

Évaluer le français professionnel, une affaire de professionnels et une entreprise collective

Marc DEMEUSE, Université de Mons (Belgique)

1. Introduction

Avant de nous engager dans la problématique spécifique de l'évaluation du français professionnel ou, plus exactement, du niveau de compétence en français, dans un contexte professionnel, de candidats qui se soumettent à une épreuve, je voudrais faire un petit détour par ma discipline, la psychologie et, de manière plus précise, le domaine de la psychométrie dont les développements vont nourrir cette introduction.

1.1. De la mesure à l'évaluation

Comme le signale Laurencelle (1998 : 14), « la distinction entre évaluation et mesure prend généralement peu d'importance en sciences et dans le monde des laboratoires scientifiques. Les mesures y sont pour la plupart physiques, purement matérielles, et ne donnent en elles-mêmes que peu de champ à l'interprétation subjective ». Ce n'est naturellement pas le cas dans les domaines de la psychologie, de l'éducation et, de manière générale, des sciences humaines et sociales car le contexte de la mesure et le sens qui lui est donné sont prépondérants.

Dans la plupart des cas, ce n'est pas un objet visible ou matériel, aussi petit soit-il, qu'il s'agit de mesurer, comme dans le domaine des sciences de la vie et de la terre, mais un construit, par exemple, l'intelligence, pour les collègues psychologues, ou le niveau de compétence en français professionnel, pour les spécialistes de ce champ. En fait, c'est très souvent une série de productions qui sont prises en compte. À partir de ces constats, une évaluation est réalisée, c'est-à-dire l'estimation d'une valeur.

La valeur ainsi déterminée est attribuée à un trait qui est supposé relativement stable chez l'individu qui a fourni les productions prises en compte. En fait, ce ne sont pas ces productions, en elles-mêmes, qui intéressent l'évaluateur, mais ce qu'il est possible d'en

inférer à propos de ce trait, en d'autres mots, le niveau de maîtrise qui est atteint par l'individu.

Pour Laurencelle (1998 : 12), « en éducation et dans les autres sciences humaines, l'évaluation peut et doit compléter la mesure, puisque souvent la valeur potentielle de celle-ci dépend entièrement du contexte dans lequel on la fait intervenir ». Mais, il convient de bien distinguer le travail d'instrumentation et les méthodes de mesure de l'interprétation qui pourra être faite du mesurage, même si la construction de l'instrument et donc la nature des tâches, par exemple, est déjà un parti pris. Pour cet auteur canadien, « comme toute autre fonction d'information, la mesure qui exprime d'une façon ou d'une autre la « grandeur » d'un objet, prend toute sa signification dans un contexte ; elle est alors interprétée et exploitée dans un but, social, éducatif, économique, scientifique. C'est cette action globale qu'on désigne par « évaluation » (ibid : 12).

Si donc l'évaluation est bien plus que la mesure, c'est parce qu'elle s'y ajoute, qu'elle s'en nourrit. Une mauvaise mesure – nous verrons quelles sont les qualités d'une « bonne mesure » – ne peut conduire qu'à une mauvaise évaluation, une évaluation biaisée.

1.2. Evaluer des apprentissages, c'est prendre en compte le changement

Les choses se compliquent encore si on veut bien considérer qu'en éducation, ce qui est évalué aujourd'hui est sujet à modification, notamment à la suite d'un enseignement qui aura déclenché, c'est l'espoir de tout enseignant, un apprentissage et donc, une modification de l'état initial de l'apprenant, soit ce qui vient précisément d'être estimé.

L'évaluation du français professionnel, qui s'inscrit généralement dans une approche par compétences (Scallon, 2004, 2015), est ainsi rendue plus complexe puisqu'elle ne se limite pas à un ensemble de connaissances ou même de savoir-faire élémentaires. Il s'agit bel et bien d'évaluer quelque chose qui n'est pas directement observable, mais un potentiel dont la définition est parfois fort variable.

À côté des enseignements formels, les individus apprennent aussi de manière plus informelle, voire franchement informelle, et donc il est très probable, par exemple si je vis dans un milieu francophone, n'étant pas francophone moi-même, que j'apprendrai plus de choses que ce que mon professeur de français m'enseigne durant les heures de cours que je peux suivre chaque semaine. Il est donc très difficile d'imputer les apprentissages à un élément parfaitement circonscrit. Il est tout aussi difficile de mesurer deux fois la même

chose, tant la situation est changeante et donc, contrairement au physicien qui peut répliquer de nombreuses fois une mesure non invasive sur un objet... le spécialiste des apprentissages n'a pas cette chance !

L'utilisation d'un instrument de mesure, disons un test¹, par sa seule administration, altèrera la compétence du sujet : on parle alors d'effet de testing pour désigner l'impact que la passation d'un test peut avoir sur les performances ultérieures d'un sujet à la même épreuve ou à une épreuve proche. Si c'est plutôt heureux pour le sujet, puisque cela indique qu'il apprend aussi lors des évaluations, y compris à propos des stratégies à adopter pour répondre correctement aux questions qui lui sont proposées, cela complique singulièrement la vie de l'évaluateur.

