

Pour citer cet article : Ochelen, J.-P., & Yerly, G. (2026, prépublication). Comment des équipes enseignantes (co)construisent-elles ce qu'elles considèrent comme une évaluation de qualité lorsqu'elles doivent évaluer les apprentissages en commun ? *La Revue LEeE, Varia*(2). <https://revue.leee.online/index.php/info/article/view/281>

COMMENT DES ÉQUIPES ENSEIGNANTES (CO)CONSTRUISENT-ELLES CE QU'ELLES CONSIDÈRENT COMME UNE ÉVALUATION DE QUALITÉ LORSQU'ELLES DOIVENT ÉVALUER LES APPRENTISSAGES EN COMMUN ?

Jean-Pascal OCHELEN, Gonzague YERLY

*Version de la prépublication : septembre 2026
Évaluation ouverte et collaborative*

Résumé

Depuis une vingtaine d'années, de plus en plus de politiques éducatives promeuvent une approche collective de l'évaluation des apprentissages. Cette évolution s'inscrit dans une transformation des modes de pilotage des systèmes éducatifs, centrée sur la qualité de l'évaluation et la comparabilité des résultats. En Suisse, cette tendance se traduit par la politique « Évaluation en commun », qui amène les enseignant-es du secondaire II (élèves de 15 à 19 ans) à collaborer pour évaluer les apprentissages, sans définir explicitement les attentes en matière de qualité. Cet article analyse comment, dans ce contexte, des équipes enseignantes construisent ce qu'elles considèrent comme une évaluation commune de qualité. Cette étude compréhensive, menée dans trois établissements du secondaire II en Suisse romande, combine des entretiens auprès de 30 enseignant-es et l'observation de réunions de huit équipes consacrées à la conception d'épreuves communes. L'analyse mobilise le cadre de la validité unifiée de Messick (1989) et la multiréférentialité du jugement professionnel (Mottier Lopez & Allal, 2008). Les résultats révèlent une représentation plurielle, mais asymétrique, de la qualité, dans laquelle les dimensions liées à la mesure – validité, fidélité et pertinence – occupent une place centrale. En l'absence de standards explicites, les enseignant-es élaborent

collectivement leurs propres critères de qualité à partir de référents collectifs, institutionnels et personnels souvent hétérogènes et parfois en tension. Cette construction repose sur des discussions, des négociations et des compromis permettant d'arbitrer les tensions entre comparabilité des résultats, contraintes du travail collectif et prise en compte des contextes d'enseignement.

Mots-clés : qualité de l'évaluation ; évaluation des apprentissages ; évaluation collective ; collaboration entre enseignant-es

Abstract

Over the past two decades, a growing number of education policies have promoted a collective approach to the assessment of learning. This development forms part of a broader transformation in the governance of education systems, centred on assessment quality and the comparability of results. In Switzerland, this trend is reflected in the Évaluation en commun ("Common Assessment") policy, which requires upper-secondary teachers (students aged 15 to 19) to collaborate in assessing learning, without explicitly defining expectations regarding quality. This article examines how, in this context, teaching teams construct what they consider to be a high-quality common assessment. This comprehensive study, conducted in three upper-secondary schools in French-speaking Switzerland, combines interviews with 30 teachers and observations of meetings held by eight teams to design common assessments. The analysis draws on Messick's (1989) unified framework of validity and the multireferential nature of professional judgment (Mottier Lopez & Allal, 2008). The findings reveal a plural yet asymmetrical conception of quality, in which measurement-related dimensions - validity, reliability, and relevance - occupy a central position. In the absence of explicit standards, teachers collectively develop their own quality criteria from collective, institutional, and personal reference points that are often heterogeneous and sometimes in tension. This construction process relies on discussion, negotiation, and compromise to navigate tensions between the comparability of results, the constraints of collective work, and consideration of teaching contexts.

Keywords: assessment quality; student assessment; collaborative assessment; teacher collaboration

1. Introduction

Dans un contexte éducatif international marqué par une attention croissante portée aux performances des élèves, à la comparabilité des résultats et à la reddition de comptes (UNESCO, 2024), comprendre ce que signifie une « évaluation de qualité » constitue un enjeu majeur, tant pour les autorités éducatives que pour les acteur-rices de terrain. Longtemps pensée à travers les propriétés psychométriques des tests, la qualité de l'évaluation en classe ne se réduit toutefois pas aux caractéristiques techniques d'un instrument. Elle dépend également de l'interprétation qui est faite des résultats, de leurs usages en contexte éducatif et de leurs conséquences (Messick, 1989 ; Stiggins, 2007 ; Van der Vleuten, 2016). Traditionnellement exercée de manière individuelle (Allal & Mottier Lopez, 2014), l'évaluation des apprentissages s'inscrit aujourd'hui dans des cadres institutionnels qui encouragent, voire



prescrivent, la collaboration entre enseignant·es afin d'en améliorer la qualité (Ochelen et al., 2024 ; Yerly, 2021).

En Suisse, la politique « Évaluation en Commun » (EeC), mise en œuvre progressivement depuis 2016 au secondaire II (niveau post-obligatoire, élèves de 15 à 19 ans), amène les enseignant·es à collaborer avec leurs collègues pour l'évaluation des apprentissages. L'objectif est d'améliorer la comparabilité des résultats entre cantons et entre établissements (CDIP, 2016). Or, la logique de régulation « douce » (Maroy & Voisin, 2014) qui caractérise cette politique repose sur un faible niveau de prescription : si la collaboration est attendue, elle laisse aux équipes une large marge de manœuvre, sans attentes explicites ni critères de qualité formalisés. Dès lors, que recouvre réellement la notion de qualité de l'évaluation lorsque celle-ci est collective ? Comment les enseignant·es coconstruisent une compréhension commune de ce qui fait qualité en l'absence de standards explicites ?

Cette étude contribue à un projet de recherche financé par le Fonds national suisse de la recherche¹ et explore une dimension encore peu documentée dans la littérature : la construction collective de la qualité de l'évaluation. Elle vise à comprendre comment des enseignant·es du secondaire II coconstruisent une évaluation collective jugée de qualité. Cette recherche compréhensive (Schurmans, 2009) a été menée dans trois établissements de Suisse romande, à partir d'entretiens individuels avec 30 enseignant·es et d'observations de huit équipes. Les données ont fait l'objet d'analyses de contenu par étude de cas (enseignant·es et équipes), puis d'une analyse globale (Miles & Huberman, 2010). Notre démarche semi-inductive mobilise le concept de validité unifiée (Messick, 1989) et la multiréférentialité du jugement professionnel (Mottier Lopez & Allal, 2008).

2. Contexte : la politique suisse « Évaluation en Commun »

En Suisse, le fédéralisme confère aux cantons la souveraineté en matière d'éducation. Si des mécanismes de coordination intercantonale portés par la CDIP² ont favorisé une certaine harmonisation de la scolarité obligatoire, les cantons conservent une large autonomie au secondaire II, notamment dans l'organisation de l'évaluation certificative et la délivrance des certificats de maturité, reconnus selon des exigences communes à l'échelle suisse.

En 2008, l'enquête EVAMAR II³ (CDIP, 2019) a conclu à une qualité globalement satisfaisante de la formation au secondaire II, tout en mettant en évidence d'importantes disparités dans les pratiques évaluatives entre les établissements et au sein de ceux-ci, questionnant la comparabilité des certificats cantonaux de maturité, qui donnent directement accès aux

¹ Projet FNS n°100019_207804 « Lorsque les enseignants du secondaire post-obligatoire doivent évaluer en commun. Quels sont les processus qui contribuent ou font obstacle au développement collaboratif des pratiques d'évaluation des apprentissages ? » (Yerly, 2022-2026).

² CDIP : Conférence suisse des directeurs cantonaux de l'instruction publique, organe politique intercantonal en matière d'éducation.

³ EVAMAR II : seconde évaluation de la formation gymnasiale. L'évaluation s'est particulièrement intéressée aux prérequis dont les étudiant·es disposent pour aborder les études universitaires.



universités dans tout le pays (Eberle et al., 2011). En réponse, la CDIP a formulé plusieurs recommandations visant à améliorer la qualité de la formation et à maintenir l'accès à l'enseignement tertiaire sans examen d'entrée ni examens de maturité standardisés. La politique EeC constitue l'un de ses quatre axes prioritaires. Elle vise à améliorer la comparabilité des résultats en faisant travailler les enseignant·es en équipe pour l'évaluation des apprentissages. Selon la recommandation (CDIP, 2016), l'évaluation en commun signifie que « les enseignantes et enseignants d'un même établissement (ou d'un groupe d'établissements) élaborent conjointement des examens, les font passer, les corrigent et les notent selon des critères qu'ils ont fixés ensemble » (p. 5).

En nous appuyant sur la typologie de Maroy et Voisin (2014), nous considérons que cette politique relève d'une forme de régulation « douce ». La recommandation demeure en effet peu contraignante quant aux modalités de mise en œuvre et aux effets attendus : elle ne s'accompagne ni de standards de qualité définis, ni de prescriptions précises, ni de dispositifs formalisés de contrôle. Les cantons, les établissements et les équipes enseignantes disposent ainsi d'une large autonomie pour déterminer les modalités concrètes d'implantation. Nos travaux antérieurs montrent que cette configuration contribue à une réception et à une mise en œuvre hétérogènes de la politique (Ochelen & Yerly, 2026), ainsi qu'à des effets variables sur les pratiques (Ochelen & Yerly, 2025).

3. Cadre théorique

3.1 La qualité de l'évaluation : de la mesure au jugement professionnel

Développée dès la première moitié du 20^e siècle, la docimologie⁴ étudie systématiquement les examens et les pratiques de notation afin d'améliorer la mesure des apprentissages. La qualité de l'évaluation y est principalement appréhendée à travers trois propriétés psychométriques, qui constituent encore aujourd'hui des référents importants (De Ketele & Roegiers, 2013) :

- la pertinence désigne l'adéquation de l'évaluation aux objectifs visés ;
- la validité concerne l'adéquation entre ce que l'épreuve mesure et ce qu'elle prétend mesurer ;
- la fidélité (ou fiabilité), renvoie à la stabilité des résultats, indépendamment des conditions de correction ou des évaluateurs.

Toutefois, le développement des approches psychométriques et du *testing* se sont accompagnées d'un éloignement entre la mesure des apprentissages et les réalités de la classe, de l'enseignement et de l'apprentissage (Brown, 2022). Plusieurs cadres ont dès lors

⁴ Le terme « docimologie » a été introduit par Piéron et al. (1934, cité dans Piéron, 1963), qui la définissent comme l'étude systématique des examens, portant notamment sur les modes de notation, la variabilité inter- et intra-individuelle des examinateurs, ainsi que les facteurs de subjectivité.



