez élevée (0.82), nous a permis de confirmer ainsi la cohérence entre les critères qui permettent d'évaluer le construit, ici la compétence en PO.

Afin de répondre à notre première question de recherche – (quel est le lien entre le score à chaque critère de l'échelle analytique et le score de l'échelle holistique ?), une analyse de corrélations de Kendall (1938) a été réalisée entre chaque score des différents critères analytiques – le contenu (ci-après « Contenu »), l'étendue du vocabulaire et la correction grammaticale (ci-après « VocGr »), la fluidité et la prononciation (ci-après « Prononciation ») et le score de l'échelle holistique (ci-après « Holistique »). La figure 1 présente ces corrélations, compte tenu des résultats des productions issues de deux tâches ensemble (212 élèves).

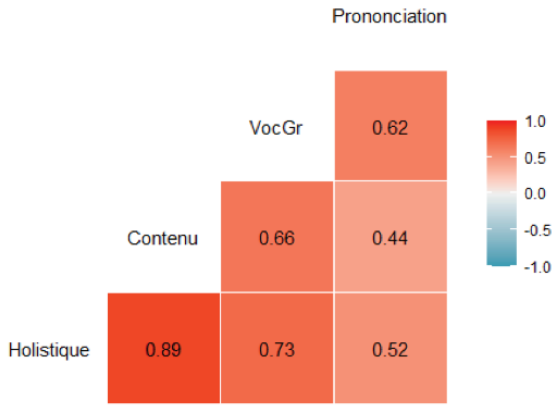
Les corrélations sont positives, importantes entre tous les scores et statistiquement significatives au seuil de 5%. Ce résultat nous permet d'observer que les deux échelles montrent un certain degré de cohérence entre elles.

Toutefois, la corrélation entre le critère « Contenu » et le score de la grille holistique est la plus élevée (0.89), ce qui suggère que la perspective holistique semble couvrir, principalement, ce critère.

Par ailleurs, les corrélations entre les différents critères de la grille analytique sont assez bonnes, ce qui suggère que les critères appartiennent au même construit et qu'ils mesurent néanmoins différentes dimensions de ce dernier.

Figure 1

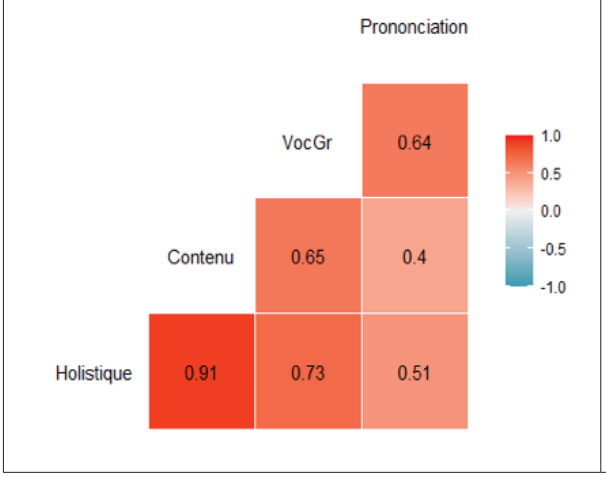
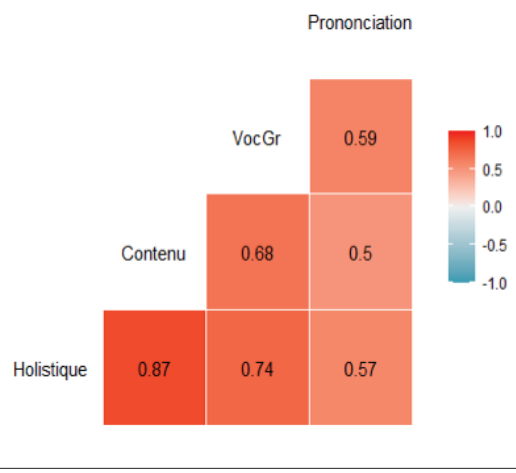
Corrélations entre les scores des critères de la grille analytique et ceux de la grille holistique pour les deux tâches ensemble.



Étant donné que le fait de considérer les deux tâches ensemble pouvait cacher des différences spécifiquement liées aux tâches, nous avons réalisé les mêmes analyses en tenant compte de deux tâches séparément. La figure 2 synthétise les corrélations observées par tâche.

Figure 2

Corrélations entre les scores des composantes de la grille analytique et ceux de la grille holistique, par tâche.