1.3. Trop ou mal évaluer peut aussi nuire aux apprentissages

Poussé à l'extrême, cet effet de testing est bien connu sous le terme « bachotage » : le candidat, à force de préparation et de tests blancs, finit par devenir un spécialiste de l'épreuve, pas nécessairement de la matière ou de la compétence qui est supposée évaluée. C'est ce que recouvre l'expression en langue anglaise « teaching to the test » – enseigner à outrance pour préparer les évaluations certificatives – et l'effet collatéral est également bien connu sous le vocable de « backwash effect »² que l'on peut simplement traduire par « effet en retour » (Parker, 2012 ; Podromou, 1995). Il s'agit d'un phénomène complexe (Cheng, Watanabe et Curtis, 2004) qui peut, dans certains cas, se traduire par la réduction du curriculum à ce qui est évalué, souvent parce que c'est ce qui est le plus facilement évaluable, et pas nécessairement ce qu'il est souhaitable d'enseigner et d'apprendre. Ce risque est d'autant plus grand que l'objet à évaluer est complexe ou que les évaluations présentent de forts enjeux³ (Au, 2007).

¹ Le terme « test » auquel il est fait référence ici possède le sens précis que lui attribue la psychologie : il s'agit d'une « situation expérimentale standardisée servant de stimulus à un comportement. Ce comportement est évalué par une comparaison statistique avec celui d'autres individus placés dans la même situation, permettant ainsi de classer le sujet examiné soit quantitativement, soit typologiquement » (Pichot, 1954 : 5). De Landsheere (1979) précise qu'il « doit être standardisé, fidèle, valide et étalonné ».

² Parfois aussi appelé « washback effect ».

³ On parle de test à forts enjeux (high-stakes test) lorsque les résultats de ce test conduisent à prendre des décisions importantes à propos de ceux qui s'y sont soumis, comme dans le cas des épreuves en langue étrangère (poursuite d'études, certification, sélection en vue d'un emploi, admission sur le territoire dans le cas d'un processus d'immigration économique...).

Le « théorème du lampadaire », que Fitoussi (2013) développe dans le domaine de l'économie, fait référence à la plaisanterie que rappelait Coluche : un ivrogne cherche ses clés sous un lampadaire. Interrogé sur le lieu où il a perdu ses clés, il répond ne pas s'en souvenir. « Mais alors, pourquoi chercher ici », lui demande un passant charitable. Et l'ivrogne de répondre : « parce que c'est le seul endroit éclairé de la rue ». Ce « théorème » illustre parfaitement la situation d'une évaluation qui, à force de vouloir en augmenter la fiabilité, finit par perdre toute validité.

Il faut cependant rester prudent : certaines informations dont on ne comprend pas immédiatement le lien avec la compétence à évaluer peuvent pourtant avoir un fort pouvoir prédictif. Dans certains cas, la recherche éclaire ce lien, en montrant par exemple pourquoi la fluidité de la lecture est aussi un bon prédictif de la compréhension de textes, comme dans un autre registre, le nombre de livres à la maison, utilisé dans PISA.

Dans la suite de cet article, nous envisagerons successivement plusieurs problèmes auxquels les concepteurs de test sont habituellement confrontés.

2. Qu'entend-on par évaluation ?

2.1. Précisons la définition

L'une des définitions les plus connues est sans doute celle de Stufflebeam et ses collègues (1980 : 48). Pour eux, l'évaluation est le « processus par lequel on délimite, obtient et fournit des informations utiles permettant de juger des décisions possibles ». Cette définition, très lapidaire, est clairement orientée par l'usage d'une collecte d'informations utiles à une prise de décision.

Laurencelle (1998 : 13) précise le cheminement de la démarche évaluative :

« Pour prendre ce jugement [évaluatif], l'évaluateur doit

- a) obtenir l'information mesurée,
- b) apprécier cette information en considérant l'instrument de mesure, la caractéristique évaluée et la personne,
- c) utiliser une référence comparative.

Dans ces conditions seulement, l'évaluation (ou l'évaluation scientifique) pourra être faite avec compétence et justice ».

Tourmen (2014), dans un texte consacré aux contributions des sciences de l'éducation à la compréhension de la pratique évaluative, propose une intéressante synthèse de nombreux modèles de l'activité évaluative. Elle résume les quatre grands éléments sur lesquels ces modèles semblent s'accorder :

- « - L'évaluation est décrite comme une activité séquencée organisée en étapes identiques ;
- Évaluer est défini comme une activité de comparaison entre ce qui est appelé des « référés » (données sur l'objet évalué) et des « référents » (normes ou informations servant de standards de comparaison) qui permettent de leur donner un sens, une valeur, à savoir de porter un jugement;
- La comparaison impliquée dans le jugement final est précédée et rendue possible par 1) une étape préalable d'explicitation ou de construction des « référents » définis par des critères et rendus mesurables par des indicateurs (stabilisés par exemple dans un référentiel), puis par 2) une autre étape préalable de recueil des données sur l'objet évalué (ou « référés ») ;
- La démarche globale de production du jugement se base sur des actes explicites, montrés, maîtrisés, en opposition aux évaluations quotidiennes, spontanées et informelles, où ces actes restent implicites et soumis à de nombreux biais potentiels. »

