Mesure et évaluation en éducation

Le jugement des examinateurs dans le cadre de l'épreuve d'expression orale du Test d'évaluation de français (TEF)

Emine Ince

Volume 44, numéro 3, 2021

L'évaluation des compétences langagières : enjeux et perspectives

Réception : 02 décembre 2021
Version finale : 22 mai 2022
Acceptation : 23 mai 2022

URI : https://id.erudit.org/iderudit/1093067ar
DOI : https://doi.org/10.7202/1093067ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Ince, E. (2021). Le jugement des examinateurs dans le cadre de l'épreuve d'expression orale du Test d'évaluation de français (TEF). *Mesure et évaluation en éducation*, *44*(3), 87–111. https://doi.org/10.7202/1093067ar

Résumé de l'article

L'épreuve d'expression orale du TEF s'effectue au moyen d'une entrevue entre un candidat et un examinateur-animateur. Or, le comportement de ce dernier peut représenter une menace à la fidélité du test. Il a été démontré que malgré les nombreuses mesures prises afin de minimiser les variabilités dans l'évaluation, des divergences sur plusieurs aspects pouvaient être présentes chez les examinateurs. Cette étude vise à observer si l'on trouve ces divergences chez les examinateurs du TEF. Ainsi, 10 participants ont pris part à la recherche et la technique de la pensée à voix haute a été utilisée. Les résultats révèlent que des divergences existent. Les examinateurs peuvent accorder une même note pour une même performance alors que leurs interprétations peuvent différer, et inversement. Certains peuvent être influencés de façon positive en raison de leur familiarité avec l'accent des candidats. D'autres peuvent faire des inférences non pertinentes pour attribuer des significations aux difficultés rencontrées par les candidats. Enfin, l'attitude de l'animateur lors de la conversation avec le candidat peut être perçue différemment et avoir une conséquence négative sur la note.

Le jugement des examinateurs dans le cadre de l'épreuve d'expression orale du Test d'évaluation de français (TEF)

Emine Ince
Université de Montréal

Mots-clés : tests de langue seconde, expression orale, examinateurs, divergences dans l'évaluation

L'épreuve d'expression orale du TEF s'effectue au moyen d'une entrevue entre un candidat et un examinateur-animateur. Or, le comportement de ce dernier peut représenter une menace à la fidélité du test. Il a été démontré que malgré les nombreuses mesures prises afin de minimiser les variabilités dans l'évaluation, des divergences sur plusieurs aspects pouvaient être présentes chez les examinateurs. Cette étude vise à observer si l'on trouve ces divergences chez les examinateurs du TEF. Ainsi, 10 participants ont pris part à la recherche et la technique de la pensée à voix haute a été utilisée. Les résultats révèlent que des divergences existent. Les examinateurs peuvent accorder une même note pour une même performance alors que leurs interprétations peuvent différer, et inversement. Certains peuvent être influencés de façon positive en raison de leur familiarité avec l'accent des candidats. D'autres peuvent faire des inférences non pertinentes pour attribuer des significations aux difficultés rencontrées par les candidats. Enfin, l'attitude de l'animateur lors de la conversation avec le candidat peut être perçue différemment et avoir une conséquence négative sur la note.

Key words: second language tests, speaking test, examiners, discrepancies in assessment

The TEF speaking test is conducted through an interview between a candidate and an interviewer-examiner, but the conduct of the latter may represent a possible threat to the reliability of the test. It has been shown that despite many measures taken to minimize variability in the assessment, discrepancies in several aspects may be present among examiners. This study aims to observe whether these discrepancies are found among TEF examiners. A total of 10 participants took part in the research and the thinking aloud technique was used. The results show that discrepancies are present. Examiners may award the same score for the same performance but with varying interpretations, and vice versa. Some may be positively influenced by their familiarity with the candidate's accent. Others may make irrelevant inferences to make sense of the difficulties faced by candidates. Finally, the interviewer's attitude when talking to the candidate may be perceived differently and have a negative impact on the score.

PALAVRAS-CHAVE: testes de segunda língua, expressão oral, examinadores, discrepâncias na avaliação

A prova de expressão oral TEF é realizada por meio de uma entrevista entre um candidato e um examinador-facilitador, mas o comportamento deste último pode representar uma possível ameaça à fidelidade do teste. Demonstrou-se que, apesar das muitas medidas tomadas para minimizar a variabilidade na avaliação, pode haver discrepâncias em vários aspetos entre os examinadores. Este estudo tem como objetivo observar se essas discrepâncias são encontradas entre os examinadores do TEF. A investigação contou com 10 participantes e foi utilizada a técnica de pensar em voz alta. Os resultados revelam que as discrepâncias estão presentes. Os examinadores podem dar a mesma nota para o mesmo desempenho, enquanto as suas interpretações podem diferir e vice-versa. Alguns podem ser influenciados positivamente devido à sua familiaridade com o sotaque do candidato. Outros podem fazer inferências irrelevantes para produzir significados para as dificuldades encontradas pelos candidatos. Por fim, a atitude do facilitador durante a conversa com o candidato pode ser percebida de forma diferente e impactar negativamente na pontuação.

Introduction

Au Canada, pour pouvoir présenter une demande de résidence permanente ou une demande de citoyenneté, les candidats doivent démontrer un niveau de compétence linguistique acceptable en anglais et/ou en français. Au Québec, le français constitue le principal facteur d'intégration socioculturelle. La sélection des candidats à l'immigration de la catégorie économique, qui est la plus importante parmi les différentes catégories d'immigrants, s'effectue en suivant une grille qui attribue des points en fonction des caractéristiques recherchées. Sur les 120 points de la grille, les candidats peuvent obtenir un maximum de 16 points pour leurs connaissances en français : sept points pour l'expression orale, sept points pour la compréhension orale, un point pour la compréhension écrite et un point pour l'expression écrite. Les tests en français sont conçus et harmonisés avec des cadres linguistiques tels que les Niveaux de compétence linguistique canadiens (NCLC) et le Cadre européen commun de référence pour les langues (CECRL) afin d'améliorer l'interprétation des scores par les différents usagers de ces scores. Le niveau B2 (utilisateur indépendant) du CECRL est le seuil minimal à partir duquel des points sont attribués.

Parmi les tests et les diplômes de compétences linguistiques proposés par le ministère de l'Immigration, de la Francisation et de l'Intégration se trouve le Test d'évaluation de français (TEF) dont l'épreuve d'expression orale est basée sur une entrevue entre un candidat et un examinateur-animateur. Par souci de validité, l'évaluation de l'expression orale est faite lors d'une performance, généralement dans le cadre d'une entrevue, car cela constitue un meilleur indicateur de la compétence de la personne évaluée dans un contexte authentique (Bachman & Palmer, 2010). Cependant, la structure de l'entrevue peut compromettre la fidélité d'un test en raison de sa nature imprévisible et créative (Bachman, 1990 ; McNamara, 1996). Dans la mesure où l'entrevue orale implique un examinateur humain et que les résultats ne sont pas objectivement vérifiables, le comportement de celui-ci peut représenter une menace possible à la fidélité d'un test (Bachman et al., 1995 ; Douglas, 1994 ; Eckes, 2011 ; Lumley & McNamara, 1997 ; Upshur & Turner, 1999).

