Cristiana CERVINI, *Au-delà de l'erreur : réflexions sur l'évaluation en langues et ses outils*, Repères DoRiF Ateliers Didactique et Recherches - Fédération Alliances Françaises d'Italie et DoRiF Università, DoRiF Università, Roma dicembre 2020, http://dorif.it/ezine

Cristiana CERVINI

AU-DELÀ DE L'ERREUR : RÉFLEXIONS SUR L'ÉVALUATION EN LANGUES ET SES OUTILS

Cristiana Cervini
Dip. di Interpretazione e Traduzione (DIT)
Université de Bologne
cristiana.cervini@unibo.it

Introduction

En contexte scolaire, et plus largement dans celui de l'apprentissage formel, notre activité éducative d'enseignement/apprentissage est parsemée de différents moments d'évaluation, chacun ayant son importance, son rôle, son but : un positionnement, pour connaître le niveau de compétence en langues ou pour la création d'un groupe-classe dont les prérequis soient bien harmonisés ; un diagnostic, pour mieux connaître les points de force et de faiblesse des élèves ; une activité de contrôle continu, tout au long du parcours d'apprentissage, ou encore, une évaluationsommative, pour montrer qu'un certain niveau a été atteint ou qu'un certain examen a été réussi ou non. Cette distinction, qui a fait l'objet de plusieurs études sur l'activité évaluative (PUREN, 2003), met en évidence le rapport bidirectionnel que les deux piliers de la didactique – enseignement/apprentissage et évaluation – entretiennent. Ces études sont souvent accompagnées de remarques concernant les caractéristiques des tests[1] et l'impact qu'ils ont sur la société et sur la vie des candidats. Il est évident que l'évaluation occupe un « espace » de plus en plus important dans les pratiques de classe et dans la vie professionnelle des enseignants. Ces derniers sont censés assumer un grand nombre de postures différentes : celle de concepteur et rédacteur des contenus (devoir qui comporte des connaissances techniques hautement spécialisées), celle de correcteur, et encore, d'évaluateur, qui note et qui décide du destin dans le domaine éducatif, des élèves. Un rôle de grande responsabilité, pour lequel parfois aucune formation spécifique n'est proposée, et qui demande donc de l'éclectisme et de la bonne volonté.

Dans cette contribution, nous nous proposons d'explorer l'évaluation dans le domaine des compétences en langues et d'ébaucher quelques concepts-clés, afin d'éclairer les pratiques quotidiennes des enseignants et être ainsi plus conscients vis-à-vis des risques de subjectivité de l'évaluation.

1. Les différentes facettes de l'évaluation en langues

Les étapes d'enseignement/apprentissage d'une langue étrangère sont jalonnées par de nombreux moments d'évaluation – initiale, intermédiaire, finale. Même si tous rentrent sous le même chapeau définitoire, ces moments sont caractérisés par des modalités et des buts très variés. Par exemple certaines activités (tâche ouverte, exercice à réponse fermée, etc.) ont le pouvoir de solliciter certains comportements langagiers et sont plus ou moins adaptées à dévoiler certaines habilités. Un exercice à réponse fermée pourra être plus difficilement un témoin fiable des capacités d'écriture des élèves (cohérence, cohésion, lexique actif…). De plus, les opérations langagières activées par les différentes typologies d'activité pourront faire ressortir ou pas certaines fautes. Selon le construit[2] de compétence en langue visé dans le test, concept au cœur des buts évaluatifs, les correcteurs pourront décider d'attribuer plus ou moins d'importance à certaines fautes par rapport à d'autres. Cela signifie que la même faute pourra être prise en considération, voire sanctionnée, de manière différente selon le construit du test et selon les buts déclarés pour chaque épreuve d'évaluation.

Ces premières réflexions mettent l'accent sur des concepts concernant le rapport entre type de tâche, compétence visée, type de notation et correction possibles, des concepts qui mériteraient d'être explorés de manière plus approfondie. Comme Tardieu (2009 : 11-13) nous le rappelle, la définition du verbe « évaluer » dans le *Trésor de la Langue Française*, contient trois notions distinctes, qui semblent être au moins partiellement en contradiction entre elles : i) déterminer, délimiter, fixer avec précision ; ii) conjecturer, faire l'estimation d'une quantité, d'une durée qui n'est pas encore vérifiable ; iii) reconnaître la valeur de, estimer.

Comment concilier donc « détermination précise », sens transmis dans le premier volet du terme, avec celui d'« estimation approximative », plus proche du troisième sens ? Si au premier abord ce sont les nuances contradictoires des définitions qui émergent, une réflexion plus approfondie dévoile que c'est plutôt l'ensemble des trois à mieux représenter la complexité des différentes formes et démarches de l'évaluation elle-même. « La première définition suggère un jugement précis, avéré, objectif, relatif à une norme » (ibid.). Depuis longtemps, les tests étiquetés comme objectifs sont proposés en opposition aux tests définis comme étant subjectifs (ROMAINVILLE, 2012). L'adjectif « objectif » est utilisé dans la plupart des cas en tant que synonyme de standardisé, tandis que « subjectif » se charge souvent de nuances négatives à cause de son association avec des tests où la personnalité de l'évaluateur, ainsi que ses préférences personnelles, jouent un rôle plus important (CECRL, 2001 : 142-143). En fait, si d'un côté il est indéniable que l'objectivité est à la base d'une évaluation équitable, de l'autre cette notion reste fictive et se révèle dans toutes ses limites (PORCELLI, 2006). Par test « objectif » on se réfère à des dispositifs à réponse fermée dont l'objectivité est poursuivie par des procédures de standardisation, notamment des validations psychométriques menées sur un large échantillon de candidats. Il s'agit donc d'un instrument de mesure qui produit des résultats sur la base de normes préétablies et de modèles de performance. La qualité de son fonctionnement est déclarée dans les coefficients de fidélité et de validité, obtenus suite aux différentes étapes de validation. Ces coefficients sont censés être toujours affichés de manière transparente et publique, par exemple dans les spécifications du test. Cependant, le concept de « norme » suscite de plus en plus de débats. Quelle est la bonne norme à suivre ? Est-ce qu'il y a une norme meilleure que d'autres ? Par rapport à quels aspects ? N'est-elle pas une représentation fictive et faussée d'un modèle de compétence, imposé par les politiques linguistiques dominantes ? (McNAMARA, 2019 ; MAURER, PUREN, 2019). En fait, il n'est pas anodin d'établir une norme valable pour différentes typologies de contextes, et juste pour tout étudiant (candidats au test), sans nuire au respect de ses caractéristiques personnelles (buts, motivation, âge, biographie langagière, etc.).

