Conférence intercantonale
instruction publique et culture
Suisse romande et Tessin

Document de travail

irdp

25.1001
NOVEMBRE 2025

L'évaluation de la compréhension et de la production de l'oral en allemand L2

Défis et contributions du numérique

Verónica Sánchez Abchi, Sophie Sieber et Alina Matei

L'évaluation de la compréhension et de la production de l'oral en allemand L2

Défis et contributions du numérique

Verónica Sánchez Abchi, Sophie Sieber et Alina Matei

Les *Rapports de recherche «Orange»* sont des documents de travail permettant de présenter l'état d'avancement d'une recherche et/ou d'un mandat en cours : délimitation d'un objet ou d'une problématique, description d'une méthodologie de recherche, présentation de résultats intermédiaires. Ils constituent également l'instrument de publication privilégié pour la valorisation et la diffusion de travaux réalisés par des stagiaires ou des assistant.es.

Cet ouvrage applique les rectifications orthographiques de 1990.

Rédaction	Verónica Sánchez Abchi, Sophie Sieber et Alina Matei
Coordination scientifique	Viridiana Marc
Coordination éditoriale	Anne Bourgoz Froidevaux
Relecture	Mathilde Ceylan
Vérification des bibliographies	Isabelle Deschenaux
Mise en page et conception graphique	Doris Penot
Photo de couverture	© Adobe Stock
Édition et diffusion	CIIP – Conférence intercantonale de l'instruction publique et de la culture de la Suisse romande et du Tessin Institut de recherche et de documentation pédagogique Case postale 556 2002 Neuchâtel – Suisse www.ciip.ch documentation@ciip.ch | +41 32 889 86 18

Table des matières

Liste des abréviations

CECR	Cadre européen commun de référence pour les langues
CIIP	Conférence intercantonale instruction publique et culture Suisse romande et Tessin
CO	Compréhension de l'oral
COMEVO	Commission évaluation des objectifs du PER
EC	Épreuves cantonales
EpRoCom	Épreuves romandes communes
Gdid	Groupe de didacticiens et didacticiennes
GRés	Groupe de résonance
IRDP	Institut de recherche et de documentation pédagogique
MER	Moyens d'enseignement romands
MSN	Mathématiques et sciences de la nature
ONAÉ	Outil Numérique d'Apprentissage et d'Évaluation
PE	Production de l'écrit
PO	Production de l'oral
PER	Plan d'études romand
PistEval	Pistes pour l'évaluation

Résumé

Inscrits dans le mandat de l'Institut de recherche et de documentation pédagogique (IRDP), les travaux relatifs à l'évaluation des *objectifs d'apprentissage* du Plan d'études romand (PER), menés sur la période quadriennale 2020-2023, visent à créer une culture commune entre les cantons romands en matière d'évaluation des apprentissages des élèves. Dans cette perspective, ils permettent de se doter d'outils d'analyse et de mettre à disposition des enseignantes et enseignants des matériaux d'évaluation pertinents, validés et fiables. Ce texte fait partie d'une publication plus complète, qui rassemble les travaux réalisés dans plusieurs disciplines pour des élèves de 8e année (11-12 ans).

Ce rapport présente les principaux résultats relatifs à l'évaluation de l'Allemand langue seconde (L2). Il rend compte du processus de conception, d'adaptation et de vérification de l'adéquation de tâches d'évaluation informatisées pour le niveau A1. Les données sont issues de la réalisation de ces tâches par environ 1 000 élèves pour l'évaluation de la *compréhension de l'oral* et de 212 élèves pour l'évaluation de la *production de l'oral*. Les analyses ont porté sur le lien entre les caractéristiques des tâches, les fonctionnalités numériques et les performances des élèves, ainsi que sur la pertinence des différents outils de correction.

Le projet a permis de mieux comprendre les avantages, les limites et les défis liés à l'évaluation sur support numérique et autonome des compétences orales dans une langue étrangère. Les tâches développées seront, à terme, mises à disposition des enseignantes et enseignants sur les PistEval, pages liées au PER en ligne, consacrées à l'évaluation.

1. Cadre général

Les travaux présentés dans ce rapport s'inscrivent dans le cadre du projet EpRoCom, initié depuis 2016 et centré sur l'évaluation des apprentissages des élèves. Plus spécifiquement, ce rapport présente les résultats obtenus lors de la dernière période quadriennale (2020-2023), en ce qui concerne la mutualisation de ressources permettant d'**évaluer les compétences** inscrites dans le Plan d'études romand (PER) (CIIP, 2010/2024).

1.1 Objectif du projet

Depuis plusieurs années, des réflexions quant à l'évaluation des apprentissages des élèves au niveau romand ont été menées par les équipes de l'Institut de recherche et de documentation pédagogique (IRDP), structure scientifique rattachée à la Conférence intercantonale de l'instruction publique et de la culture de la Suisse romande et du Tessin (CIIP). Dès l'adoption du Plan d'études romand (PER) en 2010, l'enjeu majeur a été de définir les qualités et les caractéristiques d'une **évaluation de compétences**, celle-ci ne se résumant pas à évaluer une somme de connaissances, de savoir-faire et de savoir-être, bien que ces éléments participent au développement et à l'expression des compétences des élèves inscrites dans le PER. De ce fait, la mobilisation, la sélection et la combinaison de ces différents savoirs[1], considérés comme des ressources, sont nécessaires pour accomplir une tâche complexe et ainsi développer des compétences (Allal, 1999 ; Rey et al., 2003).

Pour faire face à cet enjeu ambitieux, le projet EpRoCom s'est attaché à mettre à disposition du corps enseignant de Suisse romande des exemples de ressources évaluatives **validées** (utilisables par tous les cantons), **pertinentes** (appropriées par rapport à l'objectif visé), **fiables** (qui renseignent les savoirs objets de l'évaluation), et **basées sur les *objectifs d'apprentissage* du PER**, lesquels sont formulés en termes de compétences à développer.

Pour le programme d'activités 2020-2023, les travaux se sont concentrés sur la 8e année (élèves de 11-12 ans), dans les disciplines suivantes :

- le Français (*production de l'écrit [PE]*),
- l'Allemand langue étrangère (*compréhension de l'oral [CO]* et *production de l'oral [PO]*),
- les Mathématiques (*résolution de problèmes*).

Les concrétisations de ce projet (*cf.* chapitre 1.2) ont pour objectif de participer au développement d'une culture commune romande de **l'évaluation des *objectifs d'apprentissage* du PER**, et par conséquent de **compétences**.

[1] « Savoirs » est ici à prendre au sens large et comprend autant des savoirs encyclopédiques que des savoir-faire ou même des compétences.

1.2 Concrétisations du projet

1.2.1 Pistes pour l'évaluation (PistEval)

Dès la rentrée scolaire 2021, une des principales réalisations du projet a consisté en la création, sur des pages dédiées du site du PER, de *Pistes pour l'évaluation* (*PistEval*)[2]. Ces pages internet proposent des exemples de ressources évaluatives et des clarifications théoriques et didactiques permettant de soutenir le concept d'**évaluation de compétences** (Roth et al., 2021 ; Roth & Ruf, 2024). Elles sont destinées au corps enseignant romand, qui y accède sous identifiant.

Les pages *PistEval* contiennent actuellement, pour des élèves de 8e année, des exemples de ressources évaluatives pour la *compréhension de l'écrit* et la *production de l'écrit* en Français, ainsi que pour la *résolution de problèmes* en Mathématiques. Elles sont accompagnées d'étayages théorico-didactiques et méthodologiques, qui soutiennent d'une part le corps enseignant dans la prise en main des ressources et, d'autre part, mettent en lumière certaines réflexions essentielles à avoir lors du choix ou de l'élaboration de ressources pour évaluer les apprentissages réalisés par leurs élèves. Le statut illustratif des ressources proposées est particulièrement mis en avant, car, pour en permettre une utilisation adéquate, les enseignantes et enseignants doivent veiller à les ajuster afin qu'elles soient en adéquation avec les objectifs poursuivis, l'enseignement dispensé et les apprentissages réalisés dans leurs classes. Les réflexions mises en évidence par les étayages, idéalement réinvesties par le corps enseignant, non seulement en situation d'évaluation, mais aussi dans son enseignement, vont dans le sens d'une **évolution des pratiques évaluatives au service des apprentissages des élèves** (Roth & Ruf, 2024).

1.2.2 Outil Numérique d'Apprentissage et d'Évaluation (ONAÉ)

Au cours de 2022, le projet s'est enrichi d'un apport supplémentaire avec le développement d'une application dédiée à l'apprentissage et à l'évaluation des élèves sur support informatisé, appelée *ONAÉ* (*Outil Numérique d'Apprentissage et d'Évaluation*). Initialement conçue pour évaluer les compétences orales des élèves en Allemand langue étrangère (*compréhension de l'oral* et *production de l'oral*), *ONAÉ* présente la possibilité d'évoluer vers un outil plus polyvalent, destiné à servir le corps enseignant et les élèves. Ainsi, en complément des tâches prévues pour l'évaluation des compétences orales en Allemand, elle a aussi été utilisée, de manière exploratoire, pour évaluer la *résolution de problèmes* en Mathématiques.