Malgré les nombreuses mesures prises afin de minimiser les variabilités de l'évaluation comme l'utilisation de grilles d'évaluation, les formations et les évaluations multiples, il a été démontré que les notes des tests oraux de langue pouvaient présenter des écarts systématiques et que des difficultés à arriver à un consensus entre les examinateurs pouvaient demeurer (Bachman et al., 1995 ; Douglas, 1994 ; Lumley & McNamara, 1997 ; Upshur & Turner, 1999). Les caractéristiques des examinateurs représentent une source importante de biais et, comme ces biais sont souvent involontaires, certains chercheurs ont choisi d'utiliser une appellation plutôt neutre, soit l'effet de l'examinateur (*rater effect*) (Myford & Wolfe, 2003 ; Norman & Goldberg, 1966 ; Wesolowski, 2016).

Dans le domaine de l'évaluation de l'expression orale en langue seconde, les études empiriques sur les effets des examinateurs portent principalement sur le fait qu'ils soient locuteurs natifs ou non natifs de la langue évaluée, sur leur familiarité avec les accents des candidats, sur leur manière d'interagir avec les candidats lors de la conversation, sur les nombreuses inférences faites pour expliquer les raisons des difficultés des candidats et sur la différence entre leur raisonnement évaluatif et la note attribuée.

Des études ont comparé des examinateurs-locuteurs natifs et des examinateurs-locuteurs non natifs (de la langue évaluée) ayant les mêmes qualifications, le même nombre d'années d'expérience et ayant suivi les mêmes formations. Les conclusions montrent que, d'une part, les locuteurs natifs font des commentaires plus riches et plus précis que les locuteurs non natifs et, d'autre part, qu'ils sont plus exigeants en termes d'exactitude prosodique. Cependant, il n'y a pas de différences significatives dans l'ensemble entre les deux groupes en termes de sévérité (Brown, 1995 ; Gui, 2012 ; Kim, 2009 ; Zhang & Elder, 2011).

D'autres recherches ont tenté de déterminer s'il y avait une différence de sévérité lorsque les examinateurs étaient familiers ou non avec l'accent du candidat. Les résultats ont montré que la familiarité de l'examinateur avec l'accent du candidat avait des répercussions positives sur le score de la prononciation et de l'accent de façon significative, avec néanmoins un impact limité dans la notation globale (Carey et al., 2011 ; Hsieh, 2011 ; Huang et al., 2016 ; Huang & Jun, 2015 ; Winke et al., 2011, 2012).

La manière dont l'examinateur interagit avec les candidats peut également constituer une source de variabilité. Selon la manière de structurer les séquences de conversations, de poser des questions ou de fournir

des rétroactions, le comportement et le langage de l'examinateur peuvent avoir une incidence sur la note des candidats (Cafarella, 1994 ; Filipi, 1994 ; Lazaraton, 1996a, 1996b ; Lazaraton & Saville, 1994 ; Morton et al., 1997). Par exemple, ceux qui interagissent avec des examinateurs plutôt accommodants ont en majorité des notes plus élevées que ceux qui interagissent avec des examinateurs peu accommodants (Brown, 2003, 2005 ; Reed & Halleck, 1997).

D'autres études montrent que les examinateurs ont très souvent tendance à faire des inférences pour expliquer certains comportements de candidats ou pour justifier l'attribution de certaines notes. Par exemple, un problème de grammaire, de lexique et d'idées d'un candidat peut être expliqué par un manque de culture générale ou par une immaturité de la personnalité. Ces inférences ne sont pas pertinentes, car elles ne forment pas une base adéquate pour la formulation d'un jugement étant donné que les vraies causes des difficultés des candidats sont inconnues (Brown, 2000, 2006 ; Orr, 2002 ; Pollitt & Murray, 1993).

Enfin, en s'appuyant sur le constat que les perceptions des examinateurs sont multiples, des études ont révélé que les raisonnements évaluatifs et les notes attribuées peuvent différer. En effet, deux examinateurs peuvent attribuer la même note sur une grille d'évaluation pour une même performance, alors que leurs interprétations de la performance peuvent diverger. À l'inverse, ils peuvent percevoir une même performance de manière similaire et attribuer des notes différentes (Ang-Aw & Goh, 2011 ; Orr, 2002). Les interprétations du discours par les interlocuteurs varient en fonction des différents aspects sur lesquels ils se penchent (Douglas, 1994).

En nous basant sur ces études portant sur les effets des examinateurs, notre objectif, dans cette recherche, est d'observer s'il existe des divergences à travers le jugement des examinateurs du TEF lors de l'évaluation de l'épreuve d'expression orale. L'étude de la question est nécessaire, car la variabilité attribuable aux effets des examinateurs a des conséquences importantes sur les processus décisionnels, en particulier dans les situations de tests à enjeux élevés (Bachman et al., 1995 ; Brown, 1995 ; Engelhard & Myford, 2003 ; Lumley & McNamara, 1995 ; McNamara, 1996). Dans un test à enjeux très élevés tel que le TEF, l'incidence sociale est forte et les résultats obtenus peuvent déclencher de nouveaux événements dans la vie des candidats (Casanova & Demeuse, 2011).

Avant d'observer les divergences chez les examinateurs, il est important de rappeler que le jugement ne peut se résumer à un simple raisonnement algorithmique et qu'il s'élabore au sein d'une démarche cognitive complexe (Goasdoué & Vantourout, 2016). La partie suivante met de l'avant les différents écrits traitant de la cognition des évaluateurs.

Cadre conceptuel

L'avènement de la recherche sur la cognition des évaluateurs

Les premières études sur la cognition de l'évaluateur ont commencé à la fin des années 1800 avec les travaux du philosophe, physicien et psychologue Fechner (1897) qui visaient à rendre compte du jugement esthétique. Il revendiquait la valeur des caractéristiques observables dans une œuvre d'art comme base pour le jugement esthétique. Son travail était relié à la cognition de l'évaluateur, car il présupposait qu'une personne qui juge est capable d'analyser une création à travers un ensemble de caractéristiques. Cette tâche nécessitait alors une participation cognitive de la part de l'évaluateur (Jörgensen, 2003). Ce premier postulat a donné naissance, au 20e siècle, au concept du modèle de lentille (*lens model*) de Brunswik (1952). Ce concept consiste à attribuer à l'individu la capacité à reconnaître une sélection hétérogène et complexe dans les intrants et les extrants, en établissant de nouveaux foyers ou, simplement, en ignorant certains aspects. Cela souligne le fait que chaque personne qui évalue peut mettre l'accent sur différents aspects du stimulus afin d'arriver à un jugement.

À la même époque où Fechner (1897) s'intéressait à la nature du jugement esthétique, une étude distincte a émergé avec les travaux du statisticien Edgeworth (1890) qui a examiné les éléments de hasard pouvant affecter un score. Ce dernier était conscient du problème de l'accord interjuges et a reconnu que les différences individuelles parmi les évaluateurs pouvaient être sources d'erreur en remarquant, par exemple, que certains pouvaient être plus sévères ou plus cléments que d'autres.

Durant une grande partie du 20e siècle, aux États-Unis, le problème de l'accord interjuges a retardé l'utilisation de tests à grande échelle nécessitant des réponses construites à l'écrit et à l'oral. Le premier vrai test d'expression orale utilisé en Amérique du Nord n'a fait son apparition qu'en 1930 avec *The College Board's English Competence Examination*. Auparavant, les épreuves orales des tests étaient composées de dictées, d'exercices de prononciation, de transcriptions écrites, de réponses aux

questions orales des examinateurs (Spolsky, 1995). En raison du problème de fidélité chez les évaluateurs, les tests avec un format à choix multiples ont commencé leur vaste expansion, comme le test *Army Alpha*, introduit en 1917, destiné à la sélection des soldats de la Première Guerre mondiale. L'*Army Alpha* a par la suite pavé la voie à de nombreux tests d'admission à choix multiples dans les années 1920 (Fuess, 1950).