2.2. Il ne faut pas avoir peur de la mesure... mais il faut s'en préoccuper

Même si elle a longtemps occupé, fort inutilement, la scène française, l'opposition entre mesure et évaluation n'a plus vraiment de sens (si elle en a jamais eu !). « Mesure » ne s'oppose en effet pas à « évaluation », à la manière de « précis » qui s'opposerait à « flou » ou même « contrôle » à « évaluation ». Cette dernière opposition (Nossent, 2010) me semble d'ailleurs bien mieux analysée à travers une approche basée sur les fonctions de l'évaluation (Allal, 2008 ; Mottier-Lopez, 2015).

2.3. A quoi sert l'évaluation ? Ne pas se tromper de fonction !

La littérature est très abondante à propos des fonctions de l'évaluation et les auteurs francophones s'ingénient parfois à multiplier les termes. De manière sans doute un peu schématique, on peut distinguer quatre grandes fonctions :

- la fonction diagnostique qui permet d'analyser une situation, d'identifier des besoins... Les tests de placement (en début de formation) en sont un exemple. Ils visent à identifier les individus qui possèdent un certain niveau de compétence, par exemple pour les regrouper dans une même classe ;
- la fonction pronostique qui permet de se projeter dans l'avenir et d'estimer par exemple les « chances » de réussite lors d'une formation ultérieure, éventuellement en autorisant ou non l'accès à cette formation ;

- la fonction formative ou formatrice (de régulation) qui permet d'accompagner et de réguler les actions des enseignants et des apprenants en cours d'apprentissage ;
- la fonction certificative (ou sommative) qui permet de faire un bilan à l'issue d'une formation, de prendre des décisions en matière d'octroi d'un diplôme.

La place manque ici pour développer chacune de ces fonctions, mais il importe d'en tenir compte lors du développement d'un outil d'évaluation et de sa mise en œuvre. Plusieurs de ces fonctions sont aussi parfois attribuées à un même dispositif d'évaluation... ce qui rend la tâche sinon impossible, au moins très difficile, comme par exemple dans le cas du récent test de français destiné aux futurs enseignants à l'entrée de leur formation initiale en Belgique francophone⁴.

2.4. Pas d'évaluation sans référentiel

La définition de Tourmen (2014) a déjà évoqué la notion de référentiel. Il revêt une grande importance puisque c'est lui qui permet la comparaison entre une production et les attendus. De manière classique, on distingue ici les évaluations qui sont dites critériées (ou critérielles), c'est-à-dire les évaluations qui ne se basent pas sur la comparaison d'un individu aux autres mais où on détermine, par référence à des critères externes si l'individu maîtrise de manière satisfaisante les attendus. L'évaluation est dite normative lorsqu'au contraire la performance d'un individu est comparée à celle d'autres personnes d'un groupe de référence, d'après un même instrument (Legendre, 1993). L'approche normative, puisqu'elle repose sur des comparaisons entre individus, conduit nécessairement à des classements comprenant des « gagnants » et des « perdants » et à des effets bien documentés par la docimologie critique (effet « Posthumus », « règle des trois tiers », « courbe de Gauss », « constante macabre » d'Antibi [2003]) (Leclercq, Nicaise et Demeuse, 2004).

Par ailleurs, reconnaître que l'évaluation se nourrit de la mesure conduit à une nécessaire modestie : mesurer... c'est aussi accepter de reconnaître qu'on commet des erreurs... Cette modestie et son corolaire, qui consiste à tenter d'en estimer l'ampleur, sont indispensables dès lorsqu'il s'agit de mettre en œuvre des tests à forts enjeux, comme ceux qui permettent

⁴ Cette réforme, qui fait l'objet du Décret du 6 février 2019 définissant la formation initiale des enseignants, aurait dû entrer en vigueur dès la rentrée 2020-2021. Elle est actuellement postposée à la rentrée scolaire suivante à la suite des élections de mai 2019 et du changement de majorité parlementaire.

d'acquérir des points dans le système canadien d'accès au territoire ou en France dans le cadre de l'acquisition de la nationalité. Avec Laurencelle (1998), nous pouvons considérer que

« c'est donc d'abord un jugement, non pas seulement le résultat d'un test ou la donnée sortant d'une machine ; c'est un acte responsable d'un évaluateur (par exemple, l'enseignant), qui a des conséquences sur la vie d'une autre personne [...]. Pour faire ce jugement, il faut interpréter la mesure, le résultat, dans les contextes global et particulier dans lesquels se trouve la personne évaluée. L'évaluation porte aussi sur un objet d'évaluation plutôt que sur la personne elle-même, et le jugement doit se borner à l'aspect évalué plutôt que d'être interprété comme s'il caractérisait globalement la personne. La personne évaluée devrait avoir une connaissance claire de ce qui est évalué chez elle [...]. Le jugement évaluatif doit être basé sur quelque chose, tels un test, un appareil, un procédé défini d'avance, mais en tout cas ne pas être improvisé. Cela n'exclut pas certaines évaluations utilisant la « méthode des juges », pourvu que les aspects évalués et les critères de pondération soient préalablement et clairement définis. L'évaluation, on l'a dit, n'existe pas dans l'absolu ; c'est un jugement, ce qui implique une référence, parfois une comparaison. On jugera selon le contenu de la performance, la valeur propre du résultat (par critère intrinsèque) [évaluation critériée], ou selon l'aspect ordinaire ou exceptionnel du résultat (par critère extrinsèque) [évaluation normative] ».