L'objectif déclaré de la standardisation, comme dans le cas des tests de certification, serait de produire des résultats utiles (DOUCET, 2001), qui ne défavorisent, ou au contraire favorisent, certains candidats. Un test pourrait se révéler inutile et parfois nuisible dans plusieurs cas : i) quand ses résultats sont influencés par des facteurs cachés et incontrôlables (qui n'ont rien à voir avec la réelle compétence en langues des candidats) ; ii) quand les préférences subjectives de l'évaluateur influencent les prises de décision et les résultats produits par le test ; iii) en cas d'incohérence entre le construit du test et les contenus proposés.

À propos de la fiction qui caractériserait le concept d'objectivité dans le testing, Porcelli (2006) précisait que même les contenus d'un test « objectif » ont été choisis ou créés par un rédacteur humain, qui a peut-être opéré des sélections subjectives. De plus, il soulignait que la propension cognitive et les capacités de concentration des candidats pourraient biaiser les résultats aux tests. Surtout face à des listes de questions à réponse fermée où certains profils cognitifs pourraient être désavantagés sans que le résultat n'ait rien à voir avec les connaissances en langues. Tous ces facteurs seraient évidemment en contradiction avec la valeur d'objectivité.

Pour ce qui est de la dimension de la subjectivité, elle ressort explicitement dans la seconde définition du TLF qui « entend plutôt un jugement global, prédictif et subjectif » (TARDIEU, 2009 : 3) et dans la troisième, « qui autorise la dimension subjective, voire affective » (ibid.). La subjectivité ouvre les portes aux différentes facettes de l'interprétation d'un comportement langagier, d'une performance, d'un résultat et bien sûr à la difficulté de délimiter la compétence en langues dans sa multi-dimensionnalité et complexité. Le côté subjectif de l'évaluation est souvent associé à des tâches qui demandent une performance ; c'est le cas des productions et interactions à l'écrit ou à l'oral, dans le cadre méthodologique des approches communicatives simples, ou bien scénarisées, de type actionnel. La subjectivité se manifeste sous plusieurs facettes : les consignes, qui peuvent être interprétées de différentes façons, les attentes sur les caractéristiques de la performance, les critères de référence pour l'évaluation qui sont les plus variés. Parfois ces critères sont explicitement affichés dans des grilles ou des rubriques, parfois ils restent vagues et implicites, cachés dans la tête de l'enseignant.

Le type d'épreuve proposé (*à réponse ouverte, à réponse fermée, brève ou longue, formative ou sommative*) influence les modalités de correction et, par conséquent, les résultats finaux. Est-ce que le focus de l'activité est sur la globalité de la performance et sur le résultat obtenu suite aux actions langagières entreprises ? (par ex. mise en pratique du savoir linguistique dans une situation donnée -> *changer un ticket d'avion*) ; ou bien est-ce qu'il tombe sur la compétence ? (par ex. *comprendre un message, produire des énoncés clairs,* etc.) ou encore sur les connaissances ? (grammaticales, lexicales, culturelles -> *usage des pronoms personnels, histoire des rois de France, etc.).* C'est la différence qui existe entre tester la connaissance d'un élément (dimension parcellisée de la compétence en langue) et tester une performance langagière de type communicatif, où plusieurs dimensions vont déterminer le résultat global (compréhension de la consigne, précision lexicale, justesse syntaxique, compétence socio-pragmatique, aspects phonétiques, etc.).

Dans tous les cas, à propos de l'objectivité et de la subjectivité des résultats de l'évaluation, la pensée de Piéron est toujours actuelle et partageable : « Piéron avait calculé qu'il faudrait…127 correcteurs pour obtenir la valeur vraie d'une dissertation philosophique et 13 pour un devoir de mathématiques, pourtant réputées sciences exactes ! » (TARDIEU, 2009 : 5).

2. Le rôle de l'évaluation au fil des années

D'après une perspective historique et diachronique, le rôle attribué à l'évaluation et à ses pratiques a évolué en parallèle avec les courants méthodologiques dominants en didactique des langues (SOMMER, 2001). Les approches didactiques se sont en fait reflétées dans les caractéristiques des tests en termes de typologies d'exercices ou d'activités proposées, mode de correction et de notation, type de *feedback*, prise en compte de l'authenticité en tant que valeur. Cependant, cette tendance n'est pas à compartiments étanches. Plus précisément, si actuellement les tâches scénarisées centrées sur la production/interaction sont très exploitées dans l'évaluation en raison du fait qu'elles se collent plus fidèlement à la réalité interactionnelle et situationnelle, une grande quantité de tests récents incluent des questions à réponse unique et fermée (items de type vrai/faux, choix multiple, test de closure, etc.). Cela se produit quel que soit le but du test (positionnement, diagnostic, certification) ou son support (numérique ou papier). Parfois, dans certains tests, les approches sont mélangées : des items discrets, technique largement diffusée durant la période structuraliste, sont proposés pour tester les connaissances morphosyntaxiques, lexicales, orthographiques à côté des tâches à réponse ouverte ; en même temps, les modalités de correction et de feedback vont bien au-delà du simple 'correct/faux', pour poursuivre un but plus formatif, utile aux progrès des élèves. C'est le cas par exemple du test diagnostic multilingue DIALANG[3] où chaque item est associé à une description de la connaissance/compétence visée et

à la/aux réponse(s) correcte(s), quand plusieurs alternatives sont envisagées. Ou encore c'est aussi le cas du test SELF[4] (CERVINI, JOUANNAUD, 2016 ; MASPERI, 2011), le système d'évaluation en langues à visée formative, où chaque item est associé à une fiche d'identité qui décrit ses focalisations langagières (morphosyntaxiques, lexicales, socio-pragmatiques) et les opérations cognitives sollicitées.