Étant donné que la compréhension de la cognition de l'évaluateur était à ce moment-là intuitive, résoudre le problème de l'accord interjuges dans les tests d'admission a été une longue lutte (Bejar, 2012). Dans le contexte américain, cette lutte était motivée par le désir d'introduire des réponses construites dans les tests. Le caractère direct des réponses construites était considéré comme une qualité incontestable, car cela constituait un meilleur élément de preuve de compétence de la personne évaluée et, par conséquent, un élément important dans un argument de validité (Kane, 2006).

En 1961, l'absence d'accord entre les examinateurs a été mise en avant dans une étude majeure menée par le *Educational Testing Service*, un organisme privé américain de mesure et d'évaluation en éducation, fondé en 1947. Dans cette étude, 53 examinateurs chevronnés issus de domaines variés (professeurs d'anglais, spécialistes des sciences sociales et naturelles, écrivains, rédacteurs en chef, juristes, chefs d'entreprise, etc.) ont été invités à évaluer 300 articles sans qu'on leur impose de normes ni de critères. Par conséquent, la fidélité interexaminateurs s'est avérée très pauvre. Les examinateurs, livrés à eux-mêmes, ont valorisé différents aspects de l'écriture en s'appuyant sur des critères distincts ; ils ont alors été classés en différentes écoles de pensée (Diederich et al., 1961).

La première étude, qui traitait de la même question avec des conclusions similaires mais dans le contexte plus spécifique d'un test oral (via une entrevue) en langue seconde, a été menée en 1983 par Shohamy. Cette étude quantitative a montré le fait que les notes des candidats pouvaient varier considérablement selon les examinateurs, selon le style de discours employé (discours direct ou discours rapporté) et selon le sujet de conversation. Cette étude fait figure de pionnière parmi les études portant sur les effets des examinateurs dans les tests d'expression orale en langue seconde (Brown, 2003).

Durant les dernières décennies, des modèles théoriques de la cognition de l'évaluateur ont été explorés en langue première, notamment pour l'évaluation de l'expression écrite (Crisp, 2008, 2010 ; Freedman & Calfee, 1983 ; Sanderson, 2001 ; Wolfe, 1997) et de la communication orale (Joe

et al., 2011). La théorie du traitement de l'information constitue le fondement de ces nombreuses études sur la cognition de l'évaluateur. Cette théorie est basée sur l'idée que les humains traitent les informations qu'ils reçoivent, plutôt que de simplement répondre à des stimuli. Elle s'appuie plus spécifiquement sur des preuves tirées de recherches approfondies en psychologie cognitive et en neurosciences qui utilisent la technologie de balayage cérébral pour étudier la manière dont le traitement se manifeste dans le cerveau (Dehn, 2008).

La recherche sur la cognition des évaluateurs, spécifiquement axée sur leurs processus mentaux, se rapporte principalement à l'architecture du traitement humain de l'information (Baddeley, 2012 ; Baddeley et al., 2009 ; Gagné et al., 1993 ; Purpura, 2012), ainsi qu'aux différentes stratégies (méta)cognitives (Purpura, 2012) déployées tout au long de la notation. L'architecture du traitement humain de l'information illustre la manière dont la structure et les processus sous-jacents (par exemple, la mémoire à court terme, la mémoire de travail et la mémoire à long terme) sont impliqués dans le codage, le stockage et la récupération des informations lors de la notation. De manière complémentaire, les stratégies (méta)cognitives, par exemple, l'attention, le raisonnement, la prise de décision, la planification, peuvent être connectées à cette architecture afin de rendre compte de façon précise de ce qu'il se passe dans l'esprit des évaluateurs lors de la notation.

Dans le domaine de l'évaluation de l'expression orale en langue seconde, les modèles du traitement de l'information comme cadre de référence sont absents. Cependant, un nombre très restreint de chercheurs ont tenté de conceptualiser la cognition de l'examinateur. Ces chercheurs se sont penchés sur les concepts suivants : les approches évaluatives des examinateurs (Pollitt & Murray, 1993 ; Reed & Cohen, 2001), la cognition de l'évaluateur adaptée à un champ général (Bejar, 2012) et la cognition de l'examinateur adaptée à une situation de test d'expression orale (Han, 2016).

Les approches évaluatives des examinateurs de Pollitt et Murray (1993)

Pollitt et Murray (1993) ont relevé deux approches évaluatives contrastives auxquelles les examinateurs ont recours lors de l'évaluation de l'expression orale : une approche qui utilise un procédé dit «synthétique» et une autre qui utilise un procédé dit de «puzzle». Dans l'approche synthétique, une image holistique du candidat est formée, dérivée au préalable d'une compréhension individuelle du candidat qui est préconçue et préconstruite. Cela s'apparente à une première rencontre avec une personne inconnue lors

d'un évènement social où une image globale de la personne est tracée par quelques premières impressions. Au départ, quelques aspects de la performance servent d'indicateurs du niveau du candidat. Puis, l'examinateur compare la performance observée avec celle d'un autre candidat du même niveau qu'il a gardé en mémoire. Dans l'approche de puzzle, l'examinateur limite ses commentaires au comportement observé du candidat. La pratique consiste à noter les candidats d'après chaque énoncé observé, de nouvelles informations s'ajoutant au fur et à mesure. Il s'agit d'un mode plus objectif mais moins naturel, qui requiert un plus grand effort que le précédent, car l'examinateur doit réfléchir dans un cadre strictement évaluatif.

Les approches évaluatives de Reed et Cohen (2001)

Reed et Cohen (2001) ont mis en avant quelques aspects d'une situation d'évaluation que les examinateurs doivent garder en tête lorsqu'ils évaluent une performance linguistique. Ces aspects ont le potentiel d'affecter la nature de la performance linguistique et, par là même, son évaluation. Par exemple, l'examinateur doit être conscient du fait que la performance qu'il observe est influencée par l'attitude du candidat, par ses caractéristiques linguistiques (sa langue maternelle, la langue cible) et par les caractéristiques de la tâche (le contenu, le niveau de difficulté). De surcroît, l'examinateur doit être conscient de sa propre tendance à être sévère ou clément et de sa tendance à se laisser porter par son utilisation personnelle des critères de la grille d'évaluation. Tous ces éléments doivent être pris en compte lors de l'évaluation. D'autre part, lorsque l'examinateur doit participer à la performance avec le participant, la tâche se complexifie. L'examinateur doit simultanément jouer le rôle de l'interlocuteur et construire avec le candidat un discours conversationnel suivant un principe de coopération. L'acte d'évaluation devient alors une construction de sens qui nécessite plusieurs résolutions d'incertitudes de façon synchronisée.

Le modèle de la cognition de l'évaluateur de Bejar (2012)

Bejar (2012) a tenté de mettre en lumière les aspects de la cognition de l'évaluateur dans un modèle où il préconise, entre autres, des pratiques en matière de notation. Il propose un modèle générique, qui n'est pas propre à un domaine particulier, en deux phases : une première phase de conception de l'évaluation et une deuxième phase de notation. Dans la première phase, il souligne qu'il est important que les évaluateurs soient capables de bien comprendre les grilles d'évaluation et, surtout, qu'ils y adhèrent.

Comme les critères de notation peuvent parfois être difficiles à saisir, il est important d'être attentif aux considérations cognitives des évaluateurs, sinon l'interprétation du score risque de ne pas être valide.