3. Quelles normes et quels repères déontologiques et d'assurance qualité pour les tests ?

Il faut encore, même si ce sujet n'est pas spécifiquement abordé ici, attirer l'attention du lecteur, sur les aspects déontologiques inhérents au travail entrepris en relation avec des êtres vivants. La mesure et l'évaluation doivent être interrogées en fonction des finalités qui leur sont assignées. À ce niveau, le sujet/objet de la mesure n'est pas anodin.

Fort heureusement, il existe, dans le domaine de l'évaluation, des standards, par exemple ceux édités, en 2002, par des associations professionnelles comme *l'American Educational Research Association*, *l'American Psychological Association* ou *le National Council on Measurement in Education*.

En français, on peut, notamment, trouver un texte canadien sur le sujet intitulé « Principes d'équité relatifs aux pratiques d'évaluation des apprentissages scolaires au Canada ». Ce texte est l'œuvre d'un comité consultatif mixte, composé de représentants de *l'Association canadienne d'éducation*, de *l'Association canadienne des commissions/conseils scolaires*, de *l'Association canadienne des administrateurs scolaires*, de *la Fédération canadienne des enseignantes et enseignants*, de *l'Association canadienne d'orientation et de consultation*, de *la Canadian Association of School Psychologists*, du *Conseil canadien pour les enfants*

exceptionnels, de l'Association canadienne de psychologie et de la Société canadienne pour l'étude de l'éducation. Ce texte a été publié par la revue *Mesure et évaluation en éducation* de l'Association pour le développement de la mesure et de l'évaluation en éducation⁵.

L'International Test Commission / Commission internationale des Tests (ITC) a également promulgué un texte intitulé, dans sa version française : « Recommandations internationales sur l'utilisation des tests ». Ce texte a été publié par la Société Française de Psychologie dans un numéro spécial « hors-série » de la revue *Pratiques psychologiques* (2003)⁶. Ce texte est complété par un document relatif aux « Recommandations internationales sur les tests informatisés ou les tests distribués par Internet » (2007).

Dans le domaine des tests de langue, l'Association des organismes certificateurs en Europe (*Association of Language Testers in Europe*, ALTE), fondée en 1989, compte 34 membres⁷ et 55 affiliés institutionnels (soit 25 langues européennes) et « a établi un ensemble de critères de qualité communs qui couvrent toutes les étapes du processus d'examen linguistique » (ALTE, 2018). Son objectif : proposer des « critères [qui] peuvent être appliqués aux examens de n'importe quelle langue et [qui] garantissent la qualité des évaluations assurées » par ses membres⁸. Le label qualité d'ALTE, baptisé « Q Mark », est attribué à l'issue d'un audit basé sur les 17 standards de qualité minimale de l'association⁹.

⁵ On peut en trouver une copie en ligne : <https://docplayer.fr/55719727-Principes-d-equite-relatifs-aux-pratiques-d-evaluation-des-apprentissages-scolaires-au-canada.html> (consulté le 01/05/20).

⁶ Il est aussi disponible sur le site de la Société Française de Psychologie : <http://www.sfpsy.org/spip.php?article7> (consulté le 01/05/20).

⁷ Le Centre international d'études pédagogiques (CIEP), devenu depuis peu France Éducation International, et Le français des affaires de la Chambre de commerce et d'industrie de Région Paris Ile-de-France (CCIdF) sont les deux membres français d'ALTE.

⁸ <https://fr.alte.org/> (consulté le 01/05/20).

⁹ <https://fr.alte.org/Setting-Standards> (consulté le 01/05/20).

4. Peut-on tout évaluer ?

Si certains enseignants pensent que tout n'est pas évaluable, Decroly et son collègue Buyse, à l'origine d'une approche pédagogique s'inscrivant dans l'éducation nouvelle, sont déjà très clairs à ce sujet : « Nous pensons pour notre part que tout ce qui doit être enseigné par le maître, donc appris par l'élève, c'est-à-dire, tout ce qui est matière d'instruction, est susceptible de mesure » (Decroly et Buyse, 1929 : 11).

4.1. Si cela peut s'enseigner, cela doit aussi pouvoir s'évaluer

La question aujourd'hui n'est donc pas tant de savoir si tout s'évalue que comment évaluer. Cette question s'inscrit logiquement à la suite d'une autre : si un enseignant prétend développer certains apprentissages complexes, comment enseigner et permettre leur développement ?