Tâche 1 (effectif élèves : 104)	Tâche 2 (effectif élèves : 108)
	

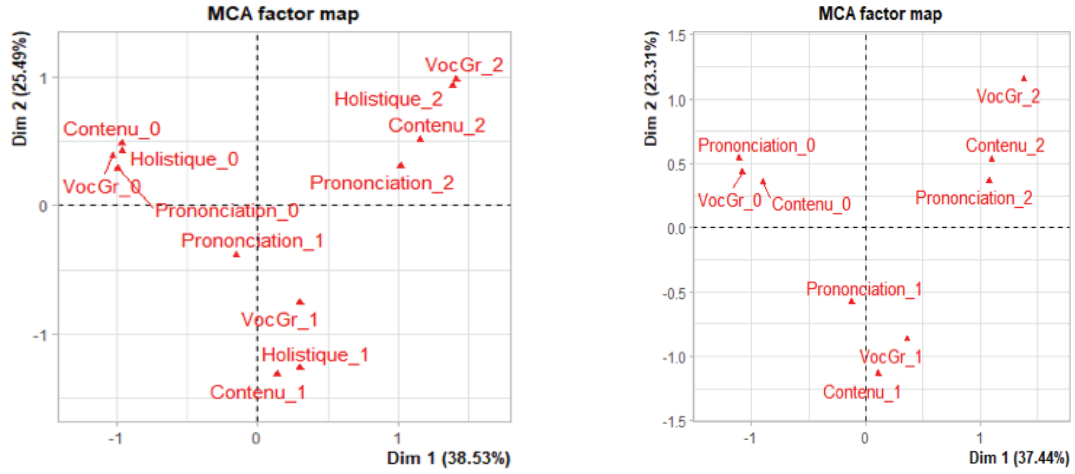
L'examen de la figure 2 fait ressortir de légères différences dans les valeurs de corrélations entre les deux tâches, mais il est toujours très clair que la corrélation entre les scores de la grille holistique et ceux du critère « Contenu » est très importante et plus élevée que les corrélations entre la valeur holistique et les autres critères de la grille analytique.

Pour répondre à notre deuxième question de recherche (si le score holistique peut remplacer l'ensemble des scores de chaque critère de la grille analytique), nous avons appliqué deux méthodes : une analyse de correspondances multiples et une régression logistique polychotomique ordinale. Les deux méthodes ont été utilisées vu le caractère ordinal des scores.

Une analyse de correspondances multiples[2] (Saporta, 2011) a été appliquée tout d'abord sur les quatre variables Holistique, Contenu, VocGr et Prononciation et l'ensemble des scores de 212 élèves. Celle-ci a mis en lumière qu'environ 64% de la variance des données est expliquée par les deux premières dimensions (voir la figure 3, gauche ; la somme des pourcentages donnés sur les deux axes est environ 64%). Ensuite, nous avons appliqué la même analyse, mais en prenant en compte seulement les variables qui correspondent à la grille analytique. Cette fois, environ 61% de la variance des données est expliquée par les deux premières dimensions (voir la figure 3, droite). Nous remarquons que la présence de la variable Holistique apporte peu d'information pour expliquer la variance des données (64% versus 61%). De plus, on note que les catégories des scores de la variable Holistique se situent très proches des catégories des scores de la variable Contenu sur la figure 3 (gauche). Ce fait suggère une redondance entre les deux variables et confirme encore une fois leur lien assez fort que nous avons indiqué auparavant sur la base de la corrélation de Kendall (0.89).

Figure 3.

Analyse de correspondances multiples sur quatre variables (à gauche) et sur trois variables (à droite).



Ensuite, trois modèles de régression logistique polychotomique ordinale[3] (Saporta, 2011) ont été réalisés. Chaque modèle a comme variable indépendante la valeur Holistique et comme variables dépendantes respectivement les valeurs associées aux critères de la grille analytique. Le but de chaque modèle est premièrement de prédire les probabilités d'obtenir un score de 0, 1 ou 2 d'une variable associée à la grille analytique (« Contenu », « VocGr » et « Prononciation » respectivement) en utilisant la variable « Holistique ». Sur la base des probabilités prédites, un deuxième objectif d'un tel modèle est de prédire dans quelle catégorie de scores (0, 1 ou 2) une valeur d'une variable de la grille analytique peut se situer quand on indique la valeur correspondante de la variable « Holistique ». Sur nos données, les modèles nous indiquent les résultats suivants : quand la variable « Contenu » est utilisée comme variable dépendante, seulement 10 valeurs parmi les 212 sont mal classées (c'est-à-dire que pour ces 10 valeurs les scores de la variable « Contenu » ne sont pas les mêmes que ceux prédits par le premier modèle, qui utilise la variable « Holistique » comme prédicteur). On obtient ensuite 56 observations mal classées quand c'est la variable « VocGr » qui est utilisée comme variable dépendante et 96 quand la variable dépendante est la « Prononciation ». On observe ainsi que la variable « Holistique » est capable de bien classer