Dans la première phase, l'auteur ajoute que la formation des évaluateurs peut être en partie conceptualisée en amenant l'évaluateur à encoder le matériel de formation dans une «grille d'évaluation mentale». Il s'agit d'un processus cognitif qui permet à l'évaluateur d'inférer une représentation des critères de notation à partir du matériel de formation. Selon l'objet à évaluer et le temps dont il dispose, l'évaluateur peut réduire les critères de notation en une représentation gérable, en indexant sa grille d'évaluation mentale uniquement sur ses connaissances de base. Cette grille est propre à chaque évaluateur et peut, par conséquent, être influencée par les attributs personnels et les antécédents de ce dernier. Toutefois, ces éléments risquent de conduire à une représentation contenant des composantes du construit non pertinentes, des biais ou des composantes n'étant pas explicitement mentionnées dans la grille.

Dans la deuxième phase, celle de la notation, l'évaluateur forme une «représentation de réponse mentale» de la réponse de la personne évaluée, qu'il met ensuite en parallèle avec sa grille mentale. Ce processus s'appuie sur des représentations de similitude et de probabilité dans l'esprit de l'évaluateur. En pratique, le processus qu'utilisent les évaluateurs pour attribuer une note peut être plus complexe puisqu'il peut introduire des informations antérieures (observations du passé, biais cognitifs) à long et à court termes dans le processus de notation.

Le modèle de la cognition de l'examinateur de Han (2016)

Han (2016) propose un modèle sur la nature de la cognition de l'examinateur dans le contexte spécifique des tests d'expression orale en langue seconde. Selon l'auteure, ce modèle reste hypothétique, car sa fonction principale est de proposer un postulat de départ servant avant tout à la future recherche sur la cognition des examinateurs.

Son modèle établit une interface entre le processus d'évaluation de l'expression orale (écoute de la performance du candidat, utilisation de la grille, notation finale), les composants de l'architecture du traitement humain de l'information qui sont activés (mémoire à court terme, mémoire de travail, mémoire à long terme) et les processus cognitifs des examinateurs qui sont exploités lors de l'évaluation (attention sélective aux caractéristiques du dialogue, traitement actif de nouveaux intrants,

etc.). Le modèle tient en compte des récepteurs sensoriels (yeux, oreilles) et intègre également un large éventail de stratégies (méta)cognitives que les examinateurs peuvent utiliser pour réguler le fonctionnement de leurs mécanismes cognitifs lors du processus de notation (comparaison, répétition, traduction, synthèse, etc.).

Tous ces éléments sont pris en compte dans les différentes étapes de représentation mentale des examinateurs, soit : former une représentation mentale de la norme et de la grille, former une représentation mentale de la réponse orale, comparer les réponses mentales et attribuer simultanément une note provisoire, examiner ou réviser la note et justifier la note finale. Ces étapes sont influencées par diverses caractéristiques inhérentes à l'examinateur (comme son expérience en évaluation, sa formation, son âge, son sexe, sa langue maternelle, son origine culturelle, son attitude à l'égard des accents des candidats, son style cognitif) et diverses variables environnementales.

Ainsi, afin de mieux cerner les examinateurs, les chercheurs mentionnés plus tôt (Bejar, 2012 ; Han, 2016 ; Pollitt & Murray, 1993 ; Reed & Cohen, 2001) ont tenté d'illustrer leurs approches évaluatives alors que d'autres ont conceptualisé leur cognition à travers des modèles. La présente recherche se situe dans la continuité des recherches recensées et, plus particulièrement, dans la continuité des modèles proposés par Bejar (2012) et par Han (2026), car elle étudie, entre autres, l'examinateur en tentant d'observer ce qu'il fait à partir du raisonnement sous-jacent à ses activités évaluatives.

Méthodologie

Cette recherche, qui a pour but d'observer les différentes formes de divergences dans le jugement des examinateurs, s'inscrit dans une perspective qualitative/interprétative, en raison de notre désir de décrire le sens que les participants attribuent à leur expérience (Savoie-Zajc, 2011). La méthode de recherche retenue est celle de la technique de la pensée à voix haute (*think aloud protocol*), mise au point afin d'accéder à ce qui se passe dans la tête des participants au moment de réaliser une tâche et, ainsi, d'expliciter les aspects cognitifs implicitement présents dans des actions.

Les participants

L'échantillon est constitué de 10 participants recrutés sur une base volontaire, et qui devaient répondre à certains critères. L'échantillonnage était intentionnel, c'est-à-dire que les éléments de la population ont été

choisis selon des critères précis afin que les éléments soient représentatifs du phénomène à l'étude. Les critères d'inclusion sélectionnés sont les suivants : être examinateur TEF certifié depuis au moins trois ans et évaluer annuellement en moyenne au moins 100 candidats de l'épreuve d'expression orale dans les centres d'examens agréés.

Leurs données d'identification sont les suivantes : sept sont des hommes et trois, des femmes (comme il n'existe aucune donnée statistique sur le genre des examinateurs du TEF, il est impossible de dire que cette répartition représente la population). Ils sont âgés entre 36 et 55 ans. Parmi ces personnes, neuf sont enseignants de français langue seconde depuis plus de 10 ans et le dixième est responsable pédagogique et a une expérience de neuf ans en enseignement du français langue seconde. Leur expérience en tant qu'examinateur TEF va de 3 à 11 ans. Ils évaluent en moyenne entre 100 et 300 candidats annuellement. Ils sont basés à Montréal et ses environs et proviennent de six centres d'examen différents. Étant donné qu'il n'existe pas d'affiliation exclusive à un centre spécifique, trois des participants travaillent pour deux centres différents.

Le déroulement de la collecte de données

Les rencontres avec les participants se sont faites individuellement, en ligne. La collecte de données n'a pas nécessité de rencontre avec des candidats, mais a été faite à l'aide d'enregistrements audio fournis par l'équipe pédagogique du concepteur du test. Ces enregistrements audios mettent en scène des entrevues authentiques de candidats interagissant avec des animateurs qui ne sont pas les examinateurs de notre étude. Chacun des dix examinateurs de notre étude a évalué les quatre mêmes candidats de niveaux différents. Il y avait donc quarante évaluations au total. Les quatre niveaux de français des candidats étaient les suivants : intermédiaire ou de survie (A2), seuil (B1), avancé ou indépendant (B2) et autonome (C1) (selon les niveaux communs de référence du CECRL). Ces niveaux ont été déterminés au préalable par l'équipe pédagogique de l'organisme.

L'épreuve d'expression orale du TEF est composée de deux sections et dure au total 15 minutes, soit 5 minutes pour la section A et 10 minutes pour la section B. Ces deux sections ont permis de compartimenter le déroulement du premier temps de notre collecte de données. Tout d'abord, l'examinateur a écouté l'enregistrement de la section A puis, une fois l'écoute terminée, il a immédiatement commencé à remplir la grille d'évaluation et à verbaliser ses pensées simultanément, c'est-à-dire qu'il a commenté les

points qu'il a attribués sur la grille d'évaluation. Dans un deuxième temps, l'examinateur a écouté l'enregistrement de la section B et, une fois l'écoute terminée, il a fini de remplir la grille d'évaluation tout en verbalisant ses pensées. L'examinateur a été invité à prendre des notes durant les écoutes.

La grille d'évaluation utilisée était la même que celle de l'évaluation de l'épreuve d'expression orale du TEF. Elle comporte cinq critères : 1) Capacité à obtenir des informations dans la section A ; 2) Capacité à présenter et à débattre dans la section B ; 3) Syntaxe ; 4) Lexique et 5) Aisance à l'oral, élocution. Les niveaux de la grille sont ceux du Cadre européen commun de références pour les langues, soit : < A1 (niveau inférieur à A1) ; A1 (niveau introductif ou de découverte) ; A2 (niveau intermédiaire ou de survie) ; B1 (niveau seuil) ; B2 (niveau avancé ou indépendant) ; C1 (niveau autonome) ; C2 (niveau maîtrise).