L'enseignant qui pense qu'il peut y parvenir est sans doute capable d'évaluer les acquisitions de ses élèves, sans quoi, comment pourrait-il se convaincre qu'il propose des actions pédagogiques adéquates... s'il ne pense pas pouvoir contribuer à ces apprentissages, la question de l'évaluation ne devrait pas réellement se poser : pourquoi évaluerait-il ce qu'il n'est pas capable d'enseigner ? Pour le dire autrement, s'il est impossible de formaliser un apprentissage en attendus, puis en moyens pour les atteindre, alors il est logique que ni un apprentissage formel, ni une évaluation ne soient possibles.

Les compétences complexes et transversales, hors sol en quelque sorte, notamment celles qui se cachent derrière les Compétences du 21^e siècle (par exemple, Bellanca et Brandt, 2010) obscurcissent à la fois le champ de l'enseignement et celui de l'évaluation en laissant croire à l'existence de compétences génériques, comme l'esprit critique ou la créativité, qui ne seraient pas enracinées dans des contenus spécifiques et qui pourraient se développer sans savoir-faire contextualisés (Bissonnette et Boyer, 2018). Elles font ainsi courir le risque de considérer que, bien qu'elles s'évaluent le plus souvent dans des contextes très abstraits, favorables aux plus éduqués à travers un cursus long et formel, elles ne s'enseignent pas et ne sont donc pas acquérables sur une base formelle. Il faut donc, pour les maîtriser, respirer l'air du temps... en fréquentant longtemps les couloirs des universités ou en bénéficiant d'un environnement familial favorable, ce qui risque bien de priver les moins favorisés de toute valorisation sociale et de confirmer la hiérarchie sociale établie. C'est un peu le retour à la

culture générale, celle qui appartient à un groupe particulier qui a su valoriser son propre mode de vie. Cette réflexion ne condamne naturellement pas à un enseignement qui se résume à un gavage sur la base de contenus réduits et sans usage. Elle plaide pour une explicitation de ce qui est enseigné et donc de ce qui est évalué, à travers un référentiel bien pensé (Figari, 1996 ; Figari et Tourmen, 2006 ; Figari et Remaud, 2014).

4.2. Encore faut-il pouvoir préciser ce qu'est le français professionnel

Le rapport avec le français professionnel me semble évident et le questionnement complètement analogue. Le référentiel qui sera utilisé et/ou construit pour mener à bien à la fois l'évaluation et la formation est une construction sociale, nourrie par une connaissance de la langue et de son utilisation en contexte, et elle doit être interrogée. Il s'agit de la première étape de toute démarche d'évaluation. C'est notamment tout l'intérêt du *Cadre européen commun de référence* du Conseil de l'Europe (2001) et de son actualisation en 2018, même s'il doit être complété par une réflexion spécifique au domaine professionnel, comme le montrent les spécialistes du français sur objectifs spécifiques. Prendre en compte ces objectifs spécifiques interroge naturellement aussi sur la possibilité d'évaluation en français général et sur l'existence d'un français standard qui s'imposerait (Charmeux, 1998). Le document du Conseil de l'Europe (2001 :4-5) précise utilement que ce dernier

« encourage toutes les personnes concernées par l'organisation de l'apprentissage des langues à fonder leur action sur les besoins, les motivations, les caractéristiques et les ressources de l'apprenant. [...] À partir de cette analyse de la situation d'enseignement/apprentissage, il est absolument essentiel de définir avec un maximum de précision des objectifs immédiatement valables au regard des besoins des apprenants et réalistes du point de vue de leurs caractéristiques et des moyens disponibles ».

Disposer d'un référentiel permet de définir ce qui sera enseigné et comment il le sera, mais cela permet aussi de préciser ce qui doit être évalué et de quelle manière. Ce référentiel aide ainsi à clarifier le contrat didactique entre l'enseignant et les étudiants et d'assurer la cohérence et la progression des enseignements/apprentissages, notamment lorsque l'enseignement se déroule pendant une période longue, durant laquelle plusieurs enseignants doivent se relayer. D'une manière générale donc, le référentiel permet d'objectiver les différents éléments du curriculum (Demeuse et Strauven, 2013).

En fonction du contexte, les évaluateurs disposent de référentiels légaux, comme les socles de compétences ou les compétences terminales, pour l'enseignement scolaire, les profils de

formation et de qualification dans l'enseignement technique et professionnel belge francophone, des programmes, des manuels scolaires, d'autres références légales, mais ils sont aussi parfois amenés à développer une démarche spécifique de construction, comme dans le domaine du français professionnel ou du français sur objectifs spécifiques, comme le montre bien, par exemple, la thèse de Beillet (2016) dans le domaine du français académique. La démarche générale implique alors une analyse des besoins, la conception d'un référentiel de tâches professionnelles, l'identification d'un référentiel langagier et, enfin, la constitution d'un référentiel d'évaluation permettant le développement d'une table de spécifications qui va permettre le développement du dispositif d'évaluation.