2 Une analyse des correspondances multiples permet d'explorer le lien entre des variables qualitatives et de réduire leur nombre en créant un nombre plus petit de variables qui résume au mieux les variables de départ.

3 Pour rappel, un modèle de régression classique consiste à mettre en relation une variable à expliquer (la variable dépendante) avec une ou plusieurs variables explicatives ou indépendantes, appelées prédicteurs. Quand la variable dépendante est qualitative, le modèle classique n'est plus adapté et il faut utiliser la régression logistique polychotomique. La régression logistique polychotomique ordinale concerne les situations où la variable dépendante est qualitative et présente plus de deux modalités (catégories) qui peuvent être ordonnées et dont on souhaite tenir compte de l'ordre. Le but d'une telle régression est de prédire les probabilités d'appartenance aux catégories de la variable dépendante, en fonction d'un ou de plusieurs variables indépendantes.

les valeurs de la variable « Contenu », mais moins bien les valeurs des variables « VocGr » et « Prononciation »[4].

De manière générale, ces résultats suggèrent que la grille holistique, bien que cohérente et compatible avec les critères analytiques, ne les couvre pas de la même manière. Par conséquent, la grille holistique n'est pas complètement interchangeable avec la grille analytique et elle ne peut pas la remplacer.

4. Discussion

Dans cet article, nous souhaitions analyser la pertinence et la complémentarité d'une grille analytique et d'une grille holistique conçues pour évaluer la PO en allemand langue étrangère (niveau A1). Pour ce faire, nous nous sommes posé deux questions de recherche qui ont guidé nos réflexions.

La première visait à déterminer quel était le lien entre le score de chaque critère de la grille analytique et le score de la grille holistique. Les résultats obtenus ont mis en évidence des corrélations statistiquement significatives entre les scores de chaque critère de la grille analytique et celui de la grille holistique. Toutefois, les valeurs des corrélations ne sont pas les mêmes pour tous les critères. En effet, en observant les résultats plus en détail, on remarque que c'est le critère de contenu qui a la valeur de la corrélation la plus élevée avec le critère holistique, en se différenciant de manière claire du reste des critères analytiques. Ainsi, la grille holistique semble correspondre principalement au critère de contenu de la grille analytique.

Ce résultat peut surprendre, étant donné que les descripteurs de la grille holistique prenaient en considération l'ensemble de facteurs présents dans la grille analytique et pas uniquement le contenu. Une explication possible de la différence observée dans les corrélations des scores réside dans une forte interdépendance entre les critères analytiques. En effet, l'évaluation des éléments de contenu (critère 1) implique également l'utilisation d'un vocabulaire suffisamment riche pour accomplir la tâche (critère 2). Il est donc possible que, lors du processus de correction, le critère du vocabulaire ait été implicitement inclus dans celui du contenu. Par ailleurs, les corrélations solides entre les différents critères analytiques suggèrent non seulement une forte cohérence interne, mais indiquent aussi que ces critères relèvent d'un même construit, tout en mesurant des dimensions distinctes de celui-ci. Ce constat peut être interprété comme un résultat très positif en faveur de la cohérence interne de la grille analytique.

Une deuxième piste d'explication, en lien avec le point précédent, concerne la manière dont les descripteurs sont présentés et interprétés. En effet, les descripteurs holistiques pour les trois niveaux utilisés commençaient par une référence à la situation de communication et aux éléments de la consigne. Dans la grille analytique, ces éléments-là (situation de communication et consigne) sont opérationnalisés par le critère de contenu. Il est possible que la première phrase de chaque descripteur holistique exerce une influence disproportionnée sur les évaluatrices et évaluateurs, influençant ainsi leur manière de coder. Il en résulterait une perception selon laquelle le respect du contenu constitue l'aspect central de la tâche, tandis que les deux autres critères sont relégués au second plan.