Pour chacun des dix participants, nous avons répété le processus d'évaluation quatre fois étant donné que la tâche consistait à évaluer quatre candidats différents. Le temps alloué était d'environ 90 minutes au total pour chaque participant.

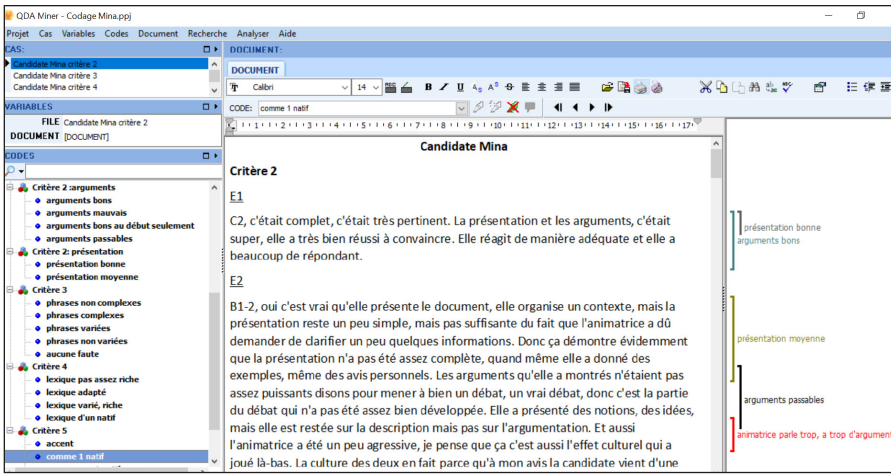
L'analyse des résultats

À la suite de la collecte des données, les rapports verbaux ont été transcrits manuellement. Les blocs de données significatives ont été découpés et mis en correspondance avec des étiquettes et des sous-étiquettes afin de faire émerger un sens. Ce travail de codage des unités de sens s'est fait à l'aide du logiciel *QDA Miner* (version 5.0) (Provalis Research, 2021).

Nous avons d'abord procédé à la préanalyse, puis à l'exploitation du matériel. La préanalyse consistait à lire et à relire les transcriptions pour saisir le sens du message, à identifier les thèmes liés aux objectifs de recherche, puis à repérer des indices permettant l'identification des thèmes afin de préparer l'étape suivante, celle de l'exploitation du matériel. Cette dernière consistait à classifier les éléments constitutifs d'un ensemble par différenciation, de les regrouper par analogie, puis à coder les unités. Pour cela, nous avons d'abord dressé une liste de rubriques émergentes, puis établit des liens entre celles-ci. Ensuite, nous les avons regroupées en sous-catégories ou en méta-catégories. Par la suite, nous avons fait des allers-retours entre les *verbatims* et notre liste de rubriques afin de voir si les interprétations concordaient avec le matériel original.

La Figure 1 présente l'émergence du schéma de codage et le processus de codage.

Figure 1
Émergence du schéma de codage et du processus de codage



Afin d'apporter plus de rigueur à notre recherche, nous avons sollicité un chercheur externe spécialisé en mesure et évaluation en sciences de l'éducation pour effectuer un travail de contre-codage des données. Comme ce processus est long et exigeant, nous l'avons limité à 40% du corpus entier plutôt qu'à l'ensemble recueilli. Cette proportion est liée à la taille du corpus et à la complexité des codages. De façon à donner des repères communs, nous avons présenté au préalable notre arbre de codage au chercheur externe. Celui-ci avait la possibilité de faire émerger de nouvelles variables ou de regarder des données existantes d'une nouvelle manière. À l'aide du logiciel *QDA Miner*, nous avons calculé l'accord interjuges et les résultats ont corroboré à plus de 90%.

Les résultats et la discussion

Les divergences dans le raisonnement évaluatif et la note

Dans l'étude, nous avons très fréquemment observé le fait que les examinateurs pouvaient accorder la même note pour une même performance alors que leurs interprétations pouvaient différer. Et inversement, ils pouvaient percevoir une performance de manière similaire et attribuer des notes divergentes. Par exemple, pour l'évaluation du critère de l'aisance

à l'oral et de l'élocution d'une candidate, nous pouvions entendre ce commentaire venant d'un examinateur : «*J'avais vraiment de la misère à comprendre (…) quand elle parle on ne comprend pas, elle a un gros accent*», puis ce commentaire venant d'un autre examinateur : «*J'avais vraiment de la difficulté à comprendre ce qu'elle dit, des fois on dirait qu'elle parlait pour elle, elle murmurait, je comprenais mal*». Les deux commentaires sont similaires, pourtant l'un a attribué A1 et l'autre B1-1. À l'inverse, pour l'évaluation de la capacité à obtenir des informations d'un candidat, nous pouvions entendre ce commentaire de la part d'un examinateur : «*Elle a posé beaucoup de questions, vraiment beaucoup de questions, y compris des questions après les interventions de l'animatrice (…) concernant l'achat d'un appartement, c'est bon*», puis ce commentaire de la part d'un autre examinateur : «*Il y a eu quand même pas mal de bonnes questions, mais pas suffisamment (…) à un moment elle a bloqué, elle n'avait plus envie de poser des questions*». Les deux commentaires sont divergents, pourtant les deux notes sont identiques : B2-1. Par ailleurs, tout au long de l'exercice de verbalisation, les disparités dans les notes étaient permanentes, les 10 examinateurs n'ont à aucun moment tous attribué la même note à un même critère pour l'ensemble des quatre candidats.

Les résultats valident ce que les études empiriques ont révélé sur la différence entre le raisonnement évaluatif et la note que les examinateurs attribuent. En effet, il a été démontré que deux examinateurs peuvent attribuer la même note sur une grille d'évaluation pour une même performance orale, alors que leurs interprétations de la performance peuvent diverger. À l'inverse, ils peuvent percevoir une même performance de façon analogue et attribuer des notes différentes (Ang-Aw & Goh, 2011 ; Douglas, 1994 ; Douglas & Selinker, 1992, 1993 ; Orr, 2002).

Les divergences dans la familiarité avec l'accent des candidats

Dans l'étude, quelques examinateurs ont pris conscience qu'ils avaient eu une perception différente des traits phoniques des candidats en raison de leur familiarité avec certaines caractéristiques. Par exemple, grâce à l'identification de l'accent d'une candidate, un examinateur a déclaré qu'il avait des origines géographiques similaires aux siennes et, par conséquent, il a affirmé qu'il la comprenait très bien par rapport aux autres examinateurs. Il a alors révélé que la note qu'il a donnée à la candidate pour le critère de l'aisance à l'oral et de l'élocution aurait pu être plus élevée de quatre échelons sur la grille d'évaluation en raison de cette familiarité.

Cependant, il a décidé de ne pas hausser sa note, car il a pris conscience de cette subjectivité. Un autre examinateur a reconnu l'origine géographique d'une candidate grâce à son accent et a évoqué les difficultés prosodiques typiques en français de ses apprenants (lors de ses classes de français langue seconde) issus de la même région qu'elle. En comparant les difficultés spécifiques de ses apprenants avec la prestation de ladite candidate, il a trouvé que celle-ci se débrouillait très bien et, par conséquent, cela l'a incité à lui donner une note plus élevée pour le critère de l'aisance à l'oral et de l'élocution. L'examinateur était toutefois conscient du fait que cette familiarité l'avait influencé de façon positive.