5. Comment assurer la qualité de tests dans le domaine du français professionnel ?

J'accompagne depuis 2000 les travaux des équipes du *français des affaires* de la Chambre de commerce et d'industrie de région Paris Ile-de-France. C'est dans le cadre d'une équipe pluridisciplinaire que les tests et les diplômes professionnels y sont développés. Nos premiers travaux en commun ont porté sur le *Test d'évaluation de français* (TEF), puis se sont progressivement intéressés aux diplômes professionnels. Dans ce cadre, il a été nécessaire de développer un processus d'assurance qualité qui devra s'intégrer aux exigences de plus en plus sévères des prescripteurs (Ministère français de l'Éducation nationale, Ministère français de l'Intérieur, le Gouvernement fédéral du Canada et le Ministère de l'Immigration, de la Francisation et de l'Intégration du Québec) et d'associations professionnelles comme ALTE, que nous avons déjà évoquée. Cela a aussi impliqué, pour le TEF par exemple, de l'aligner sur le *Cadre européen commun de référence pour les langues* (CECR) du Conseil de l'Europe, mais aussi sur les douze *Niveaux de compétence linguistique canadiens* (NCLC) et ceux de l'*Échelle québécoise des niveaux de compétence en français des personnes immigrantes adultes* (EQNCF).

Il serait ici impossible de présenter l'ensemble des travaux qui ont été menés dans le cadre de la collaboration entre l'Université de Mons et l'équipe parisienne du *français des affaires*. Je me limiterai donc à certains documents publiés de manière à illustrer la variété des travaux réalisés. Le premier, qui a déjà été mentionné, est la thèse de Beillet (2016). Cette recherche résulte d'une co-tutelle entre l'Université de Mons et l'Université d'Artois (Prof. J.-M. Mangiante). Il s'agissait de proposer un dispositif permettant d'évaluer le français académique d'étudiants non francophones à l'entrée des études universitaires en France. Ce

travail allie une réflexion théorique à une série d'expérimentations sur le terrain, avec des groupes diversifiés d'étudiants. Il a permis de proposer le prototype d'une épreuve de français sur objectifs universitaires.

Concernant le français professionnel et le développement des dispositifs d'évaluation, on peut aussi renvoyer à deux articles qui figurent dans le présent numéro de *Points communs* : le premier a été préparé par D. Casanova et F. Renaud et porte sur la conception et la validation des diplômes de français professionnel. Le second, signé par A. Mommée et A. Diogo, traite de la sélection, de la formation et du suivi des examinateurs de ces mêmes diplômes. Ce dernier texte s'inscrit dans la perspective qui avait été développée dans plusieurs contributions de notre équipe à propos de la qualité des évaluateurs, leur sélection ou leur formation à l'évaluation de productions orales ou écrites (Artus et Demeuse, 2008 ; Casanova et Demeuse, 2011, 2016) ou les outils d'évaluation qui leur sont offerts (Aw, Casanova et Demeuse, 2020).

D'autres travaux ont porté sur le passage de l'évaluation papier-crayon à une forme numérique (Desroches, Casanova, Crendal, Renaud, Mahy et Demeuse, 2005 ; Casanova, Aw et Demeuse, 2018), l'alignement sur les référentiels (Casanova, Crendal, Demeuse, Desroches et Holle, 2010 ; Demeuse, Desroches, Crendal, Renaud, Oster et Leroux, 2004 ; Demeuse, Artus, Casanova, Crendal, Desroches, et Renaud 2007) ou interrogent d'autres aspects sans doute beaucoup plus techniques, comme l'unidimensionnalité des tests (Casanova, Kaddachi, Aw et Demeuse, 2019).

Le travail se poursuit également à travers plusieurs pistes permettant d'envisager les apports de la technologie à l'évaluation de productions complexes (Casanova, Aw et Demeuse, 2020), le contrôle des réponses suspectes (Aw, Zhang, Demeuse et Casanova, 2019) ou l'influence des accents francophones sur les performances en compréhension orale (Casanova, Aw, Bourras et Demeuse 2018).

Conclusions

Le travail au sein de l'équipe du *français des affaires* de la Chambre de commerce et d'industrie de région Paris Ile-de-France présente de multiples facettes. Il répond aux standards de qualité qui s'imposent à présent en matière de tests à forts enjeux. Cela nécessite la mobilisation d'une équipe pluridisciplinaire, au-delà de la production de nouvelles versions d'un même instrument. La palette de l'offre, notamment dans le domaine

du français des affaires, continue à se développer, abordant de nouveaux domaines, mais celle-ci doit aussi faire face à de nombreux défis : généralisation de l'utilisation du numérique pour la gestion des épreuves, modification du format des interfaces destinées aux candidats, adaptation à l'évolution des exigences des prescripteurs et des référentiels, sécurisation accrue et recours à la technologie pour mieux assurer la qualité des corrections des productions complexes, tant orales qu'écrites...

Tous ces développements ont pu se réaliser avec une équipe à la taille relativement réduite, si on la compare aux équipes qui assurent par exemple le développement des tests en langue anglaise. Par ailleurs, le monde francophone de l'évaluation a longtemps été réfractaire aux développements liés à la mesure et à l'évaluation, en particulier dans le domaine de l'évaluation linguistique. Il a donc fallu aussi développer en interne des compétences spécifiques qu'il était très difficile de trouver sur le marché francophone.