ONAÉ est une application web, accessible directement en ligne via un navigateur. Son développement a été conçu pour un support numérique de type tablette, offrant ainsi de nombreux avantages (transport et mise en place facilités, usage intuitif, mobilité du dispositif, etc.). Cette application se compose de deux interfaces distinctes :

- Une **interface de gestion (*back-office*)**, qui permet d'enregistrer les informations relatives aux classes ou aux élèves, de programmer et de gérer en temps réel les passations, ainsi que d'exporter les données collectées.

[2] Les pages internet *PistEval* sont accessibles via la plateforme PER-MER :
https://bdper.plandetudes.ch/barome/francais-8/
https://bdper.plandetudes.ch/barome/maths-8/

- Une **interface élève (*frontend*)**, avec laquelle l'élève interagit. Cette interface propose plusieurs outils numériques, sélectionnés selon les besoins spécifiques de la tâche à accomplir.

Conçues pour être évolutives, ces interfaces pourront ainsi se développer progressivement, suivant les besoins des utilisatrices et utilisateurs.

1.3 Dispositif d'ensemble

Avant la mise à disposition des ressources évaluatives sur les pages *PistEval*, un **processus de validation** est indispensable pour garantir une prise en compte de la recherche et l'adéquation aux besoins de tous les cantons romands (Roth et al., 2021). Ce processus dure en moyenne quatre ans et implique différents groupes de travail qui partagent leur expertise. Ainsi, le groupe de conception de l'IRDP, chargé du projet, est soutenu par des groupes consultatifs, l'un composé d'enseignantes et d'enseignants issus de tous les cantons romands, et l'autre de didacticiennes et didacticiens pour chaque discipline concernée.

Comme convenu dès le début du projet par les organes décideurs, les travaux devaient être menés à partir des épreuves cantonales[3] (EC) produites par les cantons romands. Pour ce faire, certains axes du PER ont été priorisés, en accord avec les instances pilotant le projet. Pour la période 2020-2023, les axes priorisés ont été la *production de l'écrit* en Français, la *compréhension de l'oral* en Allemand[4], et la *résolution de problèmes* en Mathématiques. Suite à une première sélection parmi les EC, les ressources retenues ont fait l'objet d'une expertise tant scientifique, réalisée par les groupes de didacticiennes et didacticiens, que de terrain, par le groupe d'enseignantes et enseignants. Finalement, les ressources ont fait l'objet d'un processus de vérification de leur adéquation, à travers une démarche quantitative lors d'un test pilote réalisé en 2023 auprès des élèves, et/ou par une approche qualitative plus ciblée.

Pour le test pilote de 2023, afin de garantir une certaine représentativité de la population scolaire de la partie francophone de la Suisse, chacun des sept cantons romands a désigné huit classes de 8e année pour prendre part aux passations. Le test pilote s'est donc déployé auprès de **plus de 1000 élèves**, répartis dans 56 classes. Les passations se sont déroulées sur quatre périodes (de 45 ou 50 minutes selon les cantons), les deux premières étant consacrées à la partie informatisée (Allemand et Mathématiques) à l'aide du dispositif *ONAÉ* et les deux suivantes aux tâches papier-crayon (uniquement Mathématiques)[5].

Afin de soutenir la partie opérationnelle du dispositif d'ensemble, des prestataires externes ont également participé au projet :

- une équipe de développement informatique, pour maintenir et développer l'infrastructure technique des pages *PistEval* et de l'application numérique *ONAÉ* ;

[3] La terminologie « épreuves cantonales (EC) » regroupe les différentes appellations utilisées dans les cantons pour désigner les évaluations externes : évaluation cantonale (FR et GE), épreuve commune (JU), épreuve cantonale de compétences (NE), épreuve cantonale de référence (VD) et examen cantonal (VS).

[4] Une partie expérimentale pour la production de l'oral (PO) a également été prévue, mais le manque de matériau cantonal initial a nécessité l'élaboration de tâches spécifiques par le groupe de conception.

[5] Les tâches de Français ont été testées dans les classes seulement en 2019 (*cf*. Roth et al. 2021).

- des administrateurs et administratrices de test pour superviser les passations dans les classes, ainsi que des codeurs et codeuses en charge de la correction et du codage des productions des élèves.

Enfin, le dispositif a nécessité des analyses des résultats des élèves obtenus lors des tests pilotes ainsi que des processus de validation qualitatifs, permettant de déterminer si des tâches[6] d'origines cantonales différentes pouvaient s'adresser à l'ensemble des élèves de Suisse romande. L'introduction du support informatisé, par le biais de l'utilisation d'*ONAÉ*, a considérablement augmenté la quantité de données récoltées. Afin de convertir les informations brutes fournies par *ONAÉ* en données exploitables, un tri préalable suivi de plusieurs méthodes de traitement (*data process*) a été réalisé.

[6] Dans le contexte de l'évaluation, une tâche fait référence à la combinaison d'une consigne, d'une entrée et de réponses pour évaluer un objet (ALTE, 1998).

2. L'évaluation de l'Allemand L2

Pour la période 2022-2023, il a été décidé d'intégrer la discipline Allemand langue seconde (L2) dans le projet EpRoCom. L'objectif de l'enseignement des langues vivantes, dont l'allemand fait partie, est de construire une compétence communicative, ce qui implique le développement des différentes activités de la langue : parler, lire, écrire, écouter, interagir... L'évaluation des langues, au service de leur apprentissage, implique, ainsi, de pouvoir prendre en compte la maitrise et les progrès des apprenantes et apprenants dans ces différentes activités.

Dans le cadre du projet, et pour la discipline Allemand L2, il a été décidé de donner la priorité à l'évaluation des activités de *compréhension* et de *production de l'oral*. Les sections suivantes décrivent les processus de sélection, de conception et/ou d'adaptation qui ont été mis en place, afin de mettre à disposition des enseignantes et enseignants des tâches d'évaluation sur la plateforme *PistEval*.

2.1 Qu'évalue-t-on quand on évalue la compréhension de l'oral ?

La *compréhension de l'oral* en L2 suppose la construction d'une représentation mentale du texte entendu par la personne qui l'écoute. Le processus de compréhension articule des opérations de bas niveau (comme la reconnaissance des sons ou le traitement lexical et syntaxique du langage) et de haut niveau (gestion de la tâche, vérification de la cohérence, élaboration d'hypothèses). Si, en L1, ces processus de bas niveau sont automatisés lors de la compréhension d'un texte oral, en L2, en revanche, ils peuvent mobiliser une quantité excessive de ressources attentionnelles, rendant ainsi l'activité cognitivement très exigeante (pour un résumé, voir Roussel, 2020).

Étant donné que le processus de compréhension est à la fois complexe et difficilement observable, tant pour l'enseignement que pour l'évaluation, il est nécessaire de définir des tâches permettant de l'opérationnaliser et de le mesurer de manière précise et adéquate.

C'est pourquoi, dans le cas de notre projet d'évaluation de la *compréhension de l'oral* (CO), la sélection et l'adaptation des tâches ont débuté par la définition du *construit* d'évaluation. Celui-ci précise « la définition conceptuelle, théorique ou opérationnelle de ce qui est mesuré » (Bouchard et al., 2009, p. 139). Dans ce travail, il s'agit d'évaluer la CO dans un contexte précis – l'école en Suisse romande – et pour une tranche d'âge ainsi qu'un niveau spécifique : la fin du cycle 2 de l'école obligatoire (11-13 ans), visant le niveau A1 du Cadre européen commun de référence pour les langues (CECR) (Conseil de l'Europe, 2001). Dans ce contexte, la définition du construit est ainsi étroitement liée au Plan d'études romand et aux objectifs et attentes fondamentales qu'il spécifie pour l'Allemand L2, pour ce niveau :

> *PER : L2 23 — Comprendre des textes oraux brefs propres à des situations familières de communication…*
>
> *Au cours, mais au plus tard à la fin du cycle, l'élève*
>
> - *comprend, lorsque quelqu'un parle de lui-même et de sa famille, lentement, à l'aide de mots simples*
> - *comprend, dans un magasin, ce que coûte quelque chose, à condition que le vendeur fasse des efforts pour qu'il le comprenne. (CIIP, 2010/2024)*

Les tâches doivent ainsi permettre l'évaluation de la compréhension de textes – et pas uniquement de mots ou de phrases – dans un contexte communicatif proche du vécu des élèves.

L'évaluation dans le contexte scolaire doit, également, tenir compte des expériences d'enseignement et des habitudes des élèves en ce qui concerne les activités et les formats de questionnement (Buck, 2001), ainsi que des thématiques et des situations de communication qui leur sont familières. Dans notre contexte, comme il n'est pas possible de tenir compte des particularités de chaque classe de la Suisse romande et, en même temps, d'établir des généralisations pour l'ensemble du contexte romand, nous avons décidé de nous appuyer sur les tâches présentes dans les moyens d'enseignement romands (MER) existants. Ces derniers, en raison de leur statut officiel, sont en principe utilisés dans toutes les écoles publiques. Dès lors, les tâches d'évaluation doivent pouvoir prendre en considération les contextes « familiers et communicatifs » institutionnalisés dans les MER Allemand.

De même, la familiarité des élèves avec les formats de questionnement des tâches peut influencer la situation d'évaluation et, indirectement, le construit d'évaluation (Buck, 2001). Dans ce sens, et afin de réduire l'impact qu'un format inhabituel peut avoir sur l'évaluation du construit, nous avons privilégié, pour les tâches, les formats de questionnement les plus fréquents dans les différentes évaluations cantonales. Ceux-ci seront discutés dans la section méthodologique.