C'est autour des années 1950 que la phase scientifique des tests a démarré : au cœur des intérêts de l'évaluation il y avait les composantes de la langue (les mots, les traits morphosyntaxiques, les aspects formels), questionnées surtout par des phrases courtesdécontextualisées, des items discrets comme les QCM, rapides à corriger et à valider statistiquement. Le feedback le plus usuel se limitait à une distinction entre les réponses correctes et les fausses, sans viser à un impact formatif plus poussé pour les répondants. Les avantages de ces tests consistaient principalement dans la possibilité de gérer de grandes quantités d'items/questions et, plus tard, d'automatiser la correction, dans le cadre de l'évaluation assistée par ordinateur. Au cours des années 1970 on assiste à une évolution dans la conception des épreuves : dans le testing intégratif, on commence à mettre de côté les phrases morcelées et décontextualisées, pour privilégier la dimension authentique du texte, à l'écrit ou à l'oral. Les techniques utilisées sont de plus en plus élaborées : des tests de closure, des dictées, des traductions interlangues, etc. Par ces techniques plusieurs opérations langagières et cognitives sont sollicitées en même temps ; il devient donc moins évident de cerner les aspects de la compétence en langues qu'on est en train de tester. Puis, les approches communicatives à partir des années 1980, et les approches actionnelles suite à la publication du CECRL en 2001, marquent un tournant dans l'évaluation et la didactique des langues-cultures. Les descripteurs du CECRL mettent en évidence la multi-dimensionnalité et la complexité des habilités réceptives, productives, interactives, et insistent sur la valeur sociale et interculturelle de la compétence communicative et du plurilinguisme. Connaître les langues signifie pouvoir agir e interagir dans des communautés plurilingues d'acteurs sociaux. Les tâches et les activités sont de plus en plus contextualisées d'un point de vue social et culturel, selon des critères d'authenticité situationnelle et interactionnelle où la compétence pragmatique est reine. Pour évaluer les compétences en langues dans le cadre des approches communicatives et actionnelles, il faut pouvoir déterminer et expliciter les critères à suivre et leurs poids. À ce propos, surtout les tests à fort enjeu, comme c'est le cas des certifications en langues, sont toujours accompagnés de rubriques. Dans ces rubriques, les critères d'observation et jugement sont explicitement déclarés et jouent un rôle de guide pour les enseignants-évaluateurs. Ils constituent en même temps un point de repère harmonisé et un engagement de transparence vis-à-vis des évalué(e)s. En fait, même dans le cadre d'un construit de compétence actionnelle et de tâche scénarisée, personne ne nous empêche d'évaluer les connaissances morphosyntaxiques des candidats, et de leur attribuer un certain poids. La cohérence entre les finalités du test (par ex. compétences communicatives communes à toutes les situations professionnelles courantes), les types de tâche et d'item proposés (à réponse fermée, à réponse ouverte, courte ou extensive…) et les critères d'évaluation déclarés dans la grille d'évaluation constitue un véritable atout pour le respect du principe d'équité dans le testing. Depuis plusieurs années, les associations du testing en langues (ILTA, ALTE, EALTA, etc.) se sont engagées dans la mise au point de codes éthiques, à diffuser dans la communauté scientifique, afin de concevoir des dispositifs d'évaluation qui soient le plus possible équitables et justes.

3. Et à l'ère actuelle ?

À l'ère actuelle, le débat au sujet de l'évaluation et du testing est toujours très vif et, entre autres, semble tourner autour de certains points sensibles que nous allons explorer ci-dessous :

i) comment assurer qualité et sûreté dans l'évaluation à distance ? Tout récemment l'émergence Covid-19 a obligé les institutions à modifier rapidement et radicalement les pratiques adoptées pour enseigner et évaluer. Dans l'impossibilité de proposer des solutions faisables en présentiel, les efforts ont été adressés vers l'administration des dispositifs d'évaluation à distance et en ligne, sur des plateformes dédiées. Mais comment pouvoir garantir les mêmes conditions et les mêmes possibilités d'accès à tous les candidats qui ont la nécessité de soutenir un examen, parfois à fort enjeu pour leur avenir (admission universitaire, renouvellement du permis de séjour, etc.) ? Et encore, comment contrôler et éviter des tricheries et la fuite des contenus confidentiels ? Depuis des années les associations du testing ont pris une position claire dans leurs codes éthiques et dans leurs lignes directrices sur les principes d'utilité et équité. Par exemple l'association EALTA[5], dans ses lignes directrices[6], à la voix *Administration des tests*, ouvre des questions pour inviter à la réflexion et pour servir de contrôle-qualité : « Quels sont les dispositifs pour assurer la sécurité ? Les personnes chargées de l'administration des tests sont-elles formées ? L'administration du test est-elle contrôlée ? ».

Dans la même perspective, dans les lignes directrices d'ILTA[7] (2010 : 3), on affirme **:** « Test materials should be kept in a safe place and handled in such a way that no test taker is allowed to gain an unfair advantage over other test takers. Care must be taken to ensure that all test takers are treated in the same way in the administration of the test».