Ces résultats soutiennent les conclusions des études empiriques sur la familiarité des examinateurs avec l'accent des candidats dans lesquelles on constate que ce facteur influence de façon positive le score de la prononciation (Carey et al., 2011 ; Hsieh, 2011 ; Huang et al., 2016 ; Huang & Jun, 2015 ; Winke et al., 2011, 2012).

Les divergences dans les inférences non pertinentes

Dans l'étude, les examinateurs ont fait beaucoup d'inférences pour expliquer certains comportements de candidats. Par exemple, pour justifier la très grande difficulté d'une candidate à comprendre la consigne des sujets, certains ont fait allusion à son faible niveau de scolarisation, au fait qu'elle soit assez âgée (d'après leur perception de sa voix) ainsi qu'au fait qu'elle n'aurait pas suffisamment obtenu d'encadrement pour être dans de bonnes conditions pour passer le test, car l'animatrice ne lui aurait pas expliqué clairement la tâche au préalable. D'autres examinateurs ont également supposé qu'étant donné son origine culturelle, cette même candidate aurait eu de la difficulté à se projeter dans le jeu de rôle de l'épreuve. Par ailleurs, elle aurait utilisé des styles de phrases typiques de sa zone géographique qui n'étaient pas celles que demandait le test.

Globalement, les biais culturels ont souvent été évoqués afin d'expliquer certaines difficultés rencontrées par les candidats, et plus particulièrement les difficultés à argumenter et à débattre dans la conversation avec l'animateur. D'après leurs observations sur le terrain, les examinateurs constatent que cela concerne davantage les candidats d'origine asiatique et plus particulièrement de la Chine, de la Corée du Sud et du Japon. Selon eux, ces candidats sont plutôt réservés et cèdent facilement, même lorsque leur niveau de compétence langagière en français est bon.

Les résultats appuient ce que les études sur les effets des examinateurs ont démontré à propos des inférences ne donnant pas de sens au jugement. Ces études affirment que les examinateurs sont souvent incertains face aux vraies causes des problèmes vécus par les candidats (Brown, 2000, 2006 ; Pollitt & Murray, 1993). Ils peuvent alors faire des inférences pour excuser ou expliquer certains comportements ou pour justifier l'attribution de certaines notes. Or, ces inférences constituent un problème majeur, car elles ne forment pas une base adéquate pour la formulation d'un jugement.

Les divergences dans la perception de l'attitude de l'animateur

Dans l'étude, l'attitude de l'animateur avec les quatre candidats a largement été commentée, et un même animateur pouvait être perçu différemment par chaque examinateur. Par exemple, pour certains, un animateur était négligent et n'avait pas bien donné suite à l'intervention du candidat, et pour d'autres, il avait très bien agi et avait fait tout son possible pour s'adapter au candidat. Par ailleurs, plusieurs examinateurs ont constaté que quelques animateurs n'appliquaient pas les bonnes techniques d'animation et ont avoué que si ces fautes n'avaient pas été commises, ils auraient pu accorder une note plus élevée aux candidats. Le fait que l'attitude des animateurs affecte les performances des candidats fait écho aux résultats des recherches sur les effets des examinateurs traitant du propre style interactionnel des animateurs. En effet, la posture de ces derniers peut représenter une source de variabilité, les candidats ne sont alors pas traités de façon égale, car ils peuvent se retrouver face à un interlocuteur peu conciliant et ainsi être désavantagés (Brown, 2003 ; Brown, 2005 ; Brown & Hill, 1998 ; Cafarella, 1994 ; Filipi, 1994 ; Lazaraton, 1996a, 1996b ; Lazaraton & Saville, 1994 ; Morton et al., 1997 ; Reed & Halleck, 1997).

Recommandations

Malgré les formations suivies et l'utilisation d'un même outil basé sur une norme commune, les examinateurs n'attribuent pas les mêmes significations aux performances des candidats. En raison de sensibilités diverses, les mêmes réalités peuvent générer des perceptions variées. Cela conduit alors à des décisions différentes aux conséquences importantes pour les candidats à l'immigration et à la citoyenneté. Il est donc essentiel que les formations initiales et continues de l'épreuve d'expression orale du TEF se fondent sur une démarche très rigoureuse permettant de minimiser le plus possible l'hétérogénéité des points de vue. Pour cela, les formations devraient prendre en compte l'ensemble des regards évaluatifs que cette

étude a pu identifier. Il serait également pertinent de connaître les différentes études empiriques menées sur les effets des examinateurs afin de développer une meilleure prise de conscience des nombreuses variables pouvant représenter une menace à la fidélité du test.

Dans le cadre des formations continues plus spécifiquement, des mesures favorisant la standardisation pourraient consister en la tenue régulière et consciencieuse de séances d'activités d'évaluation de candidats. Les séances devraient avoir lieu de façon individuelle dans un premier temps, afin qu'il n'y ait pas d'influence mutuelle entre les examinateurs. Puis, dans un deuxième temps, les séances seraient organisées en plénière afin de trouver un consensus et d'avoir une vision commune dans l'attribution des niveaux des candidats.

Par ailleurs, et de manière plus générale, l'épreuve d'expression orale pourrait être améliorée en accordant la parole aux examinateurs. Étant donné que la voix des examinateurs est considérée comme une ressource importante qui favorise la validité des tests d'expression orale (Ducasse & Brown, 2009 ; Galaczi et al., 2012 ; Nakatsuhara et al., 2017), il serait pertinent de recueillir leur opinion au moyen de sondages à grande échelle. Les informations pourraient porter sur différents aspects de l'épreuve comme les tâches, les sujets, le format, les pistes de l'interlocuteur, la grille d'évaluation, les points problématiques de l'évaluation, l'animation de la conversation avec le candidat ainsi que les formations. Cela permettrait à l'organisme concepteur du TEF de mieux connaitre l'avis des examinateurs et d'avoir une meilleure compréhension des pratiques évaluatives utilisées sur le terrain.

Limites de cette recherche

L'emploi de la technique de la pensée à voix haute a permis de recueillir beaucoup de données. Toutefois, il faut interpréter les données obtenues avec précaution étant donné qu'elles ne proviennent que de 10 examinateurs qui ont évalué quatre candidats. Ainsi, la notation des examinateurs n'a pas coïncidé pour les évaluations orales mais, s'ils avaient évalué plus de candidats, leurs notations auraient certainement coïncidé dans certaines de leurs évaluations. De plus, le TEF est diffusé dans un réseau de 308 centres d'examen agréés dans le monde et cette recherche a été présentée uniquement en contexte québécois. Ainsi, les centres d'examen agréés auxquels sont affiliés les participants étaient tous basés dans la région de Montréal, ce qui limite la représentativité des résultats obtenus à l'échelle mondiale.

Conclusion

Fondée sur les études empiriques recensées au sujet des effets des examinateurs, cette recherche avait pour objectif de découvrir l'existence de divergences à travers le jugement des examinateurs dans le cadre de l'épreuve d'expression orale du TEF. L'analyse minutieuse des activités des examinateurs a montré que ces derniers valorisaient des aspects différents. Les résultats ont permis d'observer plusieurs formes de divergences. Celles-ci reposent sur le raisonnement évaluatif et la note, sur la familiarité avec l'accent des candidats, sur les inférences non pertinentes pour le jugement ainsi que sur la perception de l'attitude de l'animateur.