Une deuxième décision prise dans le cadre du projet (*cf.* cadre général, chapitre 1) a été de privilégier une évaluation sous format numérique. Les études sur l'évaluation numérique des langues vivantes ont mis en évidence autant d'avantages que de points critiques, notamment des questions techniques, les couts, ainsi que l'impact potentiel sur le construit évalué du transfert papier-crayon vers un support numérique (Fulcher, 2015 ; Ockey, 2007 ; Field, 2015). En Suisse, l'évaluation numérique de la CO d'une langue vivante a montré des résultats positifs pour la gestion autonome des tâches (Karges et al., 2021), bien que les élèves en difficulté puissent parfois être surchargés par cette autonomie (Roussel et al., 2008).

Dans les sections suivantes, nous décrirons, dans un premier temps, le processus de sélection et d'adaptation des tâches destinées à évaluer la CO. Dans un deuxième temps, nous présenterons les résultats de la passation de ces tâches sur format numérique et discuterons les possibles effets de ce format sur les résultats des élèves lors du test pilote organisé en 2023.

2.1.1 Les tâches d'évaluation de la compréhension de l'oral

2.1.1.1 Sélection de tâches

Dans le contexte de l'évaluation, une tâche fait référence à la combinaison d'une consigne, d'une entrée et de réponses pour évaluer un objet. Une tâche serait, par exemple, un texte à lire accompagné de questions à choix multiple, auxquelles il est possible de répondre grâce à la

consigne générale (ALTE, 1998). Les tâches qui font l'objet de cette expérience sont issues d'épreuves certificatives cantonales de type papier-crayon. Dans une première étape, un état des lieux sur l'évaluation de la CO en Suisse romande a été réalisé (Sieber, 2021), ce qui a permis de sélectionner des tâches en tenant compte de l'orientation communicative et de l'authenticité des situations présentées. Les tâches avaient été conçues pour des élèves de 8e année (fin du cycle 2 de la scolarité obligatoire, élèves de 11 à 13 ans), scolarisés en Suisse romande, et qui apprennent l'allemand comme première langue étrangère à l'école. Le niveau visé est le A1, soit un niveau débutant. Les tâches sélectionnées ont ensuite été soumises à un processus d'adaptation numérique pour leur passation sur tablette et de vérification de leur adéquation pour évaluer les objectifs visés dans le contexte suisse romand.

2.1.1.2 Adaptation au format numérique

Pour mener à bien le processus de transfert d'un format papier à un format numérique, une réflexion préalable a été menée concernant les fonctionnalités pertinentes (comme l'accès à la consigne d'une tâche en format audio), l'ergonomie de l'application numérique utilisée et les caractéristiques visuelles de la présentation des tâches (Hoffer & Marc, 2025). L'adaptation au format numérique ne devait pas changer le construit d'évaluation, autrement dit ce qu'une tâche est censée évaluer ou mesurer. Le format numérique devait par ailleurs apporter une plus-value en termes de données récoltées par rapport au support papier-crayon (Álvarez, 2016) et rendre possible un travail autonome des élèves.
En complément des outils usuels comme la « gomme » (effacer une réponse) ou la « flèche » (sélection/désélection d'objets), trois fonctionnalités spécifiques ont été intégrées :

- les consignes générales précédant les tâches avaient été enregistrées en français et pouvaient être lues par les élèves et/ou écoutées autant de fois que nécessaire ;
- les élèves pouvaient décider à quel moment lancer l'écoute du document audio ou des consignes ;
- le document audio pouvait être écouté – dans son intégralité – une ou deux fois, avec l'obligation de l'écouter au moins une fois en entier avant de passer à la tâche suivante. Il n'était pas possible d'arrêter les audios en cours d'écoute, afin de permettre l'évaluation de la compréhension de l'audio dans son intégralité et pour éviter une situation de surcharge cognitive de l'élève dans la gestion de la tâche (Roussel et al., 2008).

Les tâches ont été présentées aux élèves sur l'application *ONAÉ*, développée spécifiquement pour cette passation (Hoffer & Marc, 2025).

2.1.1.3 Vérification de l'adéquation des tâches

Les tâches adaptées ont ensuite fait l'objet d'un processus de vérification de leur adéquation, qui consistait à les soumettre à deux groupes d'expertes et d'experts : des spécialistes en didactique de l'allemand et des enseignantes et enseignants expérimentés. L'objectif était d'examiner :

a) l'adéquation des tâches pour mesurer la CO en Allemand (L2) au niveau A1, selon les objectifs du PER pour la fin du cycle 2 et les descripteurs du CECR ;

b) la pertinence du format et du contenu pour des élèves en fin de cycle 2 ;

c) la pertinence d'autres aspects, tels que les consignes, les modalités de réponse et l'utilisation d'images pour évaluer le construit.

Les tâches ont ensuite été améliorées par l'équipe de recherche du projet en fonction de leurs observations.

2.1.1.4 Caractéristiques des tâches

Finalement, 13 tâches ont été retenues pour la passation auprès des élèves : deux tâches dites d'ancrage, qui ont été passées par tous les élèves, et 11 tâches associées à trois scénarios thématiques distincts (trois tâches dans le scénario « Nouvelle école » ; quatre dans le scénario « Au parc » ; quatre dans le scénario « Fête »). Chaque tâche comprenait un fichier audio, une consigne de passation et plusieurs questions (entre trois et sept). La figure 1 présente un exemple de tâche :

Figure 1 : Exemple de tâche pour le scénario « Au parc »

Audios

Les documents audios, créés pour les épreuves cantonales, étaient adaptés au niveau A1 et alignés avec les contenus du cycle 2. Leur adéquation au niveau demandé a été validée par un groupe d'expertes et experts constitué de didacticiennes et des didacticiens de l'allemand (GDid) et à l'aide de l'outil Language Level Evaluator[7].

Consignes

Les consignes étaient formulées en français, langue de scolarisation, et non en allemand (langue cible), afin de garantir leur compréhension et ne pas interférer sur l'évaluation du construit (*cf*. Barras et al., 2016). Cette décision, basée sur la recherche et les expériences préalables, a été également soutenue par la commission de la CIIP qui accompagne le projet (COMEVO : Commission d'évaluation des objectifs du PER).

[7] Le Language Level Evaluator (LLE allemand) est un outil, conçu par une équipe interdisciplinaire, qui analyse le niveau de langue d'un texte à l'aide de divers critères mesurables. Il est disponible ici https://l-pub.com/language-level-evaluator/?lang=de

Format des questions

Les formats des 13 tâches étaient variés : six questions à choix multiple, quatre tableaux à compléter, deux tâches de classement d'informations et une question ouverte où les élèves devaient écrire un mot pertinent au clavier. La fonctionnalité numérique « glisser/déposer » (*drag and drop*) a été mise au service des tâches à choix multiple et d'association et classement d'informations.

2.1.2 Mise à l'épreuve des tâches de compréhension de l'oral

2.1.2.1 Caractéristiques de la passation

Afin de pouvoir observer comment les élèves utilisaient les fonctionnalités numériques et leur impact sur la réussite, les tâches ont été testées, en 2023, par un échantillon d'environ 1000 élèves de 8e année (11-12 ans), provenant de 56 classes de toute la Suisse romande (environ huit classes par canton). La passation a été faite sur tablette à l'aide du dispositif *ONAÉ*, qui permettait aux élèves de gérer de manière autonome avec une latitude dans le temps de passation et le nombre d'écoutes. Le même modèle de tablettes a été utilisé pour l'ensemble des élèves.

Les 13 tâches ont été organisées en cinq séries différentes. Chaque série comprenait les mêmes deux tâches d'ancrage, positionnées de manière identique pour les cinq séries, et trois tâches liées à un scénario, apparaissant dans des ordres différents d'une série à l'autre. Cette organisation a permis de tester de manière comparable les 13 tâches, sans qu'aucun élève n'ait à les résoudre toutes. Chaque tâche a été passée par environ 210 élèves, à l'exception des tâches d'ancrage qui ont été passées par la totalité des élèves.

2.1.2.2 Procédure d'analyse des données et principaux résultats

Pour chaque tâche, nous avons analysé les variables suivantes : le temps de réalisation, le nombre d'écoutes de la consigne et le nombre d'écoutes de l'audio en allemand effectués par chaque élève. Notons que ces variables sont spécifiques aux tâches et non pas aux questions posées dans les tâches. Ainsi, pour pouvoir analyser le lien entre la réussite des élèves et les différentes variables, un score par élève et par tâche a été construit, qui est égal au nombre de questions réussies par tâche. Ce score est donc une mesure de la réussite. Nous avons ensuite calculé pour chaque tâche les corrélations entre :

1. le score par tâche et le nombre d'écoutes de l'audio en allemand (une ou deux fois) ;
2. le score par tâche et le nombre d'écoutes de la consigne ;
3. le score par tâche et le temps de réalisation d'une tâche.

Les principaux résultats obtenus sont donnés ci-après.