Cet article a tenté d'offrir une réflexion nourrie par une vingtaine d'années de collaboration. Ce regard dans le rétroviseur, permet de mesurer le chemin accompli tout en entrevoyant les développements à venir. Ceux-ci impliqueront de plus en plus le traitement automatisé et l'intelligence artificielle. C'est assurément le tournant à ne pas manquer pour maintenir les produits de notre équipe au niveau d'exigence qu'ils ont pu atteindre jusqu'ici.

Pour citer l'article :

DEMEUSE, M., 2020, « Évaluer le français professionnel, une affaire de professionnels et une entreprise collective », *Points Communs*, n°48, p. 9-25.

Bibliographie

ALLAL, L., 2008, Évaluation des apprentissages, in A. VAN ZANTEN (éd.), *Dictionnaire de l'éducation*. Paris : Presses Universitaires de France, p. 311-314.

ALTE, 2018, *Guidelines for the Development of Language for Specific Purposes Tests. A supplement to the Manual for Language Test Development and Examining*. Cambridge : ALTE.

ANTIBI, A., 2003, *La constante macabre ou comment a-t-on découragé des générations d'élèves*. Toulouse : Antibi éditeurs, collection « Math'Adore ».

ARTUS, F. & DEMEUSE, M., 2008, Évaluer les productions orales en français langue étrangère (FLE) en situation de test : étude de la fidélité inter-juges de l'épreuve

d'expression orale du Test d'évaluation de français (TEF) de la Chambre de commerce et d'industrie de Paris. *Les cahiers des sciences de l'éducation*, 25-26, p.131-151.

AU, W., 2007, High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher*, vol. 36, n°5, p. 258-267.

AW, A., CASANOVA, D., & DEMEUSE, M., 2020, L'accord inter-évaluateurs dépend-il de la grille d'évaluation utilisée ? Actes du 32e colloque international de l'Association pour le Développement des Méthodologies d'Évaluation en Éducation (ADMEE-Europe). Casablanca, Maroc.

AW, A., ZHANG, A., DEMEUSE, M., & CASANOVA, D., 2019, Mise à l'épreuve de méthodes de détection de patrons de réponses suspects. Actes du 31e colloque international de l'Association pour le Développement des Méthodologies d'Évaluation en Éducation (ADMEE-Europe). Lausanne, Suisse.

BEILLET, M., 2016, *Conception d'un test d'évaluation des compétences en langue des étudiants allophones à l'entrée dans l'enseignement supérieur en France*. Thèse de doctorat non publiée, Université de Mons et Université d'Artois.

BELLANCA, J.A, & BRANDT, R. (éds.), 2010, *21st Century Skills : Rethinking How Students Learn*. Bloomington: Solution Tree.

BISSONNETTE, S., & BOYER, C., 2018, Les organismes scolaires ne devraient pas répondre aux douces sirènes des compétences du 21e siècle. *Formation et profession*, vol. 26, n° 3, p. 131-133.

CASANOVA, D., & DEMEUSE, M., 2011, Analyse des différentes facettes influant sur la fidélité d'une l'épreuve d'expression écrite d'un test standardisé de français langue étrangère. *Mesure et Évaluation en Éducation*, vol. 34, n°1, p. 25-53.

CASANOVA, D., & DEMEUSE, M., 2016, Évaluateurs évalués : évaluation diagnostique des compétences en évaluation des correcteurs d'une épreuve d'expression écrite à forts enjeux. *Mesure et évaluation en éducation*, vol. 39, n°3, p. 59-94

CASANOVA, D., AW, A. & DEMEUSE, M., 2018, Quand le numérique défie la mesure. Comment veiller la qualité de certifications en langue professionnelle au format numérique ? Actes du 30e colloque international de l'Association pour le Développement des Méthodologies d'Évaluation en Éducation (ADMEE-Europe). Belval, Luxembourg.

CASANOVA, D., AW, A., & DEMEUSE, M., 2020, Monitorer la qualité des évaluations humaines au moyen d'un correcteur automatique. Actes du 32e colloque international de l'Association pour le Développement des Méthodologies d'Évaluation en Éducation (ADMEE-Europe). Casablanca, Maroc.

CASANOVA, D., AW, A., BOURRAS, M., & DEMEUSE, M., 2018, Nos items à l'accent québécois diffèrent-ils ? Actes du 30e colloque international de l'Association pour le Développement des Méthodologies d'Évaluation en Éducation (ADMEE-Europe). Belval, Luxembourg, p. 278-289.

CASANOVA, D., CRENDAL, A., DEMEUSE, M., DESROCHES, F., & HOLLE, A. (2010). Validation empirique d'un test de français langue étrangère en regard du Cadre européen commun de référence pour les langues. Actes du 22e colloque international de l'Association pour le développement des méthodologies d'évaluation en éducation (ADMEE-Europe), Braga, Portugal.

CASANOVA, D., KADDACHI, L., AW, A. & DEMEUSE, M., 2019, Comment vérifier la dimensionnalité des outils d'évaluation. Actes du 31e colloque international de l'Association pour le Développement des Méthodologies d'Évaluation en Éducation (ADMEE-Europe). Lausanne, Suisse, p. 288-293.