Cette recherche a permis de développer une connaissance accrue et d'approfondir la compréhension de la pratique des examinateurs en portant un regard interne sur cette pratique. Nous avons tenté d'apporter des éléments favorisant une meilleure transparence et, subséquemment, une fidélité et une validité accrues de l'acte d'évaluer. Étant donné l'enjeu très élevé associé au test et compte tenu des projets de vie des candidats, il est important d'apporter du sérieux et de la rigueur aux évaluations du TEF. Les retombées de cette recherche pourraient être mises à profit par l'organisme concepteur du test afin d'améliorer la standardisation des procédures en amenant les examinateurs à être plus efficaces et plus constants dans leur évaluation. D'après plusieurs chercheurs, les organismes concepteurs et certificateurs de tests de langue qu'incarnent les professionnels de l'évaluation se doivent de montrer et de justifier les mesures qu'ils prennent afin de réduire les risques de variabilité et d'optimiser la validité des scores attribués (American Educational Research Association et al., 2014 ; Chapelle et al., 2008 ; McNamara, 1996 ; O'Sullivan & Weir, 2011 ; Saville, 2009 ; Shohamy, 2007 ; Taylor & Galaczi, 2011).

Réception : 02 décembre 2021
Version finale : 22 mai 2022
Acceptation : 23 mai 2022

LISTE DE RÉFÉRENCES

American Educational Research Association, American Psychological Association et National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* https ://www.testingstandards.net/uploads/7/6/6/4/76643089/9780935302356.pdf

Ang-Aw, H. T. & Goh, C. C. M. (2011). Understanding discrepancies in rater judgment on national-level oral examination tasks. *RELC journal, 42*(1), 31-51. https ://doi-org.proxy3.library.mcgill.ca/10.1177/0033688210390226

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing,* Oxford University Press.

Bachman, L. F., Lynch, B. & Mason, M. (1995). Investigating Variability in Tasks and Rater Judgements in a Performance Test of Foreign Language Speaking. *Language Testing*, *12*(2), 239-257. https ://doi.org/10.1177/026553229501200206

Bachman, L. F. & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Baddeley, A. (2012). Working memory : theories, models, and controversies. *Annual review of psychology, 63*, 1-29. https ://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A., Eysenck, M. W. & Anderson, M. C. (2009). *Memory*. Psychological Press.

Bejar, I. I. (2012). Rater cognition : Implications for validity. *Educational Measurement : Issues and Practice, 31*(3), 2-9. https ://doi.org/10.1111/j.1745-3992.2012.00238.x

Brown, A. (1995). The Effect of Rater Variables in the Development of an Occupation Specific Language Performance Test. *Language testing, 12*(2), 1-15. https ://doi.org/10.1177/026553229501200101

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. *IELTS Research Reports, 3*(3), 49-84.

Brown, A. (2003). Interviewer variation and the co-construction of speaking. *Language testing, 20*(1), 1-25. https ://doi.org/10.1191/0265532203lt242oa

Brown, A. (2005). Interviewer variability in oral proficiency interviews. *Language testing and evaluation, 4*, Peter Lang.

Brown, A. (2006). An examination of the rating process in the revised IELTS Speaking Test. *IELTS Research Reports, 6*. https ://search.informit.org/doi/10.3316/informit.078722747791492

Brown, A. & Hill, K. (1998). Interviewer Style and Candidate Performance in the IELTS Oral Interview. *International English Language Testing System (IELTS) Research Reports, 1.*

Brunswik, E. (1952). *The conceptual framework of psychology*. University of Chicago Press.

Cafarella, C. (1994). Assessor accommodation in the V.C.E. Italian oral test. *Australian Review of Applied Linguistics, 20*, 21-41. https ://doi.org/10.1075/aral.20.1.02caf

Carey, M. D., Mannel, R. H. & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201-219. https ://doi.org/10.1177/0265532210393704

Casanova, D. & Demeuse, M. (2011). Analyse des différentes facettes influant sur la fidélité de l'épreuve d'expression écrite d'un test de français langue étrangère. *Mesure et évaluation en* éducation, 34(1), 25-53. https://doi.org/10.7202/1024862ar

Chapelle, C. A., Enright, M. A. & Jamieson, J. M. (2008). *Building a validity argument for the test of English as a foreign language*. Routledge. https://doi.org/10.4324/9780203937891

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

Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education, 38*, 247-264. https://doi.org/10.1080/03057640802063486

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education, 36*, 1-21. https://doi.org/10.1080/03054980903454181

Dehn, M. J. (2008). *Working memory and academic learning*. John Wiley & Sons Inc.

Diederich, P. B., French, J. W. & Carlton, S. T. (1961). *Factors in judgments of writing ability* (Research Bulletin No. RB-61–15). Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.1961.tb00286.x

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing 11*(2), 125–144. https://doi.org/10.1177/026553229401100203

Douglas, D. & Selinker, L. (1992). Analysing oral proficiency test performance in general and specific purpose contexts. *System 20*, 317-328. https://doi.org/10.1016/0346-251X(92)90043-3

Douglas, D. & Selinker, L. (1993). Performance on a general versus a field-specific test of speaking proficiency by international teaching assistants. Dans C. Chapelle et D. Douglas (dir.), *A new decade of language testing research* (p. 235-256). TESOL Publications.

Ducasse, A. & Brown, A. (2009). Assessing paired orals : Raters' orientation to interaction. *Language Testing, 26*(3), 423-443. https://doi.org/10.1177/0265532209104669

Eckes, T. (2011). Introduction to many-facet Rasch measurement : Analyzing and evaluating rater-mediated assessments. Peter Lang.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society, 53*, 460-475, 644-663.

Engelhard Jr., G. & Myford, C. (2003). Monitoring Faculty Consultant Performance in the Advanced Placement English Literature and Composition Program with a Many-Faceted Rasch Model (publication n° 2003-1 ETS RR-03-01). College Entrance Examination Board. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01893.x

Fechner, G. T. (1897). *Kollektivmasslehre*. Wilhelm Engelmann.

Filipi, A. (1994). Interaction in an Italian oral test : the role of some expansion sequences. *Australian Review of Applied Linguistics, 11*, 119-136. https://doi.org/10.1075/aralss.11.06fil

Freedman, S. W. & Calfee, R. C. (1983). Holistic assessment of writing : Experimental design and cognitive theory. Dans P. Mosenthal, L. Tamor et S. A. Walmsley (dir.), *Research on writing : principles and methods* (p. 75-98). Longman.

Fuess, C. M. (1950). *The College Board, its first fifty years*. Columbia University Press.

Gagné, E. D., Yekovich, C. W. & Yekovich, F. R. (1993). *The Cognitive Psychology of School Learning*. Harper Collins College Publishers.

Galaczi, E. D., Lim, G. & Khabbazbashi, N. (2012, 1er novembre). *Descriptor salience and clarity in rating scale development and evaluation* [communication orale]. The Language Testing Forum. University of Bristol, Bristol.

Goasdoué, R. & Vantourout, M. (2016). Évaluations scolaires et étude du jugement des enseignants : pour une docimologie cognitive. Dans P. Detroz, M. Crahay & A. Fagnant (dir.), *L'évaluation à la lumière des contextes et des disciplines* (p. 141-168). DeBoeck Supérieur.

Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Language Assessment Quarterly, 9*, 186-203. https://doi.org/10.1080/15434303.2011.614030

Han, Q. (2016). Rater Cognition in L2 Speaking Assessment : A Review of the Literature. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 16*(1), 1-24. https://doi.org/10.7916/D82R53MF

Hsieh, C. N. (2011). Rater effects in ITA testing : ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 9*, 47-74.