2.1.3 Principaux résultats : évaluation de la compréhension de l'oral

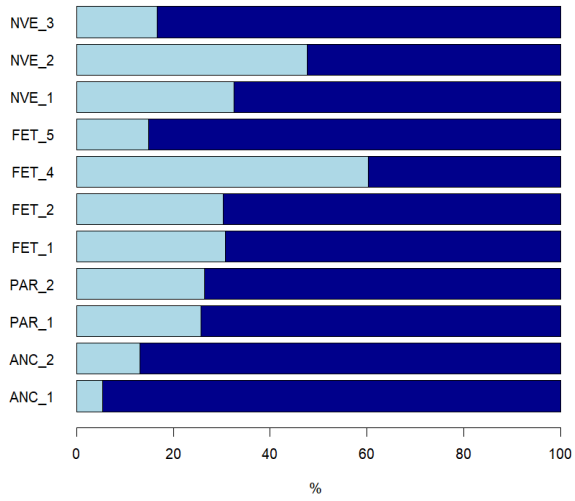
Dans cette section, nous présentons les principaux résultats concernant les analyses qui ont suivi la passation des 13 tâches finalement retenues. Dans les figures ci-dessous, les tâches sont représentées par les codes suivants :

- ANC_1 et ANC_2 pour les deux tâches d'ancrage ;
- PAR_1, PAR_2, PAR_3 et PAR_4 pour les tâches associées au scénario « Au parc » ;
- FET_1, FET_2, FET_3, FET_4 et FET_5 pour les tâches associées au scénario « Fête » ;
- NVE_1, NVE_2 et NVE_3 pour celles du scénario « Nouvelle école ».

2.1.3.1 Lien entre la gestion autonome de l'écoute de l'audio et la réussite

Pour analyser cet aspect-là, nous avons tout d'abord calculé le taux d'une ou deux écoutes par tâche (*cf*. figure 2). À savoir que deux tâches demandaient un changement de fenêtre à plusieurs reprises et, à chaque fois, les élèves pouvaient relancer l'écoute de l'audio pour des questions indépendantes. Pour ces tâches-là, le nombre d'écoutes a été nécessairement beaucoup plus élevé et, par conséquent, elles n'ont pas pu être prises en compte dans cette analyse.

Figure 2 : Taux d'écoutes de l'audio par tâche



Légende : une écoute (bleu clair), deux écoutes (bleu foncé)

Les taux d'une et de deux écoutes varient selon la tâche, sans qu'il soit possible d'affirmer qu'un taux est systématiquement plus élevé qu'un autre. Cependant, certaines tâches semblent mobiliser davantage l'attention et l'écoute des élèves : les tâches d'ancrage (ANC_1 et ANC_2) avec respectivement 95% et 87% de taux de deux écoutes, ainsi que FET_5 (85%) et NVE_3 (83%). Ces résultats invitent à examiner d'autres caractéristiques des tâches (format et formulation des questions, débit de l'audio, etc.) pour mieux comprendre ces différences.

Par ailleurs, la position des tâches d'ancrage, systématiquement situées en début de série, pourrait également jouer un rôle : au début de la passation, les élèves semblent plus enclins à écouter l'audio deux fois que lors des tâches suivantes.

Nous avons ensuite examiné la corrélation entre le score obtenu par l'élève et le nombre d'écoutes (une ou deux fois) séparément pour chaque tâche, sans trouver de lien significatif entre ces variables pour aucune des tâches. L'absence de corrélation suggère qu'écouter l'audio deux fois plutôt qu'une n'a pas de lien avec la réussite. Toutefois, comme noté, les caractéristiques des tâches et leur position pourraient avoir un lien avec le fait d'écouter l'audio une fois ou deux fois.

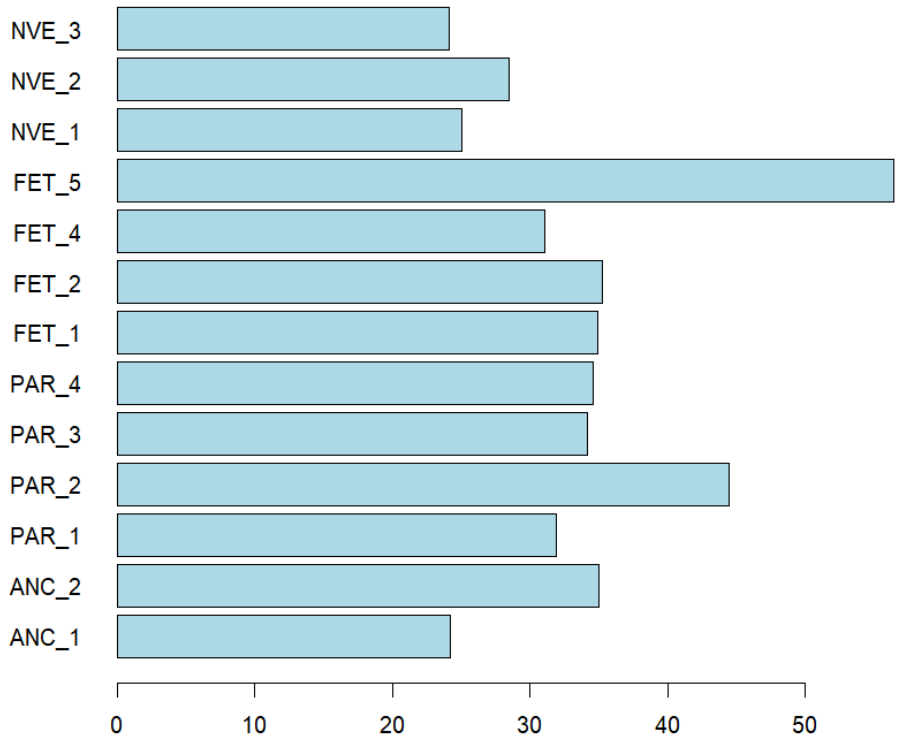
2.1.3.2 Lien entre l'écoute autonome de la consigne et la réussite de la tâche

La consigne de chaque tâche était disponible à la fois en version écrite et en version audio. Les élèves avaient la possibilité de ne pas l'écouter et de se limiter à sa lecture, ou bien d'écouter la consigne autant de fois que souhaité. Le nombre d'écoutes de la consigne par tâche varie de zéro à six, mais la valeur « zéro » reste prédominante pour toutes les tâches, indiquant que la fonctionnalité « écoute de la consigne » est peu utilisée par les élèves. Par ailleurs, aucune corrélation significative n'a été observée entre le nombre d'écoutes de la consigne et le score des élèves pour chaque tâche. Ces résultats suggèrent que les élèves préfèrent se limiter à lire les consignes ou, dans certains cas, même à les ignorer.

2.1.3.3 Lien entre le temps de réalisation et la réussite de la tâche

Comme dans d'autres recherches (par exemple, OECD, 2015), la variable « temps de réalisation » a généré des valeurs extrêmes, justifiant l'utilisation du temps médian de réalisation par tâche (en secondes), plutôt que du temps moyen. La figure 3 présente le temps médian relatif par tâche. Afin de rendre possible la comparabilité entre les tâches, le temps médian relatif a été obtenu à partir du temps médian de réalisation, exprimé en secondes, divisé par le nombre de questions par tâche.

Figure 3 : Temps médian de réalisation relatif par tâche



Pour étudier le lien entre le temps de réalisation et la réussite, nous avons calculé, pour chaque tâche, la corrélation entre les scores des élèves (comme une mesure de réussite) et le temps médian de réalisation en secondes. Comme pour les autres variables analysées, aucune corrélation significative n'a été trouvée entre les scores obtenus par les élèves et le temps de réalisation, tâche par tâche.

Bien qu'il soit difficile d'établir un lien entre le temps de réalisation et les résultats des élèves, le temps médian relatif par tâche peut fournir des indications sur les caractéristiques des tâches et leur difficulté. Comme mentionné précédemment, des éléments tels que le format des questions, leur formulation ou encore le débit de l'audio pourraient avoir un lien avec le temps de réalisation.

2.1.4 L'évaluation de la compréhension de l'oral : constats et perspectives

Ce travail sur l'évaluation de la CO avait une double visée. D'une part, il s'agissait de montrer les processus de sélection et d'adaptation de la présentation des tâches pour évaluer la CO en Allemand, au format numérique (Sánchez Abchi et al., 2025). D'autre part, il était question d'explorer l'impact de certaines fonctionnalités du format numérique sur les résultats des élèves.

En ce qui concerne la première visée, les processus mis en œuvre ont confirmé que le développement de la présentation des tâches au format numérique ne se limite pas à un simple transfert de tâches papier-crayon vers un format tablette. En effet, les activités de classement ou de tri semblent facilitées avec certaines fonctions, comme « glisser/déposer » (*drag and drop*). Pourtant, le format numérique (comme expérimenté dans ce cas) reste limité pour travailler de manière fine et stratégique la *compréhension de l'oral* en Allemand, car, avec les ressources techniques à disposition, il ne permet pas des interactions, comme c'est le cas en classe entre pairs ou avec le corps enseignant. Par ailleurs, pour chaque tâche, une analyse approfondie de l'impact des fonctionnalités sur le construit de l'évaluation devrait accompagner les processus évoqués ci-avant, afin de s'assurer que les caractéristiques des tâches sont correctement transposées en format numérique.

Quant à notre deuxième visée, les résultats n'ont pas révélé de lien entre l'utilisation des fonctionnalités numériques proposées et la réussite des élèves. Nous avions également émis l'hypothèse que la fonctionnalité « écoute des consignes » pourrait soutenir davantage les élèves ayant des difficultés de lecture. Pourtant, le faible nombre d'élèves ayant des besoins particuliers dans notre échantillon a limité la possibilité de vérifier cette hypothèse.

En outre, le nombre d'écoutes de l'audio (une ou deux fois) n'a pas non plus montré de lien avec les résultats des élèves, ce qui est conforme aux observations d'études antérieures (Pothier et al., 2000). La gestion autonome de l'audio n'a pas non plus mis en évidence un lien avec la réussite des élèves en CO, confirmant les résultats d'études précédentes (Roussel, 2011; 2014; Eberharter et al., 2023). Cela pourrait toutefois être expliqué par le fait que, si l'autonomie peut être un atout pour certains élèves, pour d'autres, notamment ceux et celles en difficulté, elle peut engendrer une situation de surcharge cognitive (Roussel et al., 2008).

Nous avons constaté que certaines tâches nécessitent une écoute répétée des consignes ou de l'audio, suggérant que cela pourrait être davantage lié aux caractéristiques intrinsèques des tâches qu'à l'impact des fonctionnalités numériques, comme les observations d'autres études l'indiquent (Bessonneau et al., 2015). Dans certains cas, les fonctionnalités numériques peuvent motiver les élèves (Seifert & Paleczek, 2022) ou réduire l'appréhension face à l'évaluation (Eberharter et al., 2023), mais la réussite semble plutôt dépendre des caractéristiques des tâches – type de question, informations requises, opérations cognitives mobilisées (Sánchez Abchi et al., 2022) – et des caractéristiques des élèves, aspects non examinés dans la présente étude.

Bien que, dans le cadre de notre expérimentation, les fonctionnalités numériques ne semblent pas apporter de valeur ajoutée à la compréhension de leur utilisation par les élèves en CO, il est possible que l'effet spécifique du format numérique soit dilué ou masqué par d'autres facteurs. En effet, les tâches présentent des caractéristiques variées (niveau de difficulté, type d'activités, opérations cognitives mobilisées), ce qui ne permet pas d'isoler clairement le lien entre les fonctionnalités numériques et les résultats des élèves. Toutefois, un lien potentiel avec la motivation et la gestion de l'anxiété ne peut être exclu et mériterait d'être exploré dans de futures recherches. Il serait également pertinent d'étudier d'autres fonctionnalités, comme l'adaptabilité des tâches.

2.2 Qu'évalue-t-on quand on évalue la production de l'oral ?

Comme pour la CO, l'évaluation de la *production de l'oral* (PO) repose sur la définition du construit d'évaluation, en fonction du niveau scolaire retenu : la fin du cycle 2 de la scolarité obligatoire. Le but était d'évaluer l'objectif ainsi formulé dans le PER : « Présentation de soi, de sa famille ou d'une tierce personne (nom, âge, provenance, domicile, école, emploi du temps, hobbys) », ce qui correspond au descripteur de PO du CECR (A1.2) « [l'apprenante ou l'apprenant] peut produire des expressions simples isolées sur les gens » (Conseil de l'Europe, 2001).

Pour ce faire, il était nécessaire de mobiliser des tâches adéquates. Dans les évaluations de PO à grande échelle, les tâches visent à susciter la production de discours variés, en interaction ou de manière individuelle, sur des thématiques pouvant être choisies ou imposées, avec ou sans temps de préparation, etc. (North, 2005 ; OECD, 2021).

Divers facteurs, comme les matériels utilisés (images, vidéos, etc.), le temps de planification, la complexité cognitive et le degré de familiarité avec la tâche, peuvent influencer la difficulté de la tâche de PO (voir synthèse chez Luoma, 2004). Bien qu'estimer le niveau de difficulté reste complexe, le nombre et la complexité des informations à inclure dans une production semblent jouer un rôle déterminant. Par ailleurs, le processus d'évaluation/correction ou notation des productions soulève plusieurs questions, notamment quant à l'importance relative de la précision linguistique, des aspects grammaticaux, de la prononciation, de l'intonation et de la fluidité par rapport à l'intelligibilité (North, 2005 ; Chavez, 2007 ; Isaacs, 2016 ; Park, 2020). Ces éléments illustrent la difficulté de concevoir des tâches d'évaluation pertinentes et adaptées à différents contextes.

Les pratiques d'évaluation de la PO ont évolué ces dernières décennies, grâce aux nouvelles technologies. Il est désormais possible d'enregistrer et d'évaluer la PO de manière semi-directe, sans interaction humaine, via des dispositifs d'enregistrement (Quian, 2009). Toutefois, cette modalité reste peu explorée pour l'évaluation de la PO en continu et s'adresse principalement à un public adulte (*cf*. Sánchez Abchi et al., 2024). De plus, il n'existe pas de recherches concluantes sur les différences entre une évaluation réalisée sur support numérique et une évaluation en présence d'un examinateur ou d'une examinatrice, ni sur l'impact du format numérique sur le construit évaluatif (Fulcher, 2015 ; Grapin & Sayac, 2022).

Dans le cadre du projet, nous examinons l'évaluation de la PO en Allemand L2 dans le contexte scolaire romand, en utilisant le format numérique et en autonomie (Sánchez Abchi et al., 2024). À cette fin, nous avons conçu deux tâches accompagnées de grilles d'évaluation. Les sections suivantes détaillent leur conception, la vérification de leur adéquation pour évaluer les objectifs

d'apprentissage du PER et leur mise à l'épreuve, ainsi que l'analyse de leur pertinence et efficacité.

2.2.1 Méthodologie

2.2.1.1 Élaboration et caractéristiques des tâches

Deux tâches d'évaluation de la PO en continu – sans interaction – ont été élaborées. Pour ce faire, nous avons d'abord examiné la littérature concernant l'évaluation de la PO (Luoma, 2004 ; Goh, 2016), ainsi que les tâches existantes dans les examens disponibles (examens du Goethe, par exemple) et dans les moyens d'enseignement officiels (*Der Grüne Max* [Endt et al., 2014] et *Junior* [Endt et al. 2017]). Nous avons tenu compte, d'une part, des objectifs d'apprentissage du PER pour l'Allemand L2, pour la fin du cycle 2, et, d'autre part, des descripteurs du CECR pour le niveau concerné (A1).

La première tâche (« Nouvelle école ») demandait à l'élève de se présenter et de présenter son école, dans le but de participer à un concours. Pour la deuxième tâche (« Au Parc »), l'élève devait se mettre dans la peau de quelqu'un qui, en excursion avec d'autres personnes, s'était perdu dans un parc d'attractions et devait donner des informations personnelles et décrire la ou les personnes l'accompagnant dans cette excursion.

Les deux tâches étaient présentées sur tablette, comme c'est le cas de l'étude sur l'évaluation de la CO. Les élèves géraient de manière indépendante l'écoute des consignes en langue de scolarisation, ce qui permettait de contourner les obstacles liés à la compréhension des instructions (Barras et al., 2016). Les élèves pouvaient réaliser jusqu'à trois enregistrements, écouter leurs productions et les valider ou recommencer.

Avant la passation à grande échelle, les tâches ont été testées auprès de 14 élèves, qui ont également été interviewés après la passation, pour améliorer les aspects techniques et pédagogiques. Les retours ont servi à ajuster les tâches.

2.2.1.2 Vérification de l'adéquation des tâches

La procédure de vérification de l'adéquation des tâches de PO – faite à l'aide de deux groupes d'expertes – a été la même que celle mise en place pour les tâches de CO.

2.2.1.3 Caractéristiques de la passation

Les deux tâches retenues ont été testées en 2023 sur un échantillon de 212 élèves de 8^{e} année (11-12 ans), provenant de 56 classes de toute la Suisse romande (environ huit classes par canton). La passation des tâches de PO avait lieu à la suite des tâches de CO sur tablette. Si ces dernières étaient passées par la totalité de la classe, la PO n'était réalisée que par quelques élèves. En effet, après avoir finalisé la partie de CO, quatre élèves de chaque classe, sélectionnés au hasard, recevaient une notification sur leur écran qui leur indiquait de réaliser la tâche de PO. En tout, 104 élèves ont ainsi passé la tâche « Au Parc » et 108 élèves la tâche « Nouvelle école ».

Les élèves concernés sortaient de la salle commune pour réaliser la tâche de PO dans une salle séparée, afin de ne pas gêner les autres élèves poursuivant le test. Dans cette salle, les élèves ont d'abord testé des actions pour se familiariser avec les fonctionnalités numériques liées à l'enregistrement, avant de réaliser les tâches de PO de manière indépendante. Une personne

adulte était toutefois présente dans la salle pour répondre à d'éventuelles questions techniques ou simplement pour rassurer les élèves.

2.2.1.4 Caractéristiques de la grille d'évaluation

Chaque PO était ensuite notée par quatre juges à l'aide de deux grilles : une grille d'évaluation analytique et une autre de type holistique.

La grille – ou échelle – analytique prenait en considération trois critères liés : le contenu, l'étendue du vocabulaire et la correction grammaticale, la fluidité et la prononciation (Sánchez Abchi et al., à paraitre). La grille analytique présentait trois bandes correspondant à différents niveaux de performance : niveau plus bas : score 0 ; niveau intermédiaire : score 1 ; niveau plus élevé : score 2. Le tableau 1 présente l'échelle analytique :

Tableau 1 : Échelle analytique

Critère/ Score	0	1	2
Contenu (tâche)	La production correspond partiellement à la consigne (plusieurs informations (plus de 3) 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 (1 à 3). 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

**Vocabulaire et correction grammaticale : étant donné le niveau A1, les deux critères ne sont pas distincts.*

La grille – ou échelle – holistique permettait d'attribuer un score global, assigné par les juges à chaque production d'après un des trois niveaux de performance qui décrivaient la performance : 0 (performance faible), 1 (performance moyenne) et 2 (performance élevée) ; chaque niveau prenant en considération un ensemble de critères. Le tableau 2 suivant présente la grille holistique :

Tableau 2 : Échelle holistique

Score	Description
2	La production est claire, bien structurée et développe tous les points de la consigne. Malgré des erreurs mineures aux niveaux 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.

2.2.2 Principaux résultats : évaluation de la production de l'oral

Les résultats s'organisent en deux sections. Dans la première partie (2.2.2.1), nous présentons les résultats des élèves qui nous permettent d'examiner les caractéristiques des tâches et de comparer leur niveau de difficulté. Dans la deuxième partie (2.2.2.2), nous présentons les résultats permettant d'analyser les deux grilles d'évaluation et leur complémentarité.

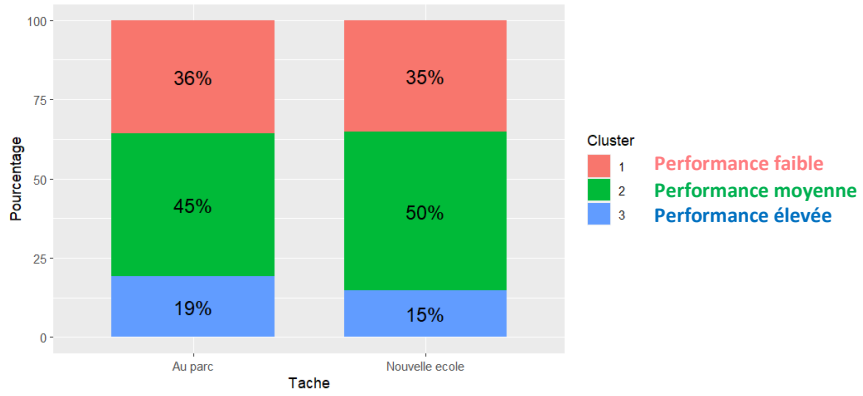
2.2.2.1 Résultats concernant la comparaison de tâches

Les deux tâches conçues pour l'expérimentation étaient jugées « adéquates » pour l'évaluation des contenus du PER par les deux groupes d'expertes et d'experts (groupe de didacticiennes et didacticiens et groupe d'enseignantes et enseignants). Cependant, la tâche « Au parc » exigeait que les élèves prennent en compte davantage d'éléments que « Nouvelle école » : en plus de la présentation personnelle, elles et ils devaient décrire une troisième personne. Par ailleurs, la thématique de la description de l'école, abordée dans « Nouvelle école », est plus courante que la description physique d'une personne dans les moyens d'enseignement romands, dont nous nous sommes inspirées pour la conception des tâches. Nous avons donc formulé l'hypothèse que la tâche « Au parc » serait plus difficile à réaliser.

Afin d'analyser les facteurs qui pourraient avoir un impact sur la difficulté des tâches, nous avons comparé les deux tâches lors de la passation auprès des élèves. Pour ce faire, les productions des 212 élèves ont été évaluées à l'aide de la grille analytique. Trois variables, correspondant aux scores de trois critères de cette grille, ont été utilisées : contenu (critère a), étendue du vocabulaire (critère b) et prononciation (critère c). Une analyse des correspondances multiples suivie d'une analyse en *clusters* (pour plus de détails, voir Sánchez Abchi et al., 2024) a ensuite été réalisée sur ces trois variables. L'analyse en *clusters* a mis en évidence trois groupes d'élèves

selon leurs performances : le *cluster* 1 (les élèves avec des performances faibles), le *cluster* 2 (les élèves avec des performances moyennes) et le *cluster* 3 (les élèves avec des performances élevées). Cette analyse a permis ensuite une comparaison des résultats des deux tâches (*cf*. figure 4).

Figure 4 : Distribution par cluster pour chaque tâche



Comme indiqué par la figure 4, les pourcentages correspondant aux trois niveaux de performance sont similaires pour les deux tâches. Ainsi, nos analyses ont permis de montrer que les deux tâches présentent une difficulté comparable, contrairement à l'hypothèse de départ. La difficulté d'une tâche semble interagir avec le niveau de maitrise de l'allemand, ce qui pourrait expliquer ces résultats : les élèves en difficulté rencontrent des obstacles dans les deux tâches, tandis que celles et ceux à l'aise en allemand réussissent indépendamment de la tâche.

2.2.2.2 Résultats concernant la comparaison des grilles

Concernant l'analyse des grilles d'évaluation, nous avons examiné les caractéristiques et la complémentarité des deux outils : la grille analytique et la grille holistique. Plus précisément, nous avons posé deux questions de recherche :

a) Existe-t-il un lien entre le score de chaque critère de l'échelle analytique et celui de l'échelle holistique ?

b) Le score holistique peut-il remplacer l'ensemble des scores analytiques ?

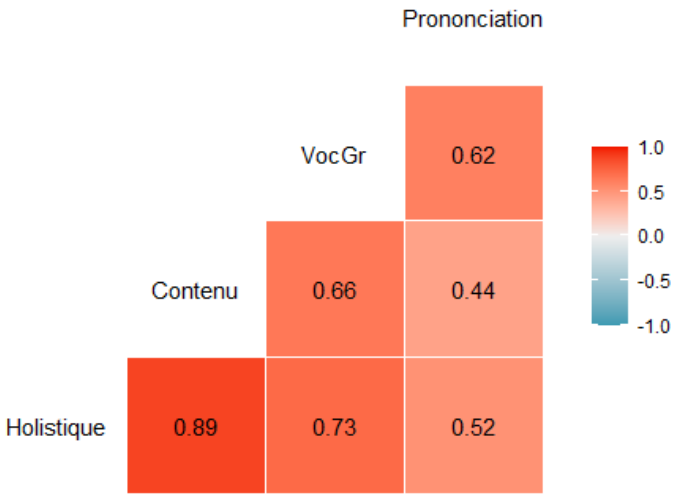
Pour y répondre, quatre chercheuses ont évalué les productions des élèves à l'aide des deux grilles. Après s'être familiarisées avec les tâches et s'être entrainées sur des productions modèles, elles ont travaillé en binômes : deux avec la grille analytique, deux avec la grille holistique. En cas de désaccord, les évaluatrices discutaient pour attribuer un score unique par grille à chaque production.

Tout d'abord, la cohérence interne de la grille analytique a été vérifiée par le coefficient de Cronbach, calculé sur les trois critères de l'échelle (toutes tâches confondues). La valeur obtenue (0,82) confirme la bonne cohérence entre les critères.

Ensuite, pour examiner le lien entre les scores des critères de la grille analytique et les scores de la grille holistique, des corrélations (de Kendall) ont été calculées. Une analyse globale sur l'ensemble des productions des 212 élèves et les deux tâches ensemble a révélé des corrélations

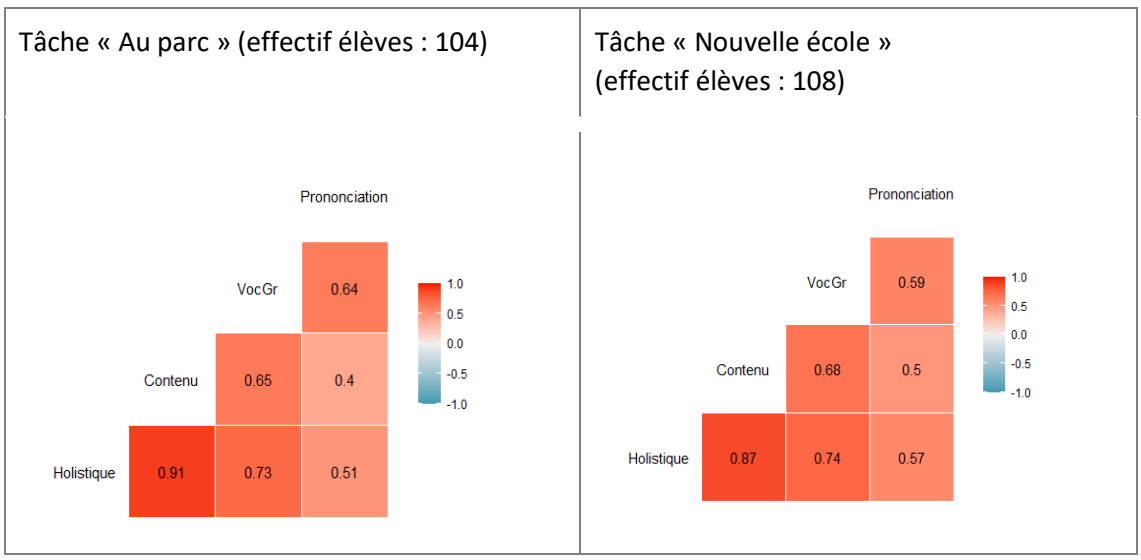
positives et significatives entre tous les scores, indiquant un lien entre les deux grilles. Ces résultats sont présentés dans la figure 5.

Figure 5. Corrélations entre les scores des critères de la grille analytique et ceux de la grille holistique pour les deux tâches ensemble.



Comme on le voit dans la figure 5, la corrélation la plus forte (0,89) concerne le critère *Contenu* et le score holistique, suggérant que l'évaluation holistique est liée principalement à ce critère. Comme l'analyse globale des deux tâches ensemble pouvait masquer des différences liées aux tâches, nous avons répété l'analyse séparément pour chacune d'elles. La figure 6 illustre ces corrélations pour chaque tâche séparément. Les résultats confirment que la corrélation entre le critère *Contenu* et le score holistique reste systématiquement la plus élevée (0,91 et 0,87 respectivement).

Figure 6 : Corrélations entre les scores des critères de la grille analytique et holistique pour les deux tâches[8]



[8] Les critères de la grille analytique considérés étaient « contenu », « étendue du vocabulaire et correction grammaticale » (VocGr) et « fluidité et prononciation » (Prononciation), chacun avec trois scores. La grille holistique présente trois scores.

Pour répondre à notre deuxième question de recherche (à savoir : le score holistique peut-il remplacer l'ensemble des scores de chaque critère de la grille analytique ?), nous avons appliqué deux méthodes adaptées au caractère ordinal des scores : une analyse de correspondances multiples et une régression logistique polychotomique ordinale (pour plus de détails, voir Sánchez Abchi et al., à paraitre). Les résultats obtenus suggèrent que la grille holistique, bien que cohérente et compatible avec les critères de la grille analytique, 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.

Par ailleurs, 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 candidates et des candidats, 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 apparait comme l'option à privilégier.

2.2.3 L'évaluation de la production de l'oral : constats et perspectives

La présente étude montre la complexité liée à la conception de tâches d'évaluation valides et pertinentes pour un contexte particulier. Notre procédure de vérification de l'adéquation des tâches de PO, à la fois grâce à des analyses d'expertes et d'experts et à la mise à l'épreuve des tâches auprès des élèves, nous a permis d'examiner les différents facteurs concernés. L'étape d'analyse a été nécessaire pour mieux saisir le construit que l'on souhaitait évaluer, ainsi que l'impact des facteurs tels que la formulation de la consigne, la présentation des informations et les implications liées au support. L'étape de comparaison des tâches via la procédure statistique, après la passation, a fourni, de son côté, des informations importantes sur les caractéristiques des tâches et nous a permis de conclure que les deux tâches semblaient comparables en termes de difficulté de contenu.

Ces tâches, jugées adéquates, peuvent ainsi être intégrées à la plateforme romande *PistEval* agrémentées d'un étayage didactique. Leur passation autonome sur ordinateur offre un gain de temps et optimise les ressources, facilitant l'évaluation de la *production de l'oral*, souvent perçue comme chronophage par le corps enseignant.

Pourtant, certains défis subsistent. L'autonomie dans l'enregistrement ne remplace pas nécessairement le temps de préparation habituel, et l'absence d'interaction peut affecter la performance, notamment chez les jeunes élèves. De plus, la familiarité avec l'outil numérique joue un rôle : bien que la prise en main ait été prévue, des écarts d'aisance subsistent. Des recherches montrent qu'une passation autonome peut induire une surcharge cognitive, défavorisant les élèves en difficulté, qui bénéficieraient davantage du soutien d'un enseignant ou d'une enseignante (Roussel, 2020).

En ce sens, il serait intéressant d'explorer l'impact que la gestion autonome de la PO peut avoir au niveau psychologique. Cet aspect, qui avait été soulevé par les groupes de validation, a été compensé par la présence d'un adulte dans la salle, qui pouvait, par sa présence, rassurer l'élève lors de la réalisation de la tâche. Pourtant, pour mieux estimer l'impact de la gestion autonome face à la tablette, il faudrait pouvoir comparer la même tâche dans des conditions de passation non numérique.

En outre, dans le cadre du projet, nous avons testé la cohérence et la complémentarité de deux grilles d'évaluation : l'analytique et l'holistique. En partant du principe que la grille holistique pourrait être plus facile à utiliser, parce qu'elle présente moins de critères, on a souhaité explorer comment elles permettaient de prendre en compte les caractéristiques des textes. Les résultats nous ont montré que, à l'heure actuelle, il ne semble pas possible de remplacer la grille analytique par l'holistique, parce que – même si les deux grilles sont cohérentes entre elles – l'analytique reprend surtout un seul critère de la grille analytique. La correction de productions orales, toujours considérée comme une difficulté par le corps enseignant, reste un défi pour l'évaluation à grande échelle.

Enfin, cette étude reste limitée : des tests à plus grande échelle sont nécessaires pour affiner ces conclusions. Néanmoins, cette expérience a permis d'élaborer un dispositif novateur de vérification de l'adéquation des tâches de PO et de mettre en lumière les enjeux liés à l'évaluation numérique de la *production de l'oral*.

3. Discussion

La mise en place d'un outil informatique pour l'évaluation des langues étrangères en Suisse constitue un apport, notamment en ce qui concerne la possibilité de garantir des situations d'évaluation similaires et comparables dans des contextes différents. Une évaluation sur support informatisé permet également de gagner du temps lors de la passation et d'avoir ensuite accès à une quantité de métadonnées susceptibles d'aider à mieux suivre les élèves, en ciblant plus précisément les remédiations utiles. En ce sens, l'évaluation sous forme numérique peut constituer une contribution intéressante à l'évaluation des compétences linguistiques à grande échelle.

Pourtant, comme nous le soulignons également dans notre travail, le numérique « tout seul » ne détermine pas les résultats d'apprentissage ou d'évaluation et son effet est souvent plutôt modeste et bien moins spectaculaire que ce que l'imaginaire social lui attribue (Roussel, 2020). Il s'agit d'un support qui doit faciliter le traitement de l'information et qui ne doit en aucun cas influencer la construction de l'évaluation. Le numérique fait partie d'un tout et interagit avec plusieurs facteurs, tels que l'enseignement, les caractéristiques de la tâche mobilisée et les compétences des élèves.

Ainsi, si le numérique peut apporter de nombreux avantages autant pour l'enseignement des langues (Tricot, 2020) que pour leur évaluation (Alvarez, 2016, pour une synthèse), il convient aussi de rappeler ses limites. Une tâche sous format numérique devrait pouvoir rendre l'expérience de passation plus efficace, mais il faut veiller à ce que la transformation du format papier en format numérique n'ait pas d'impact sur le construit évalué : il est nécessaire que ce que nous évaluons sous forme numérique soit exactement la même chose que ce que nous évaluons sous forme papier.

L'autonomie dans la gestion de la tâche est certainement un apport du numérique. En effet, la possibilité de gérer de manière indépendante et autonome le temps d'écoute, le nombre d'écoutes ou, dans le cas de la *production de l'oral*, le moment pour commencer l'auto-enregistrement et la validation de l'enregistrement, peuvent avoir un effet positif sur l'anxiété de certains élèves dans les situations d'évaluation. De manière générale, l'autonomie peut être bénéfique pour les performances des élèves (Roussel, 2011). Pourtant, il convient de rappeler que les études montrent également que les élèves en difficulté peuvent « souffrir » de la situation de surcharge cognitive qu'implique le fait de répondre à la fois à la tâche de compréhension et aux exigences de gestion autonome de l'activité (Roussel et al., 2008). En ce sens, la possibilité de travailler de manière autonome, possible grâce à une évaluation numérique, pourrait être contreproductive pour les élèves en difficulté.

Par ailleurs, les différentes fonctionnalités testées dans notre étude ne semblent pas avoir un lien clair avec une meilleure performance des élèves. En effet, les différentes caractéristiques des tâches, dans le cas de l'expérience de la compréhension, ne nous permettent pas d'identifier un effet particulier du numérique en soi. En outre, dans notre étude, l'hypothèse selon laquelle certaines fonctionnalités de l'évaluation numérique – telles que la possibilité d'écouter plusieurs fois la consigne – pourraient favoriser les élèves ayant des difficultés de lecture n'a pas non plus

pu être confirmée. En effet, aucun lien particulier n'a été trouvé entre cette fonctionnalité et une meilleure performance des élèves ayant des problèmes de lecture. À priori, nous pourrions considérer qu'il n'y a pas eu un effet du support dans la résolution de ces tâches. Cependant, la faible présence d'élèves ayant des difficultés de lecture dans notre échantillon ne nous permet pas de généraliser les résultats.

Un autre aspect souvent évoqué de manière intuitive ou naïve est le fait que le numérique a un impact positif sur la motivation des élèves pour réaliser une tâche d'apprentissage ou d'évaluation. Il convient pourtant d'être prudent avec ce type d'affirmation. D'une part, comme le souligne Tricot (2020), il faut différencier l'envie ou la satisfaction générale des élèves de la motivation, entendue comme l'engagement réel et durable des élèves pour réaliser une tâche. D'autre part, l'impact réel du support numérique sur la motivation et la performance mérite une étude spécifique. Dans les expériences futures, l'aspect motivationnel d'une évaluation sur support numérique et son lien avec les résultats devraient être étudiés avec précision.