CHARMEUX, E., 1998, *Le bon français... et les autres. Normes et variations du français aujourd'hui*. Paris : Milan éditions.

CHENG, L., WATANABE, Y., & CURTIS, A. (éds.), 2004, *Washback in Language Testing. Research Contexts and Methods*. Mahwah (NJ): Lawrence Erlbaum Associates.

CONSEIL DE L'EUROPE, 2001, *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Paris : Editions Didier.

DECROLY, O., & BUYSE, R., 1929, *Introduction à la pédagogie quantitative. Éléments de statistique appliqués aux problèmes pédagogiques*. Bruxelles : Maurice Lamertin éditeur.

DE LANDSHEERE, G., 1979, *Dictionnaire de la recherche en Éducation*. Paris : Presses universitaires de France.

DEMEUSE, M., & STRAUVEN, C., 2013, *Développer un curriculum d'enseignement ou de formation. Des options politiques au pilotage*. Bruxelles : DeBoeck.

DEMEUSE, M., ARTUS, F., CASANOVA, D., CRENDAL, A., DESROCHES, F., & RENAUD, F., 2007, Evaluation et autoévaluation en français langue étrangère : approche empirique du Cadre européen commun de référence pour les langues (CECR) du Conseil de l'Europe à travers le TEF et l'échelle d'autoévaluation DIALANG. *Journal of Educational Measurement and Applied Cognitive Sciences*, vol. 1, n°1.

DEMEUSE, M., DESROCHES, F., CRENDAL, A., RENAUD, F., OSTER, P., & LEROUX, X., 2004, L'évaluation des compétences linguistiques des adultes en français langue étrangère dans une perspective de multiréférentialisation. L'exemple du Test d'Évaluation de Français (TEF) de la Chambre de Commerce et d'Industrie de Paris (CCIP).

Actes du 17e colloque international de l'Association pour le développement des méthodologies d'évaluation en éducation (ADMEE-Europe). Lisbonne, Portugal.

DESROCHES, F., CASANOVA, D., CRENDAL, A., RENAUD, F., MAHY, C., DEMEUSE, M., 2005, Mise en équivalence des versions papier-crayon et électronique d'un test de français langue étrangère et mise en perspective des résultats par une procédure d'autoévaluation, Actes du 18e colloque international de l'Association pour le Développement des Méthodologies d'Evaluation en Education (ADMEE-Europe), Reims, 24-26 octobre 2005.

FIGARI, G., 1994, *Évaluer, quel référentiel ?* Bruxelles : De Boeck – Wesmael.

FIGARI, G., & REMAUX, D., 2014, *Méthodologie d'évaluation en éducation et formation, ou l'enquête évaluative*. Bruxelles, Belgique : De Boeck.

FIGARI, G., & TOURMEN, C., 2006, La référentialisation : une façon de modéliser l'évaluation de programme, entre théorie et pratique. Vers une comparaison des approches en France et au Québec. *Mesure et évaluation en éducation*, vol. 29, n°3, p. 5-25.

FITOUSSI, J.P., 2013, *Le théorème du lampadaire*. Paris : Les liens qui libèrent.

LECLERCQ, D., NICAISE, J., & DEMEUSE, M., 2004, Docimologie critique : des difficultés de noter des copies et d'attribuer des notes aux élèves. In M. DEMEUSE (éd), *Introduction aux théories et aux méthodes de la mesure en sciences psychologiques et en sciences de l'éducation*, Liège : Les Editions de l'Université de Liège, p. 273-292.

LAURENCELLE, L., 1998, *Théorie et techniques de la mesure instrumentale*. Sainte-Foy : Presses de l'Université du Québec.

MOTTIER-LOPEZ, L., 2015, Evaluations formative et certificative des apprentissages. Enjeux pour l'enseignement. Bruxelles : De Boeck, Collection « Le point sur... Pédagogie ».

NOSSANT, J.P., 2010, Evaluation ou contrôle : repères pour l'éducation permanente. *Institut d'histoire ouvrière, économique et sociale (IHOES), Analyse* n°63.

PAKER, T., (2013), The backwash effect of the test items in the achievement exams in preparatory classes. *Procedia – Social and Behavioral Sciences*, 70, p. 1463-1471.

PICHOT, P., 1954, *Les tests mentaux*. Paris : Presses universitaires de France, « Que sais-je ? ».

PODROMOU, L., 1995, The backwash effect: from testing to teaching. *ELT Journal*, vol. 49, Issue 1, p. 13-25.

SCALLON, G., 2004, *L'évaluation des apprentissages dans une approche par compétences*. Bruxelles : De Boeck.

SCALLON, G., 2015, *Des savoirs aux compétences. Exploration en évaluation des apprentissages*. Bruxelles : De Boeck.

STUFFELBEAM, D., et al., 1980, *L'évaluation en éducation et la prise de décision*. Ottawa: NHP.

TOURMEN, C., 2014, Contributions des sciences de l'éducation à la compréhension de la pratique évaluative, *Politiques et management public*, vol. 31, n°1, p. 69-85.