Huang, B., Alegre, A. & Eisenberg, A. (2016). A cross-linguistic investigation of the effect of raters' accent familiarity on speaking assessment. *Language Assessment Quarterly, 13*(1), 25-41. https://doi.org/10.1080/15434303.2015.1134540

Huang, B. & Jun, S. A. (2015). Age Matters, And So May Raters : Rater Differences in the Assessment of Foreign Accents. *Studies in Second Language Acquisition, 37*(4), 623-650.

Joe, J. N., Harmes, J. C. & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring : A mixed methods application to oral communication assessment. *Assessment in Education : Principles, Policy and Practice, 18*, 239-258. https://doi.org/10.1080/0969594X.2011.577408

Jörgensen, C. (2003). *Image retrieval : Theory and research*. Scarecrow Press.

Kane, M. T. (2006). Validation. Dans R. L. Brennan (dir.), *Educational measurement* (4e éd., p. 17-64). Praeger Publishers.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance : A mixed methods approach. *Language Testing, 26*(2), 187-217. https://doi.org/10.1177/0265532208101010

Lazaraton, A. (1996a). Interlocutor support in oral proficiency interviews : the case of CASE. *Language Testing, 13,* 151-72. https://doi/10.1177/026553229601300202

Lazaraton, A. (1996b). A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE). Dans M. Milanovic et N. Saville (dir.), *Performance testing, cognition and assessment : selected papers from the 15th Language Testing Research Colloquium* (p. 18-33). Cambridge University Press.

Lazaraton, A. & Saville, N. (1994). *Processes and outcomes in oral assessment* [communication orale]. 16th Language Testing Research Colloquium, Washington DC.

Lumley, T. & McNamara, T. F. (1995). Rater characteristics and rater bias : Implications for training. *Language testing, 12*(1), 54-71. https://doi.org/10.1177/026553229501200104

Lumley, T. & McNamara, T. F. (1997). The Effect of Interlocutor and Assessment Mode Variables in Overseas Assessments of Speaking Skills in Occupational Settings. *Language Testing, 14*(2), 140-56. https://doi.org/10.1177/026553229701400202

McNamara, T. F. (1996). Measuring second language performance. Longman.

Morton, J., Wigglesworth, G. & Williams, D. (1997). Approaches to the evaluation of interviewer performance in oral interaction tests. Dans G. Brindley et G. Wigglesworth (dir.), *Access : issues in English language test design and delivery* (p. 175-196). National Centre for English Language Teaching and Research.

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement : Part 1. *Journal of Applied Measurement, 4*(4), 386-422.

Nakatsuhara, F., Inoue, C., Berry, V. & Galaczi, E. (2017). Exploring the Use of Video-Conferencing Technology in the Assessment of Spoken Language : A Mixed-Methods Study. *Language Assessment Quarterly, 14*(1), 1-18. https://doi.org/10.1080/15434303.2016.1263637

Norman, W. T. & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology, 4*(6), 681-691. https://doi.org/10.1037/h0024002

Orr, M. (2002). The FCE speaking test : using rater reports to help interpret test scores. *System, 30*, 143-154. https://doi.org/10.1016/S0346-251X(02)00002-7

O'Sullivan, B. & Weir, C. J. (2011). Test development and validation. In Language testing : theories and practices. *Palgrave Advances in linguistics*, 13-32.

Pollit, A. & Murray, N. L. (1996). What raters really pay attention to? Dans M. Milanovic et N. Saville (dir.), *Performance testing, cognition and assessment : Selected papers from the 15th language research testing colloquium* (vol. 3, p. 74-91). Cambridge University Press.

Provalis Research (2021). QDA Miner (version 5.0) [logiciel] https://provalisresearch.com/fr/produits/

Purpura, J. E. (2012). *What is the role of strategic competence in a processing account of L2 learning or use?* [communication orale]. American Association for Applied Linguistics Conference, Boston, MA.

Reed, D. J. & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. *Studies in language testing. Experimenting with uncertainty, 11*, 82-96.

Reed, D. J. & Halleck, G. B. (1997). Probing above the ceiling in oral interviews : what's up there? Dans V. Kohonen, A. Huhta, A., L. Kurki-Suonio. et S. Luoma (dir.), *Current developments and alternatives in language assessment : proceedings of LTRC 96* (p. 225-38). University of Jyvaskyla and University of Tampere.

Sanderson, P. J. (2001). *Language and differentiation in examining at A level* [Thèse de doctorat non publiée]. Université de Leeds.

Saville, N. (2009). Language assessment in the management of international migration : A framework for considering the issues. *Language Assessment Quarterly, 6*(1), 17-29. https://doi.org/10.1080/15434300802606499

Savoie-Zajc, L. (2011). La recherche qualitative/interprétative en éducation. Dans T. Karsenti et L. Savoie-Zajc (dir.). *La recherche en éducation : Étapes et approches*, 3e édition (p. 123-147). ERPI.

Shohamy, E. (1983). The stability of oral proficiency assessment on the oral interview testing procedures. *Language Learning*, *33*, 527-40. https://doi.org/10.1111/j.1467-1770.1983.tb00947.x

Shohamy, E. (2007). The power of language tests, the power of the English language and the role of ELT. Dans J. Cummins et C. Davison (dir.), *International handbook of English language teaching, 11* (p. 521-532). Springer.

Spolsky, B. (1995). Measured words: The development of objective language testing. Oxford University Press.

Taylor, L. & Galaczi, E. (2011). Scoring validity. Studies in language testing 30, Examining speaking. Research and practice in assessing second language speaking. Cambridge University Press, 171-233.

Upshur, J. A. & Turner, C. E. (1999). Systematic effects in the rating of second language speaking ability: Test method and learner discourse. *Language Testing, 16(1),* 82-111. https://doi.org/10.1177/026553229901600105

Wesolowski, B. C. (2016). Exploring rater cognition: A typology of raters in the context of music performance assessment. *Psychology of Music*. https://doi.org/10.1177/0305735616665004

Winke, P., Gass, S. & Myford, C. (2011). The relationship between raters' prior language study and the evaluation of foreign language speech samples. *ETS Research Report Series, 2*, i-67. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02266.x

Winke, P., Gass, S. & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30(2)*, 231-252. https://doi.org/10.1177/0265532212456968

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*, 83-106. https://doi.org/10.1016/S1075-2935(97)80006-2

Zhang, Y. & Elder, C. (2011). Judgments of oral proficiency by non-native and native english speaking teacher raters: Competing or complementary constructs? *Language Testing, 28*(1), 31-50. https://doi.org/10.1177/0265532209360671

Annexe

Descripteurs linguistiques du CECRL pour la production orale générale

	Production orale générale
C2	Peut produire un discours élaboré, limpide et fluide, avec une structure logique efficace qui aide le destinataire à remarquer les points importants et à s'en souvenir.
C1	Peut faire une présentation ou une description d'un sujet complexe en intégrant des arguments secondaires et en développant des points particuliers pour parvenir à une conclusion appropriée.
B2	Peut méthodiquement développer une présentation ou une description en soulignant les points importants et les détails pertinents.
	Peut faire une description et une présentation détaillées sur une gamme étendue de sujets relatifs à son domaine d'intérêt en développant et justifiant les idées par des points secondaires et des exemples pertinents.
B1	Peut assez aisément mener à bien une description directe et non compliquée de sujets variés dans son domaine en la présentant comme une succession linéaire de points.
A2	Peut décrire ou présenter simplement des gens, des conditions de vie, des activités quotidiennes, ce qu'on aime ou pas, par de courtes séries d'expressions ou de phrases non articulées.
A1	Peut produire des expressions simples isolées sur les gens et les choses.

Source : Conseil de l'Europe, 2001, p. 49