Enfin, les principes de transparence, respect de la confidentialité et de l'équité sont clairement mis en évidence aussi dans le document d'ALTE[8] *Normes minimales relatives à l'établissement de profils qualité pour les examens* :

> 6. Tous les centres sont sélectionnés pour administrer votre examen conformément à des procédures claires, transparentes et dûment établies ; ils ont par ailleurs accès aux réglementations régissant les modalités d'administration. 7. Les documents d'examen sont livrés aux centres d'examen agréés en excellent état et par des moyens de transport sûrs, votre système d'administration des examens assure la sécurité et la traçabilité de tous les documents d'examen, et la confidentialité de toutes les procédures du système peut être garantie. […]. 9. Vous protégez la sécurité et la confidentialité des résultats et des certificats, et des données s'y rattachant, conformément à la législation en vigueur sur la protection des données, et les candidats sont informés de leurs droits d'accès à ces données (2007 : 1).

ii) défense du plurilinguisme et contrebalancement aux pressions de l'anglicisation : cet aspect est très complexe et cette contribution

nous permet de l'esquisser rapidement, en référence à l'évaluation et aux certifications en langues. C'est un fait indéniable que les professeurs de langues se trouvent de temps en temps, voire fréquemment, impliqués en première ligne pour défendre la valeur du plurilinguisme dans l'éducation et la certification des compétences en langues. En guise d'exemple, tout récemment des associations de professeurs de langues et civilisations étrangères[9], ainsi que la direction du CLES[10], se sont mobilisées en France contre l'arrêté du 3 avril 2020[11] relatif aux intentions d'imposer une certification en langue anglaise pour les candidats inscrits aux diplômes nationaux de licence, de licence professionnelle et au diplôme universitaire de technologie. Mais le débat n'est pas d'aujourd'hui, ou limité à la réalité francophone ; déjà en 2009 Pierre Frath parlait du paradoxe du multilinguisme en Europe, et affirmait : « malgré tout ce volontarisme politique et toutes ces connaissances pédagogiques, la diversité des langues enseignées en Europe recule, et le multilinguisme se limite le plus souvent à l'apprentissage d'un anglais d'aéroport » (FRATH, 2009 : 1). Bien évidemment, l'obligation de l'anglais à l'université aurait un impact immédiat sur le système éducatif qui, d'après les effets de retour (*backwash effects*) (ALDERSON, 1993), serait amené à détourner ses efforts et ses ressources, souvent limitées, vers la préparation des étudiants à la réussite du test. En fait l'échec à un test à fort enjeu, comme c'est souvent le cas pour les certifications linguistiques, comporterait pour les élèves, dans la plupart des cas, l'impossibilité d'avancer dans leurs propres projets. Au-delà des limites dans les ressources économiques disponibles, il y a aussi des contraintes non négligeables liées aux emplois du temps des curriculums scolaires et universitaires.

La valeur du plurilinguisme dans l'évaluation s'exprime aussi par d'autres moyens, par exemple la déclination d'un même test sur un vaste éventail de langues. C'est le cas par exemple de Dialang (ALDERSON, 2006), ou de la certification CLES (ROUVEYROL *et al.*, à paraître), ou encore de SELF (CERVINI, 2016). Tous ces dispositifs s'appuient sur une même approche méthodologique et sur un même processus de validation et calibration, mais ils sont déclinés dans plusieurs langues[12]. Enfin, mais non des moindres, nous assistons à la multiplication des projets pour l'évaluation des compétences plurilingues en intercompréhension. C'est le cas par exemple du projet EVAL-IC (2016-2019)[13]. Ces nouveautés témoignent des transformations du modèle de compétence en langues : à la base du construit du test il y a les habilités de compréhension entre langues apparentées, auxquelles viser pour une vraie communication plurilingue.

iii) l'importance de l'évaluation formative : cette question vient se greffer sur d'autres aspects centraux pour le débat actuel, comme l'utilité des tests, les effets rétroactifs de type émotionnel des pratiques évaluatives et, par conséquent, le rôle du feedback. Hadji (2012) nous met en garde contre les dérives négatives de certaines activités évaluatives, comme son effet anxiogène, l'obsession de la mesure et du chiffrage, le culte de la performance et de la compétitivité. « L'omniprésence tyrannique de la notation installe un climat de stress tel que la pression exercée sur les élèves devient contre-productive » (2012 : 1-2). La mise en valeur des finalités sociales et éducatives de l'évaluation formative, libérée de la nécessité d'attribuer des scores ou des notes finales, permettrait de décontracter le climat dans les classes de langues et de favoriser des interactions plus souples, collaboratives et transparentes entre enseignant et étudiant, ou entre étudiant/étudiant. Le concept de transparence nous semble très pertinent parce que dans l'évaluation formative les élèves bénéficient aussi de la mise à disposition de feedbacks approfondis, clairs, centrés sur les caractéristiques de chacun et parce qu'un climat collaboratif et de partage devrait réduire les essais de tricheries, dans le seul but de réussir une épreuve ou d'avoir des notes plus élevées.

> Le but prioritaire de l'évaluation formative, dans sa conception et sa mise en pratique, c'est l'amélioration des apprentissages de chaque élève. Cette activité évaluative aide aux apprentissages lorsqu'elle produit de l'information que les enseignants et les élèves peuvent utiliser comme un feedback, comme un retour, pour se situer eux-mêmes et pour modifier les activités dans lesquelles ils sont engagés. Une telle évaluation devient formative si indices et retours sont utilisés pour rencontrer les besoins de chacun des élèves (GRANGEAT, 2014 : 1).