Cette discussion nous amène à notre deuxième question de recherche : le score à la grille holistique peut-il remplacer l'ensemble des scores de chaque critère de la grille analytique ? Ici, la réponse est négative, car la grille holistique, telle que présentée, ne couvre pas l'étendue des trois critères de la grille analytique. Même si les deux grilles restent cohérentes entre elles, nous considérons que les échelles ne sont pas interchangeables.

Si l'on souhaite mettre à disposition des enseignant·e·s une grille holistique efficace, qui réduit le temps de correction, tout en permettant une évaluation fidèle du construit, il apparait nécessaire de repenser la formulation des descripteurs holistiques. Ainsi, des ajustements relatifs à la rédaction des descripteurs, à leur structure et à l'ordre de phrases doivent ainsi être envisagés. Ces aspects mériteraient d'être pris en compte dans de futures expérimentations.

En ce qui concerne l'utilisation des grilles dans le contexte étudié, la grille analytique se distingue par sa plus grande précision. Sa forte cohérence interne, combinée à sa capacité à rendre compte de différentes dimensions du construit évalué, justifie pleinement son utilisation dans le contexte scolaire en Suisse romande. Par ailleurs, comme le souligne également la littérature (Luoma, 2024), elle permet de mieux cerner les points forts et les difficultés des candidat·e·s, constituant ainsi un levier pertinent pour adapter et orienter l'enseignement de manière plus ciblée. Au vu de ces constats, pour ce type de tâches en particulier, la grille analytique apparaît comme l'option à privilégier.

[4] Des résultats plus détaillés sont disponibles auprès des auteures.

Plus largement, l'expérience présentée ici invite à une réflexion sur l'importance d'un calibrage rigoureux des instruments de pondération des productions. Un outil de notation qui ne reflète pas fidèlement le construit visé peut en effet fausser l'interprétation des résultats et, par conséquent, influencer les décisions pédagogiques ou les trajectoires futures des élèves. La principale contribution de cette expérimentation pilote réside donc dans la mise à disposition, via la plateforme *PistEval*, non seulement des tâches validées, mais également d'outils d'évaluation éprouvés et validés. Ceux-ci offrent au corps enseignant un cadre solide pour interroger et améliorer les pratiques de correction et de notation dans le contexte scolaire suisse romand.

5. Références bibliographiques

Alderson, J. C. et Banerjee, J. (2002). Language testing and assessment (Part 2). *Language teaching*, *35*(2), 79–113. https://doi.org/10.1017/S0261444802001751

Bachman, L. F. et Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The modern language journal*, *70*(4), 380–390. https://doi.org/10.2307/326817

Chavez, M. (2007). Students' and teachers' assessments of the need for accuracy in the oral production of German as a foreign language. *The Modern Language Journal*, *91*(4), 537–563. https://doi.org/10.1111/j.1540-4781.2007.00622.

Chen, J., Yang, H. et Han, C. (2022). Holistic versus analytic scoring of spoken-language interpreting: A multi-perspectival comparative analysis. *The Interpreter and Translator Trainer*, *16*(4), 558–576. https://doi.org/10.1080/1750399X.2022.2084667

Chuang, Y. (2009). Foreign language speaking assessment: Taiwanese college English teachers' scoring performance in the holistic and analytic rating methods. *Asian EFL Journal*, *11*(1), 150–173.

Conférence Intercantonale de l'instruction publique de la Suisse romande et du Tessin. (2024a). *Plan d'études romand (PER), version 3.0*. https://www.plandetudes.ch

Conférence Intercantonale de l'instruction publique de la Suisse romande et du Tessin. (2024b). *Objectifs L2, 24*. https://portail.ciip.ch/per/learning-objectives/24

Conseil de l'Europe. (2001). *Cadre européen commun de référence pour les langues; apprendre, enseigner, évaluer.* Les Éditions Didier.

Cronbach, L.J. (1951). *Coefficient alpha and the internal structure of tests*, *Psychometrika*, *16*, 297–334.

Cushing Weigle, S. (2002). *Assessing writing*. Cambridge University Press.

Fulcher, G. (2015). Assessing second language speaking. *Language Teaching*, *48*(2), 198–216. https://doi.org/10.1017/S0261444814000391

Isaacs, T. (2016). Assessing speaking. Dans D. Tsagari et Banerjee (dir.), *Handbook of second language assessment* (p. 131–146). De Gruyter Mouton.

