Université de Montréal

Le jugement des examinateurs dans le cas de l'épreuve d'expression orale du TEF

Par

Emine Ince

Département d'administration et fondements de l'éducation

Faculté des sciences de l'éducation

Thèse présentée en vue de l'obtention du grade de grade de Philosophiae Doctor (Ph.D.) en Mesure et Évaluation en éducation

Mars 2022

Université de Montréal

Faculté des sciences de l'éducation

Cette thèse intitulée

Le jugement des examinateurs dans le cas de l'épreuve d'expression orale du TEF

Présenté par

Emine Ince

A été évaluée par un jury composé des personnes suivantes

Michel Laurier
Président-rapporteur

Micheline-Joanne Durand
Directrice de recherche

Marie Thériault
Membre du jury

Beverly Baker
Examinatrice externe

RÉSUMÉ

Au Canada, chaque province et territoire possède ses propres critères de sélection des immigrants et les exigences linguistiques à des fins d'immigration ont une incidence importante sur les demandes des candidats. Au Québec, les candidats doivent avoir recours à des certifications standardisées de français afin d'attester de leurs connaissances de la langue. Parmi ces certifications, on retrouve le Test d'évaluation de français (TEF) qui évalue les quatre habiletés de la compétence langagière : l'expression écrite, la compréhension écrite, la compréhension orale et l'expression orale. L'évaluation de l'expression orale se réalise au moyen d'une entrevue entre un candidat et un examinateur, car cela constitue un meilleur indicateur de la compétence de la personne évaluée.

Cependant, étant donné que l'entrevue implique un examinateur humain, le comportement de celui-ci peut représenter une menace possible à la fidélité du test. Malgré les nombreuses mesures prises afin de minimiser les variabilités de l'évaluation, il a été démontré que l'examinateur pouvait représenter un « effet » (biais) pouvant porter sur ses caractéristiques intrinsèques et sur ses approches en matière d'évaluation.

Nous inscrivant dans une perspective qualitative/interprétative, nous avons documenté le jugement des examinateurs du TEF lors de l'évaluation de l'épreuve d'expression orale. De manière spécifique, nous avons observé s'il existait des divergences à travers leur jugement, puis nous avons brossé le portrait de leur appropriation et de leur appréciation de la grille d'évaluation. Dix participants, examinateurs TEF, ont pris part à la collecte de données au moyen d'une activité d'évaluation de candidats via la technique de la pensée à voix haute, puis en prenant part à une entrevue semi-dirigée.

Les résultats révèlent que des divergences sont présentes. En effet, les examinateurs peuvent accorder une même note pour une même performance alors que leurs interprétations peuvent différer, et inversement. De plus, certains peuvent être influencés de façon positive dans leur jugement en raison de leur familiarité avec l'accent des candidats. D'autres peuvent faire des inférences non pertinentes pour produire 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 entraîner une conséquence négative sur la note de ce dernier.

Pour ce qui est de l'appropriation et de l'appréciation des examinateurs de la grille d'évaluation, plusieurs aspects ont été relevés. Certains descripteurs ont été interprétés différemment et ont été jugés ambigus ou vagues. De plus, les échelons B1 et B2 ont été ceux posant le plus de problèmes en raison du niveau B2 qui attribue des points aux candidats ayant un projet d'émigration au Québec. Par ailleurs, beaucoup de critiques ont été émises à l'égard d'un critère à cause des deux actions distinctes qu'il contient.

Nous concluons ainsi que les points sensibles relevés dans cette recherche mériteraient une attention particulière de la part des parties prenantes concernées par le TEF afin d'amener les examinateurs à être plus constants dans leur pratique, et ainsi d'améliorer la standardisation des procédures de passation dudit test.

Mots-clés : Tests de langue seconde, TEF, expression orale, examinateurs, divergences dans l'évaluation, grille d'évaluation, technique de la pensée à voix haute

ABSTRACT

In Canada, each province and territory has its own immigrant selection criteria, and language requirements for immigration purposes have a significant impact on applicants' applications. In Quebec, candidates must use standardized French certifications to attest to their knowledge of the language. Among these certifications, we find the *Test d'Évaluation de Français* (TEF), which assesses the four skills of language proficiency: written expression, written comprehension, oral comprehension and oral expression. The assessment of oral expression is carried out by means of an interview between a candidate and an examiner, as this is a better indicator of the competence of the person being assessed.

However, since the interview involves a human examiner, his behaviour may represent a possible threat to test reliability. Despite many steps taken to minimize variability in assessment, it has been shown that the reviewer may represent an "effect" (bias) that may affect their intrinsic characteristics and approaches to assessment.

From a qualitative/interpretive perspective, we documented the judgment of TEF examiners during the assessment of the oral expression test. Specifically, we observed whether there were any discrepancies in their judgment, then we painted the portrait of their appropriation and appreciation of the rating scale. Ten participants, all TEF examiners, took part in the data collection through a candidate assessment activity via a think-aloud protocol, followed by a semi-structured interview.

The results reveal that discrepancies are present. Indeed, the examiners can give the same score for the same performance while their interpretations can differ, and vice versa. Additionally, some may be positively influenced in their judgment due to their familiarity with the candidate's accent. Others may make irrelevant inferences to make sense of the difficulties encountered by candidates. Finally, the interviewer's attitude during the conversation with the candidate may be perceived differently and have a negative impact on the latter's score.

Regarding the appropriation and appreciation of the examiners of the rating scale, several aspects were noted. Some descriptors were interpreted differently and were found to be ambiguous or vague. In addition, levels B1 and B2 were the most challenging, because success at level B2 is

required for immigration. On the other hand, much criticism has been levelled at a criterion, because of the two separate actions it contains.

We thus conclude that the sensitive points identified in this research would merit special attention from the stakeholders concerned by the TEF to encourage examiners to be more consistent in their practice, and thus improve the standardization of test implementation procedures.

Keywords : Second language tests, TEF, speaking tests, examiners, discrepancies in assessment, rating scales, think-aloud protocol

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations et des sigles

L2 = Langue seconde

FLS = Français langue seconde

TEF = Test d'évaluation de français

IELTS = *International English Language Testing System*

CECRL = Cadre européen commun de référence pour les langues

ACTFL = *American Council on the Teaching of Foreign Languages*

Remarque liminaire

Les concepts et expressions issus de l'anglais ont fait l'objet d'une traduction libre.

Remerciements

En tout premier lieu, je tiens à témoigner à Micheline-Joanne Durand, ma directrice de recherche, ma profonde gratitude pour ses précieuses directives, ses nombreux conseils, ses commentaires judicieux, sa disponibilité, ainsi que sa très grande générosité humaine. Je n'oublierai jamais le soutien moral qu'elle m'a offert et qui m'a constamment incitée à avancer pour mener à terme cette thèse.

Mes remerciements vont également à Michel Laurier pour ses éclairages critiques et pour sa grande rigueur scientifique qui m'ont beaucoup aidée, et à Marie Thériault dont les suggestions et les rétroactions ont contribué à améliorer la qualité de mon travail. Leurs encouragements m'ont apporté une source de motivation indispensable dès mon examen de synthèse.
Je remercie également Beverly Baker qui a accepté de rejoindre le jury de cette thèse en tant qu'examinatrice externe. Ses commentaires m'ont fait réfléchir sous un angle nouveau et son parcours professionnel est une source d'inspiration pour moi.

Par ailleurs, je suis très reconnaissante envers l'équipe du Français des affaires de la Chambre de commerce et d'industrie de la région Paris Île-de-France, et plus particulièrement envers Dominique Casanova, qui m'a ouvert ses portes et qui m'a fourni de nombreuses ressources essentielles.

Enfin, je tiens à remercier les dix examinateurs TEF qui ont donné de leur temps en acceptant de participer à l'exercice. Leur précieuse collaboration constitue l'essentiel de cette recherche.

INTRODUCTION

La connaissance de la langue est une condition nécessaire à l'intégration des immigrants dans leur nouvelle société d'accueil. Au Québec, le français constitue le principal facteur d'intégration socioculturelle, il est non seulement l'instrument essentiel qui permet la participation, la communication et l'interaction avec les Québécois, mais également un symbole d'identification. La sélection des travailleurs qualifiés, qui représentent la majorité des catégories d'immigrants, est constituée de 120 points. Elle s'effectue selon une grille qui attribue des points en fonction des caractéristiques recherchées. Pour leurs connaissances en français, les candidats peuvent obtenir un maximum de 16 points : 7 points pour l'expression orale, 7 points pour la compréhension orale, 1 point pour la compréhension écrite et 1 point pour l'expression écrite. Le recours aux tests et aux diplômes standardisés de compétences linguistiques est obligatoire, et parmi ces tests et diplômes, on retrouve le Test d'évaluation de français (TEF).

L'épreuve d'expression orale du TEF prend la forme d'une entrevue entre un examinateur et un candidat. On évalue non seulement les connaissances linguistiques comme la syntaxe, le lexique et les éléments prosodiques, mais également la manière dont ces connaissances sont maîtrisées en situation de communication. Du fait de sa multidimensionnalité, il est d'usage de penser que l'expression orale est l'habileté langagière la moins tangible et la plus difficile à évaluer. De plus, la structure de l'entrevue a un risque de compromettre la fidélité d'un test en raison de sa nature imprévisible, spontanée et créative (Bachman, 1990; McNamara, 1996). Malgré les nombreuses mesures prises afin de minimiser les variations de l'évaluation de l'expression orale comme l'utilisation de grilles d'évaluation critériées, les formations, les évaluations multiples, il a été démontré que les examinateurs ne se conduisaient pas de façon homogène, et par conséquent, cela peut créer des écarts dans les notes attribuées aux candidats.

En s'inscrivant dans une perspective qualitative/interprétative, cette thèse tente de documenter le jugement des examinateurs du TEF lors de l'évaluation de l'épreuve d'expression orale. De manière plus spécifique, le but est d'observer s'il existe des divergences à travers leur jugement, puis de brosser le portrait de leur appropriation et de leur appréciation de la grille d'évaluation. Cette recherche tente d'apporter des éléments allant vers une meilleure transparence, et subséquemment, vers une plus grande fidélité et validité de l'acte d'évaluer. À notre connaissance,

aucune recherche approfondie n'a été faite sur ce sujet, alors que les enjeux associés à ce test sont très élevés. En effet, les résultats obtenus par les candidats peuvent déclencher un nouveau déroulement d'événements dans leur vie et toute erreur peut conduire à un rejet de leur dossier d'immigration ou de citoyenneté.

Dans le premier chapitre, nous présentons le contexte général de notre recherche en évoquant l'intégration linguistique des nouveaux arrivants au Québec à travers les tests de L2. Nous soulignons également la complexité de l'évaluation des épreuves d'expression orale. Nous exposons ensuite le contexte spécifique de notre recherche en mettant en avant les études empiriques menées sur les effets (biais) des examinateurs dans les tests oraux de L2. Nous voyons en outre que les formations pour réduire les variabilités chez les examinateurs donnent des résultats mitigés. Enfin, nous présentons le but de notre recherche en mettant en évidence sa pertinence scientifique et sociale.

Dans le deuxième chapitre, nous définissons les concepts-clés de notre recherche. Nous commençons par définir le concept de compétence communicative, puis traitons de la contribution des référentiels en évaluation des L2. Nous continuons en définissant les notions de fidélité, de validité et de jugement, puis nous situons l'avènement de la recherche sur la cognition de l'évaluateur. Nous ouvrons une parenthèse méthodologique en évoquant la technique de la pensée à voix haute. Par la suite, nous identifions les variables qui s'interposent dans la notation, puis nous exposons certaines caractéristiques des grilles d'évaluation. Nous présentons ensuite la grille d'évaluation actuelle de l'épreuve d'expression orale du TEF, puis finalement, nous dévoilons nos deux questions de recherche.

Dans le troisième chapitre, nous exposons la méthodologie choisie afin d'établir des pistes de réponses aux objectifs posés par cette recherche. Nous définissons tout d'abord l'approche méthodologique retenue qui est la technique de la pensée à voix haute. Nous annonçons ensuite le contexte pratique en présentant notre échantillon et en précisant les modalités de la collecte de données, le milieu de la recherche ainsi que les outils choisis. Nous résumons les points essentiels à l'aide d'un tableau de synthèse, nous défendons la position de la chercheure, puis nous évoquons les considérations éthiques.

Dans le quatrième chapitre, nous présentons les résultats obtenus permettant d'établir des pistes de réponses à nos deux questions de recherche. Tout d'abord, nous analysons les résultats

correspondant à la première question de la recherche (les évaluations réalisées par les participants) en faisant un bilan des résultats saillants, et nous présentons les résultats correspondant à la deuxième question de la recherche (les réponses apportées aux questions de l'entrevue).

Dans le cinquième chapitre, nous discutons des résultats présentés au quatrième chapitre afin de répondre à nos deux questions de recherche. Nous commentons les éléments marquants et les interprétons conceptuellement. Nous présentons d'abord un résumé des résultats concernant la première question de la recherche (les diverses formes de divergences), puis nous faisons le point sur la deuxième question de la recherche (l'appropriation et de l'appréciation de la grille d'évaluation). Pour finir, nous proposons nos suggestions.

En guise de conclusion, nous résumerons les points saillants de l'ensemble de la recherche, nous présentons ses limites ainsi que ses retombées, puis nous proposons quelques pistes de recherches ultérieures.

CHAPITRE 1: PROBLÉMATIQUE

Introduction

Le présent chapitre expose la problématique sur laquelle nous fondons notre recherche. Nous situerons d'abord le contexte général en évoquant l'intégration linguistique des nouveaux arrivants au Québec à travers les tests de L2. Les caractéristiques de ces tests, comme leur harmonisation à l'échelle mondiale et les compétences langagières évaluées, notamment à l'oral, seront présentées. Nous soulignerons également la complexité de l'évaluation des épreuves d'expression orale. Dans le contexte spécifique de notre recherche, nous illustrerons les études empiriques menées sur les effets des examinateurs, autrement dit sur les biais documentés des examinateurs, dans les tests oraux de L2. Ces études se classent en deux grands types : celles portant sur leurs caractéristiques intrinsèques, et celles portant sur leurs approches en matière d'évaluation. Nous verrons en outre que les formations pour réduire les variabilités chez les examinateurs donnent des résultats mitigés. Enfin, nous présenterons le but de notre recherche en mettant en évidence sa pertinence scientifique et sociale.

1.1. Le contexte général : l'intégration linguistique des nouveaux arrivants au Québec

Généralement, choisir de résider de façon permanente en dehors de son pays natal nécessite de maitriser la langue du pays d'accueil. La question de l'intégration linguistique des migrants est un enjeu des politiques publiques dans de nombreux pays d'Europe de l'Ouest tout comme dans les pays d'immigration du Nouveau Monde. Qu'il s'agisse pour les migrants d'entrer sur le territoire, d'obtenir l'autorisation de résidence permanente ou d'en acquérir la nationalité, différentes politiques de formation linguistique sanctionnées par des tests ont été mises en place. La connaissance de la société d'accueil et de sa langue est donc une condition nécessaire à l'intégration des immigrants dans leur pays d'arrivée (Extramiana et Van Avermaet, 2010; McNamara et Shohamy, 2008).

Au Canada, les exigences linguistiques à des fins d'immigration ont une incidence importante sur les demandes des candidats et déterminent l'accès au pays. Pour pouvoir présenter une demande de séjour permanent ou une demande de citoyenneté, une majorité de candidats doivent démontrer un niveau de compétence linguistique acceptable en anglais et/ou en français (basé sur les

référentiels *Canadian Language Benchmarks* et Niveaux de compétence linguistique canadiens[1]). Chaque province et territoire possède ses propres programmes d'immigration et ses propres critères de sélection des immigrants.

1.1.1. Le contexte linguistique au Québec

Au Québec, l'apprentissage du français pour un immigrant vient appuyer le développement de son sentiment d'appartenance à la communauté québécoise. Pour les membres de la société d'accueil, le partage d'une langue commune, le français, avec les immigrants facilite l'ouverture à l'altérité (Pagé, 2011; Québec MCCI, 1990).

L'identité de la population québécoise repose non seulement sur des assises territoriales et culturelles, mais aussi linguistiques. À partir du début des années 1960, avec la Révolution Tranquille et la victoire du Parti libéral, le Québec entre dans la voie de la modernité sur les plans sociaux, politiques et économiques. Cette époque marque le début d'une volonté déterminée de maintenir l'identité francophone québécoise par la mise en place de règles visant à protéger l'emploi de la langue française (Barrats et Moisei, 2004). En 1974, la Loi sur la langue officielle, aussi appelée Loi 22, définit la première véritable politique linguistique du Québec. Trois ans plus tard, en 1977, cette loi sera modifiée et la Charte de la langue française, communément appelée Loi 101, sera promulguée. Cette loi fait du français la seule langue officielle du Québec, la langue habituelle du travail, de l'enseignement, des communications, du commerce et des affaires. La Loi 101 reconnait toutefois deux langues nationales : le français et l'anglais, et assure des droits linguistiques aux Anglophones (Durand, 2002). Parallèlement à ces changements sociaux, le Québec voit son taux de natalité et de fécondité chuter. Afin de maintenir son poids démographique dans le Canada, la province signe plusieurs accords relatifs à l'immigration avec le Gouvernement fédéral, comme l'Accord Cullen-Couture en 1978, et l'Accord Canada-Québec en 1991. Ces accords donnent les pleins pouvoirs au Québec pour déterminer ses objectifs relativement à la composition de son immigration et au nombre de personnes qu'il souhaite admettre, tout en tenant compte de sa capacité d'accueil.

[1] *Canadian Language Benchmarks* est une échelle descriptive des compétences linguistiques en anglais L2 sous la forme de 12 points de référence échelonnés sur un continuum allant d'un niveau de base à un niveau avancé. Les Niveaux de compétence linguistique canadiens sont la version traduite en français pour le FLS.

1.1.2. L'immigration au Québec

Au fil des années, la province du Québec devient une terre d'accueil et le nombre d'immigrants augmente considérablement. Le profil des immigrants se diversifie et rend le portrait du Québec beaucoup plus cosmopolite. Avant le dernier quart du XX^{e} siècle, les immigrants provenaient principalement d'Europe, et depuis, ils proviennent de tous les continents : Asie (Chine, Inde), Afrique (notamment francophone : Algérie, Maroc), Amérique Centrale (Haïti).

Selon la publication de 2019 de l'Institut de la statistique du Québec[2], la province compte 8 390 499 habitants en 2018, et 52 388 immigrants ont été admis en 2017. Parmi ces immigrants, 42 % connaissent le français. Les continents et pays principaux de naissance des immigrants sont présentés dans le tableau 1.

Tableau 1 - Nombre d'immigrants en 2017

Continents et pays principaux de naissance des immigrants	Nombres d'immigrants
Asie	22 750
▪ Chine	5 108
Afrique	14 405
▪ Algérie	2 437
Europe	8 261
▪ France	4 505
Amérique	6 868
▪ Haïti	1 931
Océanie et autres pays	104

Le Québec classe ses immigrants selon une catégorisation mise en place par le Canada en 1967. Cette catégorisation se divise en trois grandes parties : la catégorie de l'immigration économique qui se compose de travailleurs qualifiés et de gens d'affaires, la catégorie du regroupement familial, et la catégorie des réfugiés. Dans la catégorie de l'immigration économique, les travailleurs qualifiés se destinent à une activité économique (occuper un emploi, gérer une entreprise), et les gens d'affaires sont des investisseurs qui promeuvent l'expansion industrielle et la création d'emplois (Gouvernement du Canada, 2018).

[2] Source : https://www.stat.gouv.qc.ca/quebec-chiffre-main/pdf/qcm2019_fr.pdf

Au Québec, entre 2011 et 2015, ces catégories se répartissaient ainsi par rapport au nombre total des immigrants : les travailleurs qualifiés : 57,6%, le regroupement familial : 21,1%, les réfugiés : 10,2%, les gens d'affaires : 8,5%, autres : 1,2%.

C'est essentiellement la catégorie de l'immigration économique (travailleurs qualifiés et gens d'affaires) qui permet le mieux au Québec d'orienter la réponse à ses besoins. Les candidats de cette catégorie sont sélectionnés en fonction de caractéristiques qui favorisent leur insertion en emploi ou leur aptitude à réaliser un projet d'affaires. Les candidats de la catégorie du regroupement familial ainsi que les candidats à qui la qualité de réfugié a été reconnue alors qu'ils se trouvaient au Québec sont exemptés de cette sélection.

1.1.3. La grille de sélection des travailleurs qualifiés

La sélection des candidats de la catégorie de l'immigration économique s'effectue selon une grille qui attribue des points en fonction des caractéristiques recherchées. Celles-ci sont non discriminatoires au regard de la race, de la couleur, de l'origine ethnique ou nationale, de la religion, du sexe ou de l'orientation sexuelle.

La grille de sélection des travailleurs qualifiés, qui représentent la majorité des catégories d'immigrants, est constituée de 120 points. Le seuil de passage exigé est de 50 points sans conjoint, et de 59 points avec conjoint. Cette grille permet d'obtenir un certificat de sélection du Québec (CSQ) qui est la première étape de la demande de résidence permanente avant qu'elle puisse être traitée au niveau fédéral. Les caractéristiques recherchées de la grille de sélection sont, par ordre décroissant, les suivantes (Tableau 2).

Tableau 2 - Grille de sélection du Programme régulier des travailleurs qualifiés, 2018

Critères	Points maximums
Formation (niveau de scolarité et domaine de formation)	26
Connaissances linguistiques (français et anglais)	22
Caractéristiques de l'époux ou du conjoint accompagnateur (niveau de scolarité, domaine de formation, âge, connaissance linguistique)	17
Âge	16
Détention d'une offre d'emploi validée	14
Expérience professionnelle	8
Enfants	8

Séjours et famille (conjoint, père, mère, frère, sœur, fils, fille, grand-père, grand-mère)	8
Capacité d'autonomie financière	1

Source : https://cdn-contenu.quebec.ca/cdncontenu/immigration/publications/GR_Selection_Travailleurs_Qualifies.pdf?1616677921

Pour leurs connaissances linguistiques, les candidats peuvent obtenir un maximum de 22 points, soit 16 points maximum pour leurs connaissances en français et 6 points maximum pour leurs connaissances en anglais. Pour le français, les 16 points sont répartis de manière suivante : 7 points pour l'expression orale, 7 points pour la compréhension orale, 1 point pour la compréhension écrite et 1 point pour l'expression écrite. Pour l'anglais, les 6 points sont répartis ainsi : 2 points pour l'expression orale, 2 points pour la compréhension orale, 1 point pour la compréhension écrite et 1 point pour l'expression écrite.

Cette pondération correspond aux besoins du marché du travail au Québec. En effet, le gouvernement du Québec valorise davantage la connaissance du français des personnes immigrantes au moment de leur sélection, et plus particulièrement l'expression et la compréhension orales. Selon le Gouvernement du Québec (2000), cela constitue un gage d'une plus grande probabilité de rétention et permet de favoriser une intégration plus rapide au marché du travail.

Avant 2013, afin d'attester de leurs connaissances linguistiques en français et/ou en anglais, les candidats avaient deux possibilités. La première était de présenter une attestation de résultats d'un test standardisé reconnus par le Ministère, et les points dans la grille de sélection étaient octroyés en fonction des résultats. La deuxième possibilité était de présenter des preuves satisfaisantes de leurs connaissances linguistiques, telles qu'un diplôme sanctionnant des études récentes ou des documents attestant une expérience de travail en français et/ou en anglais. Depuis 2013, à la suite d'une modification dans la politique de sélection des immigrants, tous les candidats, quelles que soient leur langue maternelle et leur nationalité, sont dans l'obligation de démontrer qu'ils ont une connaissance du français au moins d'un stade intermédiaire avancé ou une connaissance de l'anglais au moins d'un stade intermédiaire afin d'obtenir des points à la grille de sélection. Le niveau intermédiaire avancé pour le français représente le niveau B2 du Cadre européen commun de références pour les langues (CECRL), et le niveau intermédiaire représente le niveau B1 pour l'anglais. Ces niveaux sont le seuil minimal à partir duquel des points sont attribués. En prenant

comme exemple le Test d'évaluation de français adapté au Québec (TEFaQ), nous pouvons avoir un aperçu détaillé des points pouvant être obtenus avec le barème suivant (Tableau 3).

Tableau 3 - Barème des points du TEFaQ selon les niveaux du CECRL

Épreuves	Niveaux de A1 à B1	Niveau B2	Niveau C1	Niveau C2
Expression orale	0 point	5 points	6 points	7 points
Compréhension orale	0 point	5 points	6 points	7 points
Expression écrite	0 point	1 point	1 point	1 point
Compréhension écrite	0 point	1 point	1 point	1 point

Source : www.lefrancaisdesaffaires.fr/tests-diplomes/test-evaluation-francais-tef/tef-quebec-tefaq

1.1.4. Les certifications en français langue seconde

Dans l'objectif d'uniformiser l'évaluation des compétences linguistiques des candidats à l'immigration au moyen d'une approche universelle, équitable et homogène, le Ministère a décidé de rendre obligatoire le recours aux tests[3] et aux diplômes standardisés des compétences linguistiques. Ces tests et diplômes standardisés en français sont les suivants :

- le Test d'évaluation de français (TEF) de la Chambre de commerce et d'industrie de Paris Île-de-France (CCIP-IDF), crée en 1998;

- le Test d'évaluation de français adapté pour le Québec (TEFaQ) de la CCIP-IDF, crée en 2007;

- le Test d'évaluation de français Naturalisation (TEF Naturalisation), de la CCIP-IDF, crée en 2012;

- le Test d'évaluation de français pour le Canada (TEF Canada) de la CCIP-IDF, crée en 2014;

- le Test de connaissance du français (TCF) de France Éducation international (FEI), crée en 2002;

- le Test de connaissance du français pour le Québec (TCFQ) de FEI, crée en 2006;

- le Diplôme d'études en langue française (DELF) de FEI, crée en 1985;

- le Diplôme approfondi de langue française (DALF) de FEI, crée en 1985.

[3] Nous définissions un test comme une « situation standardisée servant de stimulus à un comportement qui est évaluée par comparaison avec celui d'individus placés dans la même situation, afin de classer le sujet, soit quantitativement, soit typologiquement » (Pichot cité par De Landsheere, 1992, p. 295)

Le TEFaQ, le TEF Canada et le TEF Naturalisation sont des variantes du TEF, et il en est de même avec le TCFQ qui est une variante du TCF. Le TEFaQ et le TCFQ ont été adaptés spécifiquement pour les demandes d'immigration au Québec, les sujets et les items sont orientés sur la culture québécoise. Les autres tests et diplômes ont une reconnaissance plus générale, car ils sont utilisés pour émigrer, travailler ou étudier au Canada, en France ou en Suisse. Pour l'anglais, le seul test d'évaluation reconnu par le Ministère est l'*International English Language Testing System* (IELTS)[4].

Au niveau mondial, le test d'anglais IELTS est proposé dans 300 centres, et les 7 tests et diplômes en français sont proposés par environ 900 centres reconnus par le ministère de l'Immigration, de la Francisation et de l'Intégration (nommé avant septembre 2019 le ministère de l'Immigration, de la Diversité et de l'Inclusion) (ministère de l'Immigration, de la Diversité et de l'Inclusion, 2018). Rappelons que les tests situent les niveaux de compétence et que les diplômes les certifient. Cela signifie que les tests délivrent des attestations qui permettent de placer les candidats sur un continuum. Ils ne revêtent pas de caractère d'échec ou de réussite, contrairement aux diplômes qui attestent des compétences effectives par tâches accomplies et par seuils successifs. Par ailleurs, les tests ont un temps de validité limité alors que les diplômes n'ont pas de limite de validité (Anquetil et Jamet, 2010).

1.1.5. L'harmonisation des certifications en français langue seconde

Les tests et diplômes reconnus pour l'immigration au Québec sont l'unique outil d'évaluation des connaissances linguistiques des candidats, ils sont de reconnaissance nationale et internationale et se partagent le marché des langues. Afin de s'assurer que l'interprétation des niveaux des candidats est identique, quelle que soit la langue cible, le Cadre européen commun de référence pour les langues (appelé plus souvent par son acronyme CECRL ou CECR) joue un rôle primordial, car il permet la comparabilité internationale des résultats de l'évaluation et donne une base pour la reconnaissance mutuelle des qualifications en langues. Le CECRL communique au-delà des langues et dépasse les frontières nationales. Par souci de cohérence et de transparence, il a été conçu comme une norme afin de faciliter la communication et le travail en réseau des différentes

[4] L'IELTS est géré conjointement par Cambridge ESOL (*English for Speakers Of another Language*), le *British Council* et l'*IDP Education Australia (International Development Program of Australian universities and colleges).*

parties prenantes dans l'évaluation des langues secondes[5] (Conseil de l'Europe, 2001). Dans le contexte des tests de L2 standardisés à visée certificative, les différentes parties prenantes regroupent les organismes concepteurs et certificateurs des tests et diplômes, les candidats, les administrateurs du test (centres d'examen, examinateurs) et les utilisateurs finaux (bureaux d'immigration, ambassades) (Georges, 2013) (Figure 1).

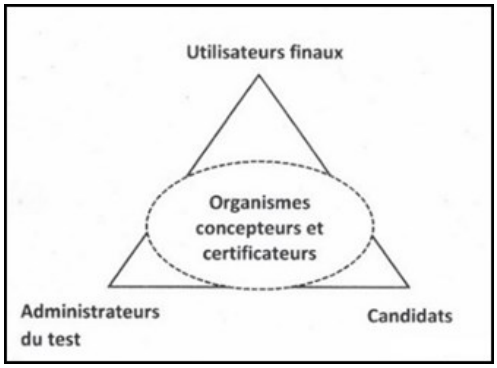


Figure 1 - Les parties prenantes en lien avec les organismes concepteurs et certificateurs de tests
Source : Georges, 2013 (Schéma adapté)

Les organismes concepteurs et certificateurs des tests et diplômes alignent leurs échelles d'évaluation sur les niveaux du CECRL, ou pour le moins, fournissent une grille de conversion. Au Québec, le ministère de l'Immigration et des Communautés culturelles (2011) met en correspondance le CECRL avec les trois principaux référentiels en évaluation des langues secondes existants en Amérique du Nord : les Niveaux de compétence linguistique canadiens (NCLC), (qui est la version traduite en français des *Canadian Language Benchmarks* (CLB)), l'Échelle québécoise des niveaux de compétence en français pour les personnes immigrantes adultes, et le *Proficiency Guidelines* de l'*American Council on the Teaching of Foreign Languages* (ACTFL).

1.1.6. Les compétences langagières évaluées

Les tests et diplômes reconnus pour l'immigration au Québec sont d'un enjeu très élevé, c'est-à-dire qu'ils ont des conséquences majeures pour les candidats (Shohamy *et al.*, 1996). Selon Messick (1981, 1994, 1996), les tests et diplômes à enjeu élevé n'existent pas de manière isolée,

[5] Dans le système scolaire et socio-politique canadien, nombreuses sont les publications où la lexie « langue seconde » recouvre indistinctement les problèmes de langue étrangère et de langue seconde (Cuq et Davin-Chnane, 2007).

mais sont liés à des variables psychologiques, sociales et politiques ayant des effets sur les programmes, l'éthique, les classes sociales, la bureaucratie, la politique, la connaissance, l'inclusion et l'exclusion. En effet, comme le soulignent Noël-Jothy et Sampsonis, (2006) et Shohamy (2005), ils vont au-delà de leur visée pédagogique, car ils sont avant tout considérés comme des instruments sociaux et politiques ayant un impact considérable sur l'éducation et l'immigration et pouvant déterminer un ordre social. Par exemple, dans la société contemporaine, ils sont utilisés pour contrôler les flux d'immigration, pour déterminer les droits de résidence et de citoyenneté, et pour règlementer l'accès aux établissements d'enseignement et l'accès au travail (McNamara, 2010). La finalité des tests et diplômes est la délivrance d'une certification reconnue socialement et professionnellement qui marque l'aboutissement d'un apprentissage en sanctionnant un niveau global de compétences langagières. L'évaluation des tests et diplômes reconnus pour l'immigration au Québec est réalisée en dehors de tout cadre d'apprentissage, elle n'est donc pas formative, car « on ne voit plus le sujet-objet évalué dans l'optique d'une construction de savoir ou d'acquisition de compétences. L'objectif didactique s'efface devant l'objectif pragmatique. L'acte d'évaluer devient acte de valider » (Ljalikova, 2004, p. 4). Par conséquent, l'évaluation est sommative et elle se termine par l'attribution d'une certification qui permet de justifier la nature de l'apprentissage (Roegiers, 2004).

Le rôle des tests et diplômes consiste à évaluer les compétences langagières, c'est-à-dire les compétences réceptives, productives et d'interaction, à travers quatre épreuves : expression orale, compréhension orale, expression écrite, compréhension orale (Noël-Jothy et Sampsonis, 2006). Cette catégorisation est basée sur le modèle des linguistes Lado (1962) et Carroll (1961, 1983) qui définit quatre habiletés de la compétence langagière. L'enseignement et l'évaluation des langues secondes s'articulent généralement selon ce découpage (Peters et Bélair, 2011). En prenant l'exemple du Test d'évaluation de français (TEF), les épreuves, les tâches demandées et les formats se présentent tels qu'illustrés dans le tableau ci-dessous (Tableau 4).

Tableau 4 - Épreuves, tâches et formats du TEF

Épreuves	Tâches	Formats
Expression orale	- Demander des informations pour un produit ou un service au téléphone - Convaincre un ami de faire une activité	Jeux de rôle avec un examinateur
Compréhension orale	- Associer des illustrations à des messages oraux - Comprendre des messages audio courts et longs - Reconnaitre différents sons	Questionnaire à choix multiples
Expression écrite	- Imaginer la fin d'un article de presse insolite - Exprimer son point de vue et le justifier en réponse à une affirmation	Papier crayon/électronique
Compréhension écrite	- Comprendre divers articles de presse - Trouver l'ordre d'un récit	Questionnaire à choix multiples

Source : Chambre de commerce et d'industrie de région Paris Île-de-France, 2018

Pour l'épreuve d'expression orale du TEF, on évalue non seulement les connaissances linguistiques comme la syntaxe, le lexique et les éléments prosodiques (prononciation, débit, etc.), mais également la manière dont ces connaissances sont maîtrisées en situation de communication. Ce que l'on teste reflète ainsi une vision de l'usage de la langue telle qu'elle peut être utilisée quotidiennement, les tâches à réaliser ne requièrent pas de connaissances culturelles particulières, ni de compétences académiques. Le candidat doit être capable de décrire, informer, raconter, développer un point de vue, argumenter, convaincre, et nuancer. En pratique, il doit pouvoir réserver une chambre d'hôtel, demander des renseignements en français, mais également échanger des expériences et des opinions afin d'avoir une relation sociale efficace avec les personnes avec lesquelles il est en contact. Le candidat doit donc être autonome dans sa pratique de la langue et être en mesure de se débrouiller face à des situations inattendues. Cette autonomie doit lui permettre d'entrer en relation avec les personnes pour des raisons personnelles, professionnelles ou académiques (Artus et Demeuse, 2007; Riba et Mavel, 2005; Tagliante et Mègre, 2008).

De ce fait, les compétences langagières orales ne se réduisent pas à la seule composante linguistique (syntaxe, lexique, prosodie), mais incluent de façon plus large le développement d'une composante sociolinguistique ainsi que d'une composante pragmatique. La composante sociolinguistique est le fait de savoir adapter la langue à l'interlocuteur. La composante pragmatique désigne la relation entre les caractéristiques d'une langue, ce que peut en faire le

locuteur et ce qu'il en fait réellement, en fonction du contexte d'usage dans lequel il est placé (Bachman, 1990).

1.1.7. La multi-dimensionnalité de la langue orale

Partant de ce constat, la langue orale s'avère être multidimensionnelle et se distingue par ailleurs de la langue écrite. La langue orale et la langue écrite constituent des modes de communication différents qui font appel à des aptitudes différentes. On peut revenir sur un texte, mais il est difficile de revenir sur une communication orale, même si elle est enregistrée sur vidéo, une partie du contexte étant absente de l'enregistrement. Selon Tochon (1997), la langue orale est fugace, discontinue, entrecoupée d'hésitations et de reprises. Elle est redondante et le mot a plus d'importance que la structure syntaxique. Le vocabulaire est également différent, plus emphatique, hétérogène et néologique. Van Den Heuvel (1985) ajoute que la langue orale se caractérise par une production sonore, c'est-à-dire par un volume, un débit, une intonation et bien d'autres nuances. Elle est aussi accompagnée « de la mimique expressive et du geste du sujet parlant qui font appel au visuel pour corriger ou préciser les nuances » (p. 47).

De manière plus analytique, l'oral se compose d'un fond (un contenu, un sens, des idées, des réflexions, des sentiments, etc.) et d'une forme qui porte le sens expressif (Rispail, 2005). La forme s'organise en fonction de plusieurs composantes (Aubert-Gea, 2005) : les composantes exécutives (habiletés cognitives qui visent à réaliser des raisonnements, des analyses des informations, des prises des décisions, etc.), les composantes linguistiques (habiletés d'appliquer les règles du code de la langue), les composantes référentielles (connaissances et savoir-faire concernant les échanges interpersonnels), les composantes socioculturelles (connaissances et représentations ethno-socioculturelles qui sont partagées par les membres d'une communauté linguistique donnée) (Boyer, 2003), et les composantes discursives (réalisations énonciatives qui sont mises en œuvre dans les situations de communications réelles, et qui reflètent l'appropriation des différents types de discours et de leur organisation en fonction des paramètres de la situation de communication dans laquelle ils sont produits et interprétés) (Moirand, 1982).

Au fil des recherches en linguistique, la capacité à utiliser la langue s'est définie en tant que compétence communicative ou compétence langagière, et a été modélisée à de nombreuses reprises par des linguistes comme Jakobson (1963), Chomsky (1965), Hymes (1972) (1984), Halliday (1976), Widdowson (1978), Canale et Swain (1980), Kramsch (1986), Bachman (1990),

Dolz (2002) et Fulcher (2003). Grâce aux travaux des premiers chercheurs et à leurs remises en question, la didactique des langues secondes a donné naissance, dès les années 1970, à l'approche communicative, une approche basée sur le principe de la compétence communicative. L'évaluation de l'expression orale a ainsi évolué au même rythme que les connaissances en linguistique et en didactique des langues secondes et s'est constamment adaptée aux nouvelles théories (Fahardy, 1983; Fulcher, 2000). Les multiples modélisations de la compétence communicative ont successivement décomposé la langue en de nombreuses composantes si interreliées que celles-ci sont devenues complexes à isoler dans le discours (Bachman, 1990; Carroll, 1983; Fulcher, 2003; Germain et Netten, 2002; Orr, 2002; Walter, 2004). Il est alors devenu difficile de mettre en place des modalités d'évaluation en expression orale, car les chercheurs ont constamment été amenés à revoir les critères de validité et de fidélité sous l'éclairage fourni par les définitions de la compétence communicative. La validité s'est longtemps définie comme la capacité d'une évaluation à mesurer réellement ce qu'elle est censée mesurer. À la suite d'une évolution du concept, elle est considérée de nos jours comme propriété de l'ensemble du processus d'évaluation, de la conceptualisation des besoins des usagers, et des conséquences de l'utilisation des évaluations (Bachman, 2005; Kane, 2013; Newton et Shaw, 2014). La fidélité, quant à elle, concerne la constance avec laquelle un test mesure ce qu'il est censé mesurer, quels que soient l'évaluateur, le moment, l'endroit et les conditions de la passation du test (Bachman, 1990; Bélair, 2007).

1.1.8. Les limites de l'évaluation de l'expression orale

Dans le cadre scolaire, il est d'usage d'affirmer que les habiletés langagières orales sont plus difficiles à administrer et à évaluer que les habiletés langagières écrites (Brown et Abeywickrama, 2010; Courtillon, 2003; De Pietro et Wirthner, 1996; Dolz et Schneuwly, 2009; Luoma, 2004; Peters et Bélair, 2011). À ce titre, Garcia-Debanc (1999) relève les nombreux obstacles à l'évaluation des habiletés langagières orales, parmi lesquels on retrouve les éléments suivants :

1) L'oral est difficile à observer et complexe à analyser : les aspects qui interviennent dans un énoncé sont nombreux (éléments linguistiques et communicatifs) et simultanés;

2) L'oral implique l'ensemble de la personne : on ne peut faire abstraction de la personne dans son ensemble quand il est question de production verbale. Voix, visage, corps sont impliqués;

3) L'oral est profondément marqué par les pratiques sociales de référence : les variables culturelles et sociales jouent un rôle important dans l'appréciation d'une production orale;

4) L'oral est volatil et ne laisse pas de trace. Dès lors, l'analyse sérieuse et l'évaluation des productions orales nécessitent des enregistrements techniquement exigeants, cela suppose donc un détour par l'écrit par le biais de transcriptions. En effet, pour analyser les aspects langagiers syntaxiques et sémantiques, l'enseignant doit avoir recours à des transcriptions. Celles-ci peuvent prendre différentes formes en fonction des enjeux poursuivis, depuis les transcriptions orthographiques jusqu'aux transcriptions phonétiques.

Dans les situations de tests de L2 standardisés à visée certificative, il est également d'usage de penser que l'expression orale est l'habileté langagière la moins tangible et la plus sujette à des variations, ce qui la rend difficile à évaluer de façon fiable (Alderson et Bachman, 2004; Luoma, 2004; O'Sullivan, 2012; Spolsky, 1995; Vidakovic et Galaczi, 2013). Logistiquement, évaluer l'oral exige aussi beaucoup de temps, car les candidats sont généralement jugés individuellement ou par pairs lors d'interactions en tête à tête avec un ou deux examinateurs[6]. Cela rend alors les tests oraux fastidieux, surtout lorsque l'on doit évaluer en masse (Fulcher, 2003; Isaacs, 2016; Ludenberg, 1929). Par ailleurs, on peut rencontrer des problèmes d'ordre pratique liés la sécurité des données et au coût, comme les salaires versés aux examinateurs (surtout lorsqu'ils travaillent en binôme) et au personnel organisateur (Malone et Montee, 2010; Taylor, 2007).

1.1.9. Les limites de l'entrevue orale

Par souci de validité, l'expression orale se doit de se manifester à travers une performance, car cela constitue un meilleur indicateur de la compétence de la personne évaluée (Kane, 2006). Dans la littérature, on cite souvent le terme de « test de performance » (*performance test*) qui se définit par un test dans lequel les habiletés du candidat à accomplir des tâches, habituellement associées aux exigences du travail ou des études, sont évaluées. Les tâches peuvent être par exemple : adopter un rôle particulier dans un jeu de rôle, décrire une photographie ou présenter des arguments à un groupe (Davis *et al.*, 1999). Dans les tests de L2, la configuration la plus courante est celle de l'entrevue individuelle en face à face entre un intervieweur-évaluateur et un candidat, aussi appelée

[6] Dans la littérature anglophone, les termes *rater* et *judge* sont utilisés. Pour les besoins de notre texte, nous utiliserons la distinction proposée par Chénier (2018) : le terme « examinateur » sera retenu pour des situations formelles de test, et le terme « évaluateur » pour des contextes plus larges, notamment pour des contextes de classe.

en anglais *Oral Proficiency Interview* (OPI)[7] (Bolton, 1987; Fulcher, 2000; Malone, 2003; McNamara, 1996; O'Loughlin, 2001; Orr, 2002; Shohamy, 1982; Vidakovic et Galaczi, 2013; Walter, 2004).

L'entrevue entre un intervieweur-examinateur et un candidat met en place une situation qui duplique de très près un contexte de la vie réelle puisqu'elle consiste en une discussion relativement non prévisible sur des sujets généraux. Même si le candidat peut prédire de manière générale les divers sujets abordés ou le rôle qu'il aura à jouer dans les jeux de rôle, il n'est pas en mesure de prédire les questions que l'on va lui poser ni la tournure que va prendre l'interaction (Brown, 2005). L'entrevue possède donc un certain degré de validité, car elle permet de donner une indication précise sur le niveau de compétence orale du candidat dans une conversation proche de l'authenticité (Brown, 2005; Clark, 1979; He et Young, 1998; Ross, 1996; van Lier, 1989; Weir, 1993). Toutefois, selon certains chercheurs, la validité de l'entrevue orale en face à face a des limites, car elle établit un rapport de pouvoir inégal entre l'intervieweur et l'interviewé. De plus, les sujets de conversation sont imposés et le système de tour de parole est fortement contrôlé. Il est donc difficile d'évaluer les compétences pragmatiques (comme le savoir-agir, le savoir-être dans la société) des candidats, car l'établissement d'un contexte qui ressemble au monde réel n'est souvent pas possible, même dans les jeux de rôle les plus sophistiqués où l'on recrée artificiellement une situation de la vie de tous les jours (Lazaraton, 2002; Liskin-Gasparro, 2003; Johnson, 2001; Johnson et Tyler, 1998; van Lier, 1989; Young, 1995; Young et He, 1998; Young et Milanovic, 1992). Ainsi, pour toutes ces raisons, la validité de l'entrevue orale a été largement débattue pendant près d'un demi-siècle (Fulcher, 2003).

Outre ce problème de validité, la structure de l'entrevue peut compromettre la fidélité d'un test en raison de sa nature imprévisible, spontanée et créative (Bachman, 1990). McNamara (1996, p. 3) affirme que « la richesse du contexte de l'entrevue orale englobe une complexité et une variabilité potentielle colossales qui peuvent facilement menacer l'équité et la généralisabilité des conclusions que l'on peut tirer des candidats ». Étant donné que l'évaluation d'une entrevue orale implique un examinateur humain et que les résultats ne sont pas objectivement vérifiables, le

[7] L'OPI est une entrevue structurée qui conduit le candidat à travers des activités qui requièrent progressivement des niveaux élevés de compétence. Elle a été conçue à la base dans les années 1940 par le *Foreign Service Institute*, une institution de formation du gouvernement américain pour les employés de la communauté des affaires étrangères. Elle a été modifiée dans les années 1970 par l'*American Council on the Teaching of Foreign Language*, une organisation américaine dédiée à l'enseignement et à l'apprentissage des langues secondes.

comportement de celui-ci peut représenter une menace possible à la fidélité d'un test (Bachman *et al.*, 1995; Douglas, 1994; Lumley et McNamara, 1997; Upshur et Turner, 1999). Malgré les nombreuses mesures prises afin de minimiser les variabilités de l'évaluation comme l'utilisation de grilles d'évaluation critériées, les formations, les évaluations multiples, il a été démontré que des écarts systématiques sont introduits dans les notes et que des difficultés à arriver à un consensus entre les examinateurs demeurent. Les examinateurs ne se conduisent pas tous de façon homogène, et par conséquent, cela crée des écarts dans les notes attribuées aux candidats.

1.2. Le contexte spécifique : les recherches empiriques menées sur les effets des examinateurs

À partir des années 1980, une multitude d'études ont traité du comportement des examinateurs en L2 en fonction de facteurs tels que leur langue maternelle, leur formation ou leur utilisation de la grille d'évaluation[8]. Ces études, appliquées aussi bien dans le domaine de l'évaluation de l'expression orale que dans celui de l'expression écrite, ont fait valoir que le jugement est plus susceptible d'être biaisé par les caractéristiques de l'examinateur que par les caractéristiques du candidat (Bachman, 2000; McNamara, 2000). Pour Wolfe et McVay (2012), les caractéristiques des examinateurs représentent donc une source importante de biais qui se définissent comme des « modèles d'évaluation qui contiennent des erreurs de mesure et qui peuvent donc poser des problèmes de validité dans les scores assignés par des humains » (p. 32). Étant donné que ces biais sont souvent involontaires, certains chercheurs ont, dès les années 60, choisi d'utiliser une appellation plus neutre qui est « effet de l'examinateur » (*rater effect*) (Norman et Goldberg, 1966; Myford et Wolfe, 2003; Wesolowski, 2016). Cette appellation est la plus courante et c'est la raison pour laquelle nous l'utiliserons dans notre texte.

1.2.1. Les effets des examinateurs

Les études les plus couramment réalisées au sujet des effets des examinateurs portent sur les différences de sévérité ou de clémence (Chénier, 2018; Myford *et al.*, 1996; Myford et Wolfe, 2000, 2003; Patz *et al.*, 2002; Wolfe, 1997; Wolfe et McVay, 2012). Ces différences représentent

[8] Nous utiliserons, de façon générale, le terme « grille d'évaluation » tout au long de notre texte bien qu'il existe d'autres termes qui spécifient le type d'échelle comme « échelle descriptive » ou qui combinent les deux comme la grille d'évaluation descriptive appelée communément « grille descriptive ».

le fait de donner, en moyenne, des notes plus élevées ou plus basses que d'autres examinateurs pour une même performance (Eckes, 2005; Myford et Wolfe, 2003). Les examinateurs peuvent donc fluctuer dans leur profil de sévérité, et ces fluctuations peuvent se produire soit de manière éparse, soit de manière plus systématique en fonction de facteurs liés au dispositif d'évaluation (sévérité différentielle selon la tâche traitée ou les critères d'évaluation) ou au référentiel d'évaluation (sévérité variable selon les niveaux de performance) (Casanova et Demeuse, 2016; McNamara, 1993, 1996).

Dans le domaine des tests d'expression orale en L2, les études empiriques sur les effets des examinateurs peuvent être classées en deux grands types : celles qui analysent leurs caractéristiques intrinsèques, et celles qui analysent leurs approches en matière d'évaluation. Les études sur les caractéristiques intrinsèques des examinateurs portent principalement sur le fait qu'ils soient locuteurs natifs ou non natifs de la langue qu'ils évaluent, sur leur familiarité avec les accents des candidats, et sur leur manière d'interagir avec les candidats. Quant aux études sur leurs approches en matière d'évaluation, elles portent sur leur rapport avec les critères d'évaluation, sur le fait de citer des critères extérieurs à la grille d'évaluation, sur les multiples inférences faites, et sur la différence entre leur raisonnement évaluatif et la note qu'ils attribuent.

Les études se rapportant aux approches évaluatives des examinateurs ont recueilli les commentaires de ces derniers grâce à la méthode de la pensée à voix haute *(Think aloud protocol)*, elles sont principalement descriptives et peu axées sur une analyse profonde de la cognition au sens psychologique du terme. La cognition des évaluateurs étant définie par Davis (2012) comme « les processus mentaux se produisant lors de la notation, à un niveau conscient ou inconscient » (p. 9).

Dans l'ensemble, la recherche sur la cognition des évaluateurs en évaluation de l'oral en L2 n'a pas encore été fondée sur une base théorique solide du traitement de l'information humaine pouvant fournir une compréhension approfondie de ce qui se passe dans l'esprit des évaluateurs lors de la notation (Dehn, 2008; Han, 2016; Purpura, 2013). Des modèles de cognition d'évaluateurs ont été explorés dans l'évaluation de l'écriture en langue première et seconde (Barkaoui, 2010, 2007; Cumming *et al.*, 1983; Kantor et Powers, 2002), mais ces modèles ne sont pas transposables à l'oral, car selon Davis (2012, p. 69) « le jugement de la performance de la langue écrite et de la langue orale représente des défis cognitifs et pratiques très différents ».

Les études empiriques recensées sur les effets des examinateurs sont présentées dans les lignes qui suivent.

1.2.1.1. Les examinateurs locuteurs natifs et locuteurs non natifs

Dans le monde de l'enseignement et de l'évaluation de la L2, la question de la langue maternelle s'avère pertinente, car la compétence des locuteurs natifs représente la norme. Le matériel pédagogique est généralement élaboré à partir de conversations et de textes de locuteurs natifs, et selon le référentiel *American Council on the Teaching of Foreign Languages* (ACTFL), pour atteindre un niveau intermédiaire, un apprenant doit être capable de se faire comprendre par un locuteur natif (*American Council on the Teaching of Foreign Languages*, 2012; Zhang et Elder, 2011).

Des études ont pris en compte cet aspect et 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. Elles sont très peu nombreuses à avoir été réalisées dans un contexte de test d'expression orale en L2, nous les avons donc regroupées avec celles menées dans un contexte de classe de L2.

Les conclusions de la grande majorité des études sont similaires. D'une part, elles démontrent que les examinateurs locuteurs natifs sont légèrement plus sévères dans les notes, mais que dans l'ensemble, les différences ne sont pas significatives entre les deux groupes en termes de sévérité. D'autre part, des orientations différentes dans l'interprétation des résultats des deux groupes apparaissent. Dans la majorité des études recensées (Brown, 1995; Gui, 2012; Kim, 2009; Zhang et Elder, 2011), on a constaté des commentaires plus riches, plus élaborés et plus précis avec une gamme plus large de termes employés chez les examinateurs natifs que chez leurs homologues non natifs. Les natifs ont été plus exigeants en termes d'exactitude prosodique et ont davantage relevé les traits phoniques du discours oral comme la fluidité et surtout la prononciation.

Chez les examinateurs non natifs, les résultats ont montré des interprétations différentes. De manière générale, on a observé une rétroaction plus pauvre et moins élaborée. Leurs commentaires ont davantage porté sur la grammaire et le vocabulaire, sur la communication non verbale (gestes et autres comportements), sur la qualité globale des performances des personnes évaluées, et sur la compréhensibilité et l'intelligibilité de leur discours oral. Selon Brown (1995), les examinateurs

non natifs sont moins aventureux dans leur évaluation et plus contraints par les critères fournis par la grille d'évaluation, alors que les examinateurs natifs interprètent les descripteurs de la grille d'évaluation de manière plus nuancée. Par ailleurs, une autre étude moins récente de Fayer et Krasinski (1987) a révélé des résultats à contre-courant de ceux mentionnés précédemment en démontrant des différences marquées en termes de sévérité entre les deux groupes d'examinateurs. Dans cette étude, les non natifs se sont montrés plus sévères, notamment envers la prononciation et les hésitations.

Malgré ces conclusions, Johnson et Lim (2009) ainsi que Kim (2009) tiennent à souligner que les examinateurs locuteurs non natifs ont leur légitimité. Ils ne se situent pas dans une catégorie isolée, étant donné qu'il n'existe pas de niveau minimum de compétence linguistique requis pour qu'ils soient considérés comme leurs homologues natifs. De plus, d'après Lazaraton (2005), les définitions proposées par les chercheurs sur l'identité du locuteur natif restent problématiques, car des facteurs psychologiques, sociologiques et politiques entrent en jeu. D'ailleurs, Zhang et Elder (2011) précisent que les différences relevées dans les études citées ne sont pas uniquement la conséquence de la langue maternelle des examinateurs, mais peuvent également s'expliquer par des différences sociales et culturelles.

1.2.1.2. La familiarité des examinateurs avec l'accent des candidats

Un autre type de recherche a tenté de déterminer s'il y avait une différence de sévérité lorsque les examinateurs sont familiers ou non avec l'accent étranger ou la langue maternelle du candidat. En linguistique et en sciences cognitives, l'idée que les accents[9] familiers de langue étrangère sont plus faciles à comprendre que les accents inconnus est soutenue par quelques auteurs comme Flowerdew (1994), Major, Fitzmaurice, Bunta et Balasubramanian (2002). De plus, la perception que les auditeurs attribuent à certaines caractéristiques de la prononciation change avec l'expérience linguistique, c'est-à-dire que plus nous sommes en contact avec une langue étrangère, plus nous devenons familiers avec celle-ci (Nittrouer *et al.*, 1993; Zhang *et al.*, 2005).

Afin d'analyser cette question, des études ont comparé deux groupes d'examinateurs : ceux étant familiers et ceux n'étant pas familiers avec l'accent des candidats. Des différences significatives

[9] Ici, nous ne décrivons pas l'accent selon les définitions habituelles des linguistes, mais comme l'ensemble des caractéristiques de prononciation liées aux origines linguistiques ou territoriales du locuteur qui permettent d'identifier sa provenance (Harmegnies, 1997).

entre les deux groupes se sont manifestées, les résultats ont été convergents et ont montré que la familiarité de l'examinateur avec l'accent du candidat a des répercussions positives sur le score de la prononciation et de l'accent. Cela s'explique par le fait que la familiarité facilite la compréhension et l'identification de la langue en question (Carey *et al.*, 2011; Huang *et al.*, 2016; Huang et Jun, 2014; Hsieh, 2011; Winke, Gass et Myford, 2011, 2012).

Par ailleurs, la familiarité est susceptible de variation en raison de l'exposition des examinateurs aux différents accents des candidats. Dans l'étude de Carey, Mannell et Dunn (2011), les résultats montrent que plus l'exposition aux accents est prolongée, plus la note de la prononciation est élevée, et inversement. C'est ce que l'on constate également dans l'étude de Huang et Jun (2014), où les examinateurs expérimentés se sont montrés plus cléments que leurs collègues débutants dans l'évaluation de la phonologie des candidats.

Toutefois, une étude de Wei et Llosa (2015) a révélé une exception à ce constat. Ces derniers n'ont trouvé aucune différence statistiquement significative entre les deux groupes d'examinateurs. Dans cette étude portant sur un test d'anglais L2 aux États-Unis, les examinateurs étaient indiens et américains et les candidats indiens. Une analyse qualitative approfondie a révélé que selon un examinateur indien, adopter un accent américain standard était considéré comme très important pour vivre aux États-Unis.

Il est important de signaler dans ces études que cette variabilité n'a pas d'effet notoire sur la compétence globale des candidats, et que les effets observés ont un impact limité dans la notation globale (Huang *et al.*, 2016). Selon Carey, Mannell et Dunn (2011) et Xi et Mollaun (2009), même si cette familiarité avec l'accent du candidat n'affecte pas beaucoup les résultats, elle ne doit pas être ignorée et doit être traitée pendant la formation des examinateurs, car elle peut compromettre l'équité, l'exactitude et l'interprétation d'une évaluation.

1.2.1.3. L'interaction entre les examinateurs et les candidats

La manière dont l'examinateur interagit avec les candidats peut également constituer une source de variabilité. Selon leur propre style interactionnel, c'est-à-dire leur 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 affecter la performance des candidats lors des tests oraux de L2 (Cafarella, 1994; Filipi, 1994; Lazaraton, 1996a, 1996b; Lazaraton et Saville,

1994; Morton *et al.*, 1997). À l'aide d'analyses conversationnelles (analyses approfondies de conversations des interlocuteurs en interaction), les chercheurs ont identifié deux grands types d'examinateurs-intervieweurs : ceux dits « faciles » et ceux dits « difficiles ». Les « faciles » ont par exemple tendance à poser des questions plus simples, à parler lentement, à surarticuler, à donner des rétroactions positives, à monter des signes d'intérêt, à être polis et souriants, à corriger les erreurs linguistiques des candidats, et à leur « mettre les mots dans la bouche ». Concernant plus spécifiquement les techniques de questions, ils sont plus enclins à poser des questions sur des informations déjà fournies par le candidat, à poser des questions ouvertes en fournissant immédiatement une ou plusieurs options de réponse, et à poser des questions d'amorce fermées juste avant de poser des questions ouvertes afin de mettre en contexte une nouvelle information et de provoquer explicitement les réponses attendues.

Les « difficiles », quant à eux, ont tendance à trop ou pas assez parler, à poser des questions plus complexes (comme des justifications d'opinion) ou moins explicites, à rire de la réponse du candidat, à insister sur un sujet de conversation alors que le candidat rencontre de la difficulté, à répéter la question de la même façon qu'elle a été posée au lieu d'avoir une stratégie de reformulation lorsque le candidat ne comprend pas. Leur discours est également empreint de plus d'interruptions et de désaccords. Globalement, ils sont moins encourageants et plutôt distants, voire « robotiques » (Brown, 2003; Brown et Hill, 1998; Lazaraton, 1996a, 1996b; Morton *et al.*, 1992; O'Sullivan et Lu, 2006; Wigglesworth et Williams, 1997).

En ce qui concerne les fréquences des divers accommodements, celles-ci ont un lien avec les complications rencontrées dans la fluidité du dialogue de l'entrevue. Les études de Lorenzo-Dus et Meara (2005), Ross (1992) et Ross et Berwick (1992, 1900) démontrent que plus le candidat présente un niveau faible, ou plus l'intervieweur rencontre des problèmes dans la discussion, plus ce dernier ajuste progressivement son niveau de langue au niveau perçu du candidat en ayant recours aux accommodements. Les candidats d'un niveau relativement élevé sont donc moins susceptibles d'être assistés.

Certains chercheurs sont d'avis que les différents types de soutien apportés au candidat jouent sur sa note finale (Bachman, 1988, 1990; Brown, 2003; Lazaraton, 1996a, 1996b; McNamara, 1996; Stansfield et Kenyon, 1992). Il a d'ailleurs été démontré dans trois études menées par Brown

(2003), Brown (2005) et Reed et Halleck (1997)[10] que ceux qui interagissent avec des intervieweurs « faciles » ont en majorité des notes plus élevées que ceux qui interagissent avec des intervieweurs « difficiles ». Néanmoins, il a été démontré le contraire dans deux études menées par McNamara et Lumley (1997) et Morton, Wigglesworth et Williams (1997). Les auteurs donnent deux explications possibles à ces résultats qui peuvent paraitre contre-intuitifs. Premièrement, ils expliquent cela par le fait que les examinateurs, par souci d'équité, ont eu tendance à offrir une compensation aux candidats parce qu'ils ont perçu l'entrevue comme étant « pauvre ». Deuxièmement, comme les examinateurs ont vu que les intervieweurs étaient « embourbés » dans l'interaction et qu'ils accordaient trop peu de temps de parole au candidat, les examinateurs ont attribué « le bénéfice du doute » aux candidats faute d'éléments de preuve de leur compétence réelle.

De manière générale, bien que les accommodements tendent vers un climat propice à une communication efficace, ils ne représentent pas toujours le modèle à suivre dans une situation de test, bien au contraire, ils risquent d'être contreproductifs. En effet, selon Lazaraton (1996b) et Ross et Berwick (1992), trop accommoder un candidat ne lui donne pas l'opportunité d'exprimer son plein potentiel, et son manque de compétence peut parfois être dissimulé. En outre, les chercheurs remarquent que les intervieweurs trop accommodants tendent à jouer leur rôle d'enseignant. Cela remet en question la validité de l'entrevue orale, car les interactions devraient refléter le plus possible des conversations de la vie quotidienne. Cette attitude de calque (enseignant-intervieweur) est appropriée dans un contexte pédagogique, mais ne devrait pas avoir lieu dans un contexte de test de langue orale.

1.2.1.4. Le rapport entre les examinateurs et les critères d'évaluation

Des études ont montré que les examinateurs ont des sensibilités très variables par rapport aux différents critères de la grille d'évaluation et qu'ils accordent une importance très différente à chaque critère. Orr (2002), dans son étude, constate que les examinateurs ont des compréhensions très variées d'un même critère, par exemple, pour le critère « gestion du discours », certains l'interprètent comme étant une « habileté à produire un discours long » et d'autres ont des interprétations très variées (toutefois, l'étude ne donne pas de détails sur ces interprétations

[10] Dans les contextes où les examinateurs ne sont pas les intervieweurs et qu'ils évaluent les candidats en différé à l'aide d'enregistrements vidéo ou audio.

variées). D'autres études ont montré que selon les niveaux des candidats, les critères auxquels les examinateurs prêtent le plus d'attention ne sont pas les mêmes, par exemple, les critères linguistiques (grammaire, vocabulaire, prononciation, débit) sont plus mis en avant pour les candidats ayant les niveaux les plus faibles, et les critères communicatifs sont plus mis en avant pour ceux ayant un niveau plus élevé (pertinence des arguments, efficacité de la communication) (Brown, 2000; Brown *et al.*, 2005; Hamilton *et al.* 1993; McNamara, 1996; Meiron, 1998; Pollitt et Murray, 1996).

En ce qui concerne le critère du vocabulaire plus spécifiquement, une étude menée par Brown, Iwashita, et McNamara (2005), montre que certains examinateurs ne s'entendent pas. Certains considèrent que la réutilisation du vocabulaire déjà employé par eux-mêmes dans les réponses des candidats est une indication d'un niveau de compétence élevé, alors que d'autres estiment que les candidats de compétence élevée devraient pouvoir paraphraser ou trouver des alternatives au vocabulaire déjà employé. Cette constatation fait écho à une étude de Brown (2000), qui a constaté que certains examinateurs accordent davantage d'importance à l'utilisation de l'autocorrection, de stratégies de clarification et de circonlocution (le fait d'exprimer une idée spécifique avec plusieurs mots plutôt que de l'évoquer directement avec peu de mots) par les candidats.

Pour Bachman (1990), dans une situation de test où l'on a recours à un entretien oral, la subjectivité demeure toujours présente, car les examinateurs doivent juger de l'exactitude de la réponse des candidats à travers leur « interprétation subjective des critères de notation » (p. 76). Plus la norme et leur propre compréhension de la norme sont éloignées, et plus il y a de l'inconstance dans la façon dont les critères sont appliqués (Bachman, 1990; Meiron et Schick, 2000; Reed et Cohen, 2001). Brown (1993) souligne que malgré le caractère explicite des descripteurs d'une grille d'évaluation et malgré la normalisation mise en place au cours des sessions de formation, les différences ne peuvent pas s'éliminer. Les examinateurs ont une perception intime de ce qui leur est acceptable et ces perceptions sont formées, dans une certaine mesure, par leur expérience antérieure.

1.2.1.5. Les critères extérieurs à la grille d'évaluation

D'autres études démontrent que les examinateurs ne tiennent pas toujours compte de la grille d'évaluation et citent des éléments extérieurs tout en évaluant les candidats. Parmi ces éléments, on retrouve par exemple des descriptions du comportement des candidats (Brown, 2006; Orr,

2002), des critiques de la grille d'évaluation, des comparaisons entre candidats, des extraits de leurs énoncés, des commentaires sur leur âge, leur genre, leurs efforts, leur langage corporel, leur contact visuel, leur degré de préparation au test (Orr, 2002), la première impression qu'ils ont eue des candidats, leur niveau de confiance (May, 2006), leur niveau de maturité, leur motivation, leur réticence à dialoguer (Pollitt et Murray, 1996), leur utilisation créative de la langue (Brown, 2006; Meiron, 1998), leur sens de l'humour (May, 2006; Meiron, 1998), leur capacité à faire face à différentes exigences dans la tâche à accomplir, leur confiance dans l'utilisation de la langue, leur voix, leur engagement dans la discussion (Ang-Aw et Goh, 2011; Brown, 2006), ainsi que leur capacité à s'en sortir linguistiquement dans des contextes variés de la vie quotidienne (Brown, 2000, 2006; Orr, 2002). Des conduites interactionnelles de la part des candidats ont également été relevées comme le fait de mettre à défi l'intervieweur, le fait d'éviter certaines questions, de faire des apartés, d'utiliser des stratégies de communication telles que l'autocorrection, de demander des clarifications, d'utiliser des périphrases, de pouvoir gérer une conversation, et d'élargir les sujets de conversation (Brown, 2000).

Dans les situations où l'examinateur n'est pas l'intervieweur[11], il a été observé par ailleurs que les examinateurs tiennent compte du comportement de l'intervieweur dans leur évaluation, notamment de leur empathie (Orr, 2002), de leur encouragement à l'égard des candidats (Pollitt et Murray, 1996), du degré de difficulté de leurs questions, du nombre de questions fermées qu'ils posent, de leur attitude de dénigrement, du caractère inapproprié des sujets qu'ils abordent (sujets délicats, ennuyeux), de leurs interruptions du discours, du temps qu'ils accordent aux candidats pour répondre, et de leurs difficultés à comprendre les arguments des candidats (Brown, 2000).

Par conséquent, pour certains chercheurs, ces attitudes montrent que les examinateurs semblent ne pas bien saisir le construit sur lequel les échelles d'évaluation sont basées (Brown, 2000, Orr, 2002, Wigglesworth, 1994). Orr (2002) relativise en ajoutant que ces commentaires périphériques ne sont pas accidentels, mais font partie intégrante du processus de pensées des examinateurs. Se concentrer sur une tâche d'évaluation tout en remarquant des éléments hors de propos est un reflet de l'activité mentale.

[11] Des situations où l'examinateur n'interagit pas avec le candidat, où l'évaluation se fait à distance à l'aide d'enregistrements audio ou vidéo de la conversation entre l'intervieweur et le candidat.

1.2.1.6. Les multiples inférences

Certaines études montrent que les examinateurs ont très souvent tendance à faire des inférences[12] qui sont typiquement utilisées pour excuser ou expliquer certains modèles de comportement des candidats ou pour justifier l'attribution de certaines notes. Par exemple, dans les études menées par Brown (2000) (2006), les examinateurs évoquent les problèmes prosodiques (débit, pause, hésitation) des candidats comme étant liés à leur nervosité, à leur timidité, à leur manque d'intérêt pour le sujet de discussion, ou simplement au fait de chercher leurs mots ou de réfléchir au contenu de leurs réponses afin de mieux structurer leurs idées. Quant aux problèmes de grammaire et de lexique, et au manque de complexité des idées des candidats, les examinateurs expliquent cela par leur manque de culture générale, par l'immaturité de leurs idées et de leur personnalité, et par leur juvénilité apparente. Dans une étude de Pollitt et Murray (1996), beaucoup de commentaires d'examinateurs font par exemple référence au manque apparent d'intelligence des candidats, à leur maturité d'esprit, ou à leur volonté ou leur réticence à dialoguer.

Les examinateurs se retrouvent souvent incertains face aux vraies causes des problèmes rencontrées par les candidats et ont de la difficulté à se décider pour émettre leurs jugements. Ces inférences représentent un problème majeur, car elles ne forment pas une base adéquate pour la formulation d'un jugement (Brown, 2000, 2006; Orr, 2002). Les inférences peuvent être vues comme des « dangers », car elles creusent l'écart entre l'information à l'état brut, c'est-à-dire l'information telle qu'elle a été recueillie, et sa traduction par l'examinateur (Roegiers, 2004).

1.2.1.7. Le raisonnement évaluatif et la note

En s'appuyant sur le constat que les perceptions des examinateurs sont multiples, des études ont révélé que leur raisonnement évaluatif et la note attribuée pouvaient différer. En effet, 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 manière similaire et attribuer des notes différentes. Par exemple, dans l'étude de Orr (2002), deux examinateurs donnent des notes avec un écart assez grand à un même candidat et émettent chacun des commentaires assez analogues qui sont les suivants : « Dans l'ensemble, sa prononciation est A1 » (niveau grand débutant selon

[12] L'inférence est le processus à travers lequel l'évaluateur vise à produire du sens, c'est-à-dire à donner une signification aux informations qu'il recueille (DeKetele et Roegiers, 2009).

l'échelle du CECRL) et « Je pensais qu'elle allait réussir, mais rien ne s'est produit pour que je change d'avis » (p. 148). On observe également des notes similaires avec des commentaires contradictoires comme : « La manière dont il pose les questions est artificielle » et « On aurait pu penser qu'il était l'intervieweur parce qu'il était si naturel » (p. 149). Dans une autre étude de Ang-Aw et Goh (2011), pour une même note, on retrouve de la part d'un examinateur des commentaires tels que « Réponses personnelles simples », puis de la part de son collègue « Très bonnes réponses personnelles » (p. 38).

Des notes quantitativement identiques n'excluent pas des différences qualitatives dans la prise de décision de l'examinateur ou dans l'interprétation du construit (Douglas, 1994; Douglas et Selinker, 1992, 1993). Étant donné que la langue orale est un phénomène multidimensionnel, les interprétations du discours par les interlocuteurs varient en fonction des différents aspects sur lesquels ils se penchent (Douglas, 1994). McNamara (1996) et Orr (2002) déclarent qu'il faut donc rester sceptique quant à l'interprétation et à la signification des résultats des tests par souci d'équité vis-à-vis des utilisateurs de tests (candidats, employeurs potentiels, personnel des admissions universitaires).

1.2.2. La formation des examinateurs

Beaucoup de chercheurs affirment qu'une façon d'assurer une certaine stabilité chez les examinateurs dans les tests oraux de L2 est de leur offrir régulièrement des formations (Barrett, 2001; Brown, 1995; Elder *et al.*, 2005; Fahim et Bijani, 2011; Fulcher, 2003; Lumley, 2002; Lumley et McNamara, 1995; McNamara, 1993, 2000; Weigle, 1998, 1999; Wigglesworth, 1993). Pourtant, des études empiriques montrent qu'elles ne suffisent pas à garantir une cohérence d'ensemble.

Selon certains chercheurs, les formations permettent avant tout d'homogénéiser les pratiques en atténuant les observations aberrantes, c'est-à-dire qu'elles réduisent les différences extrêmes chez les examinateurs étant excessivement sévères ou cléments (Lumley et McNamara, 1995; McIntyre, 1993). Mais selon d'autres chercheurs, l'efficacité des formations profite moins aux examinateurs expérimentés qu'aux examinateurs nouvellement formés. À mesure que les examinateurs acquièrent de l'expérience, ils deviennent plus précis et les accords interjuges se révèlent être plus stables et on observe alors une plus grande constance intra et inter-examinateurs (Attali, 2016; Davis, 2016; Furneax et Rignall, 2007; Knoch, 2011; Lim, 2011; O'Sullivan et Rignall, 2007;

Shaw, 2002; Shohamy *et al.*, 1992; Weigle, 1998). Toutefois, avoir acquis plus d'expérience ne signifie pas nécessairement être un meilleur évaluateur (Myford *et al.*,1996).

D'autres études sur les effets des formations montrent que des différences significatives et substantielles entre les examinateurs peuvent demeurer (Eckes, 2011; Lumley et McNamara, 1995; McNamara, 1993, 1996). Certains continuent à être plus cléments, d'autres plus sévères, et on observe par exemple des variations persistantes dans l'utilisation des grilles d'évaluation tout comme dans le raisonnement évaluatif (May, 2006; Meiron, 1998; Orr, 2002; Papajohn, 2002). Même avec des directives standardisées spécifiques, beaucoup d'examinateurs continuent à appliquer les critères selon leurs interprétations idiosyncrasiques de la norme (Taguchi, 2011; Wolfe et Chiu,1997). Lunz et Stahl (1990) affirment d'ailleurs que « les examinateurs ont souvent l'impression d'avoir des normes uniques, et il leur est difficile de modifier leurs normes » (p. 428).

Les formations ne peuvent donc pas supprimer complètement les variations chez les examinateurs. Celles-ci sont liées à l'importance qu'ils accordent à certaines caractéristiques plutôt qu'à d'autres, et à de différentes perceptions de l'objet à évaluer. De manière plus globale, les variations sont liées à plusieurs facteurs comme leur parcours de formation, leur milieu d'origine, leur expérience professionnelle antérieure, leur expérience personnelle, leur degré de sévérité, et également leurs croyances personnelles (Brown, 1995; Brown et Ahn, 2011; Chalhoub-Deville, 1995; Davis, 2016; Orr, 2002; Tajeddin et Alemi, 2014). Tous ces éléments sont autant de raisons qui provoquent une baisse de la fidélité inter-examinateurs dans l'évaluation de l'expression orale.

1.3. La pertinence de la recherche

Nous allons maintenant présenter notre recherche en situant tout d'abord le contexte, à savoir l'évaluation de l'épreuve d'expression orale du TEF. Nous mettrons en évidence sa grille d'évaluation qui représentera un outil clé, puis nous expliciterons notre but. La pertinence scientifique ainsi que la pertinence sociale de notre recherche seront ensuite développées.

1.3.1. Le contexte du TEF

La présentation du test

Le TEF (Test d'évaluation de français) est un test de référence internationale qui mesure le niveau de connaissances et de compétences en langue française. Il a été créé en 1998 par la Direction des relations internationales de l'enseignement de la Chambre de commerce et d'industrie Paris Île-de-France (CCIP). Le Français des affaires, un établissement de la CCIP, administre le TEF ainsi que d'autres certifications et assurent des formations professionnelles en français à visée professionnelle.

Au Canada, le ministère de l'Immigration des Réfugiés et de la Citoyenneté a adopté le TEF en 2002. La version pour le Québec TEFaQ (Test d'évaluation de français adapté au Québec) a été créée en 2007, puis la version canadienne TEF Canada en 2014. Dans le monde, le TEF et ses déclinaisons sont diffusés grâce à un réseau de 308 centres d'examen agréés, c'est-à-dire des établissements officiellement reconnus par la Chambre de commerce et d'industrie de la région Paris Île-de-France pour organiser la passation du TEF. Au Canada, on en dénombre 23, dont 10 dans la province du Québec.

La présentation de l'épreuve d'expression orale

L'épreuve d'expression orale du TEF est composée de deux jeux de rôle entre le candidat et l'examinateur-animateur, et est structurée en deux sections (section A et section B) qui proposent chacune des tâches authentiques de situations de communication à partir de sujets proposés au candidat (un sujet par section). Pour le candidat, l'objectif de la section A étant d'obtenir des informations à propos d'un produit ou d'un service en simulant un appel téléphonique, et les objectifs de la section B de présenter le contenu d'un document proposant une activité à un ami et d'argumenter pour le convaincre de faire l'activité.

Dans la section A, la situation est formelle, l'examinateur-animateur répond aux questions du candidat, mais de façon incomplète ou ambigüe pour amener ce dernier à demander des clarifications. Il peut toutefois apporter de courtes suggestions afin de relancer l'échange si le candidat est à court d'idées. Dans la section B, la situation est informelle, l'examinateur-animateur joue le rôle d'un ami réticent à faire l'activité proposée par le candidat, il apporte des contre arguments solides mais réfutables afin d'enrichir le débat.

Pour les deux sections, l'examinateur-animateur adapte ses interventions en fonction du niveau des candidats et des pratiques locales afin de conserver une certaine authenticité. Il doit éviter de monopoliser la parole, et doit orienter l'entrevue pour que les objectifs communicatifs puissent être évalués. Pour le guider durant l'entretien, des pistes de relances et de contre arguments lui sont fournies.

L'épreuve est organisée dans une salle d'un centre agréé et la durée totale de la séance par candidat est de 25 minutes, dont 15 minutes en présence du candidat (5 minutes pour la section A et 10 minutes pour la section B) (Demeuse et Artus, 2008). Avant chaque session d'évaluation de l'épreuve d'expression orale, les centres d'examen TEF organisent une réunion d'harmonisation en présence des examinateurs du jour, ceux-ci préparent ensemble les sujets et échangent sur les bonnes pratiques et leurs expériences.

La procédure d'évaluation exige qu'un même candidat soit évalué par deux examinateurs : un en contexte et un autre hors contexte afin d'accroitre la fidélité inter-juges. Dans la situation en contexte, l'examinateur A (qui joue également le rôle de l'animateur) est face au candidat, et dans la situation hors contexte, l'examinateur B écoute en différé l'enregistrement audio de la conversation entre l'examinateur A et le candidat. Lorsque l'on retrouve des écarts trop importants de notes sur la grille d'évaluation entre l'examinateur A et l'examinateur B, un troisième examinateur intervient afin de réaliser un arbitrage. Ce troisième examinateur est un responsable pédagogique de l'équipe du Français des affaires.

La formation suivie par les examinateurs du TEF

Une première formation en présentiel et en ligne est imposée aux nouveaux examinateurs avec un test de fin de formation. Par la suite, la même formation en ligne doit se faire de façon continue, elle est obligatoire et est imposée à l'initiative des centres d'examen TEF. Les examinateurs peuvent également s'autoformer à tout moment afin de revenir sur les modules qu'ils souhaitent revoir.

La formation en ligne est organisée en différents modules et contient des activités de mise en application. La durée totale estimée est entre 6 heures et 10 heures. Les différents modules de la formation abordent les points suivants :

- la révision des concepts-clés du CECRL pour l'évaluation de performances à l'oral;

- la présentation historique du TEF;

- la présentation de l'épreuve d'expression orale et de son déroulement (objectifs d'évaluation, grille d'évaluation utilisée, procédures permettant d'assurer des conditions de passation standardisées);

- les techniques d'animation sur la nature des réponses apportées, des relances et des arguments; les recommandations face à des cas spécifiques de candidats (francophones, faiblement scolarisés, excessivement stressés, ayant appris leur discours par cœur);

- les techniques d'évaluation : la démarche générale préconisée, les échelles du CECRL axées sur l'animation et l'évaluation de l'épreuve d'expression orale, les critères parasites de l'évaluation, l'appropriation de la grille d'évaluation;

- les entraînements à l'évaluation permettant aux examinateurs d'évaluer des candidats sur la base d'enregistrements vidéo d'épreuves authentiques passées dans des centres agréés de différents pays. Pour chaque candidat, un corrigé (une rétroaction commentée) est fourni.

La formation est validée grâce à un test de fin de formation. Le contenu du test contient trois parties : un questionnaire à choix multiples sur les objectifs d'évaluation de l'épreuve d'expression orale du TEF, un questionnaire à choix multiples sur les techniques d'animation de l'épreuve, et une évaluation de deux candidats avec la grille d'évaluation du TEF. La durée totale est d'une heure et le seuil de réussite est de 66/100. Si le participant obtient un score inférieur à 66/100, il peut solliciter une nouvelle tentative un mois après sa première tentative.

Des formations obligatoires peuvent être imposées aux examinateurs à la demande de la Chambre de commerce et d'industrie à la suite d'anomalies récurrentes constatées. Les anomalies peuvent découler par exemple de techniques d'animation, et dans cette situation, les retours sont principalement faits par les examinateurs hors contexte (ceux évaluant à l'aide d'enregistrements audio des conversations) à l'égard des examinateurs en contexte (ceux jouant le rôle de l'interlocuteur dans les interactions avec les candidats). Par ailleurs, les responsables pédagogiques peuvent identifier des anomalies résultant de grands écarts de notes, c'est-à-dire lorsque les notes d'un même examinateur dévient fortement de celles de ses collègues de façon répétitive.

Concernant les exigences de qualification des examinateurs, il est obligatoire qu'ils soient en activité régulière en tant qu'enseignant ou professionnel de FLS auprès d'un public adulte et qu'ils aient une expérience d'au moins trois ans dans le domaine. Ils doivent ainsi posséder de bonnes compétences didactiques, pédagogiques et interculturelles. De plus, il est également à préciser que le travail d'examinateur ne représente généralement pas une activité principale, mais plutôt une activité complémentaire.

1.3.2. Les fondements de la grille d'évaluation du TEF

En octobre 2018, deux changements ont été effectués dans la procédure de l'évaluation de l'épreuve d'expression orale. Ces changements sont apparus à la suite du lancement du TEF Naturalisation[13], une variante du TEF utilisé dans le cadre des démarches de naturalisation (d'acquisition de la nationalité) auprès de l'État français et suisse. Premièrement, la disposition du jury, constitué de deux examinateurs, a été modifiée. Deux examinateurs étaient auparavant face au candidat, et depuis cette date, l'un est face au candidat et l'autre évalue à distance à l'aide des enregistrements des conversations. Deuxièmement, la grille d'évaluation a été révisée puis mise à jour. Cette modification est née d'une volonté de réduire la charge cognitive des examinateurs qui désormais devaient opérer de façon individuelle, puis d'une volonté d'accroître l'indépendance des critères, car il a été constaté que les critères étaient trop nombreux et que certains se chevauchaient.

Nous présentons ici l'ancienne version, puis la nouvelle version sera exposée à la fin du chapitre suivant. La présentation de l'ancienne version de la grille nous permet d'expliquer son fondement, et notamment sa mise en correspondance avec le CECRL. Nous reviendrons ensuite sur les changements qui ont été effectués dans la nouvelle version et l'utiliserons comme support dans notre recherche empirique.

L'ancienne version de la grille d'évaluation contient 12 critères et 7 échelons (Tableau 5). Les 12 critères se répartissent ainsi : 6 critères pour les compétences communicatives (pour la section A : 1. pertinence du questionnement, 2. prise d'initiative et gestion de l'imprévu, 3. qualité des échanges; pour la section B : 4. présentation des faits, 5. qualité de l'argumentation, 6. qualité des échanges), et 6 critères pour les compétences linguistiques (pour la syntaxe : 7. complexité des

[13] Le TEF Naturalisation est entré en service en janvier 2012, il est reconnu par le ministère de l'Intérieur français et le Secrétariat d'État aux Migrations suisse.

structures, 8. cohésion du discours; pour le lexique : 9. étendue, 10. maitrise; pour l'élocution : 11. prononciation et intonation, 12. rythme et débit). Les compétences linguistiques s'appuient sur la connaissance et la capacité à utiliser la langue transversalement sur l'ensemble de la performance (Demeuse et Artus, 2008).

Tableau 5 - Ancienne version de la grille d'évaluation de l'épreuve d'expression orale du TEF

Compétences communicatives		Elémentaire			Intermédiaire		Supérieur	
Sections	Critères	0+	1	2	3	4	5	6
Section A	Pertinence du questionnement 1	Le questionnement est inapproprié ou inexistant.	Le questionnement est élémentaire et incomplet.	Le questionnement est simple sans demande de précisions.	Le questionnement est satisfaisant ; quelques demandes de précisions.	Le questionnement explore l'ensemble de la tâche.	Le questionnement est complet et précis.	Le questionnement est exhaustif et pertinent.
	Prise d'initiative et gestion de l'imprévu 2	Pas ou peu d'initiatives malgré le soutien de l'examinateur.	Est souvent sollicité pour engager les échanges et réagir aux interventions.	La prise de parole est préparée ; fait reformuler les réponses données.	Fait clarifier les informations ambiguës ; souvent troublé par des réponses inattendues.	Intervient spontanément pour demander des précisions ; réactions appropriées mais pas toujours immédiates.	Intervient spontanément et, face à l'imprévu, réagit avec justesse.	Mène avec aisance et efficacité la conversation.
	Qualité des échanges 3	Ne comprend pas les interventions de l'examinateur.	De nombreuses ruptures rendent l'échange difficile.	Les échanges sont brefs et non suivis.	Questions et réponses s'enchaînent dans les situations prévisibles.	Les échanges sont suivis même dans les situations imprévues.	Les échanges donnent lieu à une discussion soutenue.	Les échanges sont naturels et la discussion animée.
Section B	Présentation des faits 4	Les faits sont peu ou pas présentés.	L'exposé est confus ; simple lecture du document.	L'exposé est bref ; les informations présentées sont paraphrasées.	L'exposé est simple mais clair ; effort de reformulation des informations.	L'exposé est développé et structuré ; les informations sont clairement reformulées.	L'exposé est clair et complet ; les informations sont présentées avec assurance.	L'exposé est clair, précis et suscite l'intérêt ; les informations sont présentées de façon originale.
	Qualité de l'argumentation 5	Le discours est décousu ; aucune intention de convaincre.	Le discours est confus ; peu d'arguments présentés.	Le discours est contradictoire et les arguments présentés ne sont pas tous pertinents.	Le discours est cohérent; les arguments sont justes mais peu développés.	Le discours est clair ; les arguments sont pertinents et illustrés.	Le discours est bien mené ; les arguments sont convaincants et bien développés.	Le discours est juste et convaincant ; les arguments sont nuancés et appuyés.
	Qualité des échanges 6	Ne comprend pas les interventions de l'examinateur.	L'échange est difficile ; réagit à quelques sollicitations.	Les échanges sont directs et brefs ; ne réagit pas toujours aux contre-arguments.	Intervient régulièrement pour justifier son point de vue.	Soutient son point de vue mais peut être troublé par des contre-arguments.	Réagit et défend ses idées avec justesse.	Participe activement à la discussion et défend avec habileté son point de vue.

Compétences linguistiques			Elémentaire			Intermédiaire		Supérieur	
Critères			0+	1	2	3	4	5	6
Sections A et B	Syntaxe	Complexité des structures **7**	Absence de syntaxe ou phrases très incomplètes.	Les phrases sont élémentaires et répétitives.	Les phrases sont simples et stéréotypées.	Le discours est organisé avec des phrases simples et quelques phrases complexes ; les erreurs sont fréquentes.	Les phrases simples sont utilisées correctement ; erreurs dans les structures complexes.	Les structures sont variées et adéquates ; quelques erreurs sur les phrases les plus complexes.	Les structures complexes sont parfaitement maîtrisées et utilisées avec justesse.
			○ ○	○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○	○ ○ ○	○ ○
		Cohésion du discours **8**	Aucun enchaînement ; les mots sont juxtaposés.	Les phrases sont juxtaposées ; quelques connecteurs élémentaires sont utilisés.	Les connecteurs sont simples et utilisés de manière répétitive.	Les connecteurs courants (temporels et logiques) sont utilisés correctement.	Les connecteurs et les articulateurs sont variés mais pas toujours adéquats.	Les enchaînements syntaxiques sont variés et les liens logiques sont appropriés.	Les idées s'enchaînent de manière claire et fluide; les liens logiques sont variés et utilisés avec justesse.
			○ ○	○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○	○ ○ ○	○ ○
	Lexique	Etendue **9**	Mots isolés essentiellement.	Le répertoire lexical est très limité ; emploi répétitif des mots.	Le répertoire lexical est limité, les périphrases sont fréquentes.	Le répertoire lexical est plus large mais les périphrases restent nécessaires.	Le répertoire lexical est assez large ; utilise tous les moyens à sa disposition pour pallier ses lacunes.	Le répertoire lexical est étendu ; les tournures idiomatiques sont utilisées correctement.	Le répertoire lexical est très étendu ; maîtrise différents registres de langue.
			○ ○	○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○	○ ○ ○	○ ○
		Maîtrise **10**	Le vocabulaire employé n'est pas pertinent.	Le vocabulaire employé est très approximatif.	Les erreurs et les confusions sont fréquentes dans l'expression d'idées complexes.	Le vocabulaire concret est correct ; nombreuses lacunes dans le lexique abstrait.	Le vocabulaire employé est juste dans l'ensemble; quelques confusions.	Le vocabulaire employé est précis et pertinent.	Le vocabulaire employé est nuancé et parfaitement adapté.
			○ ○	○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○	○ ○ ○	○ ○
	Elocution	Prononciation et intonation **11**	Les erreurs sont systématiques et la compréhension difficile.	Les erreurs sont fréquentes et la compréhension difficile.	Effort pour respecter les codes mais les erreurs gênent encore la compréhension.	Quelques erreurs affectent encore les échanges mais l'ensemble est compréhensible.	Peu d'erreurs sont commises ; aucune gêne dans la compréhension.	La prononciation et le schéma intonatif sont proches de la langue authentique.	Très proche de la langue authentique ; un léger accent peut subsister.
			○ ○	○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○	○ ○ ○	○ ○
		Rythme et débit **12**	Le rythme est très haché et le débit très lent.	Le rythme est très irrégulier ; beaucoup d'hésitations.	Le rythme est parfois irrégulier.	Le rythme est assez homogène.	Le rythme est homogène et le débit presque régulier.	Le débit est régulier.	Le débit est naturel et fluide.
			○ ○	○ ○ ○	○ ○ ○ ○	○ ○ ○ ○	○ ○ ○	○ ○ ○	○ ○

Source : Chambre de commerce et d'industrie de Paris Île-de-France, 2018

Les sept échelons de la grille d'évaluation recouvrent les niveaux du CECRL : 0+ (pour les absences d'observables ou pour le niveau inférieur à A1), 1 pour le niveau A1 (introductif ou découverte), 2 pour le niveau A2 (intermédiaire ou de survie), 3 pour le niveau B1 (seuil), 4 pour le niveau B2 (avancé ou utilisateur indépendant), 5 pour le niveau C1 (autonome) et 6 pour le niveau C2 (maitrise). Le tableau ci-dessous (Tableau 6) fournit un portrait global de chaque niveau du CECRL afin de mieux saisir la variation graduelle.

Tableau 6 - Échelle globale des niveaux communs de compétences

UTILISATEUR EXPÉRIMENTÉ	C2	Peut comprendre sans effort pratiquement tout ce qu'il/elle lit ou entend. Peut restituer faits et arguments de diverses sources écrites et orales en les résumant de façon cohérente. Peut s'exprimer spontanément, très couramment et de façon précise et peut rendre distinctes de fines nuances de sens en rapport avec des sujets complexes.
	C1	Peut comprendre une grande gamme de textes longs et exigeants, ainsi que saisir des significations implicites. Peut s'exprimer spontanément et couramment sans trop apparemment devoir chercher ses mots. Peut utiliser la langue de façon efficace et souple dans sa vie sociale, professionnelle ou académique. Peut s'exprimer sur des sujets complexes de façon claire et bien structurée et manifester son contrôle des outils d'organisation, d'articulation et de cohésion du discours.
UTILISATEUR INDÉPENDANT	B2	Peut comprendre le contenu essentiel de sujets concrets ou abstraits dans un texte complexe, y compris une discussion technique dans sa spécialité. Peut communiquer avec un degré de spontanéité et d'aisance tel qu'une conversation avec un locuteur natif ne comportant de tension ni pour l'un ni pour l'autre. Peut s'exprimer de façon claire et détaillée sur une grande gamme de sujets, émettre un avis sur un sujet d'actualité et exposer les avantages et les inconvénients de différentes possibilités.
	B1	Peut comprendre les points essentiels quand un langage clair et standard est utilisé et s'il s'agit de choses familières dans le travail, à l'école, dans les loisirs, etc. Peut se débrouiller dans la plupart des situations rencontrées en voyage dans une région où la langue cible est parlée. Peut produire un discours simple et cohérent sur des sujets familiers et dans ses domaines d'intérêt. Peut raconter un événement, une expérience ou un rêve, décrire un espoir ou un but et exposer brièvement des raisons ou explications pour un projet ou une idée.
UTILISATEUR ÉLÉMENTAIRE	A2	Peut comprendre des phrases isolées et des expressions fréquemment utilisées en relation avec des domaines immédiats de priorité (par exemple, informations personnelles et familiales simples, achats, environnement proche, travail). Peut communiquer lors de tâches simples et habituelles ne demandant qu'un échange d'informations simple et direct sur des sujets familiers et habituels. Peut décrire avec des moyens simples sa formation, son environnement immédiat et évoquer des sujets qui correspondent à des besoins immédiats.
	A1	Peut comprendre et utiliser des expressions familières et quotidiennes ainsi que des énoncés très simples qui visent à satisfaire des besoins concrets. Peut se présenter ou présenter quelqu'un et poser à une personne des questions la concernant – par exemple, sur son lieu d'habitation, ses relations, ce qui lui appartient, etc. – et peut répondre au même type de questions. Peut communiquer de façon simple si l'interlocuteur parle lentement et distinctement et se montre coopératif.

Source : Conseil de l'Europe, 2001 : https://rm.coe.int/16802fc3a8

Les 7 échelons de la grille du TEF sont gradués en un nombre variable de sous échelons qui se répartissent de la manière suivante : 2 sous-échelons pour les échelons 0+ et 6; 3 sous-échelons pour les échelons 1, 4 et 5; et 4 sous-échelons pour les échelons 2 et 3. Cette répartition des sous-échelons s'explique par le fait qu'ils représentent les niveaux les plus courants chez les candidats/apprenants. Les auteurs du CECRL indiquent à ce sujet qu'étant donné que « la frontière

entre les niveaux est toujours un lieu subjectif, certaines institutions préfèrent des degrés larges, d'autres les préfèrent étroits. […] Les institutions peuvent développer les branches qui correspondent à leur cas jusqu'au degré de finesse qui leur convient » (Conseil de l'Europe, 2001, p. 31).

1.3.3. Le but de la recherche

Comme nous l'avons observé dans ce premier chapitre, malgré l'utilisation de procédures et d'outils d'évaluation standardisés, les tests oraux de L2 ne parviennent pas à éliminer l'ambiguïté et à assurer une constance dans l'évaluation, et ce, en partie en raison des effets des examinateurs. Notre recherche a alors deux objectifs : observer s'il existe des divergences à travers le jugement des examinateurs du TEF lors de l'évaluation de l'épreuve d'expression orale, puis brosser le portrait de l'appropriation et de l'appréciation des examinateurs de la grille d'évaluation. Notre recherche tente d'apporter des éléments allant vers une meilleure transparence, et subséquemment, vers une plus grande fidélité et validité de l'acte d'évaluer.

1.3.4. La pertinence scientifique de la recherche

À notre connaissance, très peu d'études sur les pratiques des examinateurs du TEF en ce qui a trait à l'évaluation de l'épreuve d'expression orale ont été explorées à ce jour. D'après notre recension, il existe deux études sur l'épreuve d'expression orale du TEF. La première étude porte sur l'estimation de la fidélité d'examinateurs (Demeuse et Artus, 2008) et la deuxième sur l'évolution temporelle du niveau de sévérité d'examinateurs (Chénier, 2018). Ces études sont de type quantitatif et la nôtre souhaite adopter une position épistémologique différente en ce sens qu'elle s'inscrit dans une perspective qualitative/interprétative. La présente recherche a donc pour but de pallier cette lacune.

De manière générale, le processus d'évaluation des situations d'évaluation sommative et certificative dans le domaine du FLS demeure très peu abordé. Rivière (2016) soulève à ce propos que les instruments de mesure, au même titre que les pratiques et les conditions de l'évaluation sont très peu exploitées dans les tests standardisés de FLS. Selon l'auteure, l'étude de l'interaction de ces dimensions enrichirait la compréhension de l'action et des gestes professionnels évaluatifs, car elles sont au cœur de l'acte d'évaluer.

Notre étude permettra avant tout de sensibiliser l'organisme concepteur du TEF, en l'occurrence l'équipe du Français des affaires de la Chambre de commerce et d'industrie de Paris Île-de-France, sur les pratiques évaluatives effectuées sur le terrain sous des aspects différents. D'une part, elle visera à documenter la manière dont le jugement est émis, c'est-à-dire à mieux comprendre les mécanismes sous-jacents du processus cognitif des examinateurs au regard de la grille d'évaluation. Comme les risques de fluctuations provenant des effets des examinateurs ont une incidence sur la note des candidats, il y a une nécessité de les étudier pour pouvoir les réduire. Ainsi, mieux comprendre le jugement pourra contribuer à accroitre la fidélité interjuges. D'autre part, notre étude amènera l'organisme concepteur du TEF à être plus conscient de l'utilisation concrète de la nouvelle version de sa grille d'évaluation. Cela a été le cas par exemple avec le test d'anglais standardisé IELTS. Pour l'épreuve d'expression orale de ce test (IELTS *Speaking Test*), la grille d'évaluation a été mise à jour dix ans après sa création, et dès cette mise à jour, des études ont été menées afin de mieux comprendre son utilisation réelle par les examinateurs. Ces études ont été conduites par Brown (2006) et par Merrylees et McDowell (2007), et se sont respectivement appuyées sur des analyses de *verbatim* portant respectivement sur le jugement de 12 participants, puis sur un sondage à grande échelle auprès de 151 participants. Les résultats de Brown (2006) ont été satisfaisants, tandis que ceux de Merrylees et McDowell (2007) ont été préoccupants. Dans les deux cas, les auteurs ont conclu que la bonne application de la grille d'évaluation demeurait la principale clé d'une évaluation fiable.

Les résultats de notre recherche pourront également donner lieu à une révision de la grille d'évaluation à l'avenir. Selon plusieurs chercheurs, pour concevoir des grilles d'évaluation mieux adaptées aux besoins réels et garantir une meilleure validité et fidélité, la méthode efficace serait de prendre en compte les perceptions et les rétroactions des examinateurs sur leur utilisation concrète des grilles (Brown *et al.*, 2001; Milanovic *et al.*, 1996; Orr, 2002; Pollitt et Murray, 1996; Upshur et Turner, 1995), sans toutefois y inclure toutes formes de subjectivité (Fulcher, 2003).

Finalement, les résultats obtenus de notre recherche permettront de fournir des indications précieuses pouvant être mises à profit dans le cadre des formations initiales et continues. Les éléments d'identification des aspects sensibles et variables du comportement des examinateurs pourront être réinvestis dans une méthode à suivre afin d'amener ces derniers à être plus efficaces et plus constants dans leur pratique. Généralement, les formations initiales et continues

représentent la solution pour réduire les effets de variabilité (Bonk et Ockey, 2003; Brown, 1995; Carr, 2011; Fahim et Bijani, 2011; Lumley, 2002; Weigle, 1994, 1998; Wigglesworth, 1993).

1.3.5. La pertinence sociale de la recherche

La pratique des tests de langue à des fins d'immigration nécessite la collaboration des diverses parties prenantes provenant de différents domaines : les responsables politiques ainsi que les professionnels de l'évaluation en langue. Cela permet de garantir que les résultats des tests sont utilisés de manière éthique (Saville, 2009). Les organismes concepteurs et certificateurs de tests de langue, qui incarnent les professionnels de l'évaluation en langue, se positionnent comme un centre de gravité dans la relation qu'ils entretiennent avec les différentes parties prenantes (utilisateurs finaux, examinateurs et candidats), ils se doivent alors de prendre en compte de façon équitable les attentes de chacun (Georges, 2013). Plusieurs chercheurs affirment à ce sujet que les responsables d'examens (concepteurs, comités d'examen) qui offrent des tests de langue sont sous l'obligation de démontrer 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, American Psychological Association* et *National Council on Measurement in Education* [AERA, APA, et NCME], 2014; Chapelle *et al.*, 2008; McNamara, 1996; O'Sullivan et Weir, 2011; Saville, 2009; Shohamy, 2007; Taylor et Galaczi, 2011). Par conséquent, il y a un besoin d'investir du temps et des efforts afin d'assurer la qualité des attributions des notes de la part des examinateurs, autant que dans la conception et le développement des tests (Alderson *et al.*, 1995; Kane, 2006).

Par ailleurs, il est largement admis et reconnu que la connaissance de la langue parlée dans le pays d'accueil favorise potentiellement l'intégration sociale des immigrants (Lapierre Vincent, 2004; Van Avermaet et Rocca, 2013). Ainsi, il va sans dire que « si les tests de langue deviennent un facteur de discrimination déterminant pour l'entrée [dans un pays donné], il est essentiel que tout test utilisé soit juste et corresponde à l'objectif recherché, de sorte que les immigrants ne se voient pas injustement refuser l'accès au territoire à n'importe quel stade de leur parcours en tant que migrants » (Van Avermaet et Rocca, 2013, p. 12). La variabilité due aux effets des examinateurs a ainsi des conséquences importantes sur les processus décisionnels, en particulier dans les situations de tests à enjeux élevés (Bachman *et al.*, 1995; Brown, 1995; Engelhard et Myford, 2003; Lumley et McNamara, 1995; McNamara, 1996). Dans un test à enjeu extrêmement élevé tel

que le TEF, l'incidence sociale est forte, les résultats obtenus peuvent déclencher un nouveau déroulement d'événements dans la vie des candidats. Toute erreur dans le positionnement de ces derniers peut être fatidique et conduire à un rejet erroné de leur dossier d'immigration ou de citoyenneté (Casanova et Demeuse, 2011). De ce fait, la responsabilité des examinateurs est de taille, les enjeux critiques au regard des projets de vie des candidats exigent que le sérieux et la rigueur soient apportés aux évaluations. De plus, comme les lieux de passation du TEF sont multiples à travers le monde, la diversité géographique et culturelle augmente les difficultés de contrôle et le risque d'irrégularités dans le déroulement des sessions. Ainsi, pour toutes ces raisons, il est primordial que les critères de validité et de fidélité apportent des garanties suffisantes sur la qualité du dispositif d'évaluation du TEF.

Ce premier chapitre nous a permis de préciser et d'exposer la problématique de la recherche. Il a servi à situer le contexte de la recherche et à préciser ses objectifs. Le deuxième chapitre se propose d'effectuer une recension des écrits et de définir le cadre conceptuel, ce qui tient lieu d'appui au problème de recherche et permet de déterminer la méthodologie de la recherche.

CHAPITRE 2: CADRE CONCEPTUEL

Introduction

Dans ce chapitre, nous définirons les concepts-clés de notre recherche. Nous commencerons par définir le concept de compétence communicative en présentant les modèles fondamentaux des linguistes, car le développement et l'évaluation des tests de langue reposent sur les définitions théoriques de compétence en communication. Nous aborderons ensuite la contribution des référentiels CECRL et *Proficiency Guidelines* en évaluation des L2. Leur instauration a permis la construction d'un langage international commun en facilitant la traduction des niveaux des candidats par les organismes concepteurs de tests de L2. Nous rappellerons l'origine des deux référentiels, leur évolution au fil des décennies, puis ferons part des critiques des chercheurs à leur sujet. Nous continuerons en définissant les notions de fidélité et de validité qui sont primordiales pour une évaluation juste et de qualité. Étant donné que la validité d'un test implique entre autres une bonne connaissance de l'évaluateur, et que ce concept est central dans notre recherche, nous définirons le jugement sous un angle non exhaustif, puis situerons l'avènement de la recherche sur la cognition de l'évaluateur. Nous mettrons en lumière les approches évaluatives des examinateurs et les modélisations qui ont tenté de conceptualiser leur cognition. Nous ouvrirons une parenthèse méthodologique en évoquant la technique de la technique de la pensée à voix haute qui permet de fournir des indices de l'activité cognitive des évaluateurs. Par la suite, nous identifierons des variables nuisibles, environnementales et autres, qui s'interposent dans la notation des candidats. Nous terminerons ce chapitre en exposant certaines caractéristiques des grilles d'évaluation. Nous rendrons compte de leur utilité, de leur complexité, et présenterons des approches dans leur élaboration. Nous présenterons la grille d'évaluation actuelle de l'épreuve d'expression orale du TEF en mettant en évidence les éléments mis à jour, puis finalement, nous dévoilerons nos questions de recherche.

2.1. Les modèles théoriques de la compétence communicative

Dans cette section, nous porterons un regard rétrospectif sur les fondements théoriques de la compétence communicative, puis différents modèles seront présentés : le modèle de Canale et Swain (1980), le modèle de Bachman, (1990) et le modèle du CECRL (2018). Même s'ils se

révèlent parfois approximatifs et incomplets (Beacco, 2002; Fulcher, 2003), ces modèles sont particulièrement nécessaires en évaluation puisqu'ils permettent une certaine emprise sur la langue comme objet d'évaluation. Comme l'affirment plusieurs chercheurs tels que Canale et Swain (1980), Bachman (1990), Bachman et Palmer (1996) et Fulcher (2003), avant de construire des tests permettant de mesurer l'aptitude à communiquer, il importe de définir ce que nous voulons mesurer.

2.1.1. Les fondements théoriques de la compétence communicative

Dès les débuts de la linguistique moderne, la langue a été analysée comme étant un acte individuel balisé par un ensemble de règles partagées par une communauté. Saussure, reconnu comme le précurseur de la linguistique moderne, a défini certains concepts fondamentaux. En 1916, il publie l'ouvrage « Cours de linguistique générale » où il fait une distinction entre les trois éléments suivants : la langue, le langage et la parole. Il définit la langue comme étant un phénomène social, reposant sur un système de conventions. Le langage, quant à lui, est défini comme une faculté propre à l'être humain, puis la parole comme étant un acte individuel, pouvant s'écarter de la norme et finissant par la transformer (Bronckart, 1977).

Son enseignement a influencé toute une génération de linguistes européens dits structuralistes, rassemblés de 1926 à 1940 dans le Cercle linguistique de Prague, un groupe de critiques littéraires et de linguistes influent du XXe siècle. Les membres du Cercle mettent l'accent sur le « système fonctionnel » de la langue (Bronckart, 1977, p. 139), c'est-à-dire que « d'une part qu'on ne peut comprendre aucun fait de langue sans faire référence au système auquel il appartient et d'autre part, que ce système est avant tout un système de moyens appropriés à un but » (Bronckart, 1977, p. 139-140).

En 1963, Jakobson, l'un des fondateurs du Cercle de Prague, s'est distingué par sa « théorie de la communication », aujourd'hui plus connue sous le nom de « schéma de Jakobson », où il découpe la fonction de communication en différents facteurs (Leclerc, 1989) (Figure 2). Dans ce schéma, la relation entre le destinateur (l'émetteur) et le destinataire est tantôt bidirectionnelle, dans la conversation courante par exemple quand le rôle de chacun peut s'intervertir, et tantôt unidirectionnelle, dans le cas où le destinataire a peu de possibilités d'intervention. Le message passe grâce à un contact, ou canal, qui est direct si le destinateur et le destinataire sont en présence l'un de l'autre ou indirect, s'ils ne sont pas ensemble. La communication s'effectue à travers un

contexte ou référent, appelé aussi contexte situationnel. Ce référent, commun au destinataire et au destinateur est constitué de connaissances culturelles partagées par les deux parties. Finalement, le code s'ajoute à l'ensemble du schéma de communication et est assujetti à l'ensemble de signes, linguistique ou non, partagé de façon conventionnelle par les deux acteurs du schéma (Leclerc, 1989).

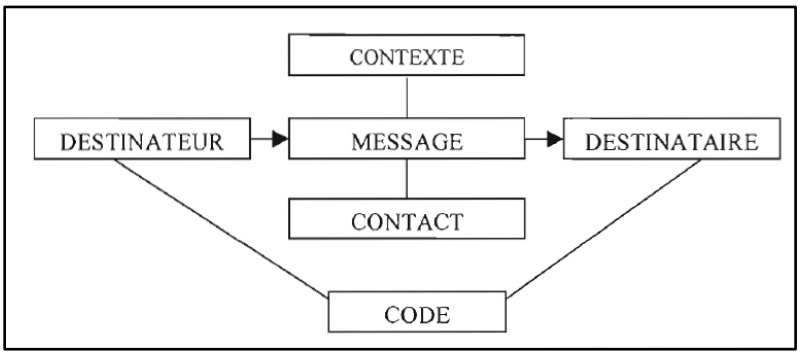


Figure 2 - Schéma général de la communication humaine de Jakobson (1963)

Vers la fin des années 1950, un nouveau courant linguistique sous le nom de « grammaire générative » se forme. Chomsky, qui se reconnaît comme continuateur de Jakobson, est à la tête de ce courant et propose une théorie du langage qui accentue sa forme créative et les propriétés innées de l'esprit humain. Il met en avant le concept de compétence grammaticale qui est la grammaire interne d'un locuteur permettant de produire des énoncés originaux, tout en respectant des règles syntaxiques précises. Cette compétence se déploie dans une performance, qui reflète les processus cognitifs du locuteur. Chomsky introduit alors les concepts de compétence et de performance (Chomsky, 1965).

Plus tard, dans les années 1970, la conception de la communication est revue par Hymes. Selon le sociolinguiste, pour communiquer entre eux, les membres d'une communauté linguistique doivent posséder à la fois un savoir linguistique (faisant référence à la compétence grammaticale de Chomsky) et un savoir sociolinguistique; c'est-à-dire non seulement des connaissances d'ordre grammatical, mais également des règles d'emploi de la langue selon la diversité des situations de communication. Il observe qu'une personne qui acquiert une compétence de communication acquiert à la fois la connaissance de la langue et l'habileté à l'utiliser en fonction de conventions sociales (Hymes, 1972, 1984).

Avant le début des années 1980, les tentatives de Hymes, et parallèlement celles d'autres chercheurs comme Halliday (1976) et Widdowson (1978), visent à préciser ce qu'implique la compétence communicative, mais ce n'est qu'en 1980 qu'apparaît l'une des plus influentes tentatives de définition, à savoir celle de Canale et Swain.

2.1.2. Le modèle de Canale et Swain (1980)

Dans le tout premier numéro de la revue *Applied Linguistics* publié en 1980, Canale et Swain font un bilan des différents modèles proposés par les chercheurs en linguistique moderne afin de décrire la langue et l'effet de ces modèles en enseignement et en évaluation des langues secondes (Tableau 7). Ils en concluent que la définition de la compétence communicative est encore au « stade embryonnaire ». Selon eux, la compétence communicative englobe une composante grammaticale, qui fait référence au code linguistique (syntaxe, morphologie, sémantique, lexique, phonologie) et une composante sociolinguistique (les règles propres à différents contextes et aux différents facteurs) qui fait référence aux différents contextes d'utilisation de la langue. Ils décident alors d'ajouter une troisième composante essentielle, la composante stratégique, qui selon eux, renvoie aux stratégies verbales et non verbales utilisées par le locuteur pour combler des lacunes dans les deux autres composantes. Trois années plus tard, Canale (1983) ajoute une quatrième composante, la composante discursive, qu'il définit comme l'habileté à relier des phrases pour produire un discours étendu qui ait du sens. Les fondements de la compétence communicative voient ainsi le jour.

Tableau 7 - Modèle de la compétence communicative de Canale et Swain (1980)

Composantes	Définitions	Exemples
Grammaticale	La maitrise du code.	La concordance des temps de verbes; La prononciation des phonèmes.
Discursive	La façon de combiner les formes grammaticales et le sens pour produire un texte; La grammaire du texte.	Le choix des arguments dans un texte qui vise à convaincre; La disposition des paragraphes dans un texte; Le choix des connecteurs logiques.
Sociolinguistique	La façon dont les énoncés sont formulés en tenant compte du contexte.	Le choix d'un niveau de langue; Le vouvoiement et le tutoiement; La compréhension d'un proverbe.
Stratégique	L'utilisation de stratégies verbales et non verbales qui visent à pallier des lacunes	Faire des gestes; Demander de répéter;

	dans les autres composantes ou à renforcer la communication.	Paraphraser; Demander d'expliquer un mot inconnu.

Source : Canale, 1983; Canale et Swain, 1980

Ce premier modèle de la compétence communicative à composantes multiples est d'une grande importance dans le domaine de l'évaluation en L2 dans les années 1980, car il a permis de mieux cerner son organisation conceptuelle (Lussier et Turner, 1995).

2.1.3. Le modèle de Bachman (1990)

Dix années plus tard, Bachman (1990) propose un autre modèle de « compétence langagière » qui prend le relais du modèle de la compétence communicative de Canale et Swain (1980) et qui devient à son tour le modèle auquel la plupart des chercheurs en linguistique se réfèrent de nos jours (Figure 3). Dans son modèle de compétence langagière, Bachman simplifie en quelque sorte le modèle de Canale et Swain (1980), puisque des quatre composantes générales (linguistique, discursive, sociolinguistique et stratégique), il n'en retient que deux qu'il considère essentielles: la composante organisationnelle et la composante pragmatique. La composante organisationnelle englobe deux sous-composantes: la composante grammaticale (le vocabulaire, la morphologie, la syntaxe et la phonologie/graphie), et la composante textuelle (la cohésion du discours, oral et écrit). La composante pragmatique, quant à elle, est divisée en deux sous-composantes: illocutoire et sociolinguistique. La sous-composante illocutoire fait référence aux différentes fonctions du langage : idéationnelle (l'échange d'information concernant les sentiments), manipulatoire (le fait d'accomplir quelque chose en le disant), heuristique (le fait de se servir de la langue à des fins d'apprentissage) et imaginative (le jeu de mots, la poésie, la blague, la métaphore). La sous-composante sociolinguistique, qui se rapproche de la composante sociolinguistique de Hymes (1984) et de Canale et Swain (1980), englobe la sensibilité aux variations dialectales, aux registres, au repérage du langage naturel et au repérage des références culturelles.

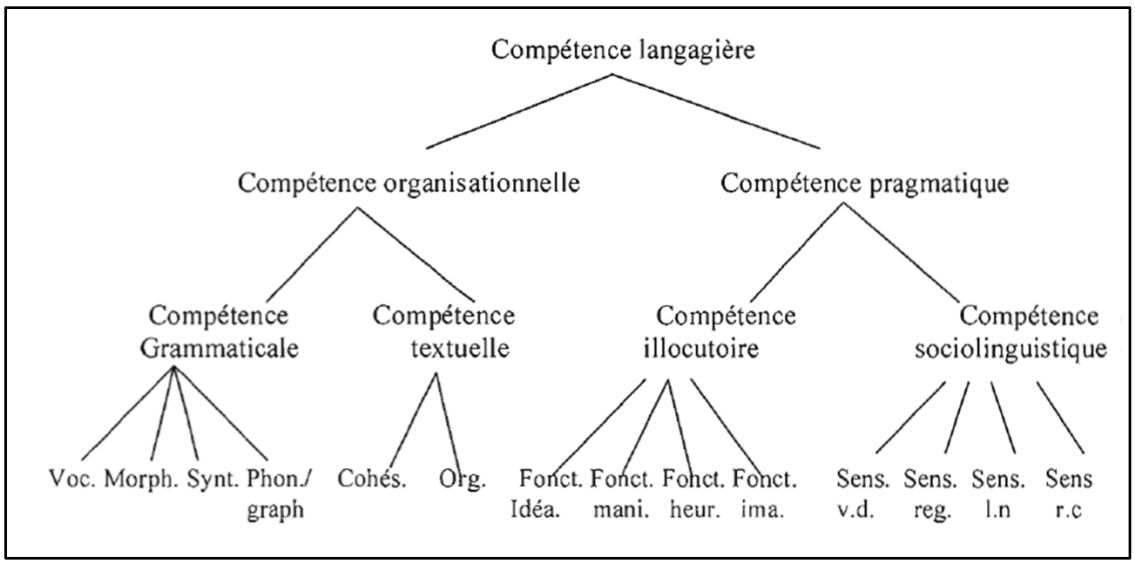


Figure 3 - Modèle de la compétence langagière de Bachman (1990)

Bachman fait une distinction entre la compétence langagière et la compétence communicative, qui est plus large et qui englobe la compétence langagière, mais aussi les connaissances sur le monde, la compétence stratégique, les mécanismes psychophysiologiques et le contexte situationnel (Figure 4). Les connaissances sur le monde représentent la somme des connaissances que l'individu possède pouvant être mobilisées lors de la communication. La compétence stratégique, quant à elle, permet de faire interagir toutes les composantes de la compétence langagière entre elles, elle se définit par la disposition du locuteur à puiser dans toutes les composantes pour exprimer un discours ayant du sens. Cette définition de la compétence stratégique est différente de celle de Canale et Swain (1980), qui eux, présentent cette dernière comme une façon de pallier les lacunes dans les autres composantes. Enfin, les mécanismes psychophysiologiques sont les processus neurologiques et psychologiques qui sont nécessaires à l'actualisation d'un message dans une forme langagière.

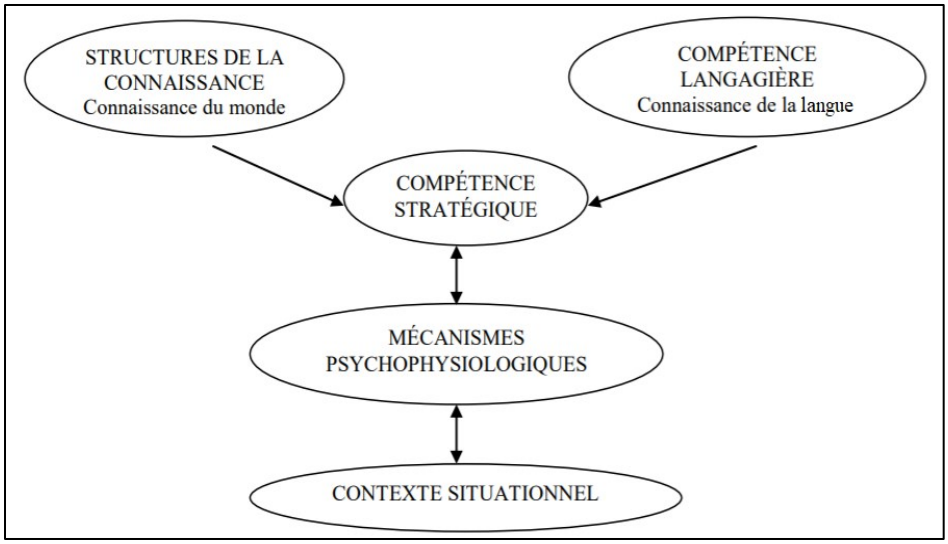


Figure 4 - La compétence langagière dans une situation de communication (Bachman, 1980)

Le modèle de Bachman a été l'objet de plusieurs validations empiriques, et par conséquent, il est d'une importance capitale dans les années 1990 (Lussier et Turner, 1995). Il a été construit dans le but d'aider à mieux évaluer les compétences en communication. Sa perspective est donc celle de l'évaluateur : « Si nous voulons développer et utiliser des tests de langue qui soient adéquats, nous devons les faire reposer sur des définitions claires des compétences à mesurer et des moyens utilisés pour observer et mesurer ces compétences » (Bachman, 1990, p. 81).

Plus tard, en 1996, Bachman s'associe à Palmer, les deux chercheurs récupèrent certaines notions de ce modèle et en propose un nouveau qui est adapté à la création de tests de langue servant à mesurer la compétence communicative (Figure 5). Leur regard sur l'utilisation de la langue se concentre sur la complexité des interactions de plusieurs composantes : la connaissance du sujet, la connaissance de la langue ainsi que les caractéristiques personnelles interagissent entre elles grâce à la compétence stratégique du locuteur. L'affect entre également en jeu dans l'interaction.

La connaissance du sujet fait référence à la connaissance du monde (Figure 4). Lors d'un test, certaines tâches qui présupposent des connaissances liées à une culture ou à un thème de la part des candidats peuvent être plus faciles pour certains et plus difficiles pour d'autres. La connaissance de la langue fait référence à la compétence langagière (Figures 3 et 4). Les caractéristiques personnelles sont les attributs individuels du candidat. Elles ne font pas partie de sa compétence langagière, mais peuvent tout de même influencer sa performance lors des tests de

langue. Les caractéristiques personnelles sont par exemple l'âge, le sexe, la nationalité, le statut socioprofessionnel, la langue maternelle, le niveau de scolarité, le style cognitif, la personnalité, le degré de préparation au test et les expériences déjà eues avec les tests. L'affect est le lien affectif ou émotionnel de la connaissance du sujet. Il peut influencer de manière positive ou négative la manière dont le candidat accomplit une tâche lorsque le sujet traité est susceptible de provoquer une charge émotionnelle (par exemple, le contrôle des armes à feu, la souveraineté nationale). On doit alors être conscient que certains sujets peuvent affecter la performance des candidats. Bachman et Palmer (1996) préconisent de prendre en compte ces composantes lors du développement d'un test de langue, car cela permet de mieux définir le construit et ainsi d'optimiser les qualités du test.

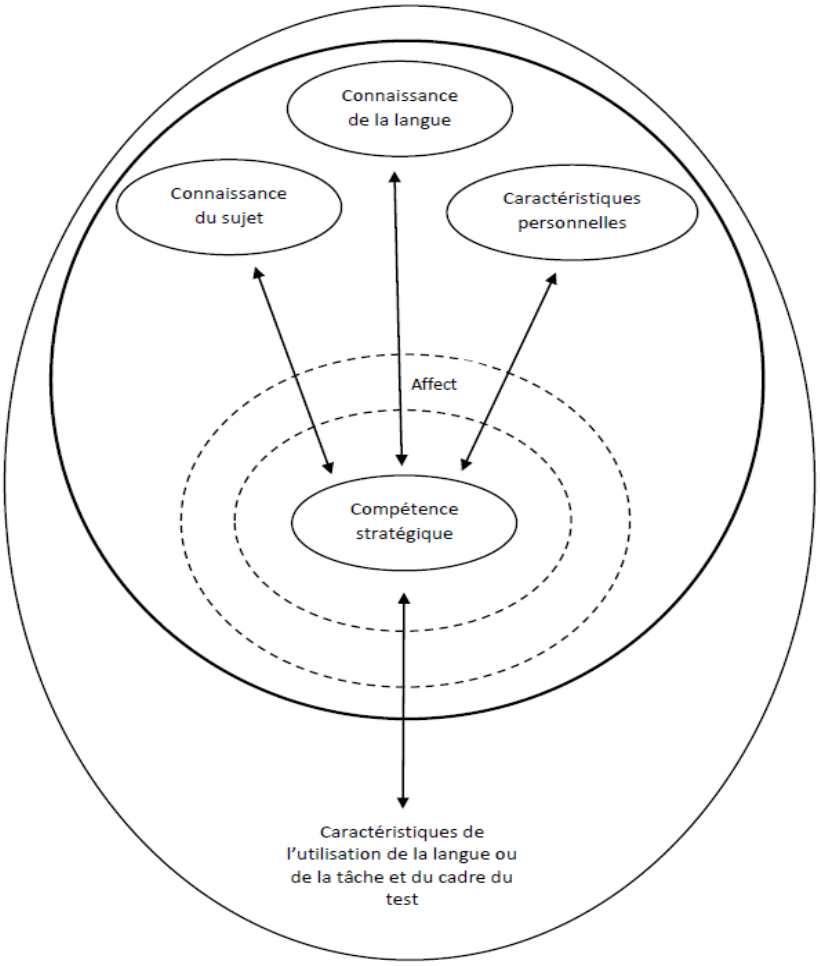


Figure 5 - Quelques composantes de l'utilisation de la langue et de la performance d'un test de langue

Source : Bachman et Palmer, 1996

Les organismes concepteurs de tests en FLS comme le Test d'évaluation de français (TEF), le Test de connaissance du français (TCF), le Diplôme d'études en langue française (DELF) et le Diplôme approfondi de langue française (DALF) affirment que les modèles de Bachman (1990) et de Bachman et Palmer (1996) leur ont servi de fondements théoriques (Demeuse *et al.*, 2010; Riba et Mavel, 2005; Tagliante et Mègre, 2008). Les auteurs du CECRL citent également les chercheurs dans leurs références (Conseil de l'Europe, 2001).

2.1.4. Le modèle du CECRL (2018)

En 2001, le Cadre européen commun de référence pour les langues, issu des travaux du Conseil de l'Europe, adopte les recherches des linguistes Canale et Swain (1981), et de Bachman (1990) et propose une perspective dite « actionnelle ». La perspective actionnelle se caractérise par une visée sociale, collaborative et citoyenne de l'utilisation et de l'apprentissage d'une langue, et se base principalement sur la réalisation de tâches collectives (Puren, 2006). En 2018, le CECRL publie un volume complémentaire et propose un schéma descriptif mis à jour qui donne une définition de la compétence langagière, fondée sur l'interaction et la co-construction du sens (Figure 6).

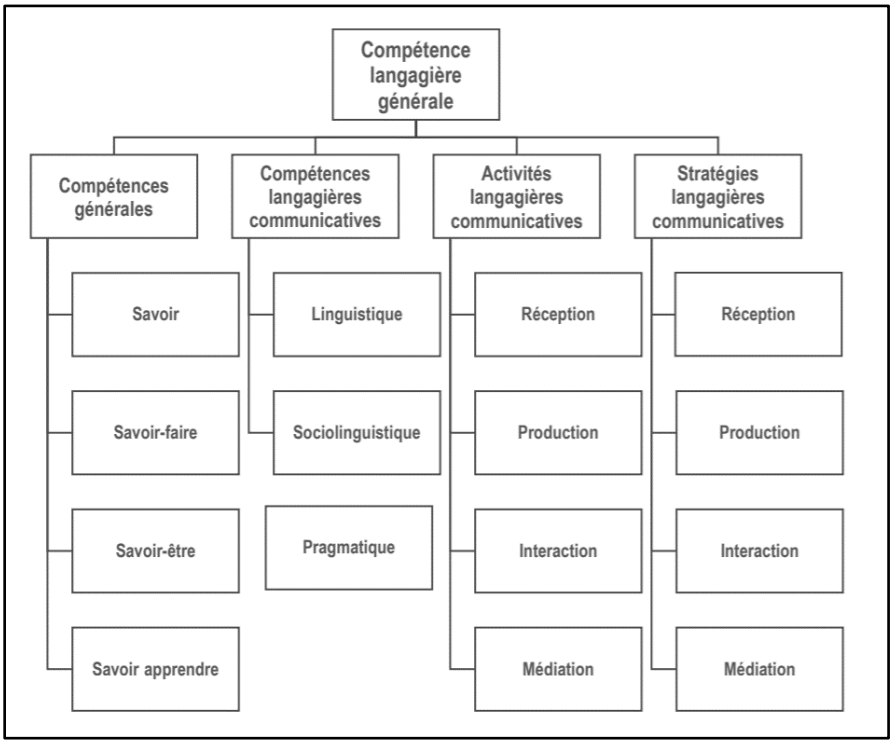


Figure 6 - Schéma descriptif de la compétence langagière générale du CECRL

Source : Conseil de l'Europe, 2018

Selon le CECRL, l'usage d'une langue, y compris son apprentissage, comprend les actions accomplies par des individus qui, comme individus et comme acteurs sociaux, développent un ensemble de compétences générales (connaissance du monde, compétence socioculturelle, compétence interculturelle, éventuellement expérience professionnelle), et notamment une compétence à communiquer langagièrement (compétences linguistique, sociologique et pragmatique[14]). Les individus mettent en œuvre les compétences dont ils disposent dans des contextes et des conditions variés, tout en se pliant à différentes contraintes, afin de réaliser des activités langagières. Ces activités permettent de traiter des données portant sur des thèmes à l'intérieur de domaines particuliers. Ils mobilisent également des stratégies (quelques stratégies générales, quelques stratégies communicatives langagières) qui conviennent le mieux à l'accomplissement des tâches à effectuer. Les activités et les stratégies langagières sont présentées selon quatre modes de communication : réception, production (monologue, rapport écrit), interaction (conversation, correspondance écrite) et médiation. La médiation est une (re)formulation qui permet de produire un résumé ou un compte rendu, par l'intermédiaire de l'interprétariat ou de la traduction, à l'intention d'un tiers qui n'a pas accès direct aux données premières.

L'usage du langage dépend donc principalement de la tâche. Par exemple, dans un contexte de déménagement, il est conseillé de communiquer, de préférence en utilisant le langage, mais celui-ci n'est pas l'objet de la tâche. De la même façon, les tâches qui demandent une communication plus sophistiquée, comme s'accorder sur une solution à un problème éthique ou se réunir sur un projet, mettent l'accent sur les résultats plutôt que sur le langage utilisé pour les réaliser (Conseil de l'Europe, 2018).

De surcroît, le CECRL distingue différentes catégories de tâches selon: les domaines dans lesquels elles se trouvent (privé, public, éducationnel, professionnel); leur nature (communicationnelle, créative, d'apprentissage, de résolution de problèmes); leur complexité (de la plus simple à la plus complexe); le traitement de textes oraux ou écrits qui lui sont associés (explications orales ou modes d'emploi, consignes); les stratégies qui doivent être mises en œuvre pour l'effectuer

[14] La compétence linguistique est le fait de connaître une langue, avec ses composantes lexicales, grammaticales, sémantiques, orthographiques, phonologiques et expressions. La compétence pragmatique est le fait de connaître les structures et articulations sémantiques des discours et messages (par exemple le développement thématique, la cohésion, les tours de parole). La compétence sociolinguistique est le fait de savoir adapter la langue à l'interlocuteur (par exemple les règles de politesse, la régulation des rapports entre statuts sociaux) (Conseil de l'Europe, 2001).

(exécution de la tâche, évitement, appel à l'aide, etc.); leur type (tâches pédagogiques simulant la vie réelle : jeux de rôles, simulations, interactions diverses, tâches métacognitives); les activités langagières qu'elles requièrent: production, réception, interaction, médiation (reformulation); et l'évaluation qui leur est associée (sous forme de contrôle ou d'évaluation formative accompagnée de critères) (Conseil de l'Europe, 2001).

Ainsi, le concept de compétence communicative/langagière en L2 a évolué au cours des années. Cette évolution conceptuelle a permis d'intégrer aux modélisations de nouvelles composantes à prendre en considération dans les instruments de mesure. Nous allons voir à présent l'importante contribution du CECRL quant à l'évaluation des L2.

2.2. L'apport du référentiel CECRL

Dans cette partie, nous nous intéresserons aux apports clés du CECRL en matière de standardisation des instruments d'évaluation, notamment aux apports des niveaux de langue et des descripteurs de capacité langagière. Nous en retracerons l'origine, puis ferons part des critiques à son encontre.

2.2.1. Les niveaux et les descripteurs de capacité langagière du CECRL

Le Cadre européen commun de référence pour les langues est mené par la Division des politiques linguistiques du Conseil de l'Europe, une instance indépendante de l'Union européenne. Le Conseil de l'Europe, doté parallèlement d'une vocation politique, a été fondé en 1949 avec vingt-deux pays. Il regroupe quarante-sept états membres et est basé à Strasbourg en France. Il a été créé pour remplir différentes missions, dont une qui touche particulièrement l'enseignement et l'apprentissage des langues et des cultures : promouvoir la prise de conscience d'une identité européenne fondée sur des valeurs communes, partagées, et dépassant les cultures particulières (Morrow, 2004).

Les grandes lignes du CECRL s'établissent dès 1991 lors d'un symposium international en Suisse sur le thème *Transparence et cohérence dans l'apprentissage des langues en Europe*, mais il faut attendre 2001 pour qu'il voie le jour. Dix-sept ans plus tard, en 2018, un volume complémentaire est publié. Des ajouts sont proposés ainsi qu'une forme d'adaptation et une standardisation en vue d'uniformiser les politiques linguistiques des pays utilisateurs.

Le CECRL est le fruit d'une réflexion scientifique menée durant près de dix ans, et animée par la volonté de trouver une transparence et une cohésion dans la grande diversité des systèmes

d'évaluation et des certifications en Europe. Il se donne comme objectifs : « de promouvoir et faciliter la coopération entre les établissements d'enseignement de différents pays; d'asseoir sur une bonne base la reconnaissance réciproque des qualifications en langues; et d'aider les apprenants, les enseignants, les concepteurs de cours, les organismes de certifications et les administrateurs de l'enseignement à situer et à coordonner leurs efforts » (Conseil de l'Europe, 2001, p. 11-12).

Le CECRL se présente sous la forme d'un document nommé *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*, disponible dans 40 versions linguistiques[15]. Il permet de fournir une base transparente, cohérente et aussi exhaustive que possible pour l'élaboration de matériels d'enseignement, de référentiels[16], d'examens, de manuels, ainsi que pour l'évaluation et l'auto-évaluation des compétences en L2 par-delà les frontières nationales et linguistiques.

Il définit six niveaux de compétence langagière (A1, A2, B1, B2, C1, C2) qui permettent de mesurer le progrès des apprenants à chaque étape de leur apprentissage (Figure 7). Cette configuration n'est pas figée, car chaque niveau autorise une subdivision en fonction des publics cibles et de leurs besoins (par exemple A2+ ou A2.2) (Conseil de l'Europe, 2001).

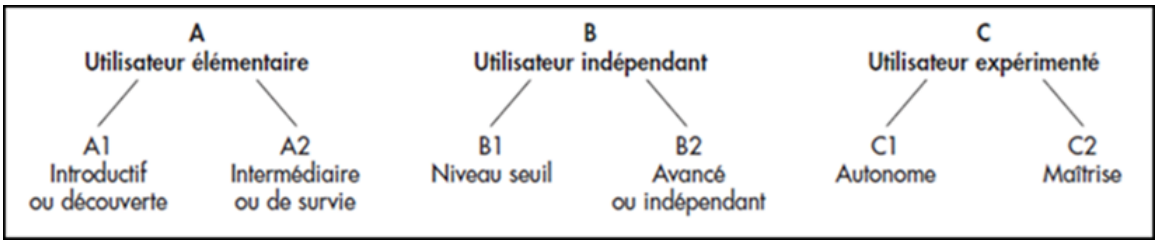


Figure 7 - Les niveaux communs de référence du CECRL

Source : Conseil de l'Europe, 2001

Ces six niveaux sont le fruit d'une réflexion progressive et collective lancée en 1913 avec le *Cambridge Proficiency Exam* (CPE), un test international sanctionnant un haut niveau de maitrise de la langue anglaise, qui définit la maitrise pratique d'une L2. Ce niveau correspond aujourd'hui au niveau le plus avancé qui est C2. Par la suite, juste avant la Seconde Guerre mondiale, le *First*

[15] albanais, allemand, anglais, arabe, arménien, basque, bulgare, catalan, chinois, coréen, croate, danois, espagnol, espéranto, estonien, finlandais, français, frioulan, galicien, géorgien, grec, hébreu, hongrois, italien, japonais, lituanien, macédonien, moldave, néerlandais, norvégien, polonais, portugais, russe, serbe (version iékavienne), slovaque, slovène, suédois, tchèque, turc et ukrainien.

[16] Selon la définition de Figari (1994), un référentiel représente un ensemble de principes directeurs, de critères ou de standards qui nous permettent de porter un jugement sur la valeur de l'objet évalué.

Certificate in English (FCE) de l'université de Cambridge a été introduit ; il est encore aujourd'hui largement considéré comme le premier niveau de compétences utile dans un contexte professionnel, et correspond au niveau B2 (Goullier, 2007). Plus tard, en 1975, le Conseil de l'Europe, sous la direction des linguistes Trim et van Ek, a défini un niveau inférieur baptisé *Threshold Level*, qui correspond aujourd'hui au niveau B1, et qui désignait à l'origine le niveau minimum des compétences de communication en L2. Autrement dit, il s'agissait de définir un niveau seuil que devait atteindre tout apprenant désireux de satisfaire aux exigences langagières de la vie de tous les jours dans un pays donné (Beacco, 2004). Cet outil de référence fournit des inventaires sur les activités langagières (ce qu'un apprenant est capable de réaliser pour accomplir une tâche) de production et de réception (par exemple : s'informer sur des attractions touristiques, comprendre des mises en garde), et sur les comportements que l'apprenant doit être capable d'adopter (par exemple : rapporter les circonstances d'un accident). *Thresold Level* a été conçu à la base pour l'apprentissage de l'anglais, puis a été adapté pour l'apprentissage du français en 1976 par Coste *et al.*, cela a donné naissance au Niveau Seuil. Par la suite, ce dernier a servi de modèle à d'autres langues occidentales comme l'allemand, l'espagnol, l'italien et le portugais (Alvarez, 1981).

Au cours des années 1990, afin de répondre à la demande des organismes de formation et des enseignants, d'autres référentiels de niveaux ont progressivement été créés : Waystage Level (niveau de survie), Vantage Level (niveau avancé), Breakthrough Level (niveau de découverte), Efficiency Level (niveau d'autonomie) et Mastery Level (niveau de maitrise) (Tableau 8). Le Conseil de l'Europe a regroupé ces six niveaux communs de référence pour le futur CECRL sorti officiellement en 2001 (Goullier, 2007). Depuis lors, pour des raisons de commodité et pour éviter les difficultés de traduction dans les différentes langues, chaque niveau porte une lettre et un chiffre.

Tableau 8 - L'origine des six niveaux de capacité langagière

Breakthrough (Découverte) devenu A1	*Waystage* (Survie) devenu A2	*Thresold* (Seuil) devenu B1	*Vantage* (Avancé) devenu B2	*Efficiency* (Autonome) devenu C1	*Mastery* (Maitrise) devenu C2

Source : Tagliante, 2005 (Tableau adapté)

Les niveaux de langue étant désormais identifiés, les capacités partielles le sont également. Pour chaque niveau de langue, le CECRL a élaboré des descripteurs génériques permettant d'expliciter

les capacités langagières (ou les « capacités de faire » « *can-do statements* ») des locuteurs. On admet, par exemple, qu'un journaliste de la presse écrite peut être à l'aise dans la compréhension de l'écrit, sans pour autant être en capacité à produire un discours à l'oral (Macaire, 2018).

Les descripteurs de capacités langagières sont élaborés à la fois de manière globale et à la fois pour:

- toutes les activités communicatives de réception, de production, d'interaction et de médiation; à l'oral comme à l'écrit;
- les stratégies de communication comme la compensation (dissimuler ses lacunes), le contrôle (revenir sur une difficulté et la reformuler), la planification (planifier ce qu'il faut dire);
- les compétences communicatives langagières : lexicales, grammaticales, orthographiques, sociolinguistiques (registres de langue), phonologiques, orthoépique (prononciation et intonation correctes dans une lecture à voix haute), sémantiques (production de sens), pragmatique (gestion des tours de parole), fonctionnelle (production d'énoncés sous différentes formes comme les questions-réponses, acceptations-refus).

Par exemple, pour le niveau B1, les descripteurs de capacités langagières sont formulés ainsi :

- Descripteur global : « Possède assez de moyens linguistiques et un vocabulaire suffisant pour s'en sortir avec quelques hésitations et quelques périphrases sur des sujets tels que la famille, les loisirs et centres d'intérêt, le travail, les voyages et l'actualité. » (Conseil de l'Europe, 2001, p. 28).
- Descripteur pour la réception orale : « Peut généralement suivre les points principaux d'une longue discussion se déroulant en sa présence, à condition que la langue soit standard et clairement articulée. » (Conseil de l'Europe, 2001, p. 55).

Ces descripteurs et leur organisation en échelles sont issus d'une recherche empirique menée entre 1993 et 1996 par les deux linguistes North et Schneider dans le cadre d'un projet du Fonds national suisse de recherche scientifique. Partant d'une analyse d'échelles de compétences existantes à l'époque, un fonds initial de descripteurs a été élaboré puis testé auprès de 300 enseignants et de 2800 apprenants du premier et du second cycle de l'école secondaire. Cet exercice a permis d'évaluer non seulement la clarté des descripteurs, mais aussi la faisabilité des catégories proposées. À l'issue de cette recherche, le CECRL a proposé au total 514 descripteurs de capacités langagières de référence, calibrés et étalonnés sur les six niveaux allant de A1 à C2 (Conseil de l'Europe, 2001; Goullier, 2007).

Les descripteurs du CECRL se sont par ailleurs inspirés d'une trentaine d'échelles de compétence langagière déjà existantes dont *Proficiency Guidelines* (de l'*American Council on the Teaching of Foreign Languages)* (1986), *Band Descriptors for the Speaking and Writing* (de l'*International English Testing System*, IELTS) (1990), et L'échelle de compétence langagière (de la fondation Eurocentres) (1993).

Il est possible pour les utilisateurs de remanier légèrement les descripteurs pour mieux les adapter au domaine concerné (recherche ; vie professionnelle ; école primaire/secondaire); de fractionner les descripteurs relativement denses en plusieurs sous-descripteurs; et d'ajouter de nouveaux descripteurs qui reflètent une orientation plus spécifique au sein du domaine concerné. Le CECRL n'a pas de caractère obligatoire et n'est pas prescriptif, il est ouvert aux adaptations et encourage l'autonomie des institutions et des personnes concernées, ces dernières restent responsables de leurs choix (Conseil de l'Europe, 2001).

2.2.2. Les outils annexes du CECRL

Comme le Conseil de l'Europe reconnait que le document du CECRL ne représente pas une ressource suffisante à l'intention des organismes certificateurs et des praticiens intéressés par les tests de langue, des outils annexes ont été créés comme :

- *Manuel pour relier les examens de langue au Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer* (Conseil de l'Europe, 2009)
- *Manuel pour l'élaboration et la passation de tests et d'examens de langues* (Conseil de l'Europe, 2011)

Ces manuels relient tous les examens de langue aux six niveaux définis et abordent des thèmes variés parmi lesquels on retrouve : la validité, la fiabilité, l'équité, la rédaction des items, la construction des tests, la passation, la correction, la notation et délivrance des résultats, le contrôle et la révision. Ils ont été conçus pour être utiles aussi bien aux concepteurs débutants qu'aux plus expérimentés. Ils présentent des principes communs qui s'appliquent aux organismes certificateurs et aux personnes impliquées dans les tests de langues de manière générale. Le public visé peut aussi bien être une grande institution préparant des tests pour des milliers de candidats dans le monde, qu'un enseignant isolé souhaitant évaluer ses élèves en classe. Les principes sont les mêmes pour les tests à fort ou à faible enjeu, seules les étapes pratiques varient (Conseil de l'Europe, 2011).

2.2.3. Les pratiques évaluatives vues par le CECRL

Le chapitre 9 du CECRL est « consacré à l'évaluation de la performance et non aux autres aspects d'un programme d'enseignement/apprentissage. » (Conseil de l'Europe, 2001, p. 135). Dans un premier temps, quelques points essentiels sont mis en avant comme : 1) l'importance des domaines d'utilisation de la langue (personnel, public, professionnel, éducationnel) et du rôle des tâches communicatives devant être accomplies; 2) l'usage des descripteurs pour le choix des critères d'évaluation; 3) l'importance des échelles dans la construction des tests et des examens. Dans un second temps, une typologie de treize pratiques évaluatives courantes mises en opposition est présentée sous forme de tableau nommé « Liste des paramètres » (Tableau 9).

Tableau 9 - Liste des paramètres des pratiques évaluatives du CECRL

1	Évaluation du savoir	Évaluation de la capacité
2	Évaluation normative	Évaluation critériée
3	Maitrise	Continuum ou suivi
4	Évaluation continue	Évaluation ponctuelle
5	Évaluation formative	Évaluation sommative
6	Évaluation directe	Évaluation indirecte
7	Évaluation de la performance	Évaluation des connaissances
8	Évaluation subjective	Évaluation objective
9	Évaluation sur une échelle	Évaluation sur une liste de contrôle
10	Jugement fondé sur l'impression	Jugement guidé
11	Évaluation holistique ou globale	Évaluation analytique
12	Évaluation par série	Évaluation par catégorie
13	Évaluation mutuelle	Auto-évaluation

Source : Conseil de l'Europe, 2001

Cette typologie est suivie d'explications assez brèves des concepts présentés, par exemple pour la première ligne du tableau : « L'évaluation du savoir (ou du niveau) est l'évaluation de l'atteinte d'objectifs spécifiques – elle porte sur ce qui a été enseigné – par voie de conséquence, elle est en relation au travail de la semaine ou du mois, au manuel, au programme. L'évaluation du savoir est centrée sur le cours. Elle correspond à une vue de l'intérieur » (Conseil de l'Europe, 2001, p. 139).

Pour certains chercheurs, cette section du CECRL ne constitue pas un outil clé en main, prêt à l'emploi. Par exemple, Rosen et Reinhardt (2010, p. 109) affirment que « ce chapitre 9 est très intéressant pour rafraîchir ses connaissances théoriques, mais ne donne que peu d'indications au niveau pratique. Il faut donc aller voir ailleurs : dans les ouvrages récemment publiés sur la question de l'évaluation ou dans un manuel qui doit permettre de relier les examens existants du Cadre et qui vise plus particulièrement les concepteurs d'examen. » Pour Tagliante (2005, p. 56) « La présentation choisie, en type 'opposé', n'est pas toujours pertinente. ». En ce qui concerne par exemple l'opposition de l'évaluation du savoir et l'évaluation de la capacité, l'auteure affirme que savoirs et performances peuvent être évalués au cours d'une même tâche langagière et communicative. Enfin, Huver (2014) qualifie ce chapitre 9 du CECRL comme étant très dichotomique, simplificateur, voire caricatural.

2.2.4. Les critiques à l'encontre du CECRL

Depuis sa large diffusion en 2001, le CECRL a acquis un respect international. Son instauration a eu un impact très important dans le monde de l'apprentissage et de l'enseignement, et surtout dans l'évaluation des langues secondes (Alderson 2002, 2005; Byrnes 2007; Coste, 2007; Figueras, 2012; Fulcher, 2008; Goullier, 2008; Jones et Saville, 2009; Little, 2007; Morrow 2004). Pourtant, des chercheurs se sont manifestés pour en relever quelques contradictions.

La critique la plus forte pointe l'absence de références aux théories sous-jacentes à l'élaboration du document, car ce dernier néglige de citer ses appuis théoriques et d'expliciter leur importance dans les propos qu'il affiche. Selon Huver et Springer (2011), le danger principal réside dans le fait de penser que les échelles représentent une hiérarchie d'acquisition d'ordre scientifique plutôt qu'une perception commune. Bien qu'elle se soit appuyée sur des panels d'experts, la démarche censée objectiver les seuils entre les niveaux de manière consensuelle n'a rien de scientifique, et par conséquent les frontières entre chaque degré restent subjectives. D'ailleurs, des auteurs du CECRL, comme North et Schneider (1998) et North (2007), admettent que les échelles de niveau et les descripteurs sont essentiellement athéoriques, et qu'elles ne sont pas basées sur les recherches en acquisition des langues secondes. North (2000, p. 573) cite par exemple que « ce qui est mis à l'échelle n'est pas nécessairement la compétence de l'apprenant, mais la perception de cette compétence par les enseignants / évaluateurs - leur cadre commun ». Ici, le terme « commun » fait référence à l'accord entre les 300 enseignants qui composaient l'échantillon de la recherche empirique lors de l'élaboration et de la classification des descripteurs du CECRL. De manière plus

générale, le terme renvoie à la façon dont les enseignants, éditeurs et concepteurs de tests européens considèrent les niveaux de langue selon les niveaux élémentaire, intermédiaire et avancé (Fulcher, 2004). Selon North (2004[17]), à travers cet esprit commun, l'on « cherche à faciliter la tâche des enseignants, des apprenants, des éditeurs et des concepteurs de tests pour communiquer à travers les langues, les secteurs éducationnels et les frontières nationales ».

Par ailleurs, on observe que le CECRL repose sur l'idée que l'apprentissage est linéaire et se construit par étapes progressives, à sens unique, et par des accumulations de savoirs, savoir-faire et savoir-être, alors que les stratégies des apprenants sont diverses et variables dans le temps (Macaire, 2018). La hiérarchisation des capacités langagières se présente sous une forme assez rigide, alors que les capacités langagières d'un niveau à un autre peuvent s'entrecroiser, un apprenant pouvant par exemple se situer à un niveau supérieur avec des éléments du degré inférieur non stabilisés (Fulcher, 2004).

En outre, l'agencement des capacités langagières d'un niveau à un autre n'indique pas comment juger l'évolution et le développement de la compétence de l'apprenant entre deux niveaux. De ce fait, il est difficile d'utiliser les échelles pour la construction des apprentissages et donc pour l'évaluation formative. Les échelles sont davantage adaptées à une évaluation sommative, car elles donnent une image globale des performances clés à un niveau donné à partir de repères et ont pour fonction d'attester la performance à un niveau donné. Elles sont considérées comme des standards normés déterminant le succès ou l'échec (Huver et Springer, 2011).

D'autres critiques, parfois virulentes, de chercheurs comme Alderson, Figueras, Kuijper, Nold, Takala, et Tardieu (2004), Forel et Gerber (2013), Fulcher (2004) (2008), Huver (2014), Weir (2005b), ainsi que les associations françaises ACEDLE (Association des Chercheurs et Enseignants Didacticiens des Langues Étrangères), ASDIFLE (Association de didactique du français langue étrangère) et Trans-Lingua (Association Travaux et Réseaux, Approches Nouvelles en Situations Interculturelles et Transnationales)[18] relèvent la concertation trop limitée avec les chercheurs-didacticiens des langues concernées. On constate par exemple que le CECRL oscille entre une intention d'harmonisation et une tendance à la normalisation, mais que cette dernière tend à prendre

[17] Citation extraite d'un article du journal *The Guardian*

[18] Ces trois associations ont initié une tribune intitulée « Le projet d'amplification du CECRL : une fausse bonne initiative du Conseil de l'Europe » en 2017. La tribune a été mise en ligne et a réuni des signatures de laboratoires et de chercheurs : https://asdifle.com/content/version-amplifi%C3%A9e-du-cecr-une-fausse-bonne-initiative-du-conseil-de-leurope-appel-%C3%A0

le dessus. Ce que l'on reproche est la domination d'une certaine conception dogmatique de la didactique des langues qui se trouve pérennisée, et qui conduit à la destruction de toute réflexion dynamique, voire d'une « confiscation du débat » d'après les termes de Huver (2017, p. 37). Selon la chercheure, la domination du CECRL a très largement contribué à stériliser la recherche, qui ne pense plus que dans ce cadre. Les chercheurs et les praticiens reprennent les fondements théoriques, les principes et les orientations méthodologiques du CECRL sans les discuter. Ainsi, comme le CECRL a été investi d'une dimension politique et que sa dimension didactique est moins crédible aux yeux des chercheurs du domaine (Comeford, 2009; Davies, 2008; Fulcher, 2004, 2008; Huver, 2014, 2017; Macaire, 2018; Maurer, 2011; Shohamy, 2001), il est souvent perçu comme un outil au service des politiques économiques néo-libérales.

Malgré les nombreuses critiques, le CECRL a occupé une vraie place de levier durant les deux dernières décennies. Dans le domaine de l'évaluation des langues secondes, il a fourni des éléments essentiels comme les échelles de niveaux et les descripteurs de capacités langagières. Ces éléments ont été utilisés comme point de départ et ont fourni un socle commun nécessaire à la conception d'une grande majorité de certifications en L2 dans le monde. Toutefois, le CECRL ne fait pas figure de pionnier, car les premières échelles décrivant les niveaux de maitrise d'une L2 ainsi que les premiers descripteurs de capacités langagières ont été développés dans les années 1950 par le *Foreign Service Institute*[19] à l'usage des militaires américains. Ils ont été révisés et adaptés par la suite par l'*American Council on the Teaching of Foreign Languages* (ACTFL)[20], ce qui a donné naissance au référentiel *Proficiency Guidelines* en 1986.

2.3. L'apport du référentiel *Proficiency Guidelines*

Dans cette partie, nous retracerons l'origine des premiers descripteurs et échelles de compétences langagières qui ont donné naissance au référentiel *Proficiency Guidelines*. Nous exposerons par la suite les controverses des chercheurs à son égard, et nous mettrons le référentiel en parallèle avec le CECRL.

[19] *Foreign Service Institute* est un institut de formation du gouvernement fédéral américain pour les employés de la communauté des affaires étrangères.

[20] *American Council on the Teaching of Foreign Languages* (ACTFL) est une organisation américaine dédiée à l'enseignement et à l'apprentissage des langues secondes.

2.3.1. L'origine des premiers descripteurs et des échelles de compétences langagières

Les premiers descripteurs de compétences langagières et les premières échelles de niveaux sont nés aux États-Unis d'une nécessité pratique lors de la Seconde Guerre mondiale, dans un contexte où le gouvernement fédéral s'apercevait que la majorité de son personnel militaire n'avait pas les compétences clés nécessaires pour communiquer dans une langue étrangère. En 1942, l'*Army Specialized Training Program* (ASTP), un programme de formation spécialisée de l'armée mis en place pour répondre aux exigences de la guerre, a été créé pour enseigner à communiquer oralement dans la langue d'un pays cible dans les domaines de l'ingénierie, de la médecine et des études régionales, tant pour les officiers subalternes que pour les soldats ayant des compétences techniques. Traitant 140 000 apprenants de 1943 à 1944, il s'agissait du premier programme de formation américain conçu pour « donner au stagiaire l'ordre de parler une langue dans un registre familier et de donner au stagiaire une solide connaissance du domaine dans lequel la langue est utilisée » (Angiolillo, 1947, p. 32). Comme l'armée américaine s'apercevait que la guerre évoluait mal en partie à cause du manque de compétences linguistiques pratiques parmi le personnel clé, elle a déclaré que l'urgence de la situation mondiale ne permettait pas d'offrir de longues heures d'apprentissage portant sur les théories de la structure grammaticale, et qu'il fallait avant tout prioriser les besoins communicatifs des situations courantes (Kaulfers, 1944). Par conséquent, promouvoir l'enseignement et l'évaluation de l'utilisation pratique de la langue est devenu la force motrice des écoles de langues du programme de formation ASTP (*Army Specialized Training Program*) (Agard et Dunkel, 1948). Des tests oraux sous forme d'entrevue ont alors été créés, mettant en avant trois tâches principales : sécuriser les services, demander de l'information et donner de l'information. La première grille d'évaluation, conçue par le linguiste Kaulfers en 1944, comportait deux grandes catégories nommées « portée de la performance orale » et « qualité (intelligibilité) de la performance orale ». La « portée de la performance orale » était composée de quatre sous catégories :

> 1) Peut faire part de quelques besoins essentiels dans des expressions ou des phrases; 2) Peut donner et sécuriser les informations routinières requises pour les voyages indépendants à l'étranger; 3) Peut extemporanément discuter de sujets communs et d'intérêts de la vie quotidienne; 4) Peut converser de manière impromptue sur n'importe quel sujet dans la limite de ses connaissances ou de son expérience (Kaulfers, 1944, p. 144).

Kaulfers a reconnu que pour certifier les compétences linguistiques pratiques, la notation devait être fondée sur des observations de la performance réelle. Il a alors entrepris d'établir un programme de recherche pour les tests de performance linguistique, mais ses premières tentatives n'ont jamais abouti en raison d'un manque de soutien et de financement en L2. Dès la fin de la Deuxième Guerre mondiale, l'impulsion pour enseigner et évaluer les langues secondes s'est estompée, puis a refait surface avec la guerre de Corée (1950-1953). Le *Foreign Service Institute* (FSI), un institut de formation du gouvernement fédéral américain, créé en 1947 pour les employés de la communauté des affaires étrangères, a repris le flambeau et s'est alors intéressé de plus près à la formation linguistique et aux tests oraux (Kramsch, 1986).

Afin d'évaluer la maitrise de la langue orale de son personnel, le FSI n'a pas repris la grille d'évaluation du linguiste Kaulfers (1944) qui s'appuyait sur des descripteurs de la performance orale, mais a élaboré intuitivement sa propre grille d'évaluation holistique. La grille, utilisée pour la première fois en 1956 et publiée en 1958, mettait en avant cinq critères: la prononciation, la compréhension, l'aisance, la grammaire et le vocabulaire. Chaque critère s'échelonnait sur six niveaux dont le plus bas qui se définissait comme « aucune habileté » (« *no ability* ») et le plus élevé comme « habileté de langue maternelle » (« *native speaking ability* ») (Adams, 1980; Sollenberger, 1978; Wilds, 1979). Par ailleurs, afin d'accroître la fidélité des notes, le FSI employait plusieurs examinateurs et avait recours à une pondération des critères pour constituer un score.

2.3.2. La naissance du référentiel *Proficiency Guidelines*

Dans les années 1960, la confiance dans les nouvelles procédures d'évaluation développées par le FSI était si élevée que celles-ci ont été adoptées et adaptées par trois organismes importants comme *Central Intelligence Agency (*CIA), l'agence de renseignement indépendante du gouvernement américain, *Defence Language Institute*, un institut de recherche et d'éducation pour les langues étrangères du département de la Défense des États-Unis, et *Peace Corps*, une agence indépendante du gouvernement américain dont la mission est de favoriser la paix et l'amitié du monde, en particulier auprès des pays en développement (Barnwell, 1996). En 1968, ces trois organismes se réunissent sous le nom d'*Interagency Language Roundtable* (ILR) pour produire une nouvelle version normalisée des différents niveaux et descripteurs de compétence langagière afin de classifier les apprenants (Jones, 1975; Lowe, 1987).

Par la suite, les niveaux et les descripteurs de l'ILR ont été révisés pour être adaptés aux besoins et aux objectifs des communautés linguistiques académiques grâce à l'implication de deux organisations : l'*Educational Testing Service* (ETS), une organisation privée de mesure et d'évaluation en éducation, et l'*American Council on the Teaching of Foreign Languages* (ACTFL) (Clark, 1988; Liskin-Gasparro, 1984a, 1984b; Lowe, 1983, 1985, 1987). L'adoption par les universités de ces niveaux et descripteurs a également été incitée par le rapport « *Strength through Wisdom : A Critique of U.S. Capability* » de la Commission des langues étrangères et des études internationales du président américain Jimmy Carter à la fin des années 1970. Parmi les recommandations du rapport figurait la mise en place de critères nationaux et d'un programme d'évaluation pour développer les tests de langue et évaluer l'apprentissage des langues secondes aux États-Unis (Chalhoub-Deville et Fulcher, 2003). C'est ainsi qu'est né en 1982 les *Provisional Proficiency Guidelines*, une version provisoire du référentiel de l'ACTFL. Par la suite, la première version complète des *Proficiency Guidelines* a été publiée en 1986, puis révisée en 1999 et en 2012 (*American Council on the Teaching of Foreign Languages*, 1986, 2012; Breiner-Sanders *et al.*, 2000).

Le référentiel *Proficiency Guidelines* touche les quatre habiletés de compétence langagière[21] (expression orale, expression écrite, compréhension orale, compréhension écrite), et identifie cinq niveaux principaux de compétence langagière : distingué, supérieur, avancé, intermédiaire, novice[22]. Pour chaque habileté de compétence langagière et chaque niveau, des descripteurs décrivent les tâches que les locuteurs sont capables d'accomplir dans des situations réelles, et dans un contexte spontané et non préparé. Les descripteurs mettent également en évidence les limites rencontrées par les locuteurs lorsqu'ils tentent d'accomplir les tâches fonctionnelles du niveau suivant. Par exemple, pour le niveau intermédiaire élevé en expression orale, les descripteurs sont formulés ainsi: « Les locuteurs […] sont capables d'accomplir avec succès des tâches à caractère social nécessitant un échange d'informations de base liées à leur travail, leur école, leurs loisirs, leurs intérêts particuliers et leurs domaines de compétence. Les locuteurs […] peuvent accomplir un nombre important de

[21] Définies par Lado (1962).
[22] Les niveaux avancé, intermédiaire et novice se subdivisent chacun en trois sous-niveaux : avancé élevé, avancé moyen, avancé bas; intermédiaire élevé, intermédiaire moyen, intermédiaire bas; novice élevé, novice moyen, novice bas.

tâches associées au niveau Avancé, mais ne sont pas capables d'accomplir la performance de l'ensemble de ces tâches de façon continue. » (ACTFL, 2012, p. 7).

2.3.3. L'évolution controversée des *Proficiency Guidelines*

Dans les années 1970, les écoles, tout comme les universités (notamment pour les certifications des enseignants de langues), ont adopté sans discernement les outils d'évaluation de l'ILR (Liskin-Gasparro, 1984b). Certains chercheurs comme Chalhoub-Deville et Fulcher (2003) et Fulcher (2003, 2016) affirment que cette popularité a rapidement conduit à la croyance manifeste, mais non fondée, selon laquelle les grilles d'évaluation en elles-mêmes possédaient « une réalité psychologique » en termes d'utilisation et de développement de la L2 (Chalhoub-Deville et Fulcher, 2003, p. 500). Par ailleurs, il a été reconnu par les chercheurs que l'histoire et les transformations des outils d'évaluation de la langue sont révélatrices de la bureaucratisation des tests de langue. Ces derniers sont perçus comme étant avant tout des outils pragmatiques au sein du gouvernement et de la bureaucratie militaire où il y a très peu de place pour la recherche. En effet, selon Adams (1980), l'élaboration des différents niveaux de compétence langagière de l'ILR à la fin des années 1960 ne s'est pas appuyée sur les recherches scientifiques pour y explorer ses hypothèses fondamentales, son utilisation et son développement. La recherche n'a porté que sur la fidélité inter-examinateurs. Le référentiel actuel *Proficiency Guidelines* n'est donc fondé sur aucune base théorique, mais plutôt sur l'expérience comme le confirment Lantolf et Frawley (1985) ainsi qu'Omaggio (1983). En effet, selon ces chercheurs, le référentiel décrit la façon dont les apprenants de langues fonctionnent de manière générale sur toute la gamme des niveaux de compétence, et n'est en aucun cas une prescription de théoriciens donnant des lignes directrices sur la manière dont les apprenants devraient fonctionner. Ces caractéristiques font écho avec les critiques à l'encontre du CECRL. Chapelle (2012, p. 42) déclare à ce sujet qu'il est trompeur de penser que les échelles de niveaux issues de ces deux cadres suivent une quelconque procédure d'acquisition des langues secondes. Selon elle, il faut les considérer de façon plus prosaïque tels des outils utiles socialement, et fonctionnant comme « des abstractions de notre expérience des trajectoires d'apprentissage des langues ».

Le référentiel *Proficiency Guidelines* est largement adopté dans le monde entier depuis les années 1970, et est considéré comme la ressource la plus appropriée pour mesurer la maitrise de la L2

(Fulcher, 2003, 2016). Il en est de même pour le CECRL depuis les années 2000, dont les descripteurs ont entre autres été tirés des *Proficiency Guidelines*.

2.3.4. Les référentiels CECRL et *Proficiency Guidelines*

Le CECRL et les *Proficiency Guidelines* représentent aujourd'hui les principaux référentiels pour l'élaboration des tests et diplômes en L2. Ils sont également utilisés pour l'élaboration de manuels et de programmes d'études, et de manière plus générale pour l'instauration de normes. Bien que les deux référentiels aient coexisté pendant près de 15 ans, très peu d'études empiriques ont permis d'établir des correspondances officielles entre eux. Grâce à l'initiative de l'*American Association of Teachers of German* (AATG), une association américaine d'enseignants d'allemand, et de l'ACTFL, une série de quatre conférences sur les alignements ACTFL-CECRL a été présentée en 2010 à l'Université de Leipzig en Allemagne. Les conférences ont réuni des experts de premier plan venus notamment des États-Unis, du Canada et d'Europe, représentant quinze organisations de quatorze pays différents, ayant reçu le soutien d'organisations américaines et européennes[23].

Cette série de conférences avait pour but principal d'établir des objectifs communs en matière de politique linguistique, comme ériger des passerelles entre les cadres et les salles de classe (relier la maitrise de la langue aux objectifs nationaux), et établir des alignements empiriques entre les tests de langue basés sur les cadres. La coopération transatlantique a donné lieu à de nombreuses publications afin de mieux informer les experts et le public sur les deux référentiels, et d'adopter des terminologies similaires. La collaboration a conduit, par exemple, au développement et la publication des descripteurs de « capacité de faire » (« *can do statements* ») qui donnent des objectifs à atteindre pour chacun des niveaux et pour chacune des quatre habiletés à l'oral et à l'écrit (expression orale, expression écrite, compréhension orale, compréhension écrite) (Tschirner, 2012; Tschirner et Bärenfänger, 2012).

Les normes internationales issues des référentiels jouent un rôle majeur pour la transparence et la qualité de l'évaluation des tests et diplômes en L2. Afin de soutenir cette démarche de qualité, il est tout autant primordial de mettre en œuvre des stratégies d'évaluation qui permettent de réduire

[23] *American Council on the Teaching of Foreign Languages (ACTFL), Conseil de l'Europe, European Centre for Modern Languages (ECML), Institute for Test Research and Test Development (ITT), University of Leipzig, Brigham Young University, American Association of Teachers of German (AATG), University of Cambridge ESOL, Goethe Institute, American Consulate General of the United States, The European Language Certificates, Gesamtverband Moderne Fremdsprachen, Language Testing International.*

l'impact des différences de jugement entre les examinateurs, et de porter un jugement qui soit le plus juste possible. Pour cela, l'évaluation de la performance orale, qui est la réponse à une tâche, doit répondre aux deux critères suivants : la fidélité et la validité (Bachman, 1990). Nous aborderons ces notions dans la section qui suit.

2.4. Les notions de fidélité et de validité

La fidélité et la validité sont des notions centrales dans les tests, car elles permettent d'apporter des garanties suffisantes sur le professionnalisme du dispositif d'évaluation. Nous aborderons ici la notion de fidélité sous un angle global, puis la notion de validité d'après des modèles adaptés à notre recherche.

2.4.1. La fidélité

La fidélité est une notion issue de la théorie classique des tests et est liée au calcul de l'erreur de mesure. Étant donné que les instruments psychométriques[24] ne sont pas parfaits, les scores obtenus contiennent inévitablement une certaine part d'erreur. La fidélité d'un instrument psychométrique représente le degré de précision et le degré de constance de ses scores. La précision se rapporte à la capacité à produire un score observé qui est le plus proche possible du score vrai (le score qui serait obtenu si l'instrument était parfaitement précis) de la personne évaluée. La constance, quant à elle, fait référence à l'obtention de résultats fortement similaires lorsqu'une personne est évaluée à l'aide du même instrument psychométrique à des moments différents dans le temps. Par conséquent, plus un instrument est fidèle, moins il contient d'erreurs de mesure, et ainsi plus le score observé (le résultat au test) est proche du score vrai de la personne évaluée (Anastasi, 1994; Bernaud, 2007).

La fidélité a donc trait à la constance avec laquelle un test mesure ce qu'il est censé mesurer. Pour qu'un test soit fidèle, il faut que la personne qui passe un même test à deux intervalles différents obtienne le même score. Cependant, les variations de la personne évaluée (humeur, maladie, fatigue) et l'environnement du test (bruit, éclairage non adapté, mobilier inconfortable) sont des facteurs qui peuvent avoir une influence sur le résultat (Bachman, 1990; Bachman et Palmer, 1996;

[24] Les instruments psychométriques sont des outils développés scientifiquement qui permettent de mesurer un concept psychologique de manière objective et standardisée. Ils peuvent prendre plusieurs formes comme par exemple des tests de rendement ou des questionnaires d'autoévaluation.

Fulcher, 2003). Afin d'optimiser la fidélité d'un test, il est nécessaire de maintenir de bonnes conditions de passation, d'assurer que la personne évaluée comprenne bien la tâche à effectuer, de la tester à plusieurs reprises, et également de procéder à une double notation (Alderson *et al.*, 1995; Genesse et Upshur, 1996).

La fidélité repose entre autres sur l'interprétation des résultats : la constance intra-évaluateur et la constance inter-évaluateurs. Premièrement, si le jugement d'un évaluateur reste stable d'une évaluation à une autre, il y a une constance intra-évaluateur. En revanche, si l'évaluateur corrige quinze copies d'affilée et que son niveau de sévérité augmente ou faiblit au fur et à mesure que son état de fatigue s'aggrave, il y a une absence de constance intra-évaluateur. Deuxièmement, si deux évaluateurs accordent le même nombre de points à un même test, la constance inter-évaluateurs est garantie. À l'inverse, si deux évaluateurs attribuent deux notes différentes à un même test, on observe un manque de constance inter-évaluateurs. Dans les deux cas, l'absence de constances intra et inter-évaluateurs affectent la fidélité d'un test.

Habituellement, plus le champ de réponse de la personne évaluée est libre, plus la constance intra-évaluateur et inter-évaluateurs est difficile à obtenir (Bachman, 1990; Gronlund, 1985; Lussier et Turner, 1995). Dans ce cas, l'utilisation de grilles d'évaluation demeure la solution pour assurer plus d'uniformité dans l'évaluation (Arter, 2010; Fulcher, 2000; Gronlund, 1985; Kaplan, 1991; Roegiers, 2004). Les items à correction objective comme les questionnaires à choix multiples, les questions « vrai ou faux », ou les questions à réponses courtes, permettent d'assurer une constance intra-évaluateur et inter-évaluateurs. En revanche, ils ne sont pas fidèles dans l'absolu du fait que la réponse peut résulter du hasard et non de la compétence réelle de la personne testée (Shohamy, 1983).

Par ailleurs, des erreurs dites systématiques peuvent augmenter ou réduire le score d'une personne en raison de facteurs étrangers au test. Par exemple, si l'on teste le niveau d'intelligence d'un étudiant anglophone avec un test dans une autre langue que l'anglais qu'il ne maitrise pas suffisamment, le niveau d'intelligence sera sans doute sous-estimé, et la sous-estimation sera relativement constante, que l'étudiant fasse le test un lundi ou un mercredi. De plus, si un étudiant est habile pour déceler les indices qui mènent à la bonne réponse dans un test, il aura tendance à obtenir un meilleur score que ce que lui permettent ses connaissances, et le jour du test n'y

changera rien. La fidélité ne tient donc pas compte de ces erreurs systématiques qui affectent la validité (Hogan, 2012).

2.4.2. La validité

Traditionnellement, la validité en évaluation a souvent été définie comme la capacité d'un test à bien mesurer ce qu'il prétend mesurer. Selon les époques et en fonction de l'arrivée de nouvelles approches, ce concept a connu des changements importants depuis son apparition au début du XX^{e} siècle. Étant un concept très largement étudié, nous retiendrons des multiples définitions existantes celle de Kane (2006), en raison de son applicabilité dans le contexte d'un test à enjeu élevé comme le TEF. Nous présenterons également une typologie proposée par Taylor (2011) de cinq concepts de validité adaptée plus spécifiquement aux tests oraux de langue seconde.

Selon Kane (2009), la démarche de validation dépend grandement du contexte d'utilisation de l'instrument à valider. L'interprétation et l'usage des scores sont des éléments à prendre en considération, il est donc essentiel de savoir si le score attribué à la personne évaluée représente bien son niveau de compétence et de prendre une décision adaptée sur la base de ce score. Le modèle de validité de Kane (2006, 2013) (Figure 8) se voit comme un modèle méthodologique qui présente une démarche à suivre pour procéder à la validation d'un instrument d'évaluation. Il est articulé autour d'une chaîne d'inférences, dont chaque maillon doit être soutenu par des arguments interprétatifs et permettre de faire le lien entre :

- les performances observées (les observations) ;
- la manière d'attribuer un score aux performances (le score observé) ;
- le caractère généralisable du score (le score univers) ;
- la signification du score (le score cible) ;
- la manière d'utiliser le score pour prendre une décision
  (Kane, Crooks et Cohen, 1999).

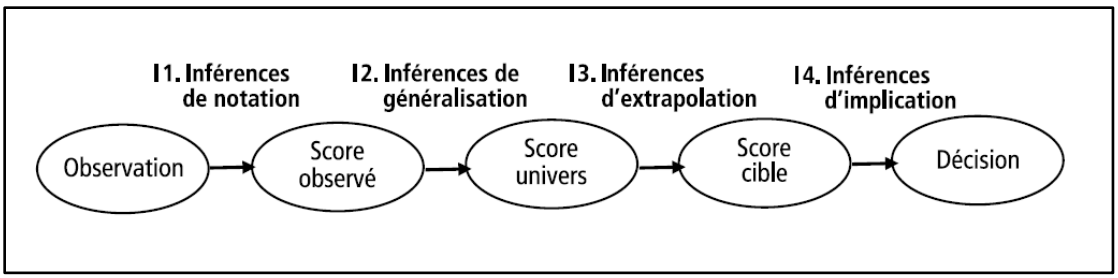


Figure 8 - Les inférences de Kane (1999)

Source : Kane, Crooks et Cohen, 1999 (Figure adaptée)

I1. Les inférences de notation se rapportent aux conditions de conception des tâches et des instruments d'évaluation, ainsi qu'à la constitution des modalités d'évaluation. La qualité de la notation peut être améliorée en mettant en œuvre des procédures d'administration standardisées et une bonne application des grilles d'évaluation. Les inférences de notation doivent attentivement documenter la démarche afin de garder des traces de chacun des arguments qui permettent de les soutenir. Plusieurs types de traces peuvent être rassemblés pour justifier la qualité de la notation, par exemple dans la notation d'une question à réponse ouverte, des données empiriques peuvent être utilisées pour vérifier la fidélité inter-examinateurs.

I2. Les inférences de généralisation se préoccupent de la fidélité et de la stabilité et des scores observés pour qu'ils puissent devenir des scores plus généraux, et non plus seulement reliés au test. Au sein de ces inférences, il est nécessaire d'identifier si des variables risquent de compromettre le processus, par exemple si de quelconques biais sont présents dans les tâches ou si les examinateurs utilisent convenablement les grilles d'évaluation ou les critères. On doit alors accumuler des preuves telles qu'un « échantillon des observations représentatif de l'univers de généralisation » (Kane, 2006, p. 24) et un assez grand nombre d'observations pour déterminer si les résultats obtenus sont représentatifs du niveau du candidat ou si ce sont des artéfacts.

I3. Les inférences d'extrapolation supposent que le score ne change pas, mais qu'il est interprété non plus comme représentatif de la performance à un sous-ensemble d'observations, mais comme étant représentatif du domaine en entier. Pour cela, l'utilisation d'informations d'autres sources sur les candidats (par exemple, les résultats à d'autres épreuves du même test ou d'un autre test) est nécessaire pour établir des corrélations ou faire des comparaisons. Les inférences

d'extrapolation sont à la fois basées sur des preuves de nature analytique et peuvent aussi être très informelles et reposer principalement sur l'expérience.

I4. Les inférences d'implication ont pour objectif de soutenir la crédibilité des interprétations des résultats et des décisions qui en découlent. Elles s'interrogent sur les conséquences de ces décisions et par conséquent, elles donnent lieu à des questions éthiques et sociales. Il y a donc une nécessité de fournir des preuves pour appuyer ce que l'on décide des individus sur la base de l'évaluation.

Comme nous l'avons déjà mentionné, Kane nous amène à réfléchir à la nature et à la variété des arguments pour soutenir les inférences du modèle, et propose pour cela une structure d'argumentation interprétative faisant référence au modèle d'argumentation de Toulmin (1958). Ce modèle d'argumentation est constitué d'une déclaration qui définit ce qui doit être établi, d'hypothèses qui ont le potentiel de la soutenir et de justifications qui explicitent la façon dont elles le font. L'argumentaire peut inclure des hypothèses de différentes natures et des justifications alternatives qui peuvent conduire à réfuter la déclaration.

Le modèle de validité de Kane est ainsi pertinent pour le TEF. D'abord en raison de la généralisation des scores, car ce que l'on teste dans l'épreuve d'expression orale doit pouvoir refléter une vision de l'usage de la langue telle qu'elle peut être utilisée dans la société d'accueil. De plus, les conséquences de l'évaluation sont importantes, car des décisions administratives et politiques découlent directement des scores obtenus.

Taylor (2011) propose dans la revue *Studies in language testing, Examining Speaking* une typologie, basée sur un modèle de Weir (2005a), de cinq concepts de validité adaptée particulièrement aux tests oraux de L2 (Figure 9). La validité est divisée en cinq types : 1) la validité cognitive (*cognitive validity*); 2) la validité de contexte (*context validity*); 3) la validité conséquentielle (*consequential validity*); 4) la validité liée aux critères (*criterion-related validity*); 5) la validité des scores (*scoring validity*).

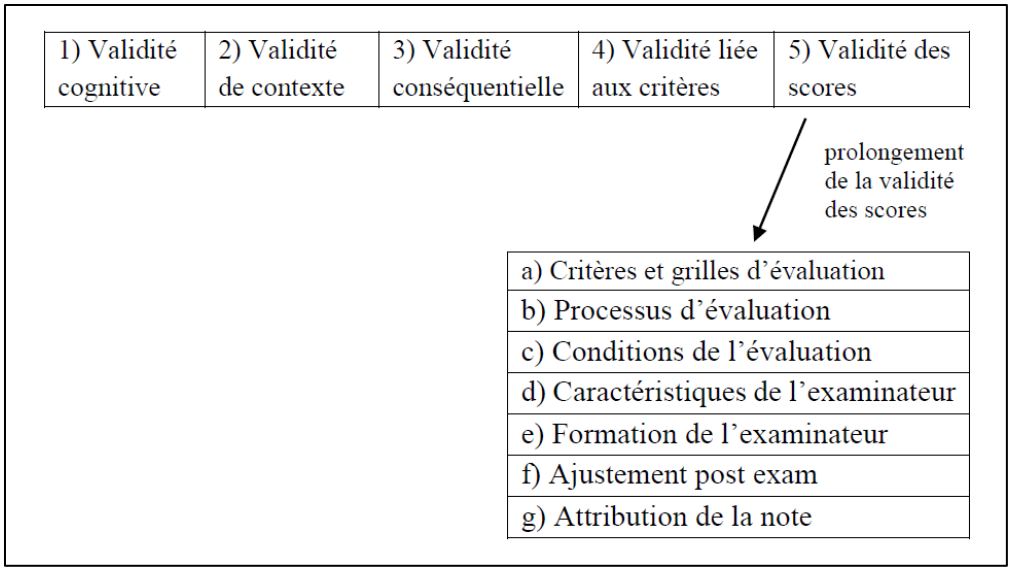


Figure 9 - Les types de validité adaptés aux tests oraux de langue seconde

Source : Taylor, 2011 (Figure adaptée)

1) La validité cognitive concerne la mesure dans laquelle la tâche parvient à susciter de la part du candidat une série de processus cognitifs qui ressemble à ceux employés dans les activités langagières de la vie courante. Les processus sont classés à travers différents niveaux selon les exigences cognitives qu'ils imposent au candidat. Par ordre décroissant, on retrouve la conceptualisation (production des idées), l'encodage grammatical (construction d'un modèle syntaxique pour exprimer les idées), l'encodage phonologique (récupération des formes phonologiques dans la mémoire), l'encodage phonétique (schématisation automatique des représentations phonologiques puis articulation) et l'autocontrôle (autoévaluation globale de l'objectif de la communication) (Field, 2011; Levelt, 1989).

2) La validité de contexte s'applique lorsque toutes les conditions nécessaires au bon fonctionnement d'un test de performance orale sont réunies afin que tous les candidats aient les mêmes chances de démontrer leurs capacités. Ces conditions sont les suivantes : l'administrateur du test doit définir clairement les tâches, les formats (monologue, dialogue), leur ordre, le temps imparti, la pondération des différentes parties du test. La tâche demandée au candidat doit permettre à ce dernier de faire appel à toutes ses ressources linguistiques et communicatives, les sujets traités doivent lui être familiers, le mode du discours (argumentatif, descriptif) doit lui être annoncé (Galaczi et ffrench, 2011). L'équipe administrative doit s'assurer de mettre en place de

bonnes conditions d'examen (salles adéquates, respect de la sécurité) (Anastasi, 1988; Weir, 2005a), et d'informer les candidats des critères sur lesquels ils sont évalués (AERA/APA/NCME[25], 1999). En outre, les caractéristiques linguistiques de l'examinateur/interlocuteur doivent être prises en considération (débit de la parole, accent, clarté de l'articulation, longueur du discours). La contribution de ce dernier à l'interaction doit être standardisée autant que possible (Weir, 2005a).

3) La validité conséquentielle porte sur les conséquences sociales de l'interprétation d'un test. Cette forme de validité met en lien les buts des tests de langue avec les valeurs des institutions et de la société de manière plus générale (Hawkey, 2011). La validité conséquentielle est un concept hérité de Messick (1989) qui estimait que la validité d'un instrument était non seulement liée à sa valeur intrinsèque, mais aussi à l'usage que l'on en fait et à l'effet qu'il produit. La validité conséquentielle porte également sur la notion de *washback*, c'est-à-dire l'influence d'un examen sur l'enseignement, l'enseignant, l'apprenant, l'apprentissage, les programmes et le matériel (Alderson et Wall, 1993; Weir, 2005a).

4) La validité liée aux critères concerne la mesure dans laquelle les scores des tests sont en corrélation avec des critères de performance externes adéquats ayant déjà fait leurs preuves (Khalifa et Salamoura, 2011). Par exemple lorsque l'on relie un test à un cadre de référence standard externe à titre de comparaison (Anastasi, 1988; Khalifa et Weir, 2009; Messick, 1989). La validité liée aux critères se divise en deux types : concourante et prédictive. La validité concourante implique la comparaison des scores d'un test avec une mesure extrinsèque pour des candidats ayant passé le test à peu près au même moment (Bachman, 1990). La validité prédictive implique également la comparaison des scores d'un test avec une mesure extrinsèque, mais pour des candidats ayant passé le test à des moments différents (Alderson *et al.*, 1995).

5) La validité des scores permet de garantir une certaine fidélité au test. Elle s'applique lorsque les scores sont basés sur des critères appropriés, qu'ils sont exempts d'erreurs autant que possible et qu'ils sont stables au fil du temps (Taylor et Galaczi, 2011). La validité des scores s'applique également lorsque les scores attestent d'un accord consensuel, et qu'ils inspirent confiance comme

[25] *American Educational and Research Association / American Psychological Association / National Council for Measurement in Education*

étant des indicateurs fiables de prise de décision (Khalifa et Weir, 2009; Shaw et Weir, 2007; Weir, 2005a).

Weir (2005a) propose une extension à la validité des scores. Il identifie plusieurs paramètres devant être pris en considération pour que les scores lors d'un test d'expression orale soient valides: a) les critères et les grilles d'évaluation; b) le processus d'évaluation; c) les conditions de l'évaluation; d) les caractéristiques de l'examinateur; e) la formation de l'examinateur; f) l'ajustement *post exam*; g) l'attribution de la note.

a) Les paramètres les plus fondamentaux à prendre en considération sont les critères d'évaluation qui vont de pair avec la grille d'évaluation. Plusieurs auteurs spécialisés dans le domaine des tests de L2 mettent l'accent sur l'importance cruciale de l'utilisation de grilles d'évaluation pour évaluer une performance en expression écrite et en expression orale (Alderson *et al.*, 1995; Bachman et Palmer, 1996; McNamara, 1996).

b) Le processus d'évaluation concerne la nature des prises de décision opérationnalisées par l'examinateur. Shaw et Weir (2007) insistent sur le fait que ce processus devient plus complexe lorsque l'examinateur est directement impliqué dans l'interaction avec le candidat ou lorsqu'il y a plus d'un candidat à évaluer simultanément.

c) Les conditions de l'évaluation dans lesquelles l'examinateur doit émettre son jugement sur la qualité de la performance orale du candidat peuvent être affectées par bon nombre de facteurs. Ces facteurs sont de différentes dimensions : temporelles, spatiales, psychologiques et environnementales. Par exemple, dans les tests oraux à grande échelle, les niveaux de variations liés à la nature du cadre (taille de la pièce, meubles, température ambiante) doivent être contrôlés afin d'être les plus constants possible, bien que cela soit impossible d'être dupliqué de manière absolument identique. Comme les conditions et les circonstances d'un test ont le potentiel d'avoir un impact sur le processus de l'évaluation et sur la fidélité des scores, il incombe aux concepteurs de tests de mettre en œuvre des procédures d'agencement qui soient les plus rigoureuses et les plus standards possible (Taylor et Galaczi, 2011).

d) Les caractéristiques personnelles de l'examinateur telles que son âge, son genre, sa langue maternelle, son expérience culturelle, son degré de sévérité ou de clémence, et son

style conversationnel peuvent créer un « effet examinateur » (ou un « effet interlocuteur » dans le cas où l'examinateur interagit avec le candidat). Selon Galaczi et ffrench (2011), les domaines comme la sociolinguistique, l'acquisition des L2 ainsi que l'évaluation des L2 affirment sans équivoque que de telles caractéristiques chez un individu peuvent affecter la qualité d'une interaction. Les caractéristiques individuelles de l'examinateur façonnent la manière dont celui-ci interprète et applique les critères et la grille d'évaluation, et cela se reflète dans ses pratiques évaluatives (McNamara, 1996). Comme tout être humain, l'examinateur est un individu à part entière qui apporte avec lui ses propres particularités à l'évaluation d'une tâche (Taylor et Galaczi, 2011).

e) La formation est le mécanisme par lequel l'impact des fluctuations chez les examinateurs peut être limité. Les formations sont bénéfiques, car elles rendent les examinateurs plus constants dans leurs attributions de notes. Cependant, elles ne sont pas suffisantes, car elles n'éliminent pas toutes les différences individuelles (McNamara, 1996).

f) Une procédure d'ajustement *post exam* peut être entreprise afin d'assurer l'exactitude et la fidélité des notes. Une série de vérifications statistiques peut être réalisée afin d'identifier les sources potentielles d'anomalies liées aux activités des examinateurs (aux effets des examinateurs).

g) Pour un test qui comprend plusieurs épreuves, l'attribution de la note finale doit s'assurer qu'elle représente de façon fiable la moyenne de toutes les épreuves individuelles.

Les concepts de validité et de fidélité sont complémentaires, car la fidélité est une condition sine qua non de la validité. Si les résultats d'un test sont mal interprétés, ou s'ils sont impossibles à interpréter, le test est conséquemment « non valide », quelle que soit la rigueur avec laquelle il a été conçu. Ainsi, un test valide est nécessairement fidèle alors qu'un test fidèle n'est pas nécessairement valide (Bachman, 1990; Genesse et Upshur, 1996).

Les cinq types de validité que nous venons d'énoncer sont essentiels à un bon fonctionnement d'un test reposant sur des réponses orales construites. Par ailleurs, la validité repose également sur la connaissance de la cognition de l'évaluateur pour plusieurs raisons. Tout d'abord, parce que la cognition de l'évaluateur se doit d'être compatible avec le construit (AERA/APA/NCME, 1999), ensuite parce qu'elle influence le développement des tests, étant donné que la variation de la note

dépend du traitement cognitif de l'évaluateur (Purpura, 2013), et enfin parce que le sens et les inférences fondées sur les scores attribués se doivent d'être justifiés (Bejar, 2012; Kane, 2006). Nous traiterons donc de la cognition de l'évaluateur dans la section suivante.

2.5. La cognition de l'évaluateur

Dans cette section, nous définirons tout d'abord le concept de jugement dans le champ de l'évaluation. Bien que ce concept soit très large, nous proposerons des définitions non exhaustives compatibles avec notre thème de recherche. Nous mettrons ensuite en lumière l'avènement de la recherche traitant de la cognition des évaluateurs, puis décrirons la manière dont les chercheurs ont tenté de la conceptualiser à travers différentes approches évaluatives et différents modèles. Nous évoquerons également la technique de la pensée à voix haute qui est une technique permettant de saisir l'activité cognitive des évaluateurs.

À l'instar de Bejar (2012), nous englobons ici dans la notion de cognition de l'évaluateur/examinateur deux dimensions principales : ses caractéristiques (caractéristiques intrinsèques et approches en matière d'évaluation) et ses processus mentaux.

2.5.1. Le jugement des évaluateurs

Toute évaluation est, comme l'étymologie le rappelle, une référence à une valeur et fondamentalement, l'évaluation consiste à porter un jugement de valeur sur un quelconque objet. Le jugement humain est toujours une comparaison d'un élément avec un autre. Dans un contexte d'examen, l'on compare un travail avec un modèle théorique, cela implique alors qu'une sorte de point de référence soit nécessaire pour émettre un jugement (Laming, 2004).

Le jugement ne peut constituer un but en soi, il est produit en vue d'un usage social qui se matérialise par une décision ou un ensemble de décisions. De ce fait, on peut considérer que le véritable produit de l'évaluation est la décision finale (y compris une possible non-décision). Cette interprétation est soutenue par De Ketele et Roegiers (1993) pour qui évaluer signifie recueillir des informations suffisamment pertinentes, valides et fiables, puis examiner le degré d'adéquation entre ces informations et des critères adéquats aux objectifs fixés au départ (ou ajustés en cours de route) en vue de prendre une décision.

Scriven (1980) a été le premier théoricien multidisciplinaire à s'investir dans la recherche d'une définition des composantes qui caractérisent l'acte spécifique d'évaluer. Il propose une métathéorie de l'évaluation qui se décline en quatre opérations :

1. Établir les critères pour connaître les éléments ou composantes à évaluer. Ces critères sont les aspects, les attributs ou les dimensions qui caractérisent la valeur de l'objet d'évaluation. Ils servent de point de référence au jugement;

2. Établir les seuils de performance attendus pour chacun des critères. Ils constituent en quelque sorte le point de coupure entre ce qui constitue un jugement favorable ou défavorable;

3. Mesurer la performance (à l'aide d'indicateurs observables et mesurables) et la comparer aux seuils de performance établis. Il s'agit donc de l'analyse des données recueillies;

4. Synthétiser et intégrer les données afin d'être en mesure de porter un jugement sur la qualité ou le mérite de l'objet d'étude. C'est l'étape qui se trouve au cœur de la démarche évaluative.

Hurteau (2013) affirme que la quatrième opération de la métathéorie de l'évaluation de Scriven (1980)[26], soit la transformation des données en un jugement, sous-tend le recours à une argumentation. Selon elle, le modèle du philosophe Toulmin (2003) (Figure 10) peut constituer une référence adéquate, car ce dernier propose une perspective moderne de l'argumentation. Dans son modèle, Toulmin met en avant trois composantes principales qui sont les suivantes :

- Les données qui constituent en quelque sorte les motifs, les raisons, les preuves sur lesquels la déclaration s'appuie, la conclusion;
- La justification, la garantie, qui constitue d'une certaine façon une loi de passage entre les données et la déclaration en explicitant la nature du lien;
- La déclaration, la conclusion (*claim*) que l'on estime vraie.

Les trois autres composantes, qui ne sont pas toujours présentes, permettent de consolider et d'étayer l'argumentation.

- Les modalités qui précisent les conditions particulières à respecter pour que la déclaration, la conclusion, puissent être considérées comme vraies;

[26] Bien que les travaux de recherche entrepris par Scriven (1980) et Hurteau (2013) s'inscrivent en évaluation de programme, ils nous permettent de mettre en avant certains principes fondamentaux de la pratique de l'évaluation.

- Les restrictions qui signalent les éventuelles exceptions qui ne permettent pas à la déclaration d'être vraie;
- Les fondements qui viennent en renfort à la justification lorsque celle-ci est trop faible. Ils constituent pour ainsi dire la « structure profonde » du raisonnement en s'appuyant sur des textes et des savoirs reconnus.

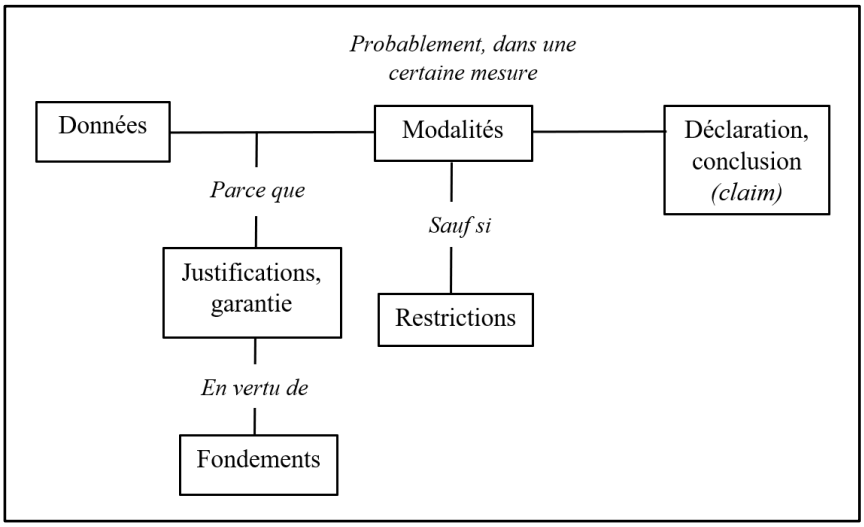


Figure 10 - Modèle d'argumentation de Toulmin (2003)

D'autres chercheurs se sont intéressés au jugement. Pour Goasdoué et Vantourout (2017), le jugement ne peut se résumer à un simple raisonnement algorithmique. Juger implique toute une « informatique cognitive » qui résulte d'opérations sophistiquées. Démontrer en mathématiques n'est pas argumenter en philosophie, comme lire un texte juridique n'est pas proposer une interprétation littéraire. Ces savoirs, du fait de leurs caractéristiques, ne suscitent pas nécessairement les mêmes raisonnements évaluatifs et reposent sur des critères de légitimité différents. Les raisonnements évaluatifs dépendent étroitement des savoirs évalués.

Par ailleurs, dans plusieurs domaines comme les tests de L2, l'évaluation de programmes, les sciences de la santé et de l'éducation, il a été largement reconnu que les expériences professionnelles antérieures, le parcours de formation, et même la représentation que l'évaluateur se fait de son métier jouent un rôle non négligeable dans la construction des pratiques et des représentations de l'évaluation. Les différentes situations que l'évaluateur a vécues participent en grande partie à la construction d'un répertoire de ressources étendu et diversifié (e.g., Brown, 2005; Cumming, 1990; Gauthier *et al.*, 2016; Hurteau *et al.*, 2012; Huver et Springer, 2011; Jorro, 2000; Laming, 2004;

Romainville, 2011; Savoie-Zajc, 2013; Schleifer et Hurteau, 2012). À ce sujet, Grinnell (2009) décrit les paradigmes en sciences comme un ensemble de croyances et de valeurs partagées par les membres de la communauté scientifique comme une façon acceptable et établie de résoudre les problèmes. Selon le chercheur, le jugement scientifique dépend de l'expérience personnelle et de la personnalité du chercheur. Lipman (1992) partage la même position en déclarant que le jugement est une représentation de celui qui le pose, car il reflète la vision du monde de ce dernier et son impression d'une réalité ou d'une situation.

Le jugement s'élabore au sein d'une démarche cognitive et fait partie d'un processus complexe composé de plusieurs opérations diverses et réunies entre elles. Parmi ces opérations, on retrouve entre autres des conceptualisations, des questionnements, des perceptions et des intuitions (Angers, 2010). D'après Lumley (2002, 2005), la tâche de l'évaluateur est de concilier son impression des données recueillies (en prenant en compte leurs caractéristiques spécifiques) avec les libellés de la grille d'évaluation afin de produire un ensemble de scores. Mais lorsque l'évaluateur doit traiter un travail atypique, c'est-à-dire qui ne peut être traité avec la grille d'évaluation, cela le force à développer ses propres stratégies de jugement. Ce faisant, l'évaluateur est fortement influencé par sa première impression du travail qu'il doit évaluer. Grainger, Purnell et Kipf (2008) sont du même avis. Ces chercheurs ont étudié le comportement de nombreux professeurs d'université en éducation et affirment que ces derniers évaluent de manière intuitive l'ensemble d'un travail en premier lieu, et l'adaptent en second lieu aux critères de leur propre grille d'évaluation. D'après Stake et Schwandt (2006), le jugement évaluatif devrait s'appuyer sur des informations qui relèvent autant de la légitimité scientifique que des perceptions. Ces chercheurs introduisent les notions de qualité mesurée (*quality as measured*) et de qualité appréciée (*quality as experimented*). La première notion réfère aux données quantitatives et qualitatives qui documentent la démarche évaluative et qui contribuent ainsi à la légitimité scientifique. La deuxième notion réfère aux préoccupations, opinions et attitudes de l'ensemble des détenteurs d'enjeux, et émerge en cours de processus. La connaissance de l'évaluateur demeure donc un sujet très complexe et est non « réductible à une formule » (Sadler, 1989, p. 124).

2.5.2. 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. Fechner (1897) 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 était 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 XXe siècle au concept du *lens model* (modèle de lentille) de Brunswik (1952). Ce concept consiste à attribuer à l'individu la capacité à reconnaître une sélection hétérogène et complexe dans les intrants et extrants, en établissant de nouveaux foyers, ou simplement en ignorant certains aspects. Cela souligne le fait que chaque personne qui évalue peut potentiellement 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 grâce aux travaux du statisticien Edgeworth (1890) qui a examiné les éléments de hasard pouvant affecter un score. Edgeworth (1890) était très conscient du problème de l'accord inter juge, et a reconnu que les différences individuelles parmi les évaluateurs pouvaient être « source d'erreur[27] », en remarquant par exemple que certains pouvaient être plus sévères et d'autres plus cléments.

Durant une grande partie du XXe siècle aux États-Unis, le problème de l'accord inter juge a retardé l'utilisation de tests à grande échelle nécessitant des réponses construites à l'écrit comme 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*[28]. 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 large 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 ouvert la voie à de nombreux tests d'admission à choix multiples dans les années 1920 (Fuess, 1950).

[27] Dans la théorie de la mesure, un manque d'accord parmi les évaluateurs ou une instabilité chez les évaluateurs constitue une « source d'erreur ». Celle-ci fait référence à une partie du score du candidat qui dévie du « vrai score ».
[28] *The College Board's English Competence Examination* était destiné aux étudiants non américains désirant poursuivre des études dans les universités aux États-Unis.

Étant donné que la compréhension de la cognition de l'évaluateur était à ce moment-là intuitive, résoudre le problème de l'accord inter juge 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 argument de validité (Kane, 2006).

En 1961, l'absence d'accord entre les examinateurs a été mise en avant à travers une étude majeure menée par *Educational Testing Service*[29]. 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, et par conséquent, la fidélité inter-examinateurs s'est avérée très pauvre. Les examinateurs, se retrouvant 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 traitant de la même question avec des conclusions similaires, mais dans le contexte plus spécifique d'un test oral (via une entrevue) en L2 a été menée en 1983 par Shohamy. Cette étude quantitative a mis en avant 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 L2 (Brown, 2003). Les études subséquentes ayant traité des effets des examinateurs-évaluateurs dans ce domaine ont déjà été documentées dans notre problématique.

2.5.3. Vers un modèle théorique de la cognition de l'évaluateur en langue seconde

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 (e.g., Crisp, 2008, Crisp 2010; Freedman et Calfee, 1983; Pula et Huot, 1993; Sanderson, 2001; Vaughan, 1991; Wolfe, 1997), de la lecture (Crisp, 2012) 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

[29] *Educational Testing Service* est une grande organisation privée américaine de mesure et d'évaluation en éducation fondée en 1947.

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 à partir de recherches approfondies en psychologie cognitive et en neurosciences, où la technologie de balayage cérébral est utilisée 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[30]) sont impliqués dans le codage[31], le stockage et la récupération[32] des informations lors de la notation. De manière complémentaire, les stratégies (méta)cognitives, qui sont 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 écrite en L2, un nombre restreint d'études a tenté de présenter le processus cognitif des évaluateurs en prenant en compte les stratégies (méta)cognitives pertinentes impliquées dans la notation de dissertations (e.g., Barkaoui, 2007, 2010; Cumming, 1990; Cumming *et al.*, 2002; Lumley, 2002; Milanovic *et al.*, 1996; Sakyi, 2000; Smith, 2000). Bien que ces études dressent un état des lieux en répertoriant une gamme d'activités mentales, les modèles du traitement de l'information comme cadre de référence demeurent absents (Dehn, 2008; Han, 2016; Purpura, 2013).

Dans le domaine des tests d'expression orale en L2, un nombre très restreint de chercheurs ont tenté de conceptualiser la cognition de l'examinateur. Ces chercheurs se sont penchés sur les

[30] La mémoire à court terme correspond à la rétention temporaire de l'information en cours de traitement. La mémoire de travail est une partie de la mémoire à court terme, elle permet de réaliser des manipulations cognitives sur des informations maintenues temporairement. C'est un système de mémoire transitoire impliquant simultanément les opérations de stockage et les opérations de traitement (comprendre une phrase ou la construire, calculer de tête). La mémoire à long terme permet de retenir, de manière illimitée, une information sur des périodes très longues.
[31] Le codage est à la phase d'acquisition des informations.
[32] La récupération est le processus au moyen duquel le sujet retrouve et restitue les informations mémorisées.

concepts suivants : les approches évaluatives des examinateurs (Pollitt et Murray, 1996; Reed et 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).

2.5.3.1. Les approches évaluatives des examinateurs de Pollitt et Murray (1996)

À travers une étude empirique dans le cadre du test d'anglais L2 *Certificate of Proficiency in English* (CPE), Pollitt et Murray (1996) 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 la première approche évaluative utilisant un procédé dit « synthétique », une image holistique du candidat est formée, et cette image est dérivée au préalable d'une compréhension individuelle du candidat qui est préconçue et préconstruite. Cela est semblable à une première rencontre d'une personne inconnue lors d'un événement 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 la performance observée est alors comparée avec celle d'un autre candidat du même niveau que l'examinateur a gardé en mémoire. La deuxième approche évaluative, quant à elle, utilise un procédé dit de « puzzle » où l'examinateur limite ses commentaires au comportement observé du candidat. La pratique consiste à noter les candidats d'après chaque énoncé observé, puis de nouvelles informations s'ajoutent au fur et à mesure. Il s'agit d'un mode plus objectif, mais moins naturel qui requiert un plus grand effort, car l'examinateur se doit de réfléchir dans un cadre strictement évaluatif.

2.5.3.2. Les approches évaluatives de Reed et Cohen (2001)

Étant une tâche ardue, l'évaluation de l'expression orale en situation de test requiert des inférences et des résolutions d'incertitudes. Reed et Cohen (2001) illustrent, à travers la figure 11, 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 que la performance qu'il observe est influencée par l'attitude du candidat, par ses caractéristiques linguistiques (sa langue maternelle, son niveau dans la langue cible), et par les caractéristiques de la tâche (le contenu, le niveau de difficulté). De surcroît, l'examinateur doit avoir un degré de « conscience de soi » sur sa propre tendance à être sévère ou clément, ainsi que sur sa tendance à

se laisser porter vers une 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'opération d'évaluation devient alors une construction de sens qui nécessite de multiples résolutions d'incertitudes de façon synchronisée.

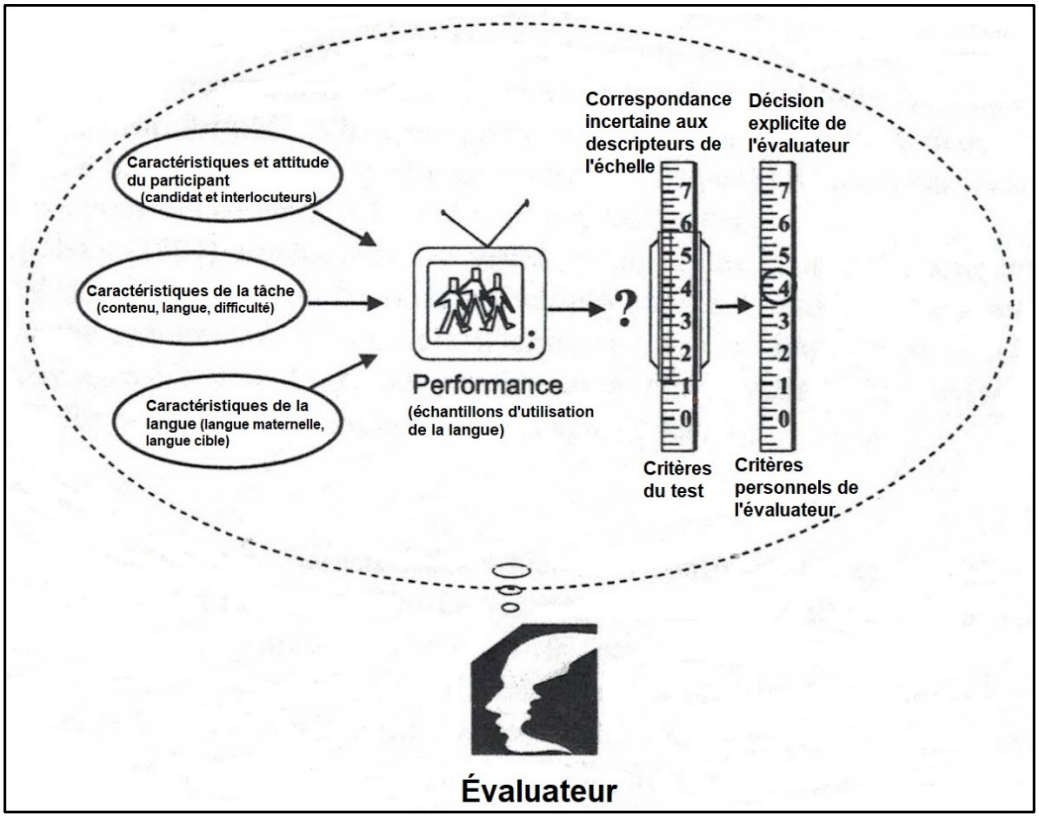


Figure 11 - L'examinateur utilisant les critères du test et ses critères personnels

Source : Reed et Cohen, 2001

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

À travers un modèle descriptif du processus d'évaluation (Tableau 9), Bejar (2012) tente de mettre en lumière les aspects de la cognition de l'évaluateur. Son modèle est basé sur la recherche, et préconise entre autres des pratiques en matière de notation. Par ailleurs, il est générique, c'est-à-dire qu'il n'est pas propre à un domaine particulier.

Dans le tableau, l'auteur distingue deux phases : une première phase de conception de l'évaluation et une deuxième phase de notation. Dans la phase de conception de l'évaluation, il souligne qu'il faut être attentif aux considérations cognitives des évaluateurs à ce stade, car il est important que ces derniers soient capables de bien comprendre les grilles d'évaluation (Joe *et al.*, 2011), et surtout qu'ils y adhèrent (Wolfe *et al.*, 1998). Même lorsque les critères de notation sont adaptés au construit et qu'ils sont acceptés des évaluateurs, ils peuvent tout de même être difficiles à saisir. Dans ce cas, l'interprétation du score qui est proposée risque d'être invalide.

Dans la phase de conception de l'évaluation, l'auteur ajoute que la formation des évaluateurs peut être en partie conceptualisée en amenant l'évaluateur à encoder[33] le matériel de formation dans ce que l'auteur appelle « une grille d'évaluation mentale ». Celle-ci est un processus cognitif qui permet à l'évaluateur d'inférer une représentation vigoureuse 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 qui soit gérable en indexant sa « grille d'évaluation mentale » uniquement sur ses connaissances de base. La « grille d'évaluation mentale » est propre à chaque évaluateur, et elle peut, par conséquent, être influencée par ses attributs personnels et ses antécédents. Toutefois, les attributs personnels et les antécédents risquent de conduire à une représentation qui contient par inadvertance des composantes du construit non pertinentes, des biais[34] ou des composantes qui ne sont pas explicitement mentionnées dans la grille.

Dans la seconde phase, qui est la phase de notation, l'évaluateur forme « une représentation de réponse mentale » de la réponse de la personne évaluée qui est ensuite mise en parallèle avec sa grille mentale. Ce processus peut prendre une multitude de formes selon les chercheurs ayant déjà étudié la question. Par exemple, pour Cumming (1990), la notation est vue comme un processus de comparaison entre une représentation de sens (construite à partir du travail de la personne évaluée) et une représentation mentale (construite à partir d'un référentiel ou à partir d'une représentation mentale déjà existante de ce que devrait être un travail idéal). Selon Bejar, c'est un processus qui, sommairement, se penche sur des représentations de similitude et de probabilité dans l'esprit de l'évaluateur. En pratique, le processus utilisé par les évaluateurs pour attribuer une note peut être

[33] Encoder signifie percevoir de nouvelles informations et les transformer en des représentations significatives.
[34] Le concept de biais sera défini ultérieurement.

plus complexe puisque celui-ci peut introduire des informations antérieures (observations du passé, biais cognitifs) à long terme et à court terme dans le processus de notation.

Tableau 10 - Modèle descriptif du processus d'évaluation mettant en avant la cognition de l'évaluateur

Phase de conception de l'évaluation	**Phase de notation**

• Le processus de conception de l'évaluation identifie les preuves des différents niveaux de performance auxquels on fait appel.
• **Les grilles d'évaluation** sont formulées afin de formaliser les échelons pertinents de la performance.
• Les items et les tâches sont prétestés et évalués, est-ce qu'ils recueillent les preuves demandées?
• Si oui, collectez **les références et les instruments de mesure** afin de former les évaluateurs.
• Recrutez et formez les évaluateurs en utilisant la grille d'évaluation, les références et les instruments de mesure.
• Les évaluateurs forment **une grille d'évaluation mentale** basée sur la formation.
• Idéalement, à la suite de la formation, tous les évaluateurs vont évaluer de façon équivalente, mais leurs antécédents ou d'autres facteurs pourraient mener à **des effets de l'évaluateur**.
	• L'évaluateur prend connaissance d'un travail et forme **une représentation de réponse mentale**.
	• L'évaluateur compare **la similitude** de cette représentation avec la grille d'évaluation mentale.
	• S'appuyant sur cette comparaison, l'évaluateur, après réflexion, attribue à la réponse une catégorie de note.
	• La note qu'il attribue à une réponse spécifique dépend de :
	  • La véritable qualité de la réponse,
	  • La qualité de sa grille d'évaluation mentale,
	  • La qualité de la représentation qu'il a formée pour ladite réponse,
	  • Les **informations antérieures** qu'il a accumulées durant la notation,
	  • L'état de l'évaluateur, notamment la fatigue,
	  • Les conditions environnementales,
	  • La nature des réponses notées précédemment.

Source : Bejar, 2012

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

Han (2016) propose un modèle unifié sur la nature de la cognition de l'examinateur dans le contexte spécifique des tests d'expression orale en L2. Elle prend principalement appui sur le modèle de Bejar (2012) ainsi que sur deux modèles de Purpura (2012) adaptés spécifiquement pour les L2 : l'un sur l'architecture du traitement humain de l'information des candidats (Figure 13 en annexe), et l'autre sur les compétences et les stratégies cognitives des candidats à chaque étape du traitement de l'information (Figure 14 en annexe). Le modèle de Han s'appuie entre autres

sur les nombreuses données empiriques issues de la recherche sur les effets des examinateurs, sur certains cadres et théories récents concernant le processus de notation en L2, et sur l'architecture du traitement humain de l'information. 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.

Le modèle de Han (Figure 12) établit une interface entre le processus d'évaluation de l'expression orale, les composants de l'architecture du traitement humain de l'information qui sont activés, et les processus cognitifs des examinateurs qui sont invoqués à chaque étape de la notation. Le modèle intègre également un large éventail de stratégies (méta)cognitives que les examinateurs peuvent déployer pour réguler le fonctionnement de leurs mécanismes cognitifs lors du processus de notation. Ce modèle délimite la manière dont les intrants de l'évaluation (c'est-à-dire la grille d'évaluation, les exemplaires[35] et les réponses orales en L2) pourraient être captés par les récepteurs sensoriels et pris en charge de manière sélective, puis initialement traités dans la mémoire à court terme. Le modèle délimite ensuite la manière dont la mémoire de travail décode et encode les informations des intrants de l'évaluation, la manière dont elle récupère et active différents types de connaissances à partir de la mémoire à long terme, de sorte que tous les types d'information puissent être réorganisés et que les représentations mentales de la grille d'évaluation et les réponses en L2 puissent être formées.

Par la suite, le modèle décrit la manière dont ces représentations mentales sont comparées et mises en contraste pour produire des scores provisoires dans la mémoire de travail. On retrouve la manière dont les scores sont examinés ou révisés et justifiés à travers un processus de notation itératif utilisant toutes les composantes cognitives (la mémoire sensorielle[36], la mémoire à court terme, la mémoire de travail, la mémoire à long terme). Toutes ces étapes sont régies par une gamme de stratégies (méta)cognitives (par exemple, le traitement, le contrôle) qui sont invoquées et soumises aux influences de diverses caractéristiques inhérentes à l'examinateur (comme son expérience en

[35] Han (2016) emprunte le terme « exemplaire » de Davis (2012) qui signifie un échantillon de discours utilisé à titre d'exemple représentant les caractéristiques typiques des réponses pour chaque niveau de la grille d'évaluation. Les exemplaires sont généralement utilisés pour les formations durant lesquelles les évaluateurs les écoutent afin d'avoir une idée de ce à quoi devrait ressembler une réponse typique pour un niveau donné.

[36] La mémoire sensorielle garde pendant un très court laps de temps l'information sensorielle, c'est-à-dire, les sons, les images, les odeurs.

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 de diverses variables environnementales.

Avant d'aborder les variables environnementales impliquées dans la note, nous ouvrons ici une parenthèse afin d'évoquer la technique de la pensée à voix haute qui est une technique méthodologique permettant de fournir des indices de l'activité cognitive des évaluateurs.

2.5.4. La cognition de l'évaluateur via la technique de la pensée à voix haute

De récentes études ont commencé à montrer un intérêt pour l'application de modèles de traitement cognitif des évaluateurs comme la technique de la pensée à voix haute (*think aloud protocol*). Étant donné que les approches statistiques ne permettent pas de bien saisir le processus de prise de décision, la technique de la pensée à voix haute s'avère être efficace, car elle permet de mieux comprendre comment l'évaluateur attribue ses scores et peut ainsi expliquer les effets des évaluateurs (Brown, 2000). Cette technique implique l'usage d'enregistrements audio ou vidéo de comportements professionnels qui sont ensuite utilisés pour aider le participant-évaluateur à se souvenir des pensées qu'il a eues au moment d'agir. Les activités de verbalisation sont enregistrées, transcrites, puis analysées pour identifier les processus de prise de décision employés et les aspects les plus marquants (Green, 2009).

L'utilisation de la technique de la pensée à voix haute a largement été utilisée afin d'étudier et d'élaborer des modèles de processus d'évaluation de productions écrites en L2 (e.g., Barkaoui, 2011; Cumming *et al.*, 2001, 2002; Lumley, 2002, 2005; Sakyi, 2000; Weigle, 1999), mais un nombre limité d'études portant sur l'évaluation de productions orales en L2 ont adopté cette méthode. Par exemple, une étude de Brown (2000) explore les pratiques évaluatives des examinateurs lors des entrevues orales du test IELTS. Une étude de Orr (2002) s'intéresse au processus qui mène à la prise de décision de la notation des examinateurs du test d'anglais FCE *Speaking test* (*First Certificate in English*). Brown, Iwashita et McNamara (2005) analysent les aspects sur lesquels les examinateurs se penchent lors de l'évaluation des performances orales du test *English for academic purposes* (EAP). Brown (2006) tente de comprendre la manière dont les examinateurs du test IELTS interprètent les grilles d'évaluation et la manière dont ils les utilisent. Ang-Aw et Goh (2011) explorent la nature et la portée des divergences chez les évaluateurs du test *Singapore-Cambridge General Certificate of Education Ordinary Level* destiné aux élèves du secondaire dont la langue maternelle est ou n'est pas l'anglais. Tanrıverdi-Köksal et Ortaçtepe

(2017), quant à eux, étudient l'influence des connaissances antérieures des niveaux de compétence des étudiants sur les comportements évaluatifs des enseignants-évaluateurs lors des entretiens oraux en anglais L2 en milieu universitaire.

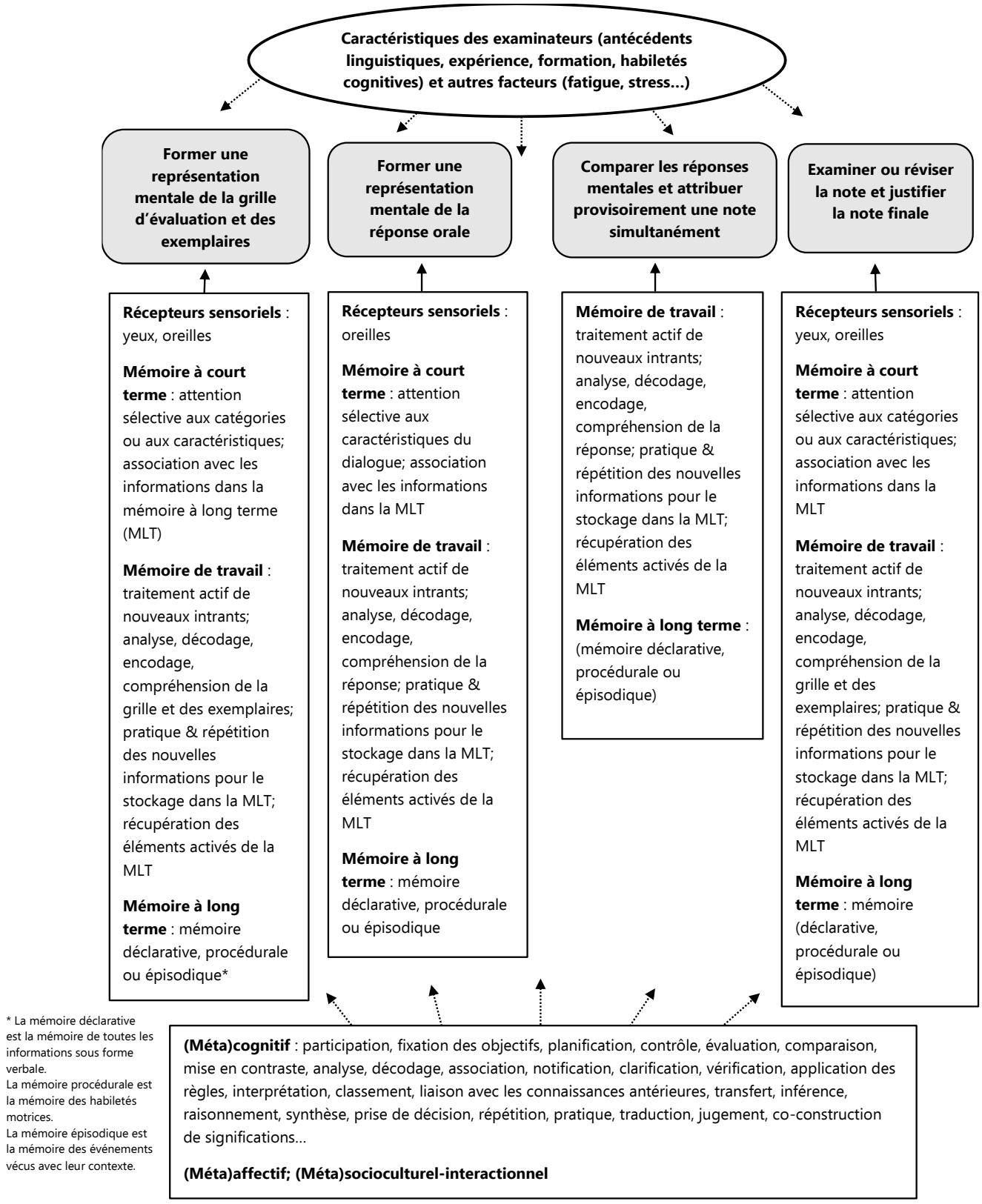


Figure 12 - Modèle hypothétique du processus cognitif de l'évaluation des réponses orales en langue seconde

Source : Han, 2016

2.6. Les variables impliquées dans la note

Les variables environnementales sont très diversifiées, elles interfèrent indirectement dans la note finale du candidat, tout comme de nombreux autres biais. Nous les décrirons dans les lignes qui suivent.

2.6.1. L'environnement du test

Lors d'un test, les notes attribuées aux candidats durant une performance peuvent être affectées par des combinaisons de multitudes variables imprévisibles issues de différentes sources. Ces variables sont par exemple les compétences du candidat, la tâche qui met en œuvre la performance du candidat, les conditions de sa performance, etc. (Fulcher, 2003; Kenyon, 1992; McNamara, 1995; Skehan, 1998, 2001). À propos des conditions d'un test, Powers *et al.* (2003) ont mené une étude sur les effets du bruit lors d'une passation de test. Les chercheurs ont analysé deux groupes de candidats qui passaient simultanément deux épreuves différentes dans des locaux voisins : un groupe qui passait une épreuve orale semi-directe via un système automatisé et qui parlait à voix haute avec des microphones, puis un groupe qui passait une épreuve de lecture. Les résultats ont montré, sans surprise, que les conditions du test perturbaient l'effort de concentration des candidats qui passaient l'épreuve de lecture. Leur niveau d'anxiété était si élevé que cela a eu un impact négatif sur leurs notes. À ce propos, selon les standards *Standards for Educational and Psychological Testing* (1999) (des organismes AERA, APA et NCME), il est stipulé que dans des situations de test, le bruit, les perturbations dans les locaux, les températures extrêmes, le manque de luminosité, les espaces de travail inappropriés, et le matériel illisible sont à éviter.

À travers un cadre conceptuel (Figure 13), Milanovic et Saville (1993) illustrent les variables impliquées dans la note finale du candidat lors d'une situation de test d'expression écrite ou orale. Ces variables sont les suivantes : a) Les conditions de l'examen : l'endroit, la durée, le moment; b) Les tâches : leurs formats (jeu de rôle, dissertation), les thèmes des sujets traités et le temps de préparation; c) Les critères d'évaluation : leur nombre et leur nature; d) Les conditions d'évaluation : le nombre d'examinateurs, l'interaction (face à face avec l'examinateur, interaction avec d'autres candidats); e) La formation : celle que le concepteur de l'examen fait suivre aux examinateurs; f) Les candidats : leur âge, sexe, nationalité, langue maternelle, personnalité, etc.; g) Les examinateurs : leur expérience professionnelle, formation antérieure, personnalité, familiarité avec la langue ou l'origine du candidat, etc.

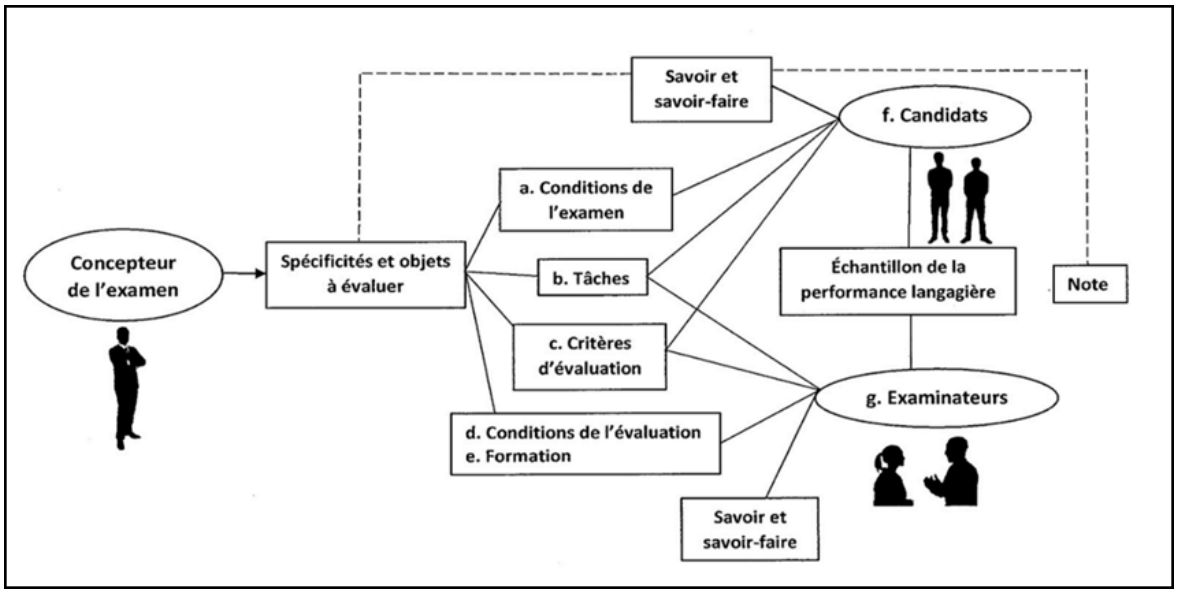


Figure 13 - Les variables impliquées dans la note finale du candidat

Source : Milanovic et Saville, 1993 (Schéma adapté)

2.6.2. Les biais

Dans une situation de test oral, certains biais sont également susceptibles de déformer la notation. Une mesure biaisée est une mesure qui conduit systématiquement à des résultats inexacts dus à des défectuosités inhérentes à l'instrument utilisé, à la façon d'utiliser un instrument ou d'interpréter les résultats (De Landsheere, 1992).

Les études scientifiques portant sur les biais se sont développées dans des contextes de tests à enjeu élevé dans lesquels les questions de fidélité et d'équité étaient primordiales (Zumbo, 2007). Les premières méthodes de détection de biais sont apparues durant les années 1960-1970. À cette époque, les chercheurs parlaient de détection de l'incidence d'item (situation où un item serait plus difficile pour un groupe d'individus que pour un autre) ou de biais d'item (certains facteurs, tels qu'une mauvaise formulation de la question, provoquent une différence de difficulté entre les groupes d'individus) (Magis *et al.*, 2010). Par la suite, est apparue la locution statistique « fonctionnement différentiel d'items » (FDI) qui se définit comme une situation où des groupes de personnes (par exemple, selon le genre, la langue maternelle, l'âge) ayant la même habileté estimée ne répondent pas de la même façon à un item. Plusieurs chercheurs ont entrepris des études au sujet du fonctionnement différentiel d'items dans le cadre des tests de langues. Par exemple, Kunnan (1990) a analysé les résultats obtenus à un test de classement en anglais L2 et des

différences liées aux origines culturelles et au sexe des répondants ont été mises en évidence. Une autre étude de Pae (2004) sur les résultats d'étudiants coréens à un test de compréhension écrite en anglais L2 a montré que les items traitant des émotions tendaient à être mieux réussis par les filles, alors que les garçons réussissaient mieux avec les items à contenu logique.

Lorsqu'un test est mis sur pied, il peut toujours subsister des items qui risquent de favoriser un groupe d'individus plutôt qu'un autre. Il est alors fortement suggéré de prendre en considération les recherches effectuées sur le sujet, puis de tenter d'identifier et d'éliminer les items à risque afin de présenter un instrument qui soit le plus juste possible (Pichette *et al.*, 2011).

Dans une situation de test oral, il existe des effets parasites venant biaiser le processus de notation. Jorro (2000) et Tagliante (2005) en ont relevé certains qui s'organisent autour de la typologie suivante : les effets dus à l'organisation de l'épreuve, les effets dus aux candidats, et les effets dus aux pratiques éducatives et à la personnalité de l'examinateur.

1. Les effets dus à l'organisation de l'épreuve :
   - L'effet d'ordre : l'examinateur peut être plus sévère à la fin de la session qu'au début, ou plus clément après une pause;
   - L'effet de contraste : lorsqu'un candidat de niveau débutant passe directement après un candidat de niveau avancé;
   - L'effet de fatigue[37] : l'examinateur est moins performant après avoir évalué une longue liste de candidats;
2. Les effets dus aux candidats :
   - L'effet de halo : lorsque l'examinateur est influencé par sa première impression du candidat, lorsqu'il le surnote si ce dernier est très souriant et plein d'entrain, bien habillé, etc., et à l'inverse, lorsque l'examinateur le sousnote car il est distant et réservé par exemple;
   - L'effet de contamination : lorsque l'examinateur a tendance à surnoter un candidat ayant un niveau avancé, ou sousnoter un candidat ayant un niveau bas.
3. Les effets dus aux pratiques éducatives et au niveau de sévérité de l'examinateur :

[37] Ling *et al.* (2014) ont observé la fatigue des examinateurs lors de l'épreuve orale du test d'ALS TOEFL. Ils déclarent que bien que les examinateurs se fatiguent en moyenne au bout de six heures, ces derniers préfèrent travailler huit heures, car cela correspond à une journée de travail à temps plein et est plus avantageux financièrement.

- L'effet choc positif : lorsque l'examinateur augmente la note du candidat, car ce dernier a une seule idée lumineuse dans une production qui est moyenne, voire médiocre;
- L'effet choc négatif : lorsque l'examinateur tolère les nombreuses erreurs puis baisse la note du candidat, car ce dernier fait la même erreur toutes les trois phrases;
- L'effet « goutte d'eau » : lorsque l'examinateur tolère les nombreuses erreurs tout au long de la production, et que la vingtième erreur fait « déborder le vase »;
- L'effet de stéréotypie : lorsque les évaluations menées en début de formation constituent une référence pour l'examinateur, qui par la suite, a des difficultés à évaluer différemment;
- L'effet de flou : lorsque l'examinateur ne comprend pas très bien les critères d'évaluation car ceux-ci ne sont pas définis avec précision;
- L'effet de surnotation : lorsque l'examinateur est clément et qu'il attribue facilement des notes élevées;
- L'effet de sousnotation : lorsque l'examinateur est sévère et qu'il attribue difficilement des notes élevées;
- L'effet de tendance centrale : lorsque l'examinateur n'exploite pas tous les échelons de la grille d'évaluation, lorsqu'il n'utilise jamais le niveau le plus bas ou le niveau le plus haut, de peur de passer pour un examinateur trop clément ou trop sévère.

Ces effets parasites jouent sur l'objectivité et menacent de déformer le jugement. Afin de réduire, voire d'éliminer ces biais, les examinateurs doivent en prendre conscience. Ils doivent savoir qu'ils existent et apprendre à reconnaitre les situations propices à leur apparition, ce travail nécessite de prendre du recul sur ses propres pratiques (Tagliante, 2005). Minimiser les nombreux effets de la subjectivité se révèle donc être une opération difficile étant donné que des êtres humains évaluent des phénomènes humains. Le recours à la grille d'évaluation permet alors de canaliser une partie de cette subjectivité.

2.7. Les grilles d'évaluation

Dans cette section, nous traiterons de l'utilité des grilles d'évaluation en présentant différents types. Nous évoquerons ensuite les principes reposant sur la construction d'une grille efficace ainsi que des approches utilisées provenant des principaux référentiels en L2 : *Proficiency Guidelines* et le CECRL. Enfin, nous présenterons l'actuelle grille d'évaluation de l'épreuve d'expression

orale du TEF, qui servira de support dans notre recherche, en mettant en évidence les différences entre l'ancienne et la nouvelle version.

2.7.1. L'utilité des grilles d'évaluation

Les grilles d'évaluation permettent de porter un jugement sur la qualité d'une production qui ne peut être jugée tout simplement bonne ou mauvaise comme dans le cas d'une question à correction objective (Scallon, 2004). Davies *et al.* (1999, p. 153-154) définissent les grilles d'évaluation pour la compétence langagière de manière suivante :

> Une grille d'évaluation permettant de décrire la compétence langagière est composée d'une série de niveaux établis sur lesquels la performance d'un apprenant est jugée. À l'instar d'un test, une grille fournit une définition opérationnelle d'un construit linguistique telle une compétence. […] Les niveaux sont communément caractérisés en fonction de ce que les sujets sont capables de faire avec la langue (réalisation de tâches et de fonctions) et en fonction de leur maitrise des composantes de la langue (vocabulaire, syntaxe, fluidité et cohésion).

Les grilles d'évaluation ont été créées pour répondre à un besoin d'uniformité dans la mesure des compétences (de Fonteney, 1991 ; Kaplan, 1991) en permettant aux évaluateurs de ne pas laisser les résultats au hasard. En effet, il a été démontré que les grilles d'évaluation accroissent la constance des jugements (Bendig, 1953; Fulcher, 2000; Gronlund, 1985; Kaplan, 1991; Matell et Jacoby, 1971), car elles permettent à l'évaluateur de porter un « jugement guidé ». « L'expression « jugement guidé » décrit la situation dans laquelle on met en œuvre une approche de l'évaluation qui transforme l'impression en jugement raisonné. » (Conseil de l'Europe, 2001, p. 143). En l'absence de grille d'évaluation, la dispersion des notes est beaucoup plus importante. Les évaluateurs, livrés à eux-mêmes, ont de fortes chances de mal maitriser, voire pas du tout, les aspects à prendre en compte dans l'évaluation et même à définir des critères non pertinents (Bøhn, 2015; Goasdoué et Vantourout, 2017; Roegiers, 2004).

2.7.2. Les deux types de grilles d'évaluation

Il existe plusieurs types de grilles d'évaluation, mais celles utilisées le plus couramment pour évaluer la compétence langagière sont celles dites holistiques et analytiques (Bachman, 1990; Fulcher, 2003). Les grilles holistiques ne fournissent qu'une seule cote et permettent d'apprécier une performance d'un point de vue d'ensemble. Les critères sont regroupés par niveaux auxquels

on compare la performance. Selon Roegiers (2004), un critère se définit comme « un élément auquel on se réfère pour porter une appréciation, un jugement : un principe, un caractère, un modèle, une valeur, […] un élément de communication entre la personne qui évalue et la personne qui est évaluée. C'est un langage sur lequel les acteurs s'entendent pour évoquer le niveau des acquis, ou la qualité des acquis » (p. 70-71).

Les grilles analytiques sont constituées d'une série de critères et d'échelons. L'évaluation holistique procède différemment de l'évaluation analytique, en ce sens que l'on cherche à porter un jugement global plutôt que de faire une analyse détaillée (Arter, 2010; Arter et Chappuis, 2006; Carr, 2011). Ces deux types de grilles contiennent des aspects positifs, tout comme des aspects négatifs.

Selon Bejar (2012), Davies *et al.*, (1999), Hamp-Lyons (1991) et Luoma (2004), les grilles holistiques sont rapides, car elles simplifient l'émission de jugement en évitant une analyse approfondie de la part des évaluateurs, et en facilitant leur concentration sur les caractéristiques évoquées par la grille. De plus, elles sont surtout commodes dans les opérations d'évaluation à grande échelle où l'on doit évaluer un large volume de candidats.

Néanmoins, d'après Elliot (2005), la notation holistique est contre-intuitive, car une impression générale et rapide ne peut remplacer l'analyse approfondie d'une réponse construite. Les grilles holistiques sont donc vues comme étant des instruments trop simples, incapables de distinguer les diverses composantes de la compétence communicative. Elles ne peuvent alors pas démontrer les forces et les faiblesses qui caractérisent la performance d'un participant. Kaplan (1991) fait remarquer que deux participants peuvent se retrouver sur le même échelon sans toutefois se ressembler. De ce fait, la grille holistique n'est pas adaptée à la rétroaction corrective (rétroaction de l'enseignant à l'égard des erreurs d'un apprenant), et n'est pas d'un grand secours pour décrire précisément la compétence communicative des individus.

Comme les grilles de type analytique contiennent plusieurs dimensions, celles-ci sont les mieux adaptées pour évaluer les multiples composantes de l'expression écrite et de l'expression orale. Elles aident à aiguiser le jugement de l'évaluateur, car elles fournissent des informations sur les forces et les faiblesses de la personne évaluée. Elles sont donc mieux adaptées à l'évaluation formative et diagnostique (Arter, 2010). Bien qu'elles comportent des critères distincts, il arrive quelquefois que ces derniers ne fonctionnent pas de manière indépendante. En effet, certains

critères sont parfois fortement liés, et peuvent même se chevaucher. Pour ces raisons, les évaluateurs peuvent mal les distinguer ou les interpréter différemment (exemple : « gestion du discours » et « communication interactive ») (Orr, 2002; Taylor et Galaczi, 2011). Par ailleurs, il y a un risque de surcharge cognitive avec ce type de grille, car l'évaluateur doit gérer conjointement plusieurs critères afin d'arriver à une décision.

2.7.3. La complexité de l'élaboration d'une grille d'évaluation

Selon plusieurs chercheurs (Arter, 2010; Carr, 201; Green et Hawkey, 2012; Isaacs, 2016; O'Sullivan, 2012), concevoir un bon instrument d'évaluation nécessite de trouver un juste équilibre entre une grille commode, sans tomber dans une trop grande simplification, et une grille élaborée. D'une part, une grille devrait être facile à utiliser : elle devrait contenir des descripteurs clairs avec un nombre raisonnable de critères pour faciliter la tâche de l'évaluateur et rendre son jugement le moins subjectif possible. D'autre part, les critères et échelons devraient être en nombre suffisant pour pouvoir refléter la myriade d'éléments sur lesquels les évaluateurs se penchent lorsqu'ils doivent prendre une décision.

D'après Roegiers (2004), on aurait spontanément tendance à croire qu'un grand nombre de critères permet d'évaluer davantage d'aspects, mais la pratique montre que l'on perd en découragement et en dépendance de critères ce que l'on aurait pu gagner en fidélité. Le CECRL mentionne que quatre ou cinq critères commencent à provoquer une charge cognitive pour les évaluateurs, et que sept est un seuil psychologique à ne pas dépasser (Conseil de l'Europe, 2001). Bejar (2012) et Luoma (2004), quant à eux, suggèrent un maximum de cinq à six critères. Mais pour Taylor et Galaczi (2011), il n'y a pas de nombres prédéfinis de critères dans l'absolu, car tout dépend des autres tâches que doit effectuer l'évaluateur : s'il doit juste évaluer, participer à la conversation avec le participant, gérer le matériel technique ou effectuer des tâches administratives. Quant au nombre d'échelons, la question du nombre idéal semble non résolue selon Preston et Colman (2000). Isaacs et Thomson (2013) résument la situation avec la citation de Miller (1956, p. 81) : « Le nombre magique sept, plus ou moins deux : quelques limites à notre capacité du traitement de l'information ».

Étant donné que les critères permettent de préciser ce qui doit être observé, il est important, selon Fulcher (2003), d'arriver à définir le plus clairement possible le construit qu'il faut évaluer. Par exemple, si l'évaluation porte uniquement sur la composante grammaticale, il faut définir le mieux

possible cette composante, de même que s'il est question d'évaluer la capacité stratégique, il faut alors décrire cette dernière clairement. Le tout est de définir un construit cohérent avec ce que l'on cherche à évaluer. Shohamy et Walton (1992) insistent sur le caractère explicite des énoncés, surtout pour les critères portant sur l'aspect communicatif qui ont plutôt tendance à s'enchevêtrer. Selon eux, plus l'on s'éloigne d'une description purement linguistique et plus le degré de doute s'accroit. Lumley (2005) évoque la tension entre le « bon ordre simplifié de la grille d'évaluation » (p. 248), qui sous-représente la complexité d'une performance en L2, et les réactions illimitées des évaluateurs par rapport à la performance, qui peuvent être désordonnées et complexes. Selon lui, les évaluateurs sont face à un défi qui consiste à unir leurs impressions idiosyncrasiques, intuitives ou non linéaires d'une performance avec les composantes d'une grille d'évaluation. Ce constat est partagé par Meiron et Schick (2000) et Reed et Cohen (2001) qui arguent que les évaluateurs ont la lourde tâche d'interpréter et d'appliquer les grilles d'évaluation. Ces derniers doivent trouver le juste milieu entre la norme et leur propre compréhension de la norme. Ils ne peuvent mener à bien leur travail que si leurs instruments d'évaluation sont explicites, et s'ils reflètent avec précision la définition du construit.

2.7.4. Les deux approches dans l'élaboration des grilles d'évaluation

Dans le contexte scolaire, selon Chadwick (1858), les premières grilles d'évaluation ont été conçues dans les années 1830. Les grilles avaient deux objectifs : informer l'enseignant sur les progrès des apprenants à des fins formatives, et générer des données pour la reddition de comptes des écoles. Les données étaient alors utilisées par les parents afin de sélectionner les écoles pour leurs enfants, et ainsi de « maximiser les bénéfices de leur investissement ». Dans le domaine des tests de L2, nous avons mentionné précédemment que les premières grilles d'évaluation, qui ont servi à la conception du référentiel *Proficiency Guidelines*, sont nées dans le contexte militaire américain de la Seconde Guerre mondiale, et que selon le linguiste Kaulfers (1944), le *Foreign Service Institute* a élaboré intuitivement sa propre grille d'évaluation sans prendre en compte les travaux déjà réalisés du linguiste. Quant aux descripteurs et échelles du CECRL, leur conception est issue d'une recherche empirique (auprès de 300 enseignants et de 2800 apprenants) menée entre 1993 et 1996, mais n'a pas été fondée sur des descriptions de la compétence langagière qui avaient antérieurement été validées empiriquement. Deux auteurs du CECRL, North et Schneider (1998), déclarent que « la plupart des grilles pour les compétences langagières semblent en fait avoir été produites de manière pragmatique par des appels à l'intuition, la culture pédagogique

locale et les grilles auxquelles l'auteur a eu accès » (p. 220). Nous observons ainsi que le référentiel de l'ACTFL a été élaboré de manière intuitive, et celui du CECRL de manière empirique.

À partir de cette observation, Fulcher (2003) s'est intéressé à la manière dont une grille d'évaluation en L2 peut être développée selon deux approches : l'approche dite intuitive et l'approche dite empirique. Il donne une description détaillée du processus de chacune de ces deux approches.

L'approche intuitive : tout d'abord, un enseignant expérimenté ou un organisme concepteur de tests élabore une grille d'évaluation en rapport avec une grille déjà existante, un programme scolaire, ou une analyse de besoins. Des consultants peuvent être sollicités afin d'obtenir des rétroactions sur l'utilité de la grille. Ensuite, un petit comité d'experts échange les différents points de vue, puis se met d'accord sur les terminologies des descripteurs et des niveaux de la grille. Enfin, les utilisateurs expérimentent la grille d'évaluation, l'ajustent et la peaufinent. Ainsi, après un certain temps, les utilisateurs comprennent intuitivement le sens des niveaux en rapport avec les échantillons de performance des participants.

L'approche empirique : cette approche requiert tout d'abord l'analyse d'une performance suscitée lors de la réalisation d'une tâche. Cette analyse porte sur la description des caractéristiques clés de la performance pouvant être observée afin de faire des inférences vis-à-vis du construit. Par la suite, des experts saisissent des échantillons oraux ou écrits des performances, et les divisent en bonnes et moins bonnes performances. Les raisons justifiant ces divisions sont rapportées. Dans cette approche, beaucoup de descripteurs sont préalablement collectés séparément de la grille, puis sont calibrés dans un ordre de difficulté par les experts. Finalement, les descripteurs sont séquencés afin de créer la grille d'évaluation.

Les grilles d'évaluation sont essentielles dans les tests oraux, car elles sont des opérationnalisations du construit que le test est censé mesurer. Or, appliquer les principes de base afin d'élaborer une grille explicite et appropriée ne permet pas de garantir qu'elle sera utilisée de la façon dont les développeurs d'un test l'ont prévu. Les variations dans l'application des grilles, c'est-à-dire dans l'interprétation des niveaux, des critères et des descripteurs par les examinateurs, ayant suivi les mêmes formations, ont été exposées précédemment. Après avoir présenté ces principes de base sur les grilles d'évaluation, nous nous orientons désormais vers la grille actuelle de l'épreuve d'expression orale du TEF qui servira à guider les examinateurs dans notre recherche.

2.7.5. L'actuelle grille d'évaluation de l'épreuve d'expression orale du TEF

Comme nous l'avons décrit précédemment, les niveaux de la grille d'évaluation du TEF sont arrimés aux six niveaux du CECRL auxquels est ajouté un niveau inférieur pour les absences d'observables. La grille est de type analytique et se compose de critères communicationnels et de critères linguistiques (Demeuse et Artus, 2008). La nouvelle version, mise à jour en octobre 2018, renferme cinq critères et sept échelons (Tableau 11). Comparativement à l'ancienne version, le nombre de critères a diminué de manière considérable, passant de douze à cinq. Les deux premiers critères sont communicationnels et se basent respectivement sur les deux sections de l'épreuve : « 1. Section A - Capacité à obtenir des informations » et « 2. Section B - Capacité à présenter et débattre ». Les trois critères suivants sont linguistiques : « 3. Syntaxe », « 4. Lexique », « 5. Aisance à l'oral, élocution ».

Le nombre d'échelons, recouvrant les niveaux du CECRL, est resté inchangé, mais celui des sous-échelons a été réduit, comme nous le montre le tableau 12. Dans la nouvelle version, la répartition se présente ainsi : 1 sous-échelon pour les niveaux <A1, A1 et C2; 2 sous-échelons pour les autres niveaux : A2, B1, B2 et C1. Dans l'ancienne version, on retrouvait 2 sous-échelons pour les échelons 0+ et 6; 3 sous-échelons pour les échelons 1, 4 et 5; et 4 sous-échelons pour les échelons 2 et 3.

Tableau 11 - Nouvelle version de la grille d'évaluation de l'épreuve d'expression orale du TEF

Critères	< A1	A1	A2	B1	B2	C1	C2
1 SECTION A Capacité à obtenir des informations	Absence d'observables.	Questionnement élémentaire et limité. La conversation dépend entièrement de l'examinateur.	Questionnement simple et très général. Quelques demandes de clarification ou de reformulation. Les échanges sont peu suivis.	Questionnement satisfaisant. Fait préciser et développer certaines informations incomplètes ou ambigües.	Questionnement approprié. Réagit avec assurance aux réponses données, même dans les situations imprévues. Les échanges sont suivis.	Questionnement complet et précis. Intervient avec justesse. La conversation est soutenue.	Questionnement exhaustif et pertinent. Mène la conversation de façon naturelle et efficace.
	☐	☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐
2 SECTION B Capacité à présenter et débattre	Absence d'observables.	Présentation sommaire, simple lecture. Donne son avis de façon très simple. Les échanges sont difficiles et très limités.	Présentation très simple, lecture et paraphrase. Quelques éléments simples et répétitifs pour convaincre. Les échanges sont brefs.	Présentation simple et claire, effort de reformulation des informations. Arguments clairs mais peu développés. Intervient régulièrement et justifie son point de vue.	Présentation claire et détaillée. Arguments illustrés ou détaillés. Défend ses idées de manière claire et déterminée. Les échanges sont suivis.	Présentation développée et structurée. Arguments variés et bien développés. Intervient avec justesse. La conversation est soutenue.	Présentation limpide qui suscite l'intérêt. Arguments nuancés, complexes. La conversation est riche, naturelle et animée.
	☐	☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐
3 Syntaxe	Absence d'observables.	Structures et phrases élémentaires, souvent mémorisées.	Phrases simples et stéréotypées. Erreurs systématiques mais l'ensemble est compréhensible.	Discours organisé avec quelques phrases complexes courantes. Les erreurs sont fréquentes mais le sens général est clair.	Phrases simples et phrases complexes courantes utilisées correctement. Les erreurs ne conduisent pas à des malentendus.	Emploi adéquat d'une grande variété de structures. Les erreurs sont rares et n'affectent pas les échanges.	Emploi constant d'une grande variété de structures très bien maîtrisées.
	☐	☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐
4 Lexique	Absence d'observables.	Répertoire lexical élémentaire et répétitif.	Répertoire lexical restreint aux situations quotidiennes et familières.	Répertoire lexical plus large permettant de s'adapter à la situation, à l'aide de périphrases ou d'emprunts à d'autres langues.	Répertoire lexical assez large. Les confusions ou approximations ne gênent pas la communication.	Répertoire lexical vaste et bien maîtrisé. Les lacunes sont compensées sans effort apparent.	Répertoire lexical riche et nuancé, adapté et très bien maîtrisé.
	☐	☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐
5 Aisance à l'oral, élocution	Absence d'observables.	Pauses très nombreuses. Discours peu intelligible, la compréhension est difficile et demande beaucoup d'efforts.	Pauses et faux-démarrages fréquents. Discours compréhensible mais nécessite efforts ou demandes de répétitions.	Hésitations fréquentes mais le discours est clair. Des erreurs de prononciation peuvent parfois gêner la compréhension.	Le débit est assez régulier, l'intonation claire et naturelle. L'accent ne gêne pas la compréhension.	Le discours est globalement naturel et fluide. L'intonation est la plupart du temps adaptée à la situation.	Le discours est constamment fluide, naturel et sans effort, comme un natif. L'intonation est adaptée aux propos et à la situation.
	☐	☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐

Source : Chambre de commerce et d'industrie de région Paris Île-de-France, 2018

Tableau 12 - Comparaison entre les sous-échelons de l'ancienne et de la nouvelle version de la grille d'évaluation du TEF

Niveau	Nombre de sous-échelons de l'ancienne version de la grille d'évaluation	Nombre de sous-échelons de la nouvelle version de la grille d'évaluation
<A1	2	1
A1	3	1
A2	4	2
B1	4	2
B2	3	2
C1	3	2
C2	2	1

À ce jour, aucune communication officielle au sujet de cette grille n'a été publiée. Nous pouvons cependant affirmer que ce nouveau format avec un nombre réduit de critères correspond à ce que les chercheurs suggèrent, c'est-à-dire un maximum de cinq à sept critères, qui évitent un risque de surcharge cognitive (Bejar, 2012; Conseil de l'Europe, 2001; Luoma, 2004). La simplification de la grille d'évaluation a été faite par des regroupements de catégories de critères, comme nous le montre le tableau 13. Par exemple, pour la section A, on ne dissocie plus la pertinence du questionnement, de la prise d'initiative et la gestion de l'imprévu, et de la qualité des échanges, mais on regroupe désormais ces trois critères en un seul : la capacité à obtenir des informations.

Tableau 13 - Comparaison entre les critères de l'ancienne version et de la nouvelle version de la grille d'évaluation du TEF

Critères de l'ancienne version de la grille d'évaluation	Critères de la nouvelle version de la grille d'évaluation
Section A : 1. Pertinence du questionnement 2. Prise d'initiative et gestion de l'imprévu 3. Qualité des échanges	1. Section A – Capacité à obtenir des informations
Section B : 4. Présentation des faits 5. Qualité de l'argumentation 6. Qualité des échanges	2. Section B – Capacité à présenter et débattre
Syntaxe : 7. Complexité des structures 8. Cohésion du discours	3. Syntaxe

Lexique : 9. Étendue 10. Maitrise	4. Lexique
Élocution : 11. Prononciation et intonation 12. Rythme et débit	5. Aisance à l'oral, élocution

Nous observons que ce regroupement donne lieu à des critères plus distincts. D'une part, ce fonctionnement indépendant des critères permet aux examinateurs de mieux les distinguer les uns des autres et évite ainsi des différences d'interprétation et d'éventuels chevauchements. D'autre part, l'indépendance des critères empêche également de pénaliser deux fois un candidat qui rencontre des faiblesses avec un même aspect (Orr, 2002; Roegiers, 2010; Taylor et Galaczi, 2011).

Les énoncés de la grille d'évaluation du TEF ont été élaborés à partir de la banque d'exemples de descripteurs issue du CECRL. Le tableau ci-dessous (Tableau 14) renvoie, pour chaque critère de la grille d'évaluation, aux échelles du CECRL concernées et précise également les objectifs d'évaluation à considérer.

Tableau 14 - Alignement entre les critères d'évaluation du TEF et les échelles du CECRL

Critères d'évaluation du TEF	Échelles du CECRL	Objectifs d'évaluation
1 SECTION A Capacité à obtenir des informations	Interaction orale générale p.61 Comprendre un locuteur natif p.62 Conversation p.62 Obtenir des biens et services p.66 Échange d'information p.67 Coopérer p.71 Faire clarifier p.71 Tours de parole p.97	Capacité à recueillir des informations pertinentes Capacité à faire préciser/développer les informations reçues Respect de la situation de communication, des tours de parole
2 SECTION B Capacité à exposer et débattre	Production orale générale p.49 Interaction orale générale p.61 Comprendre un locuteur natif p.62 Conversation p.62 Obtenir des biens et services p.66 Échange d'information p.67 Coopérer p.71 Tours de parole p.97 Développement thématique p.97	Capacité à présenter une situation, à se l'approprier, la contextualiser Capacité à développer ses idées, par des exemples, des arguments secondaires, des détails et précisions Respect de la situation de communication, des tours de parole

3 Syntaxe	Contrôle et correction p.54 Étendue linguistique générale p.87 Correction grammaticale p.90 Souplesse p.97 Cohérence et cohésion p.98	Utilisation de la langue orale, dans des situations de la vie quotidienne Capacité à reformuler et clarifier ses interventions si nécessaire Capacité à s'adapter à son interlocuteur
4 Lexique	Compensation p.54 Contrôle et correction p.54 Étendue linguistique générale p.87 Étendue du vocabulaire p.88 Maîtrise du vocabulaire p.89 Souplesse p.97	
5 Aisance à l'oral, élocution	Maîtrise du système phonologique p.92 Aisance à l'oral p.100	Prononciation, intelligibilité du discours Intonations, accents phrastiques Degré d'aisance, de spontanéité, d'effort à communiquer en français

Source : CCI Paris Île-de France

2.8. Les objectifs de la recherche

Dans cette section, nous ferons la synthèse de notre recension du chapitre 2, puis nous présenterons nos questions de recherche.

2.8.1. La synthèse de la recension

Comme nous l'avons vu dans ce chapitre, les tâches des tests oraux de L2, tout comme leurs outils d'évaluation, reposent d'une part sur des fondements théoriques qui ont défini la compétence à communiquer à l'oral, et d'autre part, sur les référentiels internationaux, notamment sur le CECRL. Ce dernier a fourni une base commune en évaluation des langues secondes en définissant des niveaux en termes de descripteurs par souci d'homogénéité. En outre, pour que les tests de langue apportent des garanties suffisantes sur la qualité du dispositif d'évaluation, ils se doivent d'être fidèles et valides. Cela permet de contrôler au mieux la subjectivité et ainsi de pouvoir porter un jugement qui soit le plus juste possible. Toutefois, porter un jugement ne représente pas un simple acte technique, mais constitue une opération complexe faisant intervenir diverses facettes. Il est alors difficile de garantir un traitement équitable et une appréciation juste des compétences à mesurer des candidats dans une situation de test oral. Par ailleurs, s'ajoutent à cela des biais qui

sont liés à l'environnement du test (conditions, tâches). Afin de mieux cerner les examinateurs-évaluateurs, certains chercheurs ont tenté d'illustrer leurs approches évaluatives et d'autres ont conceptualisé leur cognition à travers des modèles, mais on remarque qu'il en existe très peu. Notre recherche se situe dans la continuité des recherches recensées, car nous participons à l'étude de l'examinateur, dont la finalité est d'apporter une plus grande fidélité et une validité dans l'acte d'évaluer.

2.8.2. Les questions de recherche

Notre recherche vise à décrire les différentes divergences à travers le processus décisionnel des examinateurs, et également à brosser un portrait de leur appropriation et appréciation concernant la grille d'évaluation de la compétence langagière. Pour cela, nos questions de recherche sont les suivantes :

1. En considérant différents aspects du jugement évaluatif, quelles divergences pouvons-nous observer chez les examinateurs ?
2. Quel portrait peut-on dresser de leur appropriation et de leur appréciation de la grille d'évaluation de la compétence langagière ?

Nos questions de recherche étant présentées, nous préciserons dans le chapitre suivant la méthodologie retenue nous permettant de décrire les divergences dans la façon dont les examinateurs portent leur jugement, puis de faire le point sur leur appropriation et appréciation de la grille d'évaluation.

CHAPITRE 3: MÉTHODOLOGIE

Introduction

Dans ce chapitre, nous exposerons en détail la méthodologie choisie afin d'établir des pistes de réponses aux objectifs posés par cette recherche. Nous définirons tout d'abord l'approche méthodologique retenue qui est la technique de la pensée à voix haute. Nous mettrons en avant ses caractéristiques, ferons un rappel historique, puis présenterons les variantes et les limites de cette méthode. Nous annoncerons ensuite le contexte pratique en commençant par une présentation de notre échantillon, et en précisant les modalités de la collecte de données, le milieu de la recherche ainsi que les outils choisis. Nous préciserons également le déroulement de la recherche qui s'effectuera à travers deux méthodes : l'activité de verbalisation et l'entrevue semi-dirigée. Nous résumerons les points essentiels à l'aide d'un tableau de synthèse récapitulatif, nous défendrons la position de la chercheure, puis nous finirons par évoquer les considérations éthiques.

3.1. L'approche méthodologique

Notre recherche s'inscrit dans une perspective qualitative/interprétative, en raison de notre désir de décrire le sens que les sujets attribuent à leur expérience (Savoie-Zajc, 2011). En effet, nous souhaitons étudier les sujets dans leur milieu naturel en essayant de donner un sens et d'interpréter les phénomènes en nous fondant sur les significations que leur apportent ces derniers (Denzin et Lincoln, 2000). Notre but n'est pas de quantifier les phénomènes observés afin d'établir des corrélations, mais plutôt de cerner la réalité telle que la vivent les sujets en essayant de pénétrer à l'intérieur de l'univers observé (Poisson, 1983). Dans notre recherche, nous faisons ainsi émerger des significations profondes élaborées par des examinateurs concernant leur savoir-évaluer à l'aide de données qualitatives.

Après avoir réalisé une recension d'un grand nombre d'écrits, principalement auprès de nombreux ouvrages, de récents *handbooks*, encyclopédies et de revues spécialisées comme *Language Testing*, *Language Assessment Quarterly* et *Studies in Language Testing*, nous avons observé que les méthodes quantitatives et mixtes dominaient largement la recherche portant sur les tests en L2. Plusieurs chercheurs spécialisés, notamment Bachman (1989, 2000), Banerjee et Luoma (1997), Brown (2000), Brown *et al.*, (2005), Lazaraton (2002), Roever et McNamara (2006), Shohamy

(1990) Taylor et Saville (2001) et van Lier (1989) déclarent que les méthodes quantitatives statistiques traditionnelles sont efficaces pour valider les tests de L2, mais qu'elles connaissent des limites. Pour ces chercheurs, il est important d'envisager des approches plus innovantes comme les méthodologies de recherche qualitative et de sortir d'une tradition psychométrique. McNamara (1997) affirme à ce sujet qu'il y a un besoin d': « inclure un autre genre de recherche sur les tests de langue qui soit plus fondamental et qui vise pleinement à nous faire prendre conscience de la nature et de l'importance de l'évaluation en tant qu'acte social » (p. 460).

3.1.1. Les caractéristiques de la technique de la pensée à voix haute

Notre recherche se veut de type qualitatif et la méthode de recherche retenue est celle de la technique de la pensée à voix haute (*think aloud protocol*). Celle-ci a été mise au point afin d'accéder à ce qui se passe « dans la tête » des sujets au moment de réaliser une tâche et ainsi d'expliciter les aspects cognitifs implicitement présents dans des actions. Par exemple, des participants regardent un extrait vidéo d'une tâche effectuée, puis disent à voix haute (verbalisent) ce à quoi ils pensaient pendant qu'ils étaient en train de réaliser la tâche. La technique de la pensée à voix haute se distingue des autres méthodes qui emploient des données verbales, car les inférences sont faites sur les processus cognitifs qui produisent la verbalisation. De cette façon, elle diffère des autres méthodes telles que l'analyse de discours ou l'entrevue qui se focalisent essentiellement sur le contenu et la structure linguistiques, ainsi que sur la formation de ce qui est énoncé (Green, 2009). En outre, la technique de la pensée à voix haute est plus susceptible de refléter ce que les sujets font réellement, plutôt que ce qu'ils croient faire, comme c'est le cas dans les entrevues et les questionnaires (Huot, 1993). Enfin, alors que les entrevues et les questionnaires fournissent des déclarations générales sur les comportements, la technique de la pensée à voix haute inspecte des exemples de comportements réels (Connor-Linton, 1995; Ericsson et Simon, 1987).

Les données recueillies de la technique, une fois transcrites, constituent un « rapport verbal », ou un « protocole verbal », qui se définit comme étant un type particulier d'introspection basé sur un traitement de l'information. Au sein du traitement de l'information, les informations sont stockées dans plusieurs zones de la mémoire avec des capacités différentes : la mémoire de travail, de capacité limitée, et la mémoire à long terme, de grande capacité. Les informations récemment

acquises par le processeur central sont conservées dans la mémoire de travail et permettent de produire le rapport verbal d'un sujet (Ericsson et Simon, 1980, 1984).

3.1.2. Le survol historique de la technique de la pensée à voix haute

La technique de la pensée à voix haute est issue de la psychologie cognitive. À l'opposé du courant cognitiviste, les psychologues du béhaviorisme considéraient qu'un apprentissage réussi se fondait sur l'assimilation d'exercices appropriés associés au renforcement des réponses correctes. Les psychopédagogues ont été portés à considérer que l'efficacité d'un apprentissage reposait sur le temps consacré à la tâche (Rosenshine, 1986). Mais ce faisant, les chercheurs négligeaient l'activité mentale de l'apprenant au moment de la tâche. La représentation du problème était hors de la portée des instruments utilisés et des approches préconisées. La technique de la pensée à voix haute, à cet égard, a permis de savoir à quoi pense l'apprenant, comment il pense au problème à résoudre et comment il gère ses connaissances (Peterson et Swing, 1982). La technique a été un instrument privilégié de la psychologie cognitive avant de devenir un instrument de la professionnalisation en formation initiale et continue (Tochon, 1996).

À l'origine, cette technique avait pour but de susciter, chez des apprenants, le rappel des processus mentaux qu'ils avaient activés en classe lors d'une tâche d'apprentissage antérieure. Selon Tochon, (1996), Benjamin Bloom aurait été le premier à stimuler le rappel de cognitions interactives à partir d'enregistrements, afin d'aider des apprenants à perfectionner leur exposé ou leur argumentation lors d'une discussion (Bloom, 1953). Le chercheur enregistrait ses apprenants au cours de discussions et d'exposés, puis leur soumettait l'enregistrement, immédiatement après la tâche. Il s'en servait comme stimulus pour obtenir un rapport des processus mentaux propres à la tâche étudiée. Outre l'aspect « recherche » dans l'utilisation que faisait Bloom de cette technique, un aspect « formation » était clairement présent dans sa démarche.

Dans la même lignée, deux pionniers dans la recherche, Clark et Peterson (1976), ont présenté des études sur les processus mentaux des enseignants, notamment sur leur prise de décision, en cours d'action. Puis successivement, toute une série d'études sur le sujet a été publiée. D'après Tochon (1996), l'intérêt pour les pensées interactives des enseignants a propulsé la technique de la pensée à voix haute au rang des méthodologies privilégiées de la recherche en éducation. En effet, il est difficile d'accéder aux cognitions interactives des enseignants en action, tout comme dans les

autres professions de l'interaction. Les techniques d'explicitation postactives des décisions prises dans l'action sont parfois insuffisantes et manquent de validité.

Lorsque l'enseignant se place devant l'écran du moniteur vidéo et qu'il se revoit en action, les pensées qu'il avait eues en tête pendant telle ou telle action lui reviennent en mémoire plus facilement. Ainsi, la verbalisation de ses pensées permet d'accéder à des représentations bien contextualisées de la tâche professionnelle. Les indices fournis par l'enregistrement suffisent, en principe, à faire revivre rétrospectivement l'épisode afin d'obtenir une reconstitution fiable des processus mentaux sous-jacents à l'action enregistrée.

La technique de la pensée à voix haute a d'abord été exploitée dans des études relatives à l'apprentissage, notamment en lecture et en écriture. Par la suite, elle s'est généralisée à toutes sortes de situations d'apprentissage : l'apprentissage adulte d'une profession, l'apprentissage de l'enseignement, les apprentissages disciplinaires et professionnels. Par exemple, De Groot (1965), psychologue et amateur d'échecs, a commencé à utiliser cette technique pour étudier le jeu des joueurs professionnels dans les compétitions d'échecs internationales. D'après Tochon (1996), c'est probablement la période des premières utilisations systématiques des protocoles verbaux dans la recherche sur la compétence professionnelle experte, en vue de créer des systèmes experts dans diverses professions.

Depuis plusieurs décennies, la technique de la pensée à voix haute est largement utilisée dans les sciences humaines, notamment en psychologie, en sociologie, en anthropologie, en didactique des mathématiques et en didactique des langues pour :

- aider les consultants cliniques à améliorer leurs prestations (Kagan *et al.*, 1963);
- explorer les différences entre les experts et les novices résolvant des problèmes de sciences politiques (Voss *et al.*, 1983);
- examiner les interactions entre différentes catégories de comportement dans la résolution de problèmes mathématiques (Montague et Applegate, 1993; Yee, 2008);
- identifier les processus cognitifs qui différencient les bons apprenants des moins bons apprenants (Green et Gilhooly, 1900 a, 1990 b; Thorndyke et Stasz, 1980);
- identifier les stratégies et les problèmes de personnes adultes rencontrant des difficultés en lecture (Berne, 2004);

- décrire les activités d'écriture des candidats du test d'anglais TOEFL (*Test of English as a Foreign Language*) (Barkaoui, 2015);
- distinguer les évaluateurs inexpérimentés des évaluateurs expérimentés dans des évaluations de dissertations (Barkaoui, 2010; Cumming, 1990; Weigle, 1994).

Dans le domaine de l'évaluation des langues premières et des langues secondes, la technique de la pensée à voix haute est généralement utilisée pour étudier et construire des modèles de processus de notation, à l'écrit comme à l'oral (Brown, 2005; Cumming *et al.*, 2002; Green, 2009; Huot, 1993; Kormos, 1998; Lumley, 2005; Wolfe *et al.*, 1998).

3.1.3. Les variantes de la technique de la pensée à voix haute

La technique de la pensée à voix haute regroupe essentiellement trois variantes : 1. la verbalisation concomitante à la tâche; 2. la verbalisation rétrospective; 3. la verbalisation rétrospective assistée. Nous les avons définies dans le tableau ci-dessous (Tableau 15).

Tableau 15 - Les trois variantes de la technique de la pensée à voix haute

La technique de la pensée à voix haute	
Noms des variantes	Fonctions
1. La verbalisation concomitante à la tâche	- consiste à demander aux sujets de dire tout haut ce qui se passe dans leur tête et ce qu'ils font durant la réalisation d'une tâche (Ericsson et Simon, 1993; Gufoni, 1996; Lemaire, 1999; van Someren *et al.*, 1994).
2. La verbalisation rétrospective (également nommée le rappel stimulé)	- présuppose la possibilité de rappeler à la mémoire de travail des informations traitées et stockées dans la mémoire; - implique l'usage d'enregistrements audio ou vidéo, car elle fait revivre une situation du passé, et sollicite les commentaires des sujets à propos du déroulement de la tâche après sa réalisation;

	- peut s'effectuer à partir de quelques secondes suivant la tâche jusqu'à plusieurs jours après sa réalisation (Boyer, 1997; Calderhead, 1981; Préfontaine et Fortier, 1997; van Someren *et al.*, 1994).
3. La verbalisation rétrospective assistée	- est similaire à la précédente, mais concède qu'un soutien est parfois nécessaire pour aider le sujet à se rappeler le mieux possible de la séquence de ses actions (Dionne, 1996; Gufoni, 1996).

La verbalisation concomitante à la tâche s'avère plus appropriée pour analyser les productions écrites, mais inadéquate pour évaluer les interactions orales, car si le sujet tente de rapporter verbalement tout ce qu'il pense pendant qu'il est en interaction, il perturbe inévitablement le contexte. Dès lors, la méthode privilégiée pour obtenir des informations sur le processus cognitif pendant une interaction demeure la verbalisation rétrospective (Dunn et Lozinski, 2005; Lyle, 2002; Ericsson et Crutcher, 1991; Ericsson et Simon, 1993; Lumely, 2000, cité par Brown, 2005; Yinger, 1986, cité par Trudel *et al.*, 1996).

3.1.4. Les limites de la technique de la pensée à voix haute

La technique de la pensée à voix haute est inhérente à la cognition humaine et est basée sur l'assertion selon laquelle elle est vue comme étant un compte rendu précis d'informations provenant d'un individu qui s'attèle (ou s'est attelé) à une tâche particulière (Green, 2009). Cependant, elle contient des limites quant à la validité et la fiabilité scientifiques des données. L'idée de validité s'applique « si l'information qui est capturée dans les rapports verbaux correspond réellement à l'information à laquelle on a prêté attention à mesure qu'une tâche est effectuée » Green (2009, p. 10). La fiabilité, quant à elle, se réfère à « la probabilité que des rapports verbaux similaires soient produits par le même individu soumis à des tâches identiques ou très similaires » (Green, 2009, p. 11).

Ainsi, on ne peut prétendre que ce qui est verbalisé de la part des sujets corresponde vraiment à la réalité, les rapports verbaux peuvent être altérés par une mémoire imprécise et contenir des

« événements mentaux fabriqués (involontairement) » (Gass et Mackay, 2000, p.106). Ils peuvent ne pas se baser sur une vraie introspection, mais sur « des théories a priori, implicites et causales » ou peuvent être provoqués par des stimuli particuliers (Nisbett et Wilson, 1997, p. 231). En d'autres termes, il y a toujours des risques d'ajouts, de reconstruction, de déformation ou d'oubli d'informations (Dionne, 1996; Nisbett et Wilson, 1977). À ce propos, Ericsson et Simon (1993) soulignent que des informations inexactes sont fournies de la part des sujets lorsqu'ils sont amenés à faire part d'un raisonnement réfléchi, et non lorsqu'ils sont simplement amenés à faire part de leurs pensées. Lorsque les sujets verbalisent leurs pensées en eux, la verbalisation reflète les processus cognitifs qui se sont produits pendant la tâche. Toutefois, si les participants sont invités à présenter une quelconque analyse, cela crée des activités mentales supplémentaires qui affectent leurs verbalisations et qui entraînent des rapports verbaux problématiques. Pour éviter cela, Ericsson et Simon (1994) suggèrent de demander aux participants de verbaliser uniquement ce à quoi ils pensaient pendant l'événement, et non pas les raisons pour lesquelles ils pensaient de telle ou telle manière.

L'autre obstacle que l'on peut rencontrer avec la technique de la pensée à voix haute est le risque d'être confronté à des sujets « avares cognitifs ». Cette caractéristique fondamentale, qui est propre à notre fonctionnement cognitif, représente une solution de facilité lorsque nous ne sommes pas tentés par la réflexion ou lorsque nous voulons économiser nos ressources mentales (Fiske, 2008). Enfin, lorsque la verbalisation est assistée par une tierce personne, le soutien offert ne garantit pas aux sujets qu'il stimule la résurgence des actions mentales ou une reconstruction de celles-ci (Tochon, 1996).

Lorsque l'on verbalise après la tâche, l'intervalle de temps entre l'achèvement de la tâche et le début du rapport verbal revêt une importance particulière. Si la verbalisation est produite dès la fin de la tâche, beaucoup d'informations demeurent encore présentes dans la mémoire de travail. Mais si l'espace de temps entre l'achèvement de la tâche et la production du rapport verbal est trop long, le processus de récupération est assurément faillible. À cause de la mémoire à court terme, de précieuses informations peuvent disparaitre (Green, 2009; Nisbett et Wilson, 1997; Someren *et al.*, 1994). À ce sujet, Bloom (1954) a examiné l'exactitude des rapports verbaux et souligne que ces derniers sont fiables tant qu'ils sont effectués peu de temps après la tâche. De manière plus

détaillée, il déclare que le rapport verbal est très précis (à 95%) dans les 48 heures suivant la tâche, mais que cette précision chute d'environ 65% au bout de deux semaines.

3.2. L'échantillon

Les participants de notre échantillon proviennent de plusieurs centres d'examen TEF agréés de Montréal et ses environs. Étudier des terrains différents nous permet d'avoir un portrait plus global de la situation et de réduire entre autres les risques de probabilité d'obtenir des données semblables.

Au départ, nous avons contacté dix centres d'examen et avons eu des réponses favorables de quatre centres. Nous avions prévu de sélectionner douze participants, puis nous en avons obtenu dix en définitive. Les dix participants sont des examinateurs certifiés issus de quatre différents centres d'examen, mais trois d'entre eux proviennent conjointement de deux centres différents. À ce propos, tous les examinateurs certifiés peuvent exercer leurs fonctions dans plusieurs endroits, car il n'existe pas d'affiliation exclusive à un centre spécifique.

Comme chacun des dix examinateurs a évalué quatre candidats de niveaux différents à l'aide d'enregistrements sonores, nous disposions de quarante évaluations au total. La taille de notre échantillon a été le résultat de l'évolution de notre recherche et a été limitée par le recueil d'informations nécessaires. L'échantillon relève des besoins de notre recherche, de l'atteinte de la saturation des catégories d'informations et de notre jugement (Deslauriers, 1991). Selon Lincoln et Guba (1985), que la taille soit grande ou petite importe peu, car le but de l'échantillonnage en recherche qualitative est d'obtenir le maximum d'informations en produisant de nouveaux faits.

Les dix participants ont été recrutés sur une base volontaire et devaient répondre à certains critères. L'échantillonnage était donc 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 (Fortin et Gagnon, 2016). De cette manière, nous avons trouvé les personnes les plus susceptibles de fournir des données riches en information par rapport au problème étudié (Patton, 2002). Les critères d'inclusion sélectionnés étaient les suivants : être examinateur TEF certifié depuis au moins trois ans et évaluer annuellement en moyenne au moins cent candidats de l'épreuve d'expression orale dans les centres d'examen agréés. La composition de l'échantillon est présentée dans le tableau 16.

Tableau 16 - Composition de l'échantillon

Nombre total de participants	10
Nombre de centres d'examen	4
Nombre de candidats à évaluer par participant	4
Critères d'inclusion des participants	- être examinateur TEF certifié depuis au moins trois ans; - évaluer en moyenne au moins 100 candidats par an dans les centres d'examens agrées.

Dans un premier temps, nous avons contacté par courriel les responsables des centres d'examen retenus afin de les informer du projet et de les solliciter afin qu'ils invitent leurs examinateurs à participer à la recherche. Une lettre de sollicitation présentant la requête a été prévue à cet effet. Lorsque les participants ont accepté de participer à la recherche, ils ont reçu une lettre d'information présentant les objectifs et les modalités des tâches à accomplir, puis un formulaire de consentement leur a également été remis.

Afin d'établir une relation de confiance avec les participants, il a également été essentiel d'insister sur le fait que notre recherche n'était pas en lien direct avec la Chambre de commerce et d'industrie de Paris Île-de-France, et que notre objectif n'était pas de contrôler ni de porter un quelconque jugement sur leurs pratiques évaluatives. Nous avons donc fait en sorte de créer un climat favorable, dépourvu de méfiance, par une attitude empreinte de courtoisie, d'écoute et d'intérêt sincère.

3.2.1. Les données des participants

Avant le début de chaque rencontre, un questionnaire ayant pour but de brosser un portrait sociodémographique des participants a été envoyé. Les données d'identification sont les suivantes : leur sexe, leur tranche d'âge, leur(s) profession(s) principale(s), leur ancienneté dans le domaine de l'enseignement du FLS, leur ancienneté en tant qu'examinateur TEF, le nombre moyen de candidats évalués annuellement et les noms des centres d'examen auxquels ils se rattachent. Pour respecter le caractère confidentiel des données, les noms des participants ainsi que les noms des centres d'examen ont été remplacés par des codes. Les examinateurs sont nommés E1, E2, E3, E4, E5, E6, E7, E8, E9 et E10. Le genre masculin est utilisé sans discrimination et dans le seul but

d'alléger le texte. Ainsi, cette brève description de la population de notre étude a permis de dresser un profil global en rendant compte de certaines caractéristiques. Le tableau ci-dessous (Tableau 17) présente la fiche d'identification des participants.

Tableau 17 - Fiche d'identification des participants

Examinateur	Sexe	Tranche d'âge	Profession(s) principale(s)	Ancienneté dans le domaine du FLS	Ancienneté comme examinateur TEF	Nombre moyen de candidats évalués par an	Centre(s) d'examen
E1	H	51-55 ans	Enseignant en FLS	17 ans	9 ans	200	A
E2	H	36-40 ans	Enseignant en FLS et anglais L2	12 ans	7 ans	200	A
E3	H	46-50 ans	Enseignant en FLS	16 ans	3 ans	150	A et D
E4	H	51-55 ans	Enseignant en FLS	30 ans	7 ans	300	A et E
E5	F	41-45 ans	Enseignant en FLS, espagnol L2 et littérature française	20 ans	11 ans	220	B
E6	H	36-40 ans	Enseignant en FLS et responsable TEF	11 ans	8 ans	180	B
E7	F	41-45 ans	Enseignant en FLS	11 ans	7 ans	150	B
E8	H	36-40 ans	Conseiller pédagogique et responsable TEF	9 ans	7 ans	100	C
E9	F	36-40 ans	Enseignant en FLS	15 ans	10 ans	150	C et F
E10	F	36-40 ans	Enseignant en FLS	12 ans	4 ans	100	D

Dans ce tableau, nous observons que sept de nos participants sont des hommes et trois, des femmes. La tranche d'âge se situe entre 36 et 55 ans. Neuf sont enseignants de FLS depuis au moins plus de dix années, et un est responsable pédagogique et a une expérience de neuf ans en enseignement du FLS. Leur expérience en tant qu'examinateur TEF va de trois à onze ans et ils évaluent en moyenne entre 100 et 300 candidats annuellement. Ils proviennent de quatre centres d'examen différents, cependant, six centres sont représentés dans le tableau, car trois d'entre eux proviennent de deux centres différents.

3.3. La collecte de données

Initialement, les rencontres devaient se faire individuellement en présentiel dans les salles de travail fermées des différentes bibliothèques municipales et universitaires de Montréal. Mais en raison de la pandémie de COVID-19, les rencontres avec les participants se sont faites individuellement en ligne. Le choix du moment de la rencontre a été laissé aux participants.

La collecte de données n'a pas nécessité de rencontres avec des candidats, mais a été faite à l'aide d'enregistrements audio fournis par la Chambre de commerce et d'industrie de Paris Île-de-France.

Ces enregistrements audio mettent en scène des entrevues authentiques de l'épreuve d'expression orale du TEF, c'est-à-dire des dialogues entre un candidat et un animateur. Ces dialogues n'ont pas été réalisés par les participants eux-mêmes, mais par des animateurs anonymes que l'on ne retrouve pas dans les formations suivies par les examinateurs. Les quatre niveaux de français des candidats sont 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 estimés au préalable par l'équipe pédagogique du Français des affaires de la Chambre de commerce et d'industrie de Paris Île-de-France. Aucune information sociodémographique n'est disponible pour les candidats et les animateurs. Un bref portrait des quatre candidats ainsi que de leur animateur respectif a été dressé (Tableau 18). D'après leur voix et le genre utilisé dans les conversations, les candidats sont trois femmes et un homme. Les pseudonymes Nora, Jane, Mina et Rayan leur ont été attribués. Quant aux animateurs, ils sont au nombre de trois, deux femmes et un homme. Ils ont été nommés par les lettres A, B et C.

Tableau 18 - Portrait des candidats et des animateurs

Pseudonyme du candidat	Sexe du candidat	Niveau estimé du candidat	Animateur assigné	Sexe de l'animateur
Nora	F	A2	A	F
Jane	F	B1	B	F
Mina	F	B2	B	F
Rayan	H	C1	C	H

3.3.1. Les outils de la collecte de données

Nous avons collecté nos données de recherche en deux temps, d'abord au moyen de la technique de la pensée à voix haute, puis d'une entrevue semi-dirigée.

Pour la technique de la pensée à voix haute, nous disposions de la grille d'évaluation et des enregistrements audio. Nous avons préalablement procédé à une mise à l'essai des outils de collecte de données avec l'aide de deux participants autres que ceux ciblés pour la recherche proprement dite. Compte tenu des observations de ces derniers, nous avons apporté certaines modifications et avons ainsi pu vérifier si nos outils de collecte prévus permettaient effectivement d'obtenir les informations recherchées en une durée prédéterminée.

3.4. Le déroulement de la recherche

La collecte de données a été réalisée durant l'été 2020. Chaque rencontre a duré environ deux heures au total, une pause a été accordée sur demande.

3.4.1. La technique de la pensée à voix haute

Le premier temps de notre recherche était sous-tendu par notre première question de recherche : En considérant différents aspects du jugement évaluatif, quelles divergences pouvons-nous observer chez les examinateurs ?

Nous avons eu recours à la technique de la pensée à voix haute afin de recueillir les pensées des examinateurs. Cela nous a permis « d'expliciter [leurs] processus mentaux interactifs (cognitifs et métacognitifs) et de les catégoriser de manière objective » (Tochon, 1996, p. 477). La verbalisation a légèrement été assistée étant donné que nous avons veillé à maintenir le continuum de la parole durant toute la procédure. Nous avons laissé s'exprimer les participants et leur avons fourni au

besoin quelques commentaires de rappel, sans toutefois orienter leurs propos, ni mettre au défi leur appropriation des pratiques évaluatives. Comme il y avait un minimum d'interventions de notre part, la validité de cette démarche a eu plus de chance d'être maximisée (Green, 2009).

L'épreuve d'expression orale du TEF dure au total 15 minutes et est composée de 2 sections, 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 (5 minutes), puis lorsque l'écoute s'est 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. Nous rappelons que plus la verbalisation se produit aussitôt après l'activité, plus elle est précise (Gass et Mackey, 2000).

Dans un deuxième temps, l'examinateur a écouté l'enregistrement de la section B (10 minutes) et lorsque l'écoute s'est terminée, il a fini de remplir la grille d'évaluation tout en verbalisant ses pensées. L'examinateur était invité à prendre des notes durant les écoutes. Les enregistrements audio se sont effectués à l'aide d'une tablette numérique, d'un ordinateur portable ainsi que d'un téléphone cellulaire. Nous avons également pris des notes de nos réflexions en aparté.

Pour chacun des dix participants, nous avons répété ce processus quatre fois étant donné que la tâche consistait à évaluer quatre candidats différents. Pour cette activité de verbalisation, le temps alloué était d'une heure trente environ au total pour chaque participant. Les quatre enregistrements des quatre candidats différents n'ont pas été écoutés dans le même ordre par tous les participants, car cela permettait d'éviter certains biais comme l'effet d'ordre (lorsque que l'on évalue de façon plus sévère ou plus clémente à la fin d'une session d'évaluation qu'au début) et l'effet de fatigue. L'ordre de passage des candidats a été le suivant (Tableau 19).

Tableau 19 - Ordre de passage des candidats selon les examinateurs

E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
1. Jane	1. Rayan	1. Rayan	1. Nora	1. Nora	1. Mina	1. Nora	1. Mina	1. Jane	1. Mina
2. Rayan	2. Nora	2. Nora	2. Mina	2. Mina	2. Jane	2. Jane	2. Nora	2.Rayan	2. Jane
3. Nora	3. Mina	3. Jane	3. Jane	3. Rayan	3. Rayan	3. Rayan	3. Rayan	3. Mina	3.Nora
4. Mina	4. Jane	4. Mina	4. Rayan	4. Jane	4. Nora	4. Mina	4. Jane	4. Nora	4. Rayan

Selon Smagorinsky (1994), verbaliser ses pensées intérieures n'est pas une tâche qui s'effectue machinalement, cela peut s'avérer être un exercice difficile étant donné que tous les participants n'y sont pas habitués. Une formation pratique est donc indispensable. Toutefois, Gass et Mackey (2000) soulignent que la formation contient un inconvénient, car elle risque d'influencer le participant en lui faisant penser qu'il doit évaluer les performances des candidats à l'instar du chercheur, c'est-à-dire en utilisant les mêmes termes que ce dernier, et cela se produit surtout dans des contextes où l'évaluation est standardisée. Pour éviter cela, les auteurs recommandent un minimum de formation avec des consignes très générales, mais claires à la fois. Ainsi, avant de commencer notre exercice de verbalisation, chaque participant a suivi une courte formation pendant environ une minute. Les directives ont été les plus brèves possible afin de minimiser les biais. Par ailleurs, comme le suggèrent Ericsson et Simon (1993) et Gass et Mackey (2000), nous avons évité de leur demander d'effectuer un quelconque raisonnement élaboré, nous leur avons plutôt demandé de verbaliser les pensées qu'ils avaient eues lors de l'opération d'évaluation. Comme le suggèrent les chercheurs, les questions étaient plus proches de la formulation « À quoi pensiez-vous ~ ? » que de la formulation « Pouvez-vous expliquer pourquoi ~ ? ». Toutefois, comme il est parfois difficile sur le terrain d'éviter que les commentaires des participants glissent vers l'analyse, nous les avons conservés afin de préserver l'intégrité des données.

Cet exercice de verbalisation requis par les examinateurs leur était quelque peu familier, car avant octobre 2018, la procédure d'évaluation standardisée du TEF exigeait qu'ils évaluent en binôme et simultanément un même candidat, puis qu'ils justifient les notes attribuées lors des accords interjuges. Les deux examinateurs devaient arriver à un consensus en faisant la moyenne de chacune de leur note respective, et pour ce faire, ils devaient se prononcer chacun à voix haute en étayant leurs arguments. En revanche, comme nous venons de le mentionner précédemment, nous ne souhaitions pas que les examinateurs se livrent à un raisonnement trop élaboré pour qu'ils ne se dirigent pas vers une analyse, mais nous avons voulu qu'ils évoquent uniquement leurs pensées au fur et à mesure qu'ils effectuent la tâche demandée. D'autre part, la configuration de l'exercice de verbalisation requise par les examinateurs dans notre recherche réplique une situation authentique d'évaluation du TEF, car la nouvelle procédure exige qu'une partie des examinateurs évalue les candidats en contexte et qu'une autre partie évalue hors contexte.

À la suite de l'activité de verbalisation, une courte entrevue semi-dirigée composée de cinq questions a été utilisée pour recueillir davantage d'informations.

3.4.2. L'entrevue semi-dirigée

L'entrevue consiste en une interaction verbale animée de façon souple par le chercheur et fournit au sujet l'occasion d'exprimer ses sentiments et ses opinions sur le thème traité (Fortin et Gagnon, 2016; Savoie-Zajc, 2004). L'objectif ici est de permettre au participant d'exprimer une opinion consciente sur son appropriation et son appréciation de la grille d'évaluation. Cette entrevue se situe en continuité avec l'activité de verbalisation, et permet d'avoir une compréhension plus approfondie du phénomène étudié en enrichissant l'analyse des résultats proposée dans cette recherche. Nous avons mené une entrevue de type semi-dirigée, c'est-à-dire située entre l'entrevue non dirigée où le participant contrôle le contenu, et l'entrevue dirigée où l'intervieweur exerce un contrôle maximum (Fortin et Gagnon, 2016).

Le but de l'entrevue a été de brosser le portrait de l'appropriation et de l'appréciation des examinateurs de la grille d'évaluation à travers cinq questions qui sont les suivantes :

1. Quel(s) critère(s) est(sont) plus facile(s) à évaluer selon vous ? Justifiez.
2. Quel(s) critère(s) est(sont) plus difficile(s) à évaluer selon vous ? Justifiez.
3. En quoi les descripteurs sont-ils appropriés selon vous ?
4. Quel(s) échelon(s) vous posent le plus de problèmes ? Justifiez.
5. Avez-vous déjà utilisé l'ancienne grille d'évaluation? Si oui, entre l'ancienne grille d'évaluation et la nouvelle, laquelle considérez-vous la plus facile à utiliser ? Justifiez.

Comme cette dernière question traite de la nouvelle version et de l'ancienne version de la grille d'évaluation, nous avons présenté les deux grilles aux participants.

Le temps alloué pour cette entrevue était de vingt minutes. Les questions n'ont pas été envoyées à l'avance aux participants et ont été posées oralement lors de la rencontre. Cela a permis que les réponses soient les plus spontanées possible avec le moins d'influence externe (Robert, 1988), et cela a donné la chance aux participants de formuler librement leurs réponses avec leurs propres termes (Blais, 1992; Daunais, 1992; Dörnyei, 2003; Fortin *et al.,* 1996; Mackey et Gass, 2005). Notons que bien que les questions ouvertes soient riches en renseignements grâce à la grande liberté de réponses qu'elles octroient, elles peuvent donner lieu à des réponses simples, vagues ou

difficiles à interpréter, car le temps de réponse accordé est assez court pour favoriser une participation optimale (Blais, 1992; Dörnyei, 2003).

Pour formuler ces questions, nous nous sommes inspirés du questionnaire de Brown paru en 2006 dans son article *An examination of the rating process in the revised IELTS Speaking Test*, et du questionnaire de Merrylees et McDowell paru en 2007 dans leur article *A survey of examiner attitudes and behavior in the IELTS oral interview*. Le questionnaire de Brown (2006) accompagne une étude qualitative réalisée en Asie et en Océanie portant sur le raisonnement évaluatif de 12 examinateurs du test d'expression orale IELTS *Speaking test*. Les questions de recherche sont les suivantes : « 1. Comment les examinateurs interprètent-ils les grilles et quelles caractéristiques de la performance orale sont saillantes dans leurs jugements ? 2. Est-il facile pour les examinateurs de différencier les différents niveaux de performance en relation avec la grille d'évaluation ? 3. Quels problèmes identifient les examinateurs lors de leurs prises de décision ? ». Le questionnaire de Merrylees et McDowell (2007), quant à lui, étudie les comportements évaluatifs de 151 examinateurs de l'épreuve d'expression orale du test d'anglais standardisé IELTS. Les chercheurs ont mené une enquête dans huit pays différents en Asie et en Océanie, à l'aide d'un questionnaire de satisfaction composé de 39 items. Les items portaient sur le format de l'examen, sur les grilles d'évaluation critériées et sur les phases de l'entrevue.

Nos questions ont été soumises pour rétroaction auprès de quatre juges experts (deux professeurs de l'Université de Montréal et deux responsables au sein de l'administration du TEF) pour que leur pertinence et leur clarté soient vérifiées, et pour qu'il y ait le moins de modifications à apporter sur la forme et le contenu par la suite. En outre, nous leur avons demandé d'ajouter leurs suggestions et leurs commentaires qu'ils jugeront appropriés. À ce sujet, van der Maren (1995) suggère de mettre à l'essai la formulation des questions, non seulement auprès d'experts, mais aussi auprès d'individus typiques de la population cible.

3.4.3. Les outils d'analyse

Nous avons effectué l'analyse de nos données durant l'automne 2020. Le corpus à partir duquel ont été élaborées les analyses est constitué de transcriptions des rapports verbaux et des entretiens individuels. Lors de la transcription, tous les éléments vains et non pertinents ont été intentionnellement éliminés afin d'éviter d'alourdir le volume du corpus. Les transcriptions ont été faites manuellement, c'est-à-dire sans logiciel de transcription automatique, puis les blocs de

données significatives ont été découpés et mis en correspondance avec des étiquettes et des sous-étiquettes afin de faire émerger du sens. Ce travail de codage des unités de sens s'est fait à l'aide du logiciel d'analyse qualitative QDA Miner (version 5.0).

L'analyse de nos données s'est organisée autour de trois étapes chronologiques : la préanalyse, l'exploitation du matériel, ainsi que le traitement des résultats, l'inférence et l'interprétation. (Wanlin, 2007).

Notre première étape, celle de la préanalyse, consistait à lire et 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, c'est-à-dire l'exploitation du matériel (Wanlin, 2007).

La deuxième étape, celle de l'étape de l'exploitation du matériel, consistait à classifier les éléments constitutifs d'un ensemble par différenciation puis regroupement par analogie (Bardin, 1977), puis à coder les unités.

Pour cela, nous avons d'abord dressé une liste de rubriques émergentes, puis établi des liens entre celles-ci, nous les avons regroupées en sous-catégories ou en méta-catégories. Ce processus était itératif, nous avons fait des allers-retours entre les verbatim et notre liste de rubriques afin de voir si les interprétations concordaient avec le matériel original. Un tableau de rubriques et de sous-catégories a été produit pour l'exercice de verbalisation ainsi que pour l'entrevue (Tableau 20).

Tableau 20 - Liste des rubriques et des sous-catégories ayant guidé l'étape d'exploitation du matériel

Exercice de verbalisation	
Rubriques	Sous-catégories
Section A	Complétude des questions; Qualité des questions; Gestion de l'imprévu; Qualité de l'échange; Arrêt des questions; Référence aux critères de la langue.
Section B	Présence de présentation; Absence de présentation; Qualité de la présentation; Présence de débat; Absence de débat; Richesse des arguments; Variété des arguments; Qualité de l'échange; Référence aux critères de la langue; Vouvoiement.
Syntaxe	Complexité des structures; Variété des structures; Quantité d'erreurs; Ampleur des erreurs; Phrases élémentaires; Phrases mémorisées; Capacité d'autocorrection; Niveau d'un locuteur natif.

Lexique	Variété du lexique; Confusion du lexique; Lexique lié à un thème particulier; Mots élémentaires; Emprunts d'autres langues; Niveau d'un locuteur natif.
Aisance à l'oral, élocution	Qualité de la prononciation; Vitesse du débit; Intensité de l'accent; Type d'intonation; Hésitations; Manque d'articulation; Intelligibilité générale; Familiarité prosodique; Niveau d'un locuteur natif.
Grille d'évaluation	Absence d'observables; Descripteurs ambigus; Descripteurs non compréhensibles; Descripteurs non correspondants; Descripteurs insuffisants; Souhait de case entre 2 échelons; Commentaires sans lien avec les descripteurs.
Animateur	Parle beaucoup/trop vite; Aurait expliqué la consigne; N'aurait pas expliqué la consigne; Comprend le candidat; Ne comprend pas le candidat; Pose trop de questions; Répond avec trop de précision; Relance le candidat; Corrige le candidat; Anime bien l'échange; Anime mal l'échange; Est expérimenté; N'est pas expérimenté.
Candidats	Origine culturelle/géographique; Degré de compréhension de la consigne/du sujet; Niveau de scolarité; Capacités diverses; Incapacités diverses; Stress; Âge; Assurance; Manque d'assurance; Dans de mauvaises conditions de passation; Profil atypique; Difficile à évaluer; Réussite du niveau B2; Capacité à se débrouiller dans la vraie vie.
Sujets des documents	Biais culturel; Biais lié à l'âge.
Autres	Interprétation différente d'un extrait de dialogue; Comparaison avec d'autres candidats; Examinateur aurait pu accorder une autre note; Incertitude de la note; Révision de notes; Durée trop longue ou courte de la conversation.
Entrevue	
Rubriques	Sous-catégories
Question 1 Critère 1	Simplicité; Rapidité; Prévisibilité; Nombre de questions.
Question 1 Critère 5	Concret; Renvoi à la langue; Mélodieux; Premier critère observé.
Question 1 Critère 3	Limites bien marquées; Renvoi à la langue; Rapidité.
Question 2 Critère 2	Comparaison avec l'ancienne grille; 2 éléments à évaluer; Difficulté culturelle des candidats; Manque d'idées des candidats; Candidats intermédiaires; Durée; Pas de prise de notes.
Question 2 Critère 4	Candidats francophones; Difficulté des descripteurs.
Question 2	Candidats francophones.

Critère 1	
Question 3 Descripteurs	Un mot fait la différence; Incomplets; Vagues; Regroupement de 2 actions; Trop similaires; Manque d'exemples; Référence au CECRL.
Question 4 Échelons B1/B2	Enjeux importants; Candidats entre 2 niveaux; Problème de langue des candidats; Examinateur peu expérimenté; Descripteurs trop formels.
Question 4 Échelons A1/C2	1 seul sous-échelon.
Question 5 Ancienne grille	Plus juste; Trop longue; Critères/descripteurs peu clairs; Ancienne procédure d'évaluation.
Question 5 Nouvelle grille	Rapide; Plus de doutes; Moins précise; Problème critère 2; Pondération des échelons; Trop de synonymes; Éléments regroupés; Suggestions d'amélioration; Conformité avec la procédure d'évaluation; Évaluation en présentiel vs à distance.

La troisième étape reposait sur le traitement, l'interprétation et l'inférence. Dans le cadre de cette recherche, nous n'avons pas réalisé d'opérations statistiques. Nous avons uniquement généré les fréquences de codages afin d'avoir un aperçu plus visuel des idées principales avancées. À la suite de cela, nous avons avancé des interprétations à propos des objectifs prévus et concernant d'autres découvertes imprévues, puis nous avons proposé des inférences. L'interprétation et l'inférence se feront dans des chapitres subséquents.

Selon des chercheurs (Gohier, 2004, Huberman et Miles, 1991; Van der Maren, 1995), on a longtemps reproché à l'analyse qualitative d'être subjective et la validité des analyses, c'est-à-dire leur crédibilité, leur stabilité et leur fiabilité, a alors été remise en question. Ainsi, afin d'apporter une démarche rigoureuse 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 de nos données. Comme ce processus est long et exigeant, nous nous sommes limités à 40 % du corpus entier plutôt qu'à l'ensemble recueilli. De façon à nous donner des repères communs, nous avons présenté au préalable notre arbre de codage au chercheur externe, et celui-ci avait la possibilité de faire émerger de nouvelles variables ou de regarder des données existantes d'une nouvelle manière. Lors de l'accord inter juge, les résultats ont corroboré à plus de 90%.

3.5. La synthèse de la méthodologie

Dans ce tableau de synthèse (Tableau 21), nous récapitulons les éléments essentiels de notre méthodologie : les objectifs spécifiques de recherche, les participants, les outils de collecte et les outils d'analyse.

Tableau 21 - Tableau de synthèse de la méthodologie

Participants	Objectifs spécifiques de recherche	Outils de collecte	Outils d'analyse
- Dix examinateurs TEF certifiés issus de différents centres d'examen TEF certifiés de Montréal.	1. Documenter les divergences lors des prises de décision des examinateurs.	1. Activité de verbalisation à l'aide de la nouvelle version de la grille d'évaluation et de quatre enregistrements audio.	- Transcription des données de façon manuelle. - Codage des données pertinentes à l'aide du logiciel QDA Miner.
	2. Brosser le portrait de l'appropriation et de l'appréciation des examinateurs de la grille d'évaluation.	2. Entrevue semi-dirigée contenant cinq questions à l'aide des deux grilles d'évaluation (la nouvelle et l'ancienne version).	

3.6. La position de la chercheure

Le chercheur demeure un sujet social indissociable de ses propres contextes et porteur de valeurs et de finalités à l'origine de ses travaux (Gephart, 1988). Plus la distance entre celui-ci et son terrain diminue, pour en arriver à disparaître, plus la question de la neutralité se pose en termes éthiques (Brasseur, 2012). Dans notre cas, la position de la chercheure (notre position) pourrait constituer un biais étant donné qu'elle est elle-même examinatrice TEF et qu'elle fait partie de la collectivité. Ce biais risque alors d'être préoccupant en raison de son effet potentiel sur la signification des résultats (Fortin et Gagnon, 2016). Cependant, il y a d'une part une absence d'enjeux personnels ou servant des intérêts particuliers, et d'autre part, la chercheure ne connait pas tous les participants à l'étude et son intervention auprès de ces derniers reste très minime.

Selon les termes de Junker (1960), elle peut être considérée comme chercheur participant-observateur. Cette position se situe entre l'observateur qui est neutre, hors du milieu, et le participant à part entière qui s'implique activement dans la recherche. La plupart des chercheurs participants-observateurs essaient de garder un équilibre entre l'engagement et la neutralité (Deslauriers, 1991). Cette neutralité se définit par la capacité d'empathie du chercheur lui permettant de « se décentrer par rapport à lui-même… » (Jodelet, 2003, p. 149) suivant un processus « d'acculturation à l'envers » (Laplantine, 1996, p. 20). De cette manière, la chercheure se place en toute lucidité dans une quête de non-influence totale, et accède à la réalité perçue par les sujets en la reliant à ses propres grilles de compréhension (Brasseur, 2012).

3.7. Les considérations éthiques

Selon Fortin et Gagnon (2016), le principe éthique le plus important dans les études menées auprès des êtres humains demeure la capacité d'une personne à donner son consentement après avoir reçu et bien compris toute l'information relative à sa participation à l'étude. Un formulaire de consentement formulé dans un langage accessible a donc été remis aux participants afin qu'ils prennent connaissance des tenants et aboutissants de l'étude. Comme stipulé dans le certificat d'éthique que nous avons obtenu (CEREP-20-022-D), nous nous sommes assurés de bien les informer de la politique de confidentialité des résultats de la recherche, de l'anonymat de leur identité et du centre d'examen auquel ils se rattachent. Nous les avons également informés de leur droit de mettre fin à tout moment à leur participation s'ils le jugent nécessaire.

D'autre part, les participants ont été informés du désagrément que la recherche allait entrainer, c'est-à-dire le temps consacré à la recherche, soit environ deux heures. En contrepartie, l'avantage qu'ils en ont retiré est leur contribution à l'avancement des connaissances au sujet de l'évaluation du français dans les situations de tests à des fins d'immigration et de citoyenneté visant à réduire les risques de variabilité et ainsi à optimiser la validité des scores attribués.

CHAPITRE 4: RÉSULTATS

Introduction

Dans de ce chapitre, nous présenterons les résultats obtenus afin d'établir des pistes de réponses à nos deux questions de recherche. En premier lieu, nous analyserons les évaluations réalisées par les examinateurs sur les performances de chaque candidat un à un, et critère par critère. Les parties principales traiteront des notes et des commentaires, de l'attitude de l'animateur, des éléments extérieurs à la grille d'évaluation, ainsi que des propos en lien avec les descripteurs de la grille d'évaluation. Les titres de ces parties seront similaires pour chacun des quatre candidats. Un bilan des résultats saillants sera fait au fur et à mesure pour chaque évaluation de candidat. En second lieu, nous présenterons les réponses apportées aux cinq questions de l'entrevue.

4.1. L'évaluation de la candidate Nora

Dans cette section, nous présenterons l'évaluation des 5 critères concernant la candidate Nora.

4.1.1. L'évaluation du critère 1 : section A « Capacité à obtenir des informations »

La consigne demandée à la candidate Nora dans la section A est de téléphoner à une école de langues afin d'obtenir des informations générales.

4.1.1.1. Les notes et les commentaires

Pour cette section correspondant au critère 1 « Capacité à obtenir des informations », la répartition des scores est la suivante : deux examinateurs ont accordé le score inférieur à A1 (< A1), six le score A1, et deux le score A2-1 (Tableau 22).