3.1 Pratiques d'évaluation formative au DIT

L'évaluation formative joue un rôle essentiel dans la formation des futurs interprètes à l'université de Bologne (M2 en interprétation, *corso di laurea magistrale in interpretazione*), par une valorisation constante et continue de l'autoévaluation et de l'évaluation collaborative entre pairs. De quelle manière ? Dans les cours où nous sommes en première ligne, de nature théorique et pratique à la fois (linguistique pour interprètes, communication institutionnelle et langues de spécialité), les étudiants sont souvent impliqués dans des activités de reformulation intralinguistique, italien-italien, à l'oral. L'écoute active d'un discours public, de nature institutionnelle ou politique, peut être suivie par une « interprétation consécutive » ou bien par une synthèse guidée des concepts principaux. Selon le type de tâche prévu, la prise de note est permise ou interdite. Dans un contexte synchrone (qui peut être présentiel en classe ou à distance sur plateforme), pour chaque « interprète » qui restitue son discours, deux autres rôles sont engagés en première ligne : un *tuteur de soutien* qui peut intervenir en cas de blocage pendant les activités de reformulation, et un *tuteur évaluateur* qui est censé donner un feedback ponctuel et motivé à l'interprète. Cette triade, répétée maintes fois, permet d'engager activement une grande partie du groupe-classe et de mettre au centre les impressions et les réflexions des participants. Il s'agit donc d'une implication collégiale qui a le double effet positif de garder un certain dynamisme et de décontracter les tensions souvent associées à des activités de performance ou l'élève s'expose publiquement. À la base de ce travail de réflexion et d'évaluation collective, il y a une grille d'évaluation de référence, qui a été co-construite ensemble, en amont de l'activité. Cette grille détaille deux aspects principaux à considérer pour l'hétéroévaluation ou l'autoévaluation de l'activité en cours : usage de la langue (par ex. qualité du lexique, clarté d'exposition, fluidité, cohésion, etc.) et fidélité entre discours original de départ et discours cible. De cette manière, les feedbacks sont plus utiles et plus cohérents et une démarche de soutien proactif, de type linguistique et de socio-affectif à la fois, s'installe (QUINTIN : 2011).

Des approches similaires, qui prévoient la co-construction des grilles de référence et l'évaluation collaborative entre pairs, pourraient être adoptées aussi dans un contexte d'enseignement/apprentissage des langues à l'école. Quelques expressions de timidité ou de résistance sont à prévoir, surtout vis-à-vis de l'activité inhabituelle de commenter explicitement et franchement la proposition d'un pair. Les interactions gagnent en fluidité si les enseignants mettent tout de suite en clair les règles du jeu : il ne s'agit pas de critiquer la proposition d'un copain, mais plutôt de partager avec lui, et avec la classe, des suggestions, des conseils constructifs, et des impressions.

4. Réduire la subjectivité de l'évaluation ? Suggestions des pratiques de « standard setting » à utiliser en classe

D'après ce que nous venons de lire, on pourrait craindre que l'évaluation en classe de langues aboutisse irrémédiablement à des résultats subjectifs, donc partiels et probablement inexacts. Nous avons déjà dit que l'impartialité est, dans la plupart des situations, une aspiration, un idéal et que l'exactitude de la mesure et du score ne correspond pas forcément à une vision fidèle de la compétence de notre élève. Toutefois, en regardant attentivement le cycle de validation des tests de ALTE (2011), nous remarquons que parmi les dernières opérations à accomplir avant le déploiement d'un test il y le *standard setting* (CIZEK, 2001). Le standard setting est une étape essentielle dans la conception d'une épreuve évaluative, dans le but d'atteindre un consensus généralisable, par exemple entre différents évaluateurs, autour des concepts de niveau, de réussite, d'échec, etc.

> En éducation, le concept de standard est à l'œuvre principalement lorsqu'il faut établir des niveaux (des échelons gradués) pour différencier les apprentissages ou les performances des individus. Il se réfère, d'une part, au contenu qui fait l'objet de l'apprentissage et, d'autre part, a la performance des élèves dans l'apprentissage de ce contenu. Il existe ainsi un contenu standard ou commun à aborder (i.e. le programme), et il y a une exigence minimale à satisfaire pour réussir selon ce qui est attendu socialement. Cette exigence minimale s'appelle un standard de réussite (BLAIS, 2008 : 95).

Les méthodes pour établir des standards en éducation étaient généralement distinguées entre deux grandes catégories : celles qui sont centrées sur l'individu et celles qui sont centrées sur le test. Toutefois, plus récemment, Hambleton, Jaeger, Plake et Mills (2000) ont proposé une classification plus fine qui distingue l'examen des items et des rubriques d'attribution des scores, des méthodes qui impliquent l'examen des candidats (performance ou produits), ou bien des profils de scores (BLAIS, 2008 : 97-98). Probablement, dans le domaine de l'évaluation des compétences en langues, le plus grand effort récent de standardisation a été fait pour la publication des descripteurs du CECRL (2001, 2016) auxquels toutes les certifications linguistiques européennes sont adossées.

Lors de l'atelier pratique pour les professeurs de FLE[14], nous nous sommes confrontés sur le type d'usage que chacun d'entre nous fait du CECRL dans son activité didactique quotidienne (enseignement, préparation des contenus, évaluation). Nous avons discuté de différentes interprétations des descripteurs et de la perception de leur utilité pour nos enseignements. Ce dernier point a fait l'objet d'une véritable activité pratique, qui se proposait de mettre en évidence les variations dans la perception de qualité d'une performance en FLE. C'est à cause de ces différences que certains panelistes (terme utilisé pour indiquer les évaluateurs impliqués dans les sessions de standard setting) sont étiquetés comme plus sévères que d'autres. Les ateliers de standardisation sont finalisés à améliorer la qualité des tests et la qualité des prises de décisions, pour produire des résultats plus cohérents et harmonisés, dans le cadre d'une même institution, ou bien dans le cadre d'un même test proposé à des publics différents. En détail, par les processus de standardisation, on peut mettre au point le lien entre les échelles du CECRL et les contenus des examens (activité, tâches, items). Il s'agit d'une validation qualitative obtenue par consensus. Les utilisateurs directs et indirects du test sont impliqués dans la validation à travers une discussion guidée et modérée. Cette discussion vise l'acceptation sociale des décisions qui seront prises par les résultats du test et vise la neutralisation des biais[15] dus aux différentes pratiques et aux différentes interprétations du CECRL.