Kendall, M. (1938), *A New Measure of Rank Correlation*, *Biometrika*, 30(1/2), 81–89.

Kim, H. J. (2015). A Qualitative Analysis of Rater Behavior on an L2 Speaking Assessment, *Language Assessment Quarterly*, *12*(3), 239–261. https://doi.org/10.1080/15434303.2015.1049353

Laveault, D. et Grégoire, J. (2014). *Introduction aux théories des tests en sciences humaines*. De Boeck Université.

Lenz, P. et Barras, M. Does teaching chunks and fluency make a difference in migrants' language learning? Dans J.C. Beacco, H.-J. Krumm, D. Little et P. Thalgott (dir.), *The Linguistic Integration of Adult Migrants / L'intégration linguistique des migrants adultes: Some lessons from research / Les enseignements de la recherche* (p. 195–200) De Gruyter Mouton. https://doi.org/10.1515/9783110477498-026

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

Ma, W. (2022). What the analytic versus holistic scoring of international teaching assistants can reveal: Lexical grammar matters. *Language Testing, 39*(2), 239–264. https://doi.org/10.1177/02655322211040020

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

North, B. (2005). Assessing Spoken Performance in relation to the Common European Framework of Reference. *Babylonia,* 2(5), 46–49.

Ounis, M. (2017). A comparison between holistic and analytic assessment of speaking. *Journal of Language Teaching and Research*, *8*(4), 679–688. http://dx.doi.org/10.17507/jltr.0804.06

Park, M. S. (2020). Rater Effects on L2 Oral Assessment: Focusing on Accent Familiarity of L2 Teachers, *Language Assessment Quarterly*, *17*(3), 231–243. https://doi.org/10.1080/15434303.2020.1731752

Roth, M., Ruf, I., Sánchez Abchi, V., Soussi, A. et Weiss, L. (2021). EpRoCom : dispositif romand de mutualisation de tâches évaluatives : premiers constats. *irdp FOCUS, 08.*2021(août), 1–8.

Roth, M. et Ruf, I. (2023). Ressources évaluatives pour les enseignant·es romand-es. *La Revue LEeE,8*, 1–18. ttps://doi.org/10.48325/rleee.008.02

Sánchez Abchi, V., De Pietro, J.F. et Roth, M. (2016). *Évaluer en français : comment prendre en compte la difficulté des items et des textes*. IRDP.

Sánchez Abchi, V., Roth, M. et Matei, A. (2022). Estimer la difficulté des questions en compréhension de l'écrit en français. Vérification empirique d'un modèle théorique. *EJIREF. Évaluer. Journal international de recherche en éducation et formation, 8*(1), 29–46. https://doi.org/10.48782/dzc5py67

Sánchez Abchi, V., Sieber Meylan, S. et Matei, A. (2024). Innover dans l'évaluation de la production orale en langue étrangère. Étude exploratoire sur l'évaluation de l'allemand en Suisse romande. *E-Jiref Évaluer. Journal international de recherche en éducation et formation, 10*(2), 23–42. https://doi.org/10.48782/e-jiref-10-2-23

Saporta, G. (2011). Probabilités, analyse des données et statistique. Éditions TECHNIP
Savard, D. et Sévigny, S. (2007). La méthode de conversion de la cote de rendement au collégial (cotes R) en moyenne cumulative exprimée en pourcentage. *Mesure et évaluation en éducation*, *30*(3), 99–128. https://doi.org/10.7202/1085731ar
Savard, D., Sévigny, S. et Beaudoin, I. (2008). Évaluations à grande échelle de l'écriture: lien entre le score holistique et les composantes de l'écriture. *Canadian Journal of Program Evaluation*, *22*(3), 99–119. https://doi.org/10.3138/cjpe.0022.007
Witzigmann, S. et Sachse, S. (2020). Verarbeitung von Hinweisreizen beim Beurteilen von mündlichen Sprachproben von Schülerinnen und Schülern durch Hochschullehrende im Fach Französisch. *Unterrichtswissenschaft*, *48*, 551–571. https://doi.org/10.1007/s42010-020-00076-6
Yetiş, V. A. (2017). Les grilles d'évaluation critériée pour évaluer des performances: Exemples pour la production ecrite. *Uludağ Üniversitesi Eğitim Fakültesi*, *30*(2), 683–703. http://hdl.handle.net/11452/12963

Mots clés : Évaluation de la production orale ; tâches d'évaluation ; évaluation holistique ; évaluation analytique ; grille d'évaluation

Die Bewertung der mündlichen Sprachproduktion in Deutsch S2: Welche Beurteilungsinstrumente? Vergleich einer holistischen und einer analytischen Beurteilungsraster

Zusammenfassung

Dieser Beitrag untersucht die Relevanz und die Komplementarität von Korrektur- und Beurteilungsinstrumenten für die mündliche Sprachproduktion in einer Fremdsprache. Insbesondere interessieren wir uns für den Vergleich zweier Skalen oder Rastern – einer analytischen und einer holistischen – zur Beurteilung der mündlichen Sprachproduktion in Deutsch als Fremdsprache bei Schülerinnen und Schülern am Ende der Primarstufe (Niveau A1), die in der französischsprachigen Schweiz eingeschult werden. Die mündlichen Äusserungen von 212 Schüler:innen aus der gesamten Westschweiz, die auf der Grundlage von zwei spezifischen Aufgaben erstellt wurden, wurden von vier Beurteiler:innen anhand von zwei verschiedenen Bewertungsrastern bewertet. Die Ergebnisse zeigen die Möglichkeiten und Grenzen der beiden Beurteilungsmethoden auf.

Schlagworte: Bewertung der mündlichen Sprachproduktion; Testaufgaben; holistische Beurteilung; analytische Beurteilung; Beurteilungsraster

La valutazione della produzione orale in tedesco L2: quali strumenti di correzione? Confronto tra una griglia olistica e una griglia analitica

Riassunto

Questo contributo ha lo scopo di esplorare la pertinenza e la complementarità degli strumenti di correzione e valutazione della produzione orale in lingua straniera. In particolare, ci interessiamo al confronto tra due scale o griglie di analisi – una analitica e l'altra olistica – volte a valutare la produzione orale in tedesco come lingua straniera presso alunni della scuola primaria (livello A1) scolarizzati nel contesto francofono della Svizzera romanda. Le produzioni orali realizzate da 212 studenti di tutta la Svizzera romanda, sulla base di due compiti specifici, sono state valutate da quattro giudici, utilizzando due tipi di griglie diverse. I risultati evidenziano i vantaggi e i limiti dei due metodi di valutazione.

Parole chiave: Valutazione della produzione orale; compiti di prova; valutazione olistica; valutazione analitica; griglia di valutazione

The assessment of oral language production in German L2: Which assessment tools? Comparison of a holistic and an analytical assessment grid

Abstract
This contribution aims to explore the relevance and complementarity of tools for correcting and assessing oral production in a foreign language. In particular, we are interested in comparing two scales or analysis grids – one analytical and the other holistic – designed to assess oral production in German as a foreign language among late primary school pupils (level A1) educated in the French-speaking context of French-speaking Switzerland. The oral production of 212 students from all over French-speaking Switzerland, based on two specific tasks, was assessed by four judges using two different types of grids. The results highlight the strengths and limitations of the two assessment methods.

Keywords: Assessment of oral production; assessment tasks; holistic assessment; analytical assessment; assessment grid

Verónica Sánchez Abchi est docteure en linguistique. Elle est chercheuse à l'Institut de recherche et de documentation Pédagogique à Neuchâtel et chargée d'enseignement à l'IUFE-UNIGE. Ses recherches portent sur l'évaluation et l'enseignement de langues.
Institut de recherche et de documentation pédagogique, Faubourg de l'Hôpital 43, CH-2000 Neuchâtel
E-Mail : veronica.sanchez@irdp.ch

Sophie Sieber Meylan est chargée d'enseignement en didactiques de l'allemand à la HEP Vaud (Unité Didactique des Langues et Cultures). Ses intérêts de recherche portent sur l'enseignement et l'évaluation de l'allemand comme langue étrangère.
Haute École Pédagogique Vaud, Avenue de Cour 33, CH-1014 Lausanne
E-Mail : sophie.sieber-meylan@hepl.ch

Alina Matei, docteure en statistique, est chercheuse à l'Institut de recherche et de documentation pédagogique à Neuchâtel et professeure titulaire à l'Université de Neuchâtel. Ses domaines de recherche visent la statistique appliquée et la théorie des sondages.
Institut de recherche et de documentation pédagogique, Faubourg de l'Hôpital 43, CH-2000 Neuchâtel
E-Mail : alina.matei@irdp.ch