élargi la qualité au-delà des seules propriétés de l'instrument, afin de prendre en compte le caractère multidimensionnel et contextualisé des apprentissages et de leur évaluation (Baartman et al., 2007 ; Leclercq, 2005 ; Messick, 1989 ; Van der Vleuten, 2016). La qualité dépend alors de dispositifs cohérents et contextualisés (Schuwirth & Van der Vleuten, 2011 ; Van der Vleuten et al., 2010), mais aussi du jugement professionnel mobilisé par les enseignant·es pour concevoir, mettre en œuvre et interpréter les évaluations (Allal & Mottier Lopez, 2008 ; Laveault, 2008). Dans ce texte, la qualité de l'évaluation est ainsi conçue comme située (Leduc & Béland, 2022 ; Mottier Lopez, 2021) : elle ne peut être appréciée indépendamment des contextes, des acteur·rices et des finalités de l'évaluation.

En première ligne pour concevoir et mettre en œuvre les dispositifs d'évaluation, les enseignant·es disposent souvent de connaissances et de compétences limitées dans ce domaine (Coombs et al., 2018 ; Goss et al., 2015 ; Grapin & Vantourout, 2023). Leurs pratiques sont en outre orientées par leurs représentations de l'évaluation (Xu & Brown, 2016), qui ne sont pas nécessairement alignées avec celles de leurs collègues ou avec les attentes institutionnelles. Plusieurs travaux montrent ainsi que, dans l'enseignement secondaire, ces représentations demeurent principalement centrées sur la mesure des performances et les propriétés psychométriques des épreuves (Brown, 2004 ; Brown et al., 2011 ; Fulmer et al., 2015 ; Remesal, 2011 ; Schellenkens et al., 2023).

Pour analyser ces représentations, nous mobilisons le cadre de la validité unifiée de Messick (1989). Celui-ci dépasse les approches distinguant différentes formes de validité, notamment les validités de contenu, critériée et de construit. Nous en retenons surtout l'attention portée aux conséquences de l'évaluation : la qualité de l'évaluation ne dépend pas uniquement de la solidité des interprétations, mais également des usages des résultats et de leurs effets pédagogiques, sociaux et éthiques.

3.2 La qualité de l'évaluation : entre qualité à prouver et qualité éprouvée

Évaluer des apprentissages consiste à recueillir des informations pertinentes, valides et fiables, puis à les confronter à un référent afin de prendre une décision (De Ketele, 2010). Ces référents, entendus comme un ensemble de critères, normes ou valeurs servant de point de comparaison, peuvent être structurés au niveau macro par des standards et des dispositifs institutionnels. En classe, ils sont également interprétés et ajustés par les enseignant·es en fonction des contextes, des objectifs et des caractéristiques des élèves (Figari et al., 2014). La référentialisation (Figari, 1994) désigne ainsi la démarche par laquelle les acteur·rices construisent, organisent et ajustent les référents, critères et indicateurs nécessaires à l'évaluation.

Pour rendre compte de la manière dont ces repères sont mobilisés, Mottier Lopez et Allal (2008) conceptualisent le jugement professionnel en évaluation comme une activité cognitive et socialement située. Évaluer ne consiste pas à appliquer mécaniquement des procédures, mais à apprécier une situation à partir de plusieurs cadres de référence, parfois concurrents. Le jugement professionnel s'inscrit ainsi dans une logique de multiréférentialité : il mobilise



des normes professionnelles et institutionnelles, des choix collectifs propres aux établissements, des repères liés à la microculture de la classe, ainsi que les valeurs et croyances personnelles de l'enseignant-e. Il repose donc sur un travail continu d'ajustement entre des exigences et des repères susceptibles de converger ou d'entrer en tension. Sa crédibilité dépend également de la capacité de l'enseignant-e à expliciter et à justifier ses choix (Laveault, 2008).

La qualité de l'évaluation peut dès lors être comprise comme prise dans une double exigence (Andrey et al., 2026). D'une part, elle doit être « à prouver », en répondant à des attentes institutionnelles de conformité, de comparabilité et de légitimité. D'autre part, elle est « éprouvée » dans l'action, à travers les ajustements réalisés au regard des contextes, des élèves et des contraintes de l'activité⁵. Elle se construit ainsi dans l'articulation et la négociation entre des normes prescrites et des normes élaborées dans l'activité (Mottier Lopez, 2020).

3.3 Le travail collectif comme levier de développement des pratiques évaluatives

Pour limiter certains effets indésirables des évaluations standardisées, comme le rétrécissement du curriculum ou le bachotage (Mons, 2009), de plus en plus de politiques éducatives encouragent ou imposent une approche collective de l'évaluation (Yerly, 2021). Cette évolution s'inscrit dans une transformation des modes de pilotage des systèmes éducatifs, centrée sur la qualité, la comparabilité et la régulation des pratiques (Laveault & Yerly, 2017). Dans ce texte, l'expression « évaluation collective » désigne les situations dans lesquelles des enseignant-es collaborent à différentes étapes du processus évaluatif (Allal, 2012) : définition des objets et des finalités, recueil et interprétation des informations, formulation des jugements et prise de décision.

Depuis une vingtaine d'années, plusieurs systèmes éducatifs anglophones décentralisés, notamment en Australie et en Nouvelle-Zélande, ont institutionnalisé des dispositifs de modération sociale des résultats. Fondés sur des échanges autour de standards institutionnels communs et de productions d'élèves, ces dispositifs visent à renforcer la cohérence des jugements professionnels (Adie et al., 2012 ; Laveault & Yerly, 2017 ; Wyatt-Smith et al., 2010). Ils remplissent à la fois une fonction de contrôle de la qualité des jugements (*accountability*) et une fonction de développement professionnel⁶ (*improvement*) (Maxwell, 2002).

⁵ Les expressions « qualité à prouver » et « qualité éprouvée » sont proposées par Andrey, Progin et Maulini (2026) et reprennent le titre du symposium coordonné au colloque de l'ADMEE-Europe (Bienne, 2026), dans lequel ont été présentés les premiers résultats de cette recherche.

⁶ Comme l'expliquent Laveault et Yerly (2017), sur le terrain, ces finalités (*accountability* et *improvement*) coexistent, mais ne produisent pas les mêmes effets. Les logiques d'*accountability* agissent principalement sur le produit de l'évaluation ; les logiques d'*improvement* impactent directement les pratiques et le jugement professionnel, avec des effets plus indirects et différés, au prix de démarches longues et exigeantes.



Ochelen et al. (2024) documentent d'autres dispositifs collectifs qui visent plus précisément le développement de pratiques communes d'évaluation-soutien d'apprentissage (Furtak et al., 2016 ; Hargreaves et al., 2013 ; Hermansen & Nerland, 2014 ; Stewart & Houchens, 2014) ou la co-construction d'outils d'évaluation (Buckley & Liscombe, 2020 ; Heredia et al., 2016 ; Mottier Lopez & Pasquini, 2017). Généralement soutenus par des cadres institutionnels ou des recherches collaboratives, ces dispositifs constituent des espaces de confrontation des jugements professionnels. Les enseignant·es y sont amené·es à expliciter, confronter et parfois transformer leurs conceptions et leurs pratiques, contribuant ainsi à la construction de référents négociés et d'une culture partagée de l'évaluation (Mottier Lopez et al., 2012).

Le travail collectif présente également des limites. Il peut conduire à des décisions influencées par des opinions personnelles peu alignées avec les exigences de qualité, notamment en cas d'asymétries épistémiques problématiques ou de luttes de pouvoir (Black et al., 2011 ; Van der Schaaf et al., 2012 ; Yerly, 2021), ou favoriser une application inflexible des critères et des arrangements évaluatifs susceptibles d'en fragiliser la validité et l'équité (Hipkins & Robertson, 2012 ; Nilserth & Sandlund, 2021).

4. Questions de recherche

La littérature consultée montre que la collaboration entre enseignant·es pour l'évaluation des apprentissages peut constituer un levier pour améliorer sa qualité, mais peut aussi la fragiliser. La plupart des recherches ont toutefois été menées dans des contextes où la collaboration est institutionnellement cadrée et s'appuie sur des référents formalisés. En revanche, on sait encore peu de choses lorsque les enseignant·es disposent d'une large autonomie, sans repères institutionnels explicites. Par conséquent, notre question générale de recherche est : quand des enseignant·es doivent collaborer pour l'évaluation des apprentissages, comment (co)construisent ces enseignant·es une conception commune de ce qui constitue une évaluation de qualité ? Cette question générale se divise en trois sous-questions :

- Quelles sont les représentations individuelles et collectives des enseignant·es relatives à la qualité de l'évaluation des apprentissages lorsque celle-ci est collective ?
- Quels sont les référents mobilisés par les enseignant·es, tant individuellement que collectivement, pour légitimer ou élaborer une évaluation collective jugée de qualité ?
- Par quels mécanismes les enseignant·es coconstruisent une compréhension partagée de la qualité de l'évaluation dans le cadre d'un travail en équipe ?

5. Méthodologie

5.1 Approche de recherche

Notre approche qualitative et compréhensive (Schurmans, 2009) vise à saisir le sens que les acteur·rices attribuent à leurs actions sans chercher à quantifier, comparer ou généraliser.



Cette recherche combine des études de cas et des analyses globales (Miles & Huberman, 2010) à partir d'entretiens individuels menés avec 30 enseignant·es et d'observations de huit équipes engagées dans la construction d'épreuves communes. La démarche semi-inductive mobilise le cadre de la validité unifiée de Messick (1989), adapté à notre contexte⁷, ainsi que la multiréférentialité du jugement professionnel (Mottier Lopez & Allal, 2008).

5.2 Description des établissements

Pour cette recherche, nous avons pu nous rendre dans trois établissements du secondaire II situés dans des cantons différents en Suisse romande⁸. Les principales caractéristiques de mise en œuvre de l'EeC, telles que définies par les directions, sont reprises dans le tableau 1. Les différences les plus marquées tiennent aux choix opérés par les directions : une obligation formelle dans toutes les disciplines à Guillaume Tell, une obligation limitée aux disciplines dites « éliminatoires » au Collège Le Corbusier et l'absence d'obligation à Johanna Spiry. Dans ces établissements, le type de collaboration est laissé à l'appréciation des équipes.

Tableau 1

Principales caractéristiques de mise en œuvre de l'EeC dans les établissements

	Guillaume Tell	Le Corbusier	Johanna Spiry
Mise en place par la direction	Obligation formelle dans toutes les disciplines (au moins une fois par année)	Obligation formelle pour maths, français, allemand, anglais (une fois par année)	Aucune obligation
Type de collaboration	Libre (mais un responsable d'équipe est désigné)	Libre	Libre
Contrôle par la direction	Validation de l'épreuve par la direction. Aucun outil de suivi systématique. Pas d'analyse des résultats	Validation de l'épreuve par les pairs. Aucun outil de suivi systématique. Pas d'analyse des résultats	Pas de contrôle

5.3 Recueil des données et participant·es

Entre septembre 2023 et mai 2025, 30 enseignant·es volontaires ont participé à un entretien individuel d'environ 60 minutes (tableau 2). Ils avaient répondu à un appel général présenté lors d'une réunion plénière, sans que leur identité soit communiquée à leur direction. Le guide d'entretien portait sur leurs caractéristiques professionnelles, leurs pratiques évaluatives, le fonctionnement de l'EeC dans leur établissement, leurs perceptions et usages de celle-ci, ainsi que ses effets perçus sur les pratiques, le travail collectif et l'établissement. Les

⁷ Ce cadre a été éprouvé récemment par Yerly (2017).

⁸ Afin de garantir l'anonymat des participant·es, les cantons ne sont pas spécifiés. De même, les noms des établissements sont fictifs et les prénoms attribués aux enseignant·es sont des pseudonymes.



participants, qui enseignent généralement deux disciplines pour lesquelles ils ont suivi une formation universitaire, présentent des profils disciplinaires et des niveaux d'expérience professionnelle diversifiés.

Tableau 2

Caractéristiques des 30 enseignant-es interrogés individuellement

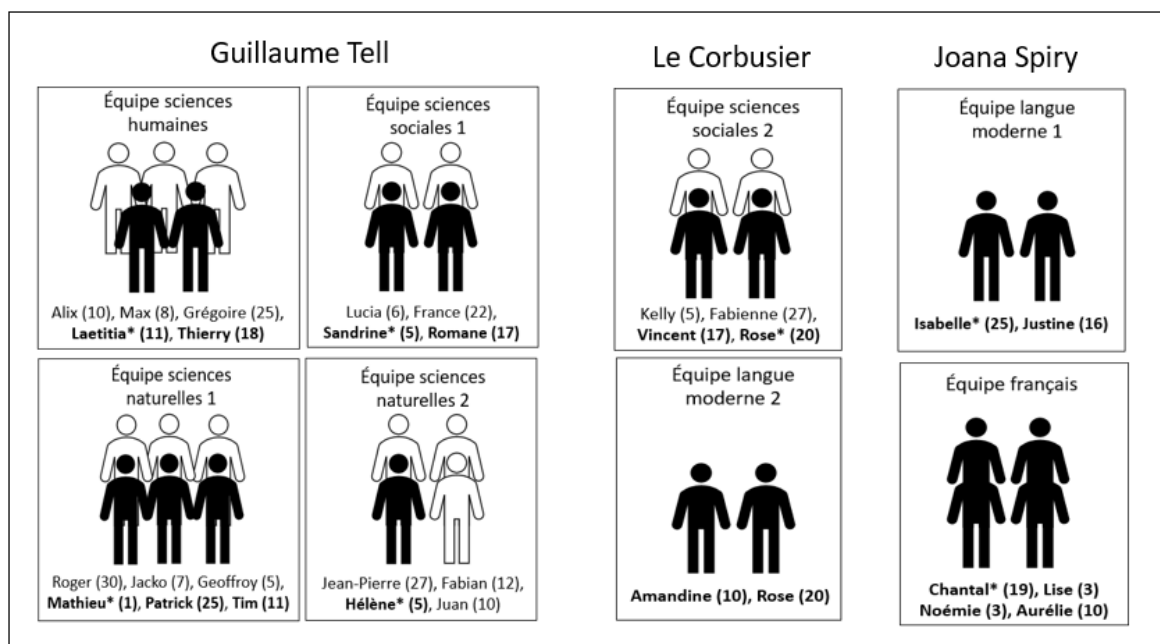
Établissements	Pseudonymes	Disciplines enseignées	Ancienneté de service (en années)
Guillaume Tell	Hélène	Biologie, chimie	5
	Jean	Physique	25
	Laetitia	Géographie, français	11
	Mathieu	Physique	1
	Romane	Psychologie, sociologie	17
	Sandrine	Psychologie, histoire	5
	Thierry	Histoire, géographie	15
	Tim	Physique, math	11
	Julie	Biologie, géographie	12
	Raphaël	Philosophie, musique	14
	Sabine	Économie, droit	5
Le Corbusier	Amandine	Biologie, espagnol	10
	Michel	Math	13
	Vincent	Psychologie, psychopédagogie	17
	Rose	Espagnol, psychologie	20
	Pierre	Français, histoire	2
	Luc	Math	18
	Philippe	Français	35
	Claire	Anglais	20
	Daniel	Français, histoire	2
	Kelly	Anglais, psychologie	6
Johanna Spiry	Aurélie	Français	10
	Chantal	Français	19
	Isabelle	Anglais	25
	Justine	Anglais, allemand	16
	Lise	Français	3
	Noémie	Français	3
	Rémy	Économie, droit	5
	Fabrice	Histoire, histoire de l'art	17
	Sabrina	Économie, droit	4

Nous avons également obtenu l'accord pour observer les réunions de travail de huit équipes (figure 1). Les astérisques indiquent les enseignant-es qui assumaient la responsabilité de l'épreuve commune au moment des observations. Lors de ces observations non participantes, le chercheur prenait également des notes. Au moins une réunion de co-construction de l'épreuve commune a été observée dans chaque équipe, à l'exception de l'équipe Sciences sociales 1, qui avait principalement travaillé par courriel. Pour cette équipe, l'observation a porté sur une réunion de débriefing organisée après la passation.



Figure 1

Les huit équipes observées: disciplines, membres de l'équipe (en gras, les participants aux entretiens individuels), ancienneté de service



5.4 Cadres d'analyse, traitement et analyse des données

Les entretiens ont été intégralement transcrits et codés à l'aide du logiciel MAXQDA. Les réunions observées ont été transcrites sous la forme de scripts minutés. Pour les deux types de données, chaque enseignant·e et chaque équipe a d'abord fait l'objet d'une analyse de cas, puis d'une analyse globale (Miles & Huberman, 2010) afin de dégager les régularités, les variations et les configurations récurrentes. Des grilles de synthèse sous Excel ont permis de mettre en relation les analyses individuelles et collectives.

Pour répondre à la première sous-question, les segments d'entretien et les épisodes de réunion portant sur la qualité de l'évaluation ont d'abord été identifiés. Ils ont ensuite fait l'objet d'une analyse catégorielle mobilisant le cadre de la validité unifiée de Messick (1989). Trois catégories ont été définies à partir de ce cadre :

- la « mesure », qui regroupe la pertinence, la validité et la fidélité de l'évaluation ;
- l'« utilité pour l'action », qui concerne l'usage de l'évaluation par les élèves et les enseignant·es ;
- les « conséquences éthiques et sociales », qui renvoient aux effets de l'évaluation sur les élèves.

Conformément à la démarche semi-inductive, l'analyse est restée ouverte à des dimensions non couvertes par ces catégories. Une quatrième catégorie, l'« acceptabilité », a ainsi émergé des données. Elle désigne le caractère praticable et acceptable du dispositif pour les élèves et les enseignant·es. L'application d'un même cadre aux entretiens et aux observations a

permis de mettre en relation les représentations exprimées individuellement et les dimensions de la qualité mobilisées dans les échanges collectifs.

Pour la deuxième sous-question, une nouvelle lecture des mêmes données a permis d'identifier les référents explicitement ou implicitement mobilisés pour qualifier ou légitimer la qualité d'une épreuve commune. Les catégories initiales, issues de la multiréférentialité du jugement professionnel (Mottier Lopez & Allal, 2008), distinguaient les référents individuels, les référents collectifs d'équipe et les référents institutionnels. Elles ont été progressivement précisées à partir des données. Les grilles de synthèse ont permis d'examiner leur récurrence, leurs combinaisons et les tensions observées entre eux.

Enfin, pour répondre à la troisième sous-question, seules les observations ont été mobilisées. Les épisodes au cours desquels les enseignant·es discutaient de la qualité de l'épreuve commune ont fait l'objet d'une analyse thématique de contenu inductive, visant à identifier les mécanismes par lesquels une compréhension partagée de la qualité se construit, ou non, au cours des interactions.

6. Présentation des résultats

Cette section est structurée selon les trois sous-questions de recherche et présente principalement les résultats des analyses globales. Les valeurs numériques indiquent des tendances, sans viser une comparaison quantitative ni mesurer l'importance des phénomènes.

6.1 Quelles sont les représentations individuelles et collectives des enseignant·es relatives à la qualité de l'évaluation des apprentissages, quand celle-ci est collective ?

L'analyse thématique a permis de repérer les éléments mobilisés par les enseignant·es pour juger de la qualité de l'évaluation, spontanément dans les entretiens ou au cours des discussions, justifications et accords observés en réunion. Le tableau 3 présente ces indicateurs et leur fréquence d'apparition.

Tableau 3

Indicateurs de la qualité de l'évaluation identifiés dans les entretiens et les réunions d'équipe

Dimensions de la qualité		L'évaluation est jugée de qualité si
Mesure (352)	Pertinence (163)	- Le curriculum enseigné correspond au curriculum évalué (51)
		- L'épreuve contient des tâches variées qui permettent d'évaluer différentes dimensions des apprentissages (39)



		- Le niveau de difficulté est adapté aux élèves (31)
		- Les enseignant-es partagent les mêmes exigences, contenus, critères, attentes (19)
		- Il y a cohérence entre le plan d'études et l'évaluation (12)
		- Elle est transparente, les élèves connaissent les critères d'évaluation et les objectifs (11)
	Validité (118)	- Les tâches doivent refléter les apprentissages/objectifs visés (37)
		- La pondération est cohérente en fonction des niveaux cognitifs (25)
		- Les consignes sont claires, lisibles, compréhensibles (Si les consignes sont floues, l'élève peut échouer pour de mauvaises raisons (20)
		- Les résultats doivent refléter le niveau réel des élèves (17)
		- Les tâches contiennent des éléments centraux du cours (pas d'éléments périphériques ou de pièges) (10)
		- L'épreuve contient des tâches authentiques contextualisées dans des situations réelles ou réalistes, portant sur des problèmes complexes (9)
	Fidélité (70)	- Les modalités de correction sont précises et rendent les résultats comparables (41)
		- Les enseignant-es harmonisent la correction, permettant d'assurer une cohérence entre les évaluateurs (20)
		- Les élèves disposent des mêmes conditions de passation (temps, aides, ressources) (9)
Conséquences éthiques et sociales (80)		- Tous les élèves sont traités de la même manière (mêmes contenus, exigences, modalités de correction, de passation) (40)
		- Le dispositif peut être adapté en fonction des conditions d'enseignement et des élèves pour assurer l'équité (flexibilité, différenciation) (18)
		- Les décisions (certification, redoublement) qui en découlent sont justes (12)
		- L'épreuve ne génère pas de stress et suscite l'engagement (10)
		- Les décisions (certification, redoublement) qui en découlent sont justes (10)
Utilité pour l'action (67)	Élèves	- Elle a une fonction informative (permet aux élèves de se situer) et formative (permet de réguler leurs apprentissages) (28)
		- Elle prépare les élèves à l'examen final (15)
	Enseignant-es	- Elle fournit des informations aux enseignant-es pour réguler leurs pratiques (10)
		- Elle soutient la prise de décisions (orientation, certification, redoublement) (7)
		- Elle est rigoureuse, légitime afin de se défendre face à des contestations (7)
Acceptabilité (63)	Élèves	- L'épreuve est faisable (longueur, nombre de questions par rapport au temps disponible) (24)
		- Les élèves ont été suffisamment entraînés (9)
	Enseignant-es	- La charge de travail est acceptable (12)
		- Les conditions de travail sont acceptables (stress, conflits) (10)
	Institution	- L'épreuve est conforme au cadre institutionnel (8)

6.1.1 Une représentation multidimensionnelle de la qualité de l'évaluation, dominée par la mesure

Les participant-es mobilisent une conception plurielle de la qualité, qui concerne l'épreuve, ses effets sur les élèves, son utilité pour les enseignant-es, les contraintes de travail et les attentes institutionnelles. Cette multidimensionnalité demeure toutefois largement



asymétrique : la mesure constitue le registre le plus souvent mobilisé pour définir, justifier ou discuter la qualité.

Dans les entretiens, les enseignant·es associent spontanément une « bonne évaluation » à sa capacité à mesurer ce qui a été enseigné et doit être appris. Les critères les plus fréquents relèvent de la pertinence (163), de la validité (118) et de la fidélité (70) : adéquation des contenus, cohérence des tâches avec les objectifs, difficulté adaptée, clarté des consignes, précision du barème et harmonisation de la correction. Dans les réunions, les échanges portent également en priorité sur les tâches et la correction : contenus demandés, formulation des consignes, attribution des points et comparabilité entre correcteur·rices.

La prédominance de la mesure éclaire un résultat central de l'étude : dans un contexte d'évaluation collective, la qualité est d'abord construite autour de ce qui peut être stabilisé, objectivé et partagé entre collègues. Les dimensions éthiques, sociales ou pragmatiques sont bien présentes, mais elles apparaissent plus souvent comme des dimensions secondaires, mobilisées pour ajuster ou nuancer les décisions prises autour de l'instrument de mesure. Cette conception de la mesure n'est cependant pas strictement techniciste ou psychométrique : elle est aussi située. La pertinence d'une tâche est appréciée au regard du plan d'études, mais aussi du travail réalisé dans les classes, des exemples utilisés, du niveau perçu des élèves et des exigences de l'équipe.

Dans nos données, la multidimensionnalité renvoie à différents niveaux : l'épreuve elle-même, les élèves, les enseignant·es et l'institution. Comme le montre l'extrait d'entretien suivant, les représentations de la qualité intègrent aussi des considérations pragmatiques, notamment en termes de charge de travail :

« Je n'ai pas trois heures à passer sur la mise en page ou sur des barèmes hyper complexes. Si je peux construire une épreuve claire sans y consacrer un temps démesuré, alors ça aussi, c'est une bonne évaluation » (Justine)

6.1.2 Des dimensions en interaction et en tension

L'analyse des données indique que les différentes dimensions s'entrecroisent dans les raisonnements des enseignant·es. Un même argument peut articuler plusieurs registres de qualité. Par exemple, le recours à des tâches jugées plus « authentiques » est justifié à la fois au regard de leur pertinence pour évaluer des compétences complexes (dimension mesure), de leur utilité pour préparer les élèves aux exigences de l'examen certificatif (utilité pour l'action) et de leur potentiel à susciter l'engagement des élèves (dimension sociale). De même, l'harmonisation entre enseignant·es (exigences, correction) est envisagée comme un moyen d'assurer à la fois l'égalité de traitement et la comparabilité des résultats.

Les représentations de la qualité de l'évaluation peuvent être contradictoires. Par exemple, dans nos données, l'égalité de traitement défendue par 21 enseignant·es (sur 30) se confronte à une autre conception de la justice portée par neuf enseignant·es, basée sur la possibilité d'adapter l'évaluation au contexte de classe. Le discours de Pierre montre que cette tension peut également se manifester chez une même personne :



« Pour moi, c'est important que tous les élèves soient évalués de la même manière, avec les mêmes critères... mais en même temps, on sait que dans certaines classes on n'a pas vu les choses de la même façon, donc il faut quand même adapter un peu » (Pierre)

6.1.3 Une reconfiguration des représentations dans le collectif

Les analyses croisées montrent un déplacement des représentations individuelles dans le collectif. Dans les entretiens, l'utilité est surtout pensée pour l'élève : se situer et connaître ses forces et faiblesses. En réunion, elle devient plus pragmatique et orientée vers les enseignant·es : disposer d'un outil commun, répartir le travail, harmoniser les exigences et faciliter la correction. Elle tend ainsi à se redéfinir selon les contraintes organisationnelles, au risque de reléguer sa fonction formative au second plan.

6.2 Sur quels cadres de référence les enseignant·es s'appuient-ils pour justifier la qualité d'une évaluation lorsque celle-ci est collective ?

Pour juger de la qualité de l'évaluation, les participant·es s'appuient sur un ensemble de cadre de référence. À partir du concept de multiréférentialité (Mottier Lopez & Allal, 2008), nous les avons classés en trois grandes catégories synthétisées et décrites dans le tableau 4.

Tableau 4

Synthèse des cadres de référence mobilisés individuellement et collectivement pour juger la qualité de l'évaluation

Catégories de référents	Modalités du référent	Description
Individuels (75)	Valeurs identitaires (18)	Valeurs professionnelles, conceptions de la justice scolaire (égalité vs équité), exigences, rigueur, bienveillance, mérite
	Expérience d'épreuves communes antécédentes (16)	Expérience accumulée grâce aux épreuves communes précédentes, erreurs récurrentes des élèves à ces épreuves
	Pratiques personnelles (15)	Pratiques évaluatives habituelles en classe, mais également pratiques pédagogiques et didactiques plus larges (qui définissent les formes d'évaluation considérées comme pertinentes)
	Repères épistémologiques-disciplinaires (14)	Conceptions de ce qu'est un « bon savoir », ce qui compte comme preuve valable, hiérarchie des compétences
	Connaissance des élèves (13)	Niveau perçu, hétérogénéité, motivation, stress des élèves



	Expertise professionnelle (9)	Savoirs professionnels, normes professionnelles intériorisées, intuitions professionnelles en lien avec les épreuves communes antérieures
Collectifs (équipes) (55)	Normes implicites et culture d'équipe (22)	Habitudes de l'équipe, règles de fonctionnement tacites
	Pratiques des collègues (18)	Pratiques de collègues dont l'expertise est reconnue et pouvant servir de référence
	Histoire collective (15)	Décisions passées faisant office de référence
Collectifs (institutionnels) (52)	Curriculum à enseigner (21)	Supports de cours, type d'exercices travaillés
	Cadres curriculaires officiels (18)	Plan d'études, compétences à certifier, grille d'évaluation institutionnelle, normes institutionnelles
	Exigences du contexte sommatif/certificatif (16)	Exigences associées à la conception d'épreuves sommatives/certificatives ; valeur sociale du diplôme, risque de contestation, nécessité de précision et de comparabilité entre classes
	Prescriptions locales explicites (5)	Directives de la direction ou du canton (obligation ou non, durée, format)
	Prescriptions locales implicites (2)	Attentes implicites de l'établissement (taux de réussite acceptable, réputation)

Le tableau 4 met en évidence la pluralité des cadres de référence, qui apparaît transversale à l'ensemble des enseignant-es interrogé-es et des équipes observées. Ceux-ci concernent non seulement l'épreuve, mais aussi leurs pratiques, expériences, valeurs et conceptions de l'enseignement. La qualité apparaît ainsi enchâssée dans l'activité. Ces référents constituent des ressources que les enseignant-es mobilisent, articulent et ajustent collectivement pour construire leurs jugements.

6.2.1 Le poids des référents individuels

Les référents individuels dominent l'ensemble du corpus (75), y compris dans les situations collectives. Les enseignant-es s'appuient donc principalement sur des ressources construites au fil de leur parcours professionnel.

« Avec l'expérience, on sait ce qui fonctionne et ce qui ne fonctionne pas »
(Daniel)

« On voit assez vite si c'est trop difficile pour eux ; on les connaît bien finalement » (Isabelle)

Ces référents introduisent cependant une variabilité dans les jugements : une même épreuve peut être appréciée différemment selon les exigences, les conceptions disciplinaires ou les valeurs de chacun-e.

6.2.2 Les référents d'équipe : un cadre partagé souvent implicite

Les référents d'équipe reposent largement sur une culture et des normes implicites (22/55), qui orientent les décisions sans être nécessairement discutées. L'histoire collective constitue



également un point d'appui susceptible d'affecter la collaboration et la qualité de l'épreuve, comme l'explique Laetitia⁹ :

« Pendant des années il y a eu des grosses tensions en français. Une collègue avait accusé un autre collègue de l'équipe d'avoir donné les questions à l'avance. Et donc là, c'est vrai que la collaboration a été plus compliquée. Ça pouvait se résumer à recycler les épreuves de l'année précédente. Cette année en géo, même si on n'était pas d'accord au début, on a produit une jolie épreuve commune » (Laetitia)

6.2.3 Les référents institutionnels : un cadre de légitimation

Les référents institutionnels fournissent des points d'ancrage relativement stabilisés concernant les contenus et les formats. Comme l'indique Luc dans son entretien : *« ce n'est pas nous qui décidons les contenus, c'est le plan d'études »*. Toutefois, la faible présence des prescriptions locales (7) suggère que les équipes conservent une marge d'autonomie importante pour définir ce qu'elles considèrent comme une évaluation de qualité.

6.3 Comment les enseignant·es construisent-ils collectivement une compréhension commune de la qualité de l'évaluation lorsque celle-ci est collective ?

Cette partie analyse les séquences interactionnelles dans lesquelles les enseignant·es évaluent, justifient ou problématisent la qualité de l'épreuve commune. Elles ont été repérées à partir de la formulation de jugements, de la mobilisation de critères d'acceptabilité et des discussions portant sur des ajustements concrets.

6.3.1 Les objets de discussion relatifs à qualité de l'épreuve

L'analyse fait apparaître 99 épisodes consacrés à la qualité de l'évaluation. Le tableau 5 regroupe les objets discutés en quatre catégories : tâches, modalités de correction, aspects organisationnels et communicationnels.

Tableau 5

Les objets de la qualité de l'évaluation discutés collectivement

Objets discutés		Description	Objectifs
Tâches (118)	Format (35)	Arbitrages/répartition entre questions ouvertes et fermées, niveaux taxonomiques	Équilibrer les types de compétences évaluées

⁹ Laetitia enseigne le français et la géographie et collabore régulièrement pour la construction d'épreuves communes dans ces disciplines.



	Formulation des consignes (28)	Précision de l'attendu (niveau cognitif, format de réponse), degré d'ouverture de la tâche	Préciser les attentes et réduire les ambiguïtés dans les réponses attendues
	Clarté de l'énoncé (21)	Compréhension du langage/vocabulaire, lisibilité de la situation, mise en page, absence d'ambiguïté, identification de ce qui est important	Garantir que la difficulté de la tâche repose sur les apprentissages visés et non sur des problèmes de compréhension
	Difficulté (18)	Niveau d'exigence des tâches, équilibre entre questions accessibles et complexes	Ajuster le niveau de l'épreuve pour qu'il reflète les apprentissages attendus
	Contenu (12)	Sélection des contenus et des tâches en lien avec les objectifs d'apprentissage	Assurer l'adéquation entre l'épreuve et le curriculum enseigné
	Ordre des questions (4)	Organisation des tâches (progressivité des niveaux de difficulté)	Faciliter l'engagement des élèves et structurer la progression dans l'épreuve
Modalités de correction (114)	Barème-pondération (65)	Répartition des points entre les questions ou parties de réponse	Assurer un équilibre entre les objectifs évalués
	Grilles d'évaluation (32)	Définition, des critères de réussite et des attendus, niveau de détails des indicateurs, degré d'harmonisation	Clarifier les attentes, garantir une cohérence inter-correcteurs, assurer la comparabilité des résultats, tout en limitant la complexité de la correction
	Correction des questions ouvertes (17)	Gestion de la diversité des réponses possibles, correction analytique vs globale	Chercher la comparabilité des résultats malgré l'ouverture des tâches
Aspects organisationnels (22)	Temps de passation – durée de l'épreuve (15)	Adéquation entre le nombre et la complexité des tâches et le temps disponible pour les élèves	Assurer la faisabilité de l'épreuve et éviter que des contraintes temporelles n'affectent les performances
	Cas particuliers (7)	Nécessité d'adapter l'épreuve pour des élèves à besoins particuliers	Assurer l'équité de l'évaluation
Communication (7)	Objectifs de l'évaluation (7)	Rendre visibles les objectifs de l'évaluation, définir le moment où ils seront communiqués	Rendre l'évaluation transparente

Lors des réunions, les discussions se concentrent principalement sur l'élaboration des tâches (118) et les modalités de correction, dont les barèmes et la pondération (114). Ces objets cristallisent les tensions entre les dimensions de la qualité. Les aspects organisationnels (22) et communicationnels (7), moins présents, semblent plutôt perçus comme des contraintes à gérer. L'analyse des 99 épisodes fait apparaître quatre mécanismes de construction d'une compréhension commune de la qualité.



6.3.2 Explicitation des référents tacites

Les enseignant·es verbalisent des critères habituellement mobilisés dans leur pratique individuelle : intentions pédagogiques, exigences, conception de la difficulté et attentes envers les élèves. Cette explicitation révèle des écarts dans leurs représentations, leurs pratiques évaluatives et leurs manières d'enseigner un même contenu. Construire une épreuve de qualité suppose ainsi de composer avec les pratiques - parfois difficilement compatibles entre-elles - de leurs collègues. En Sciences naturelles 1, par exemple, la correction d'une tâche de résolution de problème mobilisant une série de calculs, certains enseignant·es souhaitaient sanctionner chaque omission d'une unité au nom des conventions et de la rigueur scientifique. D'autres défendaient une pénalité unique, afin que l'accumulation des retraits de points ne conduise pas à un score reflétant mal la maîtrise du raisonnement (problème de validité). Dans ce cas observé, l'équipe n'est pas parvenue à s'accorder sur le principe d'une correction soit analytique, soit globale. Elle a donc conservé un barème commun, tout en laissant chacun appliquer ses pratiques habituelles de correction.

6.3.3 Confrontation et mise à l'épreuve des propositions

La construction de l'épreuve donne souvent lieu à des désaccords (70 épisodes) : les enseignant·es questionnent les propositions, demandent des justifications et remettent les choix en discussion. Les référents individuels, notamment l'expérience et la connaissance des élèves, constituent souvent des arguments difficiles à contester. Les décisions tendent néanmoins à s'aligner sur des référents collectifs afin de parvenir à des solutions acceptables pour l'équipe et compatibles avec les contraintes organisationnelles. Certains référents institutionnels s'imposent aussi comme non négociables, tels que l'obligation d'épreuves communes ou la durée imposée.

L'extrait suivant, tiré d'une réunion en Sciences naturelles 1, illustre cette confrontation autour d'une tâche proposée par Tim¹⁰ (figure 2). Les enseignant·es mobilisent leur compréhension de la situation, leur appréciation de la difficulté, leurs propres pratiques, les normes d'équipe, leur connaissance des élèves et les conventions disciplinaires. Il leur faut ainsi arbitrer entre ces référents et composer avec l'hétérogénéité des contenus enseignés, des progressions, des exemples et des pratiques. L'équipe décide finalement de remplacer l'exemple par un autre, jugé plus abordable pour les élèves.

Figure 2

Exemple de tâche ayant fait l'objet de discussions

¹⁰ Dans cet exemple, les élèves doivent comprendre que la voiture est en train de freiner. L'indice, relativement subtil, réside dans l'écrasement des amortisseurs à l'avant.



Question 4. Représentez vectoriellement toutes les forces qui agissent sur la voiture



« Moi, je préfère l'exemple avec l'avion, c'est plus simple. Et vous » ?
(Mathieu)

« Je trouve difficile celle-ci, y'a tellement de forces en jeu. Et puis, y a les forces de frottements, d'habitude on les ignore » (Jean)

« Tu mets pas les frottements ? Moi je le fais presque systématiquement »
(Tim)

« Oui mais dans les épreuves communes, on l'a jamais fait » (Jean)

« Oui mais surtout, on voit pas bien que ça freine. C'est les amortisseurs » ?
(Mathieu)

« Oui, ils sont écrasés devant. Mais vraiment, c'est simple. Ils sont en 2^e
quoi » (Tim)

« Pour nous, oui, on voit qu'elle freine mais des élèves de 15 ans qui n'ont
jamais conduit ou qui n'y connaissent rien en voiture » (Jean)

(...)

« Oui, mais tu parlais de l'avion avant et là aussi, c'est un peu la même chose.
En plus, un exercice avec un avion, j'ai déjà fait avec les miens » (Tim)

(...)

« On peut prendre la chute libre et le parachute, vous en pensez quoi ? (...)

Ok, alors, je me charge de faire le changement » (Mathieu)

6.3.4 Ajustements, compromis et reformulations

Les compromis (35 épisodes) prennent souvent la forme d'ajustements concrets : simplifier une consigne sans changer l'objectif, réduire le nombre d'items/de questions, transformer une question ouverte en question guidée, ajuster la difficulté ou la pondération, réduire l'exigence d'une question problématique, ajouter un exemple ou une aide implicite. Ils permettent de concilier plusieurs dimensions de la qualité. En Sciences humaines, l'équipe referme une question de raisonnement afin de renforcer la fidélité de la correction, au prix d'une perte de pertinence et d'authenticité. En Sciences sociales 1, un compromis par compensation est construit autour de l'utilisation de QCM. Dans l'extrait ci-dessous, ceux-ci sont critiqués par Romane et France au regard de leur validité et de leur caractère inéquitable, mais défendus pour leur faisabilité et leur acceptabilité. L'équipe accepte d'en réduire le nombre à condition d'en diminuer la pondération.

« Oui, il y avait 6 QCM » (France)

« Du coup, si c'est à refaire, tu en mettrais moins ? » (Sandrine)
« Moi, je les supprimerais. Je les trouve compliqué, les formulations sont ambiguës (...) j'ai des élèves qui étaient super bons qui ont surtout perdu des points là en fait » (Romane)
« Je suis d'accord parce que des fois, on évalue plus leur capacité à bien remplir un QCM plutôt que ce qu'ils ont réellement appris (France)
« Oui mais on discutait de la longueur de l'épreuve, et là justement, c'est plus rapide pour les élèves. (...) » (Lucia)
« Et puis pour nous aussi ! (...) et si on en met moins, ça jouerait ? » (Sandrine)
« Mouais... alors, on change la pondération » (Romane)

6.3.5 Stabilisation provisoire d'un accord

Dans 21 épisodes, les équipes parviennent à un consensus autour d'une définition commune de la qualité en s'alignant sur un même référent : objectifs d'apprentissage, difficultés anticipées des élèves ou contraintes temporelles. Dans 29 autres, l'accord se construit sans désaccord explicite, par alignement progressif, activation de règles tacites ou adoption de solutions jugées suffisantes. Ces configurations sont particulièrement visibles en Langues modernes 2, seule équipe engagée dans une collaboration pleinement volontaire, dont les membres partagent des représentations proches de la qualité.

7. Discussion des résultats

L'objectif de cette étude était de comprendre comment des enseignant·es coconstruisent une conception commune de ce qui constitue une évaluation de qualité quand celle-ci est collective. Son originalité tient au fait qu'elle porte sur une évaluation sommative (voire certificative dans certains cas), comportant des enjeux tant pour les élèves que pour les enseignant·es, et élaborée collectivement sous l'influence d'une politique pensée selon une logique de régulation « douce » (Maroy & Voisin, 2014), qui encourage la collaboration sans définir explicitement les attentes en matière de qualité de l'évaluation.

Dans la première partie de l'étude, nous avons examiné les représentations individuelles et collectives des enseignant·es. Deux résultats marquants se dégagent. D'abord, la qualité de l'évaluation est principalement envisagée sous l'angle du paradigme de la mesure. Toutefois, les dimensions liées aux conséquences et à l'utilité de l'évaluation sont également présentes, témoignant d'une conception plurielle de la qualité. Dans la deuxième partie, nous avons analysé les cadres de référence sur lesquels les enseignant·es s'appuient pour construire une évaluation collective jugée de qualité. Les résultats montrent que les enseignant·es mobilisent une pluralité de référents, majoritairement issus de l'activité individuelle mais reconfigurés dans et par le collectif. Dans une troisième partie, nous avons examiné les mécanismes interactionnels par lesquels les enseignant·es en viennent, ou non, à construire une compréhension commune de la qualité de l'évaluation.



Nos analyses montrent que la construction d'une épreuve commune jugée de qualité repose sur des discussions, des négociations et des compromis, à l'articulation de représentations individuelles et collectives, ainsi que de cadres de référence eux aussi pluriels et souvent contradictoires. Pour approfondir ces constats généraux, la discussion est organisée autour de trois questions : pour quelles raisons la qualité de l'évaluation est-elle d'abord pensée à partir de la mesure des apprentissages ? En l'absence de standards institutionnels explicites, comment les enseignant-es construisent une évaluation jugée de qualité ? Quelle est l'importance du collectif sur la définition de la qualité d'une évaluation commune ?

7.1 Pour quelles raisons la qualité d'une évaluation commune est-elle d'abord pensée à partir de la mesure des apprentissages ?

Les résultats invitent d'abord à s'interroger sur l'objet auquel se rapporte la qualité de l'évaluation collective. Dans les interactions observées, celle-ci est principalement discutée à propos de l'épreuve commune elle-même (tâches, formats, niveau de difficulté) et des modalités de correction. Si cette focalisation sur l'instrument rejoint d'autres travaux (Brown, 2004, 2011 ; Fulmer et al., 2015 ; Schellekens et al., 2023), elle tient en partie à la nature du matériau étudié. Les observations portaient sur des réunions consacrées à la conception d'épreuves communes, comme demandé par les directions : la qualité est donc logiquement travaillée au niveau de l'instrument. Elle semble également liée aux caractéristiques du dispositif. Le travail collectif suppose en effet de disposer de repères partagés pour coordonner l'activité, objectiver les désaccords et garantir la comparabilité des résultats entre classes.

De ce fait, le collectif tend à privilégier les dimensions de la qualité susceptibles d'être explicitées, discutées et justifiées auprès des collègues. Ce ne sont donc pas nécessairement la validité ou la fidélité qui occupent une place centrale, mais plutôt leur capacité à fournir des repères communs pour soutenir l'argumentation, réguler les désaccords et construire des décisions collectives. Cette tendance vers la mesure est renforcée par une contrainte propre aux épreuves communes : celles-ci doivent pouvoir être passées simultanément dans plusieurs classes, parfois différentes dans leurs rythmes, leurs pratiques d'enseignement ou leurs niveaux perçus. Les enseignant-es sont donc amené-es à construire des tâches suffisamment partagées, compréhensibles et réalisables pour l'ensemble des élèves concerné-es. Cette exigence favorise de facto une certaine standardisation de l'évaluation, dans laquelle la recherche de comparabilité et d'égalité de traitement tend à occuper une place centrale. Cette logique est particulièrement présente dans notre contexte, car l'évaluation revêt aussi une fonction certificative : dans certaines équipes observées, les résultats de l'épreuve commune contribuent directement à l'obtention du diplôme sanctionnant la fin des études secondaires. Le caractère certificatif de ces épreuves peut donc laisser moins de place à la dimension utilitaire de l'évaluation. Une fois l'épreuve passée, les élèves quittent l'établissement ou changent d'année et de professeur, ce qui réduit les possibilités de réinvestir les résultats dans une perspective de régulation des apprentissages et des enseignements. Cette situation fait écho aux observations de Yerly (2017) concernant le faible usage des résultats des évaluations externes. Les enseignant-es ne se considèrent pas



véritablement comme les destinataires de ces résultats, perçus avant tout comme appartenant aux élèves.

Cette faible orientation vers l'usage des résultats peut aussi être mise en relation avec la temporalité de la collaboration. Dans la majorité des équipes observées, le travail collectif se concentre sur la conception des épreuves et s'étend rarement à une correction collective ou à une analyse partagée des résultats, ce qui rejoint les constats d'Holmeier et al. (2017) et de Yerly (2021). Si les travaux sur la modération sociale suggèrent que les bénéfices du travail collectif résident aussi dans la confrontation des jugements à partir de productions d'élèves et dans la discussion des résultats obtenus (Colbert et al., 2012 ; Wyatt-Smith et al., 2010), la mise en œuvre de tels dispositifs dans notre contexte se heurte à des contraintes temporelles et organisationnelles qui en limitent le déploiement (Ochelen & Yerly, 2025, 2026). De même, si la qualité est principalement pensée en amont de la passation plutôt qu'à partir d'un retour collectif sur ses effets réels, les dimensions liées aux conséquences sociales de l'évaluation apparaissent plus difficilement discutables collectivement. Les équipes disposent alors de peu d'occasions pour examiner comment les épreuves communes sont vécues par les élèves et quels effets elles produisent sur leur réussite ou leurs conditions de travail. Ces observations invitent à s'interroger sur les conditions susceptibles de favoriser une collaboration plus étendue autour de l'analyse des productions et des résultats des élèves. Parmi celles évoquées par les participant·es, l'octroi de temps dédié apparaît comme un levier important. Cette possibilité demeure toutefois tributaire de contraintes organisationnelles qui dépassent largement le niveau des équipes ou des directions d'établissement.

La prédominance de la dimension « mesure » mérite toutefois d'être nuancée. Dans nos données, une part importante des représentations des enseignant·es, mais aussi de leur activité, s'organise autour du respect de la triple concordance entre les objectifs, les méthodes et l'évaluation (Leclercq et al., 2016). Si s'assurer que chaque objectif d'apprentissage déclaré soit effectivement enseigné puis évalué constitue un principe fondamental de tout dispositif de formation, nos résultats montrent que l'évaluation collective en complexifie la mise en œuvre. Dans plusieurs situations, les enseignant·es ne débattent pas uniquement de la qualité intrinsèque des tâches, mais aussi de leur compatibilité avec les contenus effectivement travaillés dans leurs classes respectives. Construire une épreuve commune ne consiste donc plus uniquement à aligner enseignement, méthodes et évaluation à l'échelle individuelle, mais aussi à rechercher une forme de concordance interindividuelle entre des pratiques d'enseignement potentiellement hétérogènes. Ainsi, cette concordance ne renvoie pas uniquement au curriculum prescrit, mais intègre les conditions réelles d'enseignement, ce qui lui confère une dimension située, proche de ce que la littérature désigne comme une forme de « validité écologique » (Leclercq, 2005). Si dans nos analyses, les discussions relatives à cette préoccupation ont été rattachées au critère « pertinence », celle-ci ne semble pas se réduire à une question de mesure : elle renvoie également à la prise en compte des conditions concrètes d'enseignement et d'apprentissage dans l'appréciation de la qualité de l'évaluation. Nos résultats suggèrent ainsi une tension entre deux manières d'envisager la qualité de l'évaluation. D'un côté, l'épreuve commune poursuit un objectif de comparabilité qui tend à privilégier des formes de standardisation reposant sur des tâches identiques administrées dans différentes classes. De l'autre, une conception plus située de l'évaluation (Mottier Lopez, 2021), qui conduit à considérer que la qualité d'une épreuve dépend également de son



adéquation aux contextes d'enseignement dans lesquels les apprentissages ont été construits (Leduc & Béland, 2022). Cette tension permet de mieux comprendre certains débats autour du choix des tâches. Lorsque les situations évaluatives proposées sont fortement simplifiées (ex. questions fermées) ou reposent sur des formats dichotomiques (ex. Vrai-Faux), les réponses produites peuvent être moins représentatives de la complexité des apprentissages visés. À l'inverse, les tâches plus complexes et contextualisées sont fréquemment perçues comme plus à même de rendre compte de leur richesse. Toutefois, cette dimension écologique entre en tension avec d'autres exigences, notamment pragmatiques. Elle suppose en effet de limiter le recours à des formats tels que les QCM, pourtant appréciés par plusieurs enseignant·es de notre étude en raison de leur facilité de correction et de leur capacité à produire des résultats comparables.

7.2 En l'absence de standards de qualité institutionnels explicites, comment les enseignant·es construisent une évaluation collective jugée de qualité ?

La politique EeC amène les enseignant·es à collaborer pour l'évaluation pour en améliorer la qualité, sans leur fournir de repères ou d'attentes claires quant à la qualité à atteindre. Cette absence de balises rend particulièrement visible le caractère foncièrement professionnel de l'activité évaluative : pour définir collectivement ce qui constitue une évaluation de qualité, les enseignant·es mobilisent des cadres de référence hétérogènes (Laveault, 2008 ; Mottier Lopez & Allal, 2008 ; Wyatt-Smith et al., 2010). Si Mottier Lopez & Allal (2008) décrivent cette multiréférentialité comme une caractéristique du jugement professionnel, nos résultats montrent qu'en contexte d'évaluation collective, celle-ci devient aussi un objet explicite de discussion et de négociation. Les processus observés relèvent d'un travail de référentialisation (Figari, 1994), par lequel les acteur·rices construisent et organisent les repères servant à apprécier une situation. Cette étude rend ainsi visibles les mécanismes concrets par lesquels les référents (jugés) de qualité sont mobilisés et mis en discussion dans le collectif, au-delà des représentations que les enseignant·es peuvent en avoir. La difficulté ne réside pas uniquement dans la coexistence de référents multiples, mais aussi dans le fait que les enseignant·es ne leur attribuent pas la même importance ni la même légitimité pour juger de la qualité d'une épreuve commune. Il ne s'agit donc plus simplement de mobiliser individuellement une pluralité de référents, mais de construire collectivement le jugement professionnel à travers des discussions et des compromis (Van der Vleuten, 2016), au cours desquels les référents, les interprétations et les critères de qualité sont confrontés et progressivement stabilisés.

Si les participant·es ne disposent pas de standards définissant explicitement la qualité de l'évaluation attendue par l'institution, leur activité ne se déroule pas pour autant sans aucun repère. Des référents institutionnels existent (ex. plans d'études, compétences à certifier), mais restent peu prescriptifs quant à ce qui constitue une évaluation de qualité. Certaines prescriptions locales encadrent les modalités organisationnelles de l'EeC (ex. le calendrier imposé ou les conditions de passation des épreuves), mais sont rarement mobilisées par les enseignant·es dans les discussions consacrées à la qualité. Cette situation peut présenter des risques. Lorsque les standards sont peu explicités institutionnellement, leur définition tend



à se déplacer vers le collectif. Les normes de qualité deviennent alors plus « vivantes », construites et ajustées dans l'interaction, mais aussi potentiellement plus instables et inégalement partagées. Les rapports de force, les traditions locales ou les conceptions dominantes de certain·es enseignant·es – pas nécessairement partagées par l'ensemble de l'équipe – peuvent alors progressivement s'imposer comme standards implicites de qualité (Yerly, 2021).

Cette situation invite aussi à interroger la portée de standards de qualité plus formalisés : leur présence garantirait-elle leur appropriation et leur usage effectif par les enseignant·es, ainsi que des pratiques de meilleure qualité ? Les travaux australiens sur la modération sociale permettent de nuancer cette hypothèse. Wyatt-Smith et al. (2010) montrent que, même en présence de standards explicites, les enseignant·es ne s'y réfèrent pas de manière directe et univoque. Ces standards deviennent eux-mêmes des objets de discussion nécessitant un travail collectif d'interprétation et de contextualisation, visant à co-construire une compréhension partagée des attentes et des critères de qualité. Nos résultats vont dans le même sens : l'absence de standards explicites ne conduit pas à une absence de critères de qualité. Les enseignant·es mobilisent une pluralité de référents pour construire et justifier leurs jugements. Leur mise en discussion collective participe à la construction de significations partagées et d'un jugement professionnel collectif et situé (Mottier Lopez et al., 2012), mais sans balises formalisées.

Nos analyses montrent que, même si les référents communs sont rares et très diversifiés (car créés localement), ils ne sont pas toujours alignés et peuvent entrer en tension au cours de la construction des épreuves. Ces tensions ne sont pas sans effets sur l'activité collective. La littérature montre qu'elles peuvent, dans certaines circonstances, fragiliser les relations entre les membres du collectif ou conduire à la marginalisation de certaines positions (Black et al., 2011 ; Ochelen & Yerly, 2025 ; Yerly, 2021). À l'inverse, lorsqu'elles sont reconnues et mises en discussion, elles peuvent constituer des occasions de réflexion, de négociation et d'apprentissages professionnels (Hermansen & Nerland, 2014 ; Mottier Lopez & Pasquini, 2017 ; Ochelen et al., soumis), ainsi que de développement du jugement professionnel en évaluation (Allal & Mottier Lopez, 2014).

Par ailleurs, la convergence des référents – par exemple lorsque les membres d'une même équipe partagent des attentes similaires quant aux critères de correction, au niveau d'exigence ou aux finalités de l'évaluation – peut favoriser la fluidité du travail collectif, l'engagement et le bien-être professionnel. Elle n'est toutefois pas sans limites, puisqu'elle peut encourager des formes de conformisme ou de standardisation des pratiques lorsque certains référents collectifs – qui ne sont pas nécessairement les « bons référents » – s'imposent progressivement comme allant de soi et sont moins fréquemment remis en question. Cela peut être le cas lorsque certaines modalités de correction ou de pondération, ou certains niveaux d'exigence, sont reconduits d'année en année sans discussion critique au sein du collectif, parce qu'ils sont devenus des normes locales parfois justifiées par l'idée que « l'on a toujours fait comme ça ».

L'enjeu principal n'est peut-être pas la définition préalable d'un référentiel unique de qualité, mais la mise en place d'espaces de discussion permettant aux enseignant·es d'explicitier et de confronter les référents qui fondent leurs jugements sur l'évaluation, afin de les ajuster et



de les aligner collectivement. Au-delà de leur fonction de coordination, ces espaces peuvent contribuer à la construction d'un sens collectif autour de l'évaluation collective et de ses finalités (Ochelen & Yerly, 2026), en permettant aux enseignant-es de discuter, d'interpréter et de négocier les critères à partir desquels la qualité est définie. Encore faut-il que les conditions soient réunies pour que ces espaces existent et soient investis. Certains discours laissent penser que le manque de temps, mais aussi l'absence perçue de plus-value de la démarche, peuvent limiter leur développement.

7.3 Quelle est l'importance du collectif sur la définition de la qualité d'une évaluation commune ?

Nos résultats rejoignent les travaux de Mottier Lopez (2020), selon lesquels la qualité de l'évaluation ne réside pas dans des critères fixes, mais dans la manière dont les normes sont interprétées, ajustées et négociées dans l'activité. Ils prolongent également la distinction proposée par Andrey et al. (2026) entre une qualité « à prouver », répondant à des attentes institutionnelles, et une qualité « éprouvée » dans l'action.

Dans sa recommandation, la CDIP (2016) ne définit pas explicitement ce que doit être une « bonne » évaluation commune. Elle associe toutefois indirectement sa qualité à la comparabilité des résultats entre classes et entre établissements, dans le prolongement des constats d'EVAMAR II sur l'hétérogénéité des pratiques évaluatives (Eberlé et al., 2011). À l'échelle institutionnelle, la qualité est donc d'abord « à prouver » : elle doit être visible, justifiable et défendable. Dans le contexte de l'EeC, cette exigence prend principalement la forme de la comparabilité et de l'harmonisation des contenus, des exigences et des modalités de correction. Mais comme le montrent nos résultats, la qualité est aussi « éprouvée ». Dans nos données, les enseignant-es ne cherchent pas uniquement à rendre les évaluations comparables. Les propositions sont également appréciées à partir de référents construits au fil de l'expérience professionnelle : connaissance des élèves, contextes d'enseignement et effets concrets des dispositifs évaluatifs. Ce qui paraît comparable à l'échelle institutionnelle n'est donc pas nécessairement ce qui paraît le plus pertinent dans une classe. Ce décalage peut s'expliquer par les positions différentes occupées par les personnes impliquées dans le système éducatif, qui n'exposent pas aux mêmes contraintes, finalités et expériences de l'évaluation (Laveault & Yerly, 2017).

Nos résultats conduisent toutefois à prolonger la distinction entre qualité « à prouver » et « éprouvée ». Lorsque les critères institutionnels de qualité sont peu formalisés, la qualité est également « à trouver ». Sa définition n'est pas donnée d'avance : les équipes doivent déterminer les critères à partir desquels l'épreuve commune pourra être considérée comme pertinente, acceptable et suffisamment juste. La qualité se construit alors dans la confrontation de référents individuels, collectifs et institutionnels, à travers des discussions, des négociations et des compromis. Enfin, pour devenir véritablement collective, la qualité doit également être « approuvée ». L'épreuve construite ne correspond pas nécessairement à l'idéal de chaque enseignant-e. Elle suppose parfois de renoncer à certaines préférences ou exigences individuelles. La qualité qui en résulte devient donc partielle : elle ne satisfait pas pleinement tous les critères portés individuellement, mais repose sur un accord jugé



suffisamment acceptable pour permettre l'action collective. Être approuvée ne signifie donc pas faire l'unanimité. Cela signifie être reconnue, acceptée et rendue possible par le collectif, dans un processus de construction partagée du sens de l'évaluation (Ochelen & Yerly, 2026) et d'ajustement des normes et des critères (Adie et al., 2012 ; Wyatt-Smith et al., 2010).

Dans l'évaluation collective, la qualité est ainsi « à prouver » auprès d'autrui, « éprouvée » dans l'action, « à trouver » dans la discussion et « approuvée » par le collectif. Ces quatre dimensions ne convergent pas nécessairement. Ce qui peut être prouvé n'est pas toujours éprouvé comme pertinent ; ce qui est trouvé dans la discussion n'est pas toujours pleinement approuvé ; et ce qui est approuvé par l'équipe ne correspond pas nécessairement à ce que chaque enseignant·e aurait privilégiée individuellement. L'importance du collectif réside précisément dans cette transformation : la qualité n'est plus seulement un critère permettant de juger l'évaluation, elle devient elle-même un objet de construction commune.

8. Conclusion, limites et perspectives

Cette recherche contribue à la réflexion sur le développement de l'évaluation collective des apprentissages entre enseignant·es, une pratique de plus en plus institutionnalisée dans les systèmes éducatifs. Son principal apport est de rendre visible un objet encore peu documenté : la manière dont des équipes d'enseignant·es définissent, négocient et légitiment ce qui constitue une évaluation sommative de qualité lorsqu'elle est commune à plusieurs classes. Elle montre également comment ces conceptions orientent les choix relatifs aux tâches, aux critères, aux modalités de correction et, plus largement, à la conception des épreuves.

L'analyse conjointe des entretiens et des observations met en évidence une conception plurielle, mais asymétrique, de la qualité. Les enseignant·es mobilisent des dimensions pédagogiques, éthiques, sociales et pragmatiques, mais les enjeux liés à la mesure demeurent centraux dans leurs discours et leurs pratiques, notamment à travers la validité, la fidélité et la comparabilité des résultats. En l'absence de repères formalisés, les enseignant·es s'appuient sur des référents individuels, collectifs et institutionnels qui entrent fréquemment en tension. La qualité résulte ainsi d'arbitrages entre des référents parfois concurrents, dans un contexte marqué par une tension entre comparabilité et prise en compte des contextes d'enseignement.

Le travail collectif transforme les conditions d'exercice du jugement professionnel : les enseignant·es doivent expliciter des référents tacites, confronter leurs interprétations, négocier des accords et construire des compromis. Il ne s'agit pas nécessairement de produire une évaluation « parfaite », mais une évaluation jugée suffisamment juste et acceptable par l'équipe. Cette recherche de compromis suppose parfois de renoncer à certaines préférences individuelles au profit d'une qualité partielle, mais compatible avec les exigences d'une épreuve sommative commune administrée à plusieurs classes. La collaboration ne conduit donc pas nécessairement à une évaluation meilleure ou plus juste, mais elle crée un espace collectif d'explicitation et de discussion des conceptions de la qualité.

Notre étude invite ainsi à considérer les limites d'une épreuve commune pour garantir la qualité de l'évaluation des apprentissages. Une telle épreuve peut soutenir certaines dimensions, en particulier l'harmonisation des exigences et la comparabilité des résultats, mais elle ne suffit pas à garantir une évaluation de qualité. Celle-ci se construit aussi dans les



pratiques de classe, au plus près des apprentissages, des élèves et des ajustements réalisés par les enseignant-es.

Cette recherche présente certaines limites. La démarche compréhensive permet d'identifier des tendances, mais non les effets réels de l'évaluation collective sur les pratiques, les apprentissages ou les parcours des élèves. Une étude longitudinale intégrant le point de vue d'autres acteur-rices — autorités, directions et élèves — permettrait de mieux comprendre ses usages et ses effets sur la qualité de l'évaluation. Nos données ne permettent pas non plus d'analyser finement les écarts entre les représentations d'une évaluation individuelle de qualité et celles mobilisées lors de la construction d'une épreuve commune, ni les renoncements auxquels les enseignant-es consentent pour parvenir à un accord collectif.

Références

- Adie, L. E., Klenowski, V., & Wyatt-Smith, C. (2012). Towards an understanding of teacher judgement in the context of social moderation. *Educational Review*, 64(2), 223–240. <https://doi.org/10.1080/00131911.2011.598918>
- Allal, L. (2012). Les exigences inconciliables des activités évaluatives et de leurs paradigmes de référence. In L. Mottier Lopez & G. Figari (Eds.), *Modélisations de l'évaluation en éducation* (pp. 181-194). De Boeck.
- Allal, L., & Mottier Lopez, L. (2014). Teachers' professional judgment in the context of collaborative assessment practice. In C. Wyatt-Smith, V. Klenowski & P. Colbert (Eds.), *Designing assessment for quality learning: The enabling power of assessment* (pp. 151–165). Springer. <https://doi.org/10.1007/978-94-007-5902-210>
- Andrey, A., Progin, L., & Maulini, O. (2026, 7 janvier – 9 janvier). *Entre qualité à prouver et qualité éprouvée, comment le travail enseignant s'apprécie-t-il effectivement ?* [Symposium]. 37^e colloque de l'ADMEE-Europe, Bienne.
- Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. (2007). Teachers' opinions on quality criteria for Competency Assessment Programs. *Teaching and Teacher Education*, 23(6), 857-867. <https://doi.org/10.1016/j.tate.2006.04.043>
- Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning ? *Assessment in Education: Principles, Policy & Practice*, 18(4), 451–469. <https://doi.org/10.1080/0969594X.2011.557020>
- Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, 11(3), 301–318. <https://doi.org/10.1080/0969594042000304609>
- Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education*, 27(1), 210–220. <https://doi.org/10.1016/j.tate.2010.08.003>
- Brown, G. T. (2022). The past, present and future of educational assessment: A transdisciplinary perspective. *Frontiers in Education*, 7, 1-8. <https://doi.org/10.3389/feduc.2022.1060633>
- Buckley, K., & Lipscombe, K. (2022). Validity and the design of classroom assessment in teacher teams. *The Australian Educational Researcher*, 49(2), 425-444. <http://dx.doi.org/10.1007/s13384-021-00437-9>
- Conférence suisse des Directeurs cantonaux de l'Instruction (CDIP). (2016). *Recommandations relatives à la garantie à long terme de l'accès sans examen aux hautes écoles avec une maturité gymnasiale*. CDIP. <https://edudoc.ch/record/121448 ?ln=fr>
- Conférence suisse des directeurs cantonaux de l'instruction publique (CDIP). (2019). *Évolution de la maturité gymnasiale : un état des lieux*. CDIP. <https://edudoc.ch/record/203995 ?ln=fr>



- Coombs, A., DeLuca, C., LaPointe-McEwan, D., & Chalas, A. (2018). Assessment literacy in teacher education: A review of research. *Cogent Education*, 5(1), 1514687. <https://doi.org/10.1080/2331186X.2018.1514687>
- Colbert, P., Wyatt-Smith, C., & Klenowski, V. (2012). A systems-level approach to building sustainable assessment cultures: Moderation, quality task design and dependability of judgement. *Policy Futures in Education*, 10(4), 386–401. <https://doi.org/10.2304/pfie.2012.10.4.386>
- De Ketele, J. M. (2010). Ne pas se tromper d'évaluation. *Revue française de linguistique appliquée*, (1), 25-37. <https://doi.org/10.3917/rfla.151.0025>
- De Ketele, J.-M., & Roegiers, X. (2013). *Méthodologie du recueil d'informations*. De Boeck.
- Eberle, F., Gehrler, K., Jaggi, B., Kottonau, J., Oepke, M., Pflüger, M., & Huber, C. (2011). *Évaluation de la réforme de la maturité 1995 (EVAMAR II)*. CDIP. https://edudoc.ch/record/101897/files/EVAMAR_II.pdf
- Figari, G. (1994). *Évaluer : quel référentiel ?* De Boeck Université.
- Figari, G., Remaud, D., & Tourmen, C. (2014). Méthodologie d'évaluation par la référentialisation. In G. Figari & D. Remaud (Eds.), *Méthodologie d'évaluation en éducation et formation* (pp. 71-103). De Boeck Supérieur.
- Fulmer, G. W., Lee, I. C. H., & Tan, K. H. K. (2015). Multi-level teacher perspectives on assessment reform. *Assessment in Education: Principles, Policy & Practice*, 22(4), 475–497. <https://doi.org/10.1080/0969594X.2015.1017445>
- Furtak, E. M., Kiemer, K., Circi, R. K., Swanson, R., de León, V., Morrison, D., & Heredia, S. C. (2016). Teachers' formative assessment abilities and their relationship to student learning: Findings from a four-year intervention study. *Instructional Science*, 44, 267-291. <https://doi.org/10.1007/s11251-016-9371-3>
- Goss, P., Hunter, J., Romanes, D., & Parsonage, H. (2015). *Targeted teaching: How better use of data can improve student learning*. Grattan Institute.
- Grapin, N., & Vantourout, M. (2023). Didactiques et formation des enseignants à l'évaluation des apprentissages. *Recherche et formation*, (103). <https://doi.org/10.4000/15hqp>
- Hargreaves, E., Berry, R., Lai, Y. C., Leung, P., Scott, D., & Stobart, G. (2013). Teachers' experiences of autonomy in continuing professional development: Teacher learning communities in London and Hong Kong. *Teacher development*, 17(1), 19-34. <https://doi.org/10.1080/13664530.2012.748686>
- Heredia, S. C., Furtak, E. M., Morrison, D., & Renga, I. P. (2016). Science teachers' representations of classroom practice in the process of formative assessment design. *Journal of Science Teacher Education*, 27(7), 697-716. <https://doi.org/10.1007/s10972-016-9482-3>
- Hermansen, H., & Nerland, M. (2014). Reworking practice through an AfL project: An analysis of teachers' collaborative engagement with new assessment guidelines. *British educational research journal*, 40(1), 187-206. <https://doi.org/10.1002/berj.3037>
- Hipkins, R., & Robertson, S. (2012). The complexities of moderating student writing in a community of practice. *Assessment Matters*, 4, 30-52. <https://doi.org/10.18296/am.0099>
- Holmeier, M., Maag Merki, K., & Hirt, C. (2017). *Gemeinsames Prüfen. Eine Fallanalyse in Gymnasien in der Schweiz*. Springer.
- Laveault, D. (2008). Le jugement professionnel : foyer de tensions et de synergies nouvelles en évaluation scolaire. *Revue suisse des sciences de l'éducation*, 30(3), 483-500. <https://doi.org/10.24452/sjer.30.3.4799>
- Laveault, D., & Yerly, G. (2017). Modération statistique et modération sociale des résultats scolaires : approches opposées ou complémentaires ? *Mesure et évaluation en éducation*, 40(2), 91–123. <https://doi.org/10.7202/1043569ar>



- Leclercq, D. (2005). ETIC PRAD et SEPIAS. Systèmes d'évaluation des performances d'individus en formation. In D. Leclercq (Ed.), *Edumétrie et docimologie pour Praticiens chercheurs* (pp. 1-8). Presses de l'université de Liège.
- Leclercq, D., Philippe, G., & Van der Vleuten C. (2016). Trois méthodes pédagogiques en formations médicale et pharmaceutique : l'APP, les tests de progression et les ECOS. *J Pharm Clin* 2016, 35(1), 7-22. <https://doi.org/10.1684/jpc.2016.0326>
- Leduc, D., & Béland, S. (2022). Rendre compte des pratiques évaluatives en arts au postsecondaire avec le regard de l'évaluation située. In S. Béland & D. Leduc (Eds.), *Regards sur l'évaluation des apprentissages en arts à l'enseignement supérieur* (pp. 27-58). Presses de l'Université du Québec.
- Maroy, C., & Voisin, A. (2014). Une typologie des politiques d'accountability en éducation: l'incidence de l'instrumentation et des théories de la régulation. *Éducation comparée. Revue de recherche internationale et comparative en éducation*, 11, 31-58. <https://doi.org/10.1866/14074>
- Maxwell, G. S. (2002). *Moderation of teacher judgements in student assessment*. Discussion Paper on Assessment and Reporting. Queensland School Curriculum Council.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, 18(2), 5-11. <https://doi.org/10.2307/1175249>
- Miles, M. B., & Huberman, A. M. (2010). *Analyse des données qualitatives*. De Boeck.
- Mons, N. (2009). Effets théoriques et réels des politiques d'évaluation standardisée. *Revue française de pédagogie*, 169, 99-140. <https://doi.org/10.4000/rfp.1531>
- Mottier Lopez, L. (2020). Processus de normalisation dans l'évaluation des apprentissages des élèves. In N. Younès, C. Gremion & E. Sylvestre (Eds.), *Évaluations, sources de synergie*. Presses de l'ADMEE.
- Mottier Lopez, L. (2021). Une évaluation située pour soutenir la capacité d'agir des élèves individuellement et collectivement. In C. Blanc & F.-X. Bernard (Eds.), *Enfants conférenciers. Une expérience éducative, sociale, culturelle* (pp. 287-298). Dunod.
- Mottier Lopez, L., & Allal, L. (2008). Le jugement professionnel en évaluation : un acte cognitif et une pratique sociale située. *Revue suisse des sciences de l'éducation*, 30(3), 465-482. <https://doi.org/10.25656/01:4229>
- Mottier Lopez, L., & Pasquini, R. (2017). Professional controversies between teachers about their summative assessment practices: A tool for building assessment capacity. *Assessment in Education: Principles, Policy & Practice*, 24(2), 228-249. <https://doi.org/10.1080/0969594X.2017.1293001>
- Mottier Lopez, L., Tessaro, W., Dechamboux, L., & Morales Villabona, F. (2012). La modération sociale : un dispositif soutenant l'émergence de savoirs négociés sur l'évaluation certificative des apprentissages des élèves. *Questions Vives*, 6, 159-175. <https://doi.org/10.4000/questionsvives.1235>
- Nilsberth, M., & Sandlund, E. (2021). On the interactional challenges of revealing summative assessments: Collaborative scoring talk among teachers and students in Swedish national tests. *Linguistics and education*, 61(2), 1-16. <https://doi.org/10.1016/j.linged.2020.100899>
- Ochelen, J. P., & Yerly, G. (2025). Faire collaborer les enseignants autour de l'évaluation des apprentissages : Quelles continuités et quelles ruptures dans les pratiques des enseignants ? *Évaluer. Journal international de recherche en éducation et formation*, 12(1), 3-28. <https://doi.org/10.48782/e-jiref-12-1-3>
- Ochelen, J.-P., & Yerly, G. (2026). Faire collaborer les enseignant·e·s pour évaluer les apprentissages des élèves : quel sens aux yeux des enseignant·e·s ? *Revue suisse des sciences de l'éducation*, 48(1). <https://doi.org/10.24452/sjer.48.1.11>
- Ochelen, J.-P., Yerly, G., & Mottier Lopez, L. (2024). Le travail collectif des enseignants pour l'évaluation des apprentissages comme norme professionnelle ? Une revue de la littérature



- pour interroger cette tendance émergente. *Mesure et évaluation en éducation* 47(1), 25-63. <https://doi.org/10.7202/1114565ar>
- Ochelen, J.-P., Mottier Lopez, L., & Yerly, G. (soumis). Les débats professionnels au cœur du travail collectif en évaluation. Analyse des débats entre enseignants dans la construction collective d'épreuves communes au secondaire en Suisse romande.
- Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment. *Teaching and Teacher Education*, 27(2), 472–482. <https://doi.org/10.1016/j.tate.2010.09.017>
- Schellekens, L. H., Kremer, W. D., Van der Schaaf, M. F., Van der Vleuten, C. P., & Bok, H. G. (2023). Between theory and practice: educators' perceptions on assessment quality criteria and its impact on student learning. *Frontiers in Education*, 8, 1-10. <https://doi.org/10.3389/feduc.2023.1147213>
- Schurmans, M.-N. (2009). *La démarche compréhensive*. Presses universitaires de France.
- Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher*, 33(6), 478–485. <https://doi.org/10.3109/0142159X.2011.565828>
- Stewart, T. A., & Houchens, G. W. (2014). Deep impact: How a job-embedded formative assessment professional development model affected teacher practice. *Qualitative Research in Education*, 3(1), 51-82. <https://doi.org/10.4471/qre.2014.36>
- Stiggins, R. J. (2007). *Assessment through the student's eyes*. *Educational Leadership*, 64(8), 22–26.
- UNESCO. (2024). *Données des évaluations d'apprentissage et enquêtes sur les compétences : Défis et solutions proposées*. Conférence de l'UNESCO sur les données et statistiques de l'éducation. Institut de statistique de l'UNESCO. <https://www.uis.unesco.org/sites/default/files/medias/fichiers/2025/08/EDS-7-Learning-outcomes-and-skills-FR-Final-WEB.pdf>
- Van der Schaaf, M., Baartman, L., & Prins, F. (2012). Exploring teacher judgement in competence-based assessment. *Assessment & Evaluation in Higher Education*, 37(5), 555–569. <https://doi.org/10.1080/02602938.2011.557147>
- Van der Vleuten, C. P. M. (2016). Revisiting programmatic assessment. *Medical Teacher*, 38(9), 885–888. <https://doi.org/10.1111/medu.12632>
- Van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & Van Tartwijk, J. (2010). A model for programmatic assessment fit for purpose. *Medical Teacher*, 34(3), 205–214. <https://doi.org/10.3109/0142159X.2012.652239>
- Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education: Principles, policy & practice*, 17(1), 59-75. <https://doi.org/10.1080/09695940903565610>
- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149–162. <https://doi.org/10.1016/j.tate.2016.05.010>
- Yerly, G. (2017). Quel impact des politiques de responsabilisation douce sur les pratiques enseignantes ? In Y. Dutercq & C. Maroy (Eds.), *Le travail et le professionnalisme enseignant face aux politiques de responsabilisation* (pp. 121-141). De Boeck supérieur.
- Yerly, G. (2021). La collaboration des enseignants autour de l'évaluation des apprentissages comme moyen de piloter le système ? Le cas de la politique « évaluation commune » dans le secondaire supérieur en Suisse. *Les dossiers des sciences de l'éducation*, 44. <https://doi.org/10.4000/dse.5322>

Jean-Pascal OCHELEN

Jean-Pascal Ochelen est doctorant aux universités de Fribourg et Genève. Sa thèse porte sur les pratiques collaboratives d'évaluation et s'inscrit dans le cadre du projet FNS dont cette étude est issue.



Gonzague YERLY

Gonzague Yerly est Maître d'enseignement et de recherche à l'université de Fribourg. Ses recherches portent principalement sur les pratiques d'évaluation en classe et sur les politiques d'évaluation en éducation.

Prépublication