Concrètement, pendant l'atelier, nous avons simulé les étapes d'un véritable processus de standardisation, en les adaptant à notre but, qui était plus souple et nous laissait plus de liberté d'action. Tous les participants, professeurs de FLE au collège ou au lycée en Italie, ont relu attentivement les descripteurs du CECRL pour la « production orale générale », niveaux pré-A1-C2 (2016 : 72) ainsi que les descripteurs sur les « aspects qualitatifs de l'utilisation de la langue parlée » (2016 : 181-182). Suite à cette lecture attentive nous avons travaillé, en groupes de trois, à la création d'une grille de référence pour évaluer des productions/interactions orales, ancrée sur l'approche communicative. Puis, afin de comparer les différentes représentations des nœuds de la compétence communicative à l'oral (selon une perspective linguistique, sociopragmatique et interactionnelle), les grilles ont circulé d'un groupe à l'autre et elles ont été analysées et intégrées. Une version unique de la grille a donc été rédigée et utilisée par tout le monde dans l'étape suivante. Lors de l'étape suivante, trois vidéos contenant des discours en FLE prononcés par des non-francophones, ont été transmises. Les participants devaient attribuer à chaque apprenant un niveau en production orale, selon les descripteurs du CECRL, et selon les critères plus précis affichés dans la grille. Une brève introduction à la vidéo était mise à disposition, avec une description synthétique de la biographie langagière des élèves, leur âge et les sujets traités dans les monologues et les dialogues[16]. Cette étape, menée individuellement, a été suivie, comme dans les ateliers de standard setting classique, d'une confrontation par petits groupes (deux enseignants) et ensuite par groupes un peu plus larges (en quatre, puis huit). Tout le monde a verbalisé les raisons les plus importantes (des fautes, des choix lexicaux intéressants, la prononciation, les reformulations, etc.) qui ont déterminé l'attribution d'un certain niveau. La prise de décision devait s'appuyer aussi sur des exemples, en positif ou en négatif, tirés des discours en interlangue. Le but final était d'atteindre un consensus et, le cas échéant, de revoir sa propre opinion, si l'évaluateur avait été trop sévère, ou au contraire, trop indulgent, par rapport aux autres collègues. Enfin, les décisions prises ont été comparées avec celles publiées dans le rapport officiel du CIEP pour le calibrage des productions orales des candidats sur les 6 niveaux de l'échelle du CECRL[17].

Cette activité à entendu montrer le côté interprétatif et subjectif lié à toute prise de décision sur l'évaluation des performances et des descripteurs. Une réflexion active de ce type peut aider à prendre conscience des contraintes et des difficultés des processus évaluatifs et de possibles influences culturelles et interculturelles (TARDIEU : 2010). Toutefois, c'est également un moyen pour décontracter les tensions qui caractérisent l'évaluation, grâce à un processus de confrontation entre pairs.

Réflexions conclusives

Le débat de la communauté scientifique autour de l'évaluation des compétences en langues est de plus en plus riche et vif et strictement influencé par les tendances des politiques linguistiques, culturelles et éducatives, au niveau européen et international. À côté des progrès théoriques et des aspects politiques qui caractérisent cette discipline, les pratiques évaluatives sont au centre des activités quotidiennes des professeurs de langues, dans tous les cycles du système éducatif, de l'école primaire à l'université. Il est donc très important de prévoir des moments de formation pour réfléchir de façon partagée sur les différents rôles que le prof de langues doit assumer dans les différents moments de l'évaluation : formative, sommative, diagnostique, initiale, intermédiaire, finale, des compétences, des connaissances, des performances, etc. Il y aussi des moyens très intéressants pour décontracter les moments dédiés à l'évaluation et les exploiter pour améliorer les apprentissages et l'esprit de collaboration entre pairs.

Références bibliographiques

ALDERSON, Charles, WALL, Dianne, « Does washback exist? », *Applied Linguistics* 14/2, 1993, p. 115-129.

ALDERSON, Charles, *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*, Londres, New York, Continuum Editions, 2006.

ALTE, *Multilingual glossary of language testing terms*, Cambridge, Cambridge University Press, 1998.

ALTE, *Manuel pour l'élaboration et la passation de tests et d'examens de langue.* Division des Politiques linguistiques. Strasbourg : Conseil de l'Europe, DG II – Service de l'éducation, 2011.

BLAIS, Jean-Guy, « Les standards de performance en éducation », *Mesure et évaluation en éducation*, *31*(2), 2008, p. 93–105.

CONSEIL DE L'EUROPE, *Cadre Européen Commun de Référence pour les Langues: apprendre, enseigner, évaluer,* Paris, Didier, 2001.