Enfin, il ressort des résultats que l'impact sur la performance n'est pas clair, bien que, d'une manière générale, l'évaluation numérique semble offrir plusieurs avantages en termes de logistique de passation et d'analyse des résultats. Pour le moment, les résultats suggèrent que ce sont les caractéristiques des tâches – la complexité des audios, la formulation des questions, les opérations cognitives mobilisées –, et pas nécessairement le format d'évaluation en soit, qui ont un impact sur la performance. Dans les études futures, une analyse particulière des caractéristiques des tâches et des audios devrait être prise en compte, ainsi que la considération des compétences initiales des élèves.

4. Références

Álvarez, M. F. (2016). Language testing in the digital era. In E. Martín-Monje, I. Elorza & B. García Riaza (eds), *Technology-enhanced language learning for specialized domains: practical applications and mobility* (pp. 83-94). Routledge.

Association of language testers in Europe (ALTE). (1998). *Multilingual glossary of language testing terms* (Local Examinations Syndicate). Cambridge University Press.

Barras, M., Karges, K., & Lenz, P. (2016). Leseverstehen überprüfen : welche Sprache für die Fragen und Antworten in den Testitems? *Babylonia, 16*(2), 13-18.

Bessonneau, P., Arzoumanian, P., & Pastor, J. M. (2015). Une évaluation sous forme numérique est-elle comparable à une évaluation de type « papier-crayon » ?. *Éducation & formations, 86-87*, 159-180.

Bouchard, M. E. G., Fitzpatrick, E. M., & Olds, J. (2009). Psychometric analysis of assessment tools used with francophone children/Analyse psychométrique d'outils d'évaluation utilisés auprès des enfants francophones. *Canadian Journal of Speech-Language Pathology & Audiology, 33*(3), 129-140.

Buck, G. (2001). *Assessing listening*. Cambridge University Press.

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.

Conférence intercantonale de l'instruction publique de la Suisse romande et du Tessin (CIIP). (2010/2024). *Plan d'études romand (PER)*. CIIP. https://www.plandetudes.ch

Conseil de l'Europe. (2001). *Un cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Didier.

Eberharter, K., Kormos, J., Guggenbichler, E., Ebner, V. S., Suzuki, S., Moser-Frötscher, D., Konrad, E., & Kremmel, B. (2023). Investigating the impact of self-pacing on the L2 listening performance of young learner candidates with differing L1 literacy skills. *Language Testing, 40*(4), 960-983.

Endt, E. [et al.]. (2014). *Der grüne Max: Deutsch für die Romandie* (5. und 6. Klasse). Klett-Langenscheidt.

Endt, E. [et al.]. (2017). *Junior: Deutsch für die Romandie* (7. und 8. Klasse). Ernst Klett Sprachen.

Field, J. (2015). *The effects of single and double play upon listening test outcomes and cognitive processing*. British Council.

Fulcher, G. (2015). Assessing second language speaking. *Language Teaching, 48*(2), 198-216.

Goh, C. C. (2016). Teaching speaking. In W. A. Renandya & H. P. Widodo (eds), *English language teaching today: linking theory and practice* (pp. 143-159). Springer.

Grapin, N., & Sayac, N. (2022). From paper-pencil to tablet-based assessment: a comparative study at the end of primary school. In J. Hodgen, E. Geraniou, G. Bolondi, & F. Ferretti (eds), Proceedings of the *twelfth congress of the european society for research in mathematics education (CERME12), 2-7 February 2012, Bozen-Bolzano, Italy* (pp. 3811-3818). University of Bozen-Bolzano and ERME.

Hoffer, G., & Marc, V. (2025). *ONAÉ – Un outil numérique au service de l´évaluation et des apprentissages : analyse scientifique*. IRDP.

Isaacs, T. (2016). Assessing speaking. In D. Tsagari & J. Banerjee (eds), *Handbook of second language assessment* (pp. 131-146). De Gruyter Mouton.

Karges, K., Lenz, P., Aeppli, T., & Barras, M. (2021). *Fremdsprachenkompetenzen nahe an der Realität testen: Szenariobasierte Testaufgaben für den Computer–eine Vertiefungsstudie.* Wissenschaftliches Kompetenzzentrum für Mehrsprachigkeit.

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

North, B. (2005). Assessing spoken performance in relation to the Common european framework of reference. *Babylonia, 2*, 46-49.

Ockey, G. J. (2007). Construct implications of including still image or video in computer-based listening tests. *Language Testing, 24*(4), 517-537.

Organisation for Economic Co-operation and Development (OECD). (2015, 15 September). *Students, computers and learning: making the connection, PISA*. OECD Publishing. https://doi.org/10.1787/9789264239555-en

Organisation for Economic Co-operation and Development (OECD). (2021). *PISA 2025: foreign language assessment framework*. OECD Publishing.

Park, M. S. (2020, 16 March). 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

Pothier, M., Iotz, A., & Rodrigues, C. (2000, 15 juin). Les outils multimédias d'aide à l'apprentissage des langues : de l'évaluation à la réflexion prospective. *Alsic, 3*(1), 137-153. https://doi.org/10.4000/alsic.1788

Quian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: affective effects on test takers. *Language Assessment Quarterly, 6*(2), 113-125.

Roth, M., Ruf, I., Sánchez Abchi, V., Soussi, A., & Weiss, L. (2021). EpRoCom : dispositif romand de mutualisation de tâches évaluatives : premiers constats : résumé. *irdp FOCUS, 08.2021.*

Roth, M., & Ruf, I. (2024, 5 mars). Ressources évaluatives pour les enseignants·es romand·es : une démarche intercantonale. *La Revue LEeE, 8*.

Roussel, S. (2011). A computer assisted method to track listening strategies in second language learning. *ReCALL, 23*(2), 98-116.

Roussel, S. (2014). *À la recherche du sens perdu : comprendre la compréhension de l'oral en langue seconde*. La clé des langues.

Roussel, S. (2020). *Apport du numérique à l'enseignement-apprentissage des langues*. Cnesco.

Roussel, S., Rieussec, A., Nespoulous, J.-L., & Tricot, A. (2008, 30 mars). Des baladeurs MP3 en classe d'allemand : l'effet de l'autorégulation matérielle de l'écoute sur la compréhension auditive en langue seconde. *Alsic, 11*(2), 7-37. https://doi.org/10.4000/alsic.413

Sánchez Abchi, V., Roth, M., & Matei, A. (2022). Estimer la difficulté des questions en compréhension de l'écrit en français : vérification empirique d'un modèle théorique. *Évaluer - Journal international de recherche en éducation et formation (e-JIREF), 8*(1), 29-46.

Sánchez Abchi, V., Sieber Meylan, S., & 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. *Évaluer - Journal international de recherche en éducation et formation (e-JIREF)*, 10(2), 23-42. https://doi.org/10.48782/e-jiref-10-2-23

Sánchez Abchi, V., Sieber, S., & Matei, A. (2025). Évaluer la compréhension orale en allemand : défis et apports du numérique. *Résonances, 9*, 42-43.

Sánchez Abchi, V., Sieber, S., & Matei, A. (à paraitre). *L'évaluation de la production orale en allemand L2 : quels outils de correction ? : comparaison d'une échelle holistique et d'une échelle analytique.*

Seifert, S., & Paleczek, L. (2022, 4 March). Comparing tablet and print mode of a german reading comprehension test in grade 3: influence of test order, gender and language. *International Journal of Educational Research, 113*.

Sieber, S. (2021). *Épreuves cantonales d'allemand à la fin du cycle 2 (8e année) : état des lieux* (rapport interne). IRDP.

Tricot, A. (2020). *Quelles fonctions pédagogiques bénéficient des apports du numérique* ? Cnesco.

Inscrits dans le mandat de l'Institut de recherche et de documentation pédagogique (IRDP), les travaux relatifs à l'évaluation des *objectifs d'apprentissage* du Plan d'études romand (PER), menés sur la période quadriennale 2020-2023, visent à créer une culture commune entre les cantons romands en matière d'évaluation des apprentissages des élèves. Dans cette perspective, ils permettent de se doter d'outils d'analyse et de mettre à disposition des enseignantes et enseignants des matériaux d'évaluation pertinents, validés et fiables. Ce texte fait partie d'une publication plus complète, qui rassemble les travaux réalisés dans plusieurs disciplines pour des élèves de 8e année (11-12 ans).

Ce rapport présente les principaux résultats relatifs à l'évaluation de l'Allemand langue seconde (L2). Il rend compte du processus de conception, d'adaptation et de vérification de l'adéquation de tâches d'évaluation informatisées pour le niveau A1. Les données sont issues de la réalisation de ces tâches par environ 1 000 élèves pour l'évaluation de la *compréhension de l'oral* et de 212 élèves pour l'évaluation de la *production de l'oral*. Les analyses ont porté sur le lien entre les caractéristiques des tâches, les fonctionnalités numériques et les performances des élèves, ainsi que sur la pertinence des différents outils de correction.

Le projet a permis de mieux comprendre les avantages, les limites et les défis liés à l'évaluation sur support numérique et autonome des compétences orales dans une langue étrangère. Les tâches développées seront, à terme, mises à disposition des enseignantes et enseignants sur les PistEval, pages liées au PER en ligne, consacrées à l'évaluation.