CONSEIL DE L'EUROPE, *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer. Volume complémentaire avec de nouveaux descripteurs*, Conseil de l'Europe, Programme des Politiques linguistiques, Division des Politiques éducatives, Service de l'Éducation, 2016 (2018 en français). En ligne : https://rm.coe.int/cecr-volume-complementaire-avec-de-nouveaux-descripteurs/16807875d

CERVINI, Cristiana, JOUANNAUD Marie-Pierre, « Ouvertures et tensions liées à la conception d'un système d'évaluation numérique multilingue en ligne dans une perspective communicative et actionnelle », *ALSIC – Apprentissage des langues et systèmes d'information et de communication*. Numéro spécial 'Des machines et des langues', 2015, http://journals.openedition.org/alsic/282

CERVINI, Cristiana, «*Approcci integrati nel testing linguistico: esperienze di progettazione e validazione in prospettiva interlinguistica», Interdisciplinarità e apprendimento linguistico nei nuovi contesti formativi. L'apprendente di lingue tra tradizione e innovazione*, Bologna, Quaderni del CESLiC, 2016, p. 64 – 85.

CIZEK, G. J. (dir.), *Setting performance standards: Concepts, methods and perspectives*. Mahwah, Lawrence Erlbaum, 2001.

DOUCET, Patrick, « Pour un test utile », *ASp* 34, 2001, http://journals.openedition.org/asp/1696

FRATH, Pierre, « Le paradoxe du multilinguisme », publié en traduction estonienne dans le numéro de SIRP du 13 mars 2009, Talinn, Estonie, http://www.sirp.ee

GRANGEAT, Michel, « Connaître les principes de l'évaluation formative », 2014, https://www.pedagogie.ac-nantes.fr/medias/fichier/principes-eval-form-grangeat_1462960655128-pdf?ID_FICHE=468651&INLINE=FALSE

HADJI, Charles, *Faut-il avoir peur de l'évaluation ?,* Bruxelles, De Boeck, 2012.

HAMBLETON, R.K., JAEGER, R.M., PLAKE, B.S., & MILLS, C.N., « Setting performance standards on complex educational assessments », *Applied Psychological measurements* (24/4), 2000.

QUINTIN Jean-Jacques, « Chapitre 4. L'efficacité des modalités d'intervention tutorale et leurs effets sur le climat socio-relationnel des groupes restreints », dans : Christian Depover éd., *Le tutorat en formation à distance.* Louvain-la-Neuve, De Boeck Supérieur, « Perspectives en éducation et formation », 2011, p. 61-86. DOI : 10.3917/dbu.depov.2011.01.0061. URL : https://www.cairn.info/le-tutorat-en-formation-a-distance--9782804163426-page-61.htm

MASPERI, Monica, « Innovalangues : Innovation et transformation des pratiques de l'enseignement-apprentissage des langues dans l'enseignement supérieur ». MESRI. ANR. Investissements d'avenir. https://hal.archives-ouvertes.fr/hal-02004250.

MAURER, Bruno, PUREN, Christian, *CECR : par ici la sortie !*, Editions des archives contemporaines, France, 2019.

McNAMARA, Tim, *Language and Subjectivity*, Cambridge, Cambridge University Press, 2019.

PORCELLI, Gianfranco, «Verifiche comode e verifiche valide», in Jafrancesco E. (a cura di), *La valutazione delle competenze linguistico-comunicative*, Roma, Edizioni Edilingua, 2006.

PUREN, Christian, « Relations entre activités d'évaluation, activités d'apprentissage et d'usage : un chantier à reprendre en didactique des langues », *L'évaluation en langue : pour qui ? pour/quoi ? comment ?*, IUFM de Rouen, 2003.

ROMAINVILLE, Marc**,** « Objectivité *versus* subjectivité dans l'évaluation des acquis des étudiants », *Revue internationale de pédagogie de l'enseignement supérieur* [En ligne], 27(2) | 2011, http://journals.openedition.org/ripes/499

ROUVEYROL, Laurent, BARDIERE, Yves, CHOUISSA, Catherine, « Le CLES, levier de la politique des langues dans l'Enseignement Supérieur », *Actes du Colloque sur la gouvernance linguistique des universités et établissements d'enseignement supérieur*, A paraître. ⟨hal-02053970⟩.

SOMMER, Sylvia, « La nécessaire interaction entre évaluation et processus d'apprentissage en langues », *ASp* [En ligne], 34, 2001, http://journals.openedition.org/asp/1706

TARDIEU, Claire, « Corriger ou évaluer ? », *Cahiers de l'APLIUT* [En ligne], Vol. XXVIII N° 3 | 2009, http://journals.openedition.org/apliut/65

TARDIEU, Claire, « Votre B1 est-il mon B1 ? L'interculturel dans les tests d'évaluation en Europe », *Les Cahiers de l'Acedle*, volume 7, numéro 2, 2010.

1
Dans cet article le mot *test* est employé en tant qu'hyperonyme pour indiquer toute typologie d'épreuve évaluative.

2
« attribut hypothétique des individus ou opération mentale qui ne peut être directement ni mesurée ni observée (par exemple, en évaluation des langues, la capacité de compréhension orale). Les tests de langue essaient de mesurer les différents construits qui sous-tendent les capacités langagières […] » (ALTE, 1998 : 216).

3
DIALANG is a freely available language diagnosis system. It was developed by a consortium of European higher education institutions with support from the European Commission's Socrates programme (1996-2004) and is now funded and maintained on a "pro bono" basis by Lancaster University (UK). DIALANG tests reading, writing, listening, grammar and vocabulary in 14 languages: Danish, Dutch, English, Finnish, French, German, Greek, Icelandic, Irish-gaelic, Italian, Norwegian, Portuguese, Spanish and Swedish. It reports your level of skill against the Common European Framework of Reference (CEFR) for language learning. Since its inception, several million DIALANG test sessions have been recorded.
http://wp.lancs.ac.uk/ltrg/projects/dialang-2-0/ (consulté le 20/05/2020).

4
SELF - http://innovalangues.fr/realisations/systeme-d-evaluation-en-langues-a-visee-formative/ (consulté le 20/05/2020).

5
« Association professionnelle réunissant des centres européens d'évaluation en langues, EALTA vise à promouvoir la compréhension des principes théoriques qui sous-tendent l'élaboration de tests et l'évaluation en langue ainsi que l'amélioration et le partage des pratiques concernant les tests et l'évaluation en Europe. Ses lignes directrices pour une bonne pratique dans l'élaboration/utilisation des tests et l'évaluation en langues sont déclinées sous différents supports et traduites en 35 langues ».
https://www.ciep.fr/produits-documentaires/repertoire-liens/langues (consulté le 20/05/2020).

6
« Pour une bonne pratique dans l'élaboration/utilisation des tests et l'évaluation en langues »,https://www.ealta.eu.org/guidelines.htm (consulté le 20/05/2020).

7
« L'ILTA est un organisme international de chercheurs et de praticiens travaillant dans le champ de l'évaluation et des tests de langues. Ses activités comprennent les fonctions suivantes : stimuler le développement professionnel de ses membres par le biais d'ateliers et de conférences ; promouvoir la publication et la diffusion d'informations relatives aux tests de langues ; assurer la fonction de chef de file dans le domaine des tests de langues ; améliorer la connaissance du grand public et soutenir le métier de certificateur ; coopérer avec d'autres organismes œuvrant dans les domaines des tests de langues, de la linguistique appliquée et de l'évaluation. L'ILTA est à l'origine d'un colloque annuel de recherche dans le domaine de l'évaluation en langues, Language Testing Research Colloquium (LTRC) ».
https://www.ciep.fr/produits-documentaires/repertoire-liens/langues(consulté le 20/05/2020).
Lignes directrices d'ILTA, https://www.iltaonline.com/page/ILTAGuidelinesforPra (consulté le 20/05/2020).

8
« ALTE vise à promouvoir le multilinguisme à travers l'Europe, et au-delà. Engagée pour la qualité et l'équité, elle établit notamment des normes communes pour toutes les étapes de l'élaboration de tests de langues, anime des formations et organise des conférences ».
https://www.ciep.fr/produits-documentaires/repertoire-liens/langues (consulté le 20/05/2020).
Normes minimales : https://www.alte.org/Materials (consulté le 20/05/2020).

9
Entre autres : APLV (Association des professeurs des langues vivantes), SAES (Société des Anglicistes de l'Enseignement Supérieur), GERAS (Groupe d'étude de recherche en anglais de spécialité), APLIUT (Association des professeurs des langues des instituts universitaires de Technologie, AFEA (Association Françaises des Études Américaines), RANACLES (RANACLES (Rassemblement National des Centres de Langues de l'Enseignement Supérieur, membre de CERCLES – Confédération européenne des centres de langues de l'enseignement supérieur).

10
Copil CLES (Comité de pilotage du CLES, la Certification de compétences en langues de l'enseignement supérieur (convention tripartite MESRI, CPU et UGA).

11
Arrêté du 3 avril 2020 relatif à la certification en langue anglaise pour les candidats inscrits aux diplômes nationaux de licence, de licence professionnelle et au diplôme universitaire de technologie
https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000041782410&dateTexte=&categorieLien=id (consulté le 20/05/2020).

12
Le test Dialang est un test diagnostic en ligne décliné en 14 langues européennes différentes ; La certification CLES est proposée en 9 langues différentes, sur trois niveaux ; SELF, test de positionnement à visée formative est disponible en 6 langues différentes.

13
« EVAL-IC a élaboré ces descripteurs (déclinés en descripteurs fins) de même qu'un protocole et des outils d'évaluation et d'attestation des compétences en intercompréhension pour les langues romanes. Il a établi, également, à l'attention de la communauté de l'enseignement-apprentissage des langues, un état des lieux des travaux sur les compétences d'intercompréhension et leur évaluation ». https://evalic.eu/leprojet/ (consulté le 20/05/2020).

14
Les journées de formation en FLE organisées par l'Alliance Française avec le DORIF sur la 'didactique de l'erreur à l'oral et à l'écrit : quelles interventions, quelle évaluation ?' se sont déroulées à Matera en 2019 et à Venise en 2020.

15
« Erreur systématique. Un test ou un item peuvent être considérés comme biaisés, si un de leurs attributs se révèle non pertinent par rapport à ce qu'ils sont censés tester et qu'ils avantagent ou désavantagent une partie des candidats. Le biais est principalement lié au sexe, à l'âge, à la culture, etc. des candidats » (ALTE, 1998 : 212).

16
« Lena a 14 ans et est en 3ème à l'école allemande de Paris où elle apprend le français en LV2. Elle apprend le français depuis 4 ans (école allemande de Tokyo et de Paris). Elle est en France depuis 1an ½. Ses parents sont allemands et elle ne parle qu'allemand à la maison. Autres langues apprises : anglais et latin. Tobias a 14 ans et est en 3ème à l'école allemande de Paris où il apprend le français en LV3. Il apprend le français depuis 2 ans (école allemande de Rome et de Paris). Il est en France depuis 3 ans. Ses parents sont allemands et il ne parle qu'allemand à la maison. Autres langues apprises : anglais et italien ». Monologue Lena : *une chanteuse*. Monologue Tobias : *un livre*. Interaction : *organisation d'une fête*.
Lien :https://www.ciep.fr/ressources/ouvrages-cederoms-consacres-a-levaluation-certifications/dvd-productions-orales-illustrant-les-6-niveaux-cecrl (consulté le 20/05/2020).

17
Séminaire interlingue pour le calibrage de productions orales en allemand, anglais, espagnol, français et italien sur les 6 niveaux de l'échelle du Cadre européen commun de référence pour les langues, 2008 : 14.
Lien : https://www.ciep.fr/conferences-colloques/archives/seminaire-inter-langues.
L'objectif du séminaire était d'identifier des séquences présentant des productions de candidats dont les compétences illustrent les niveaux du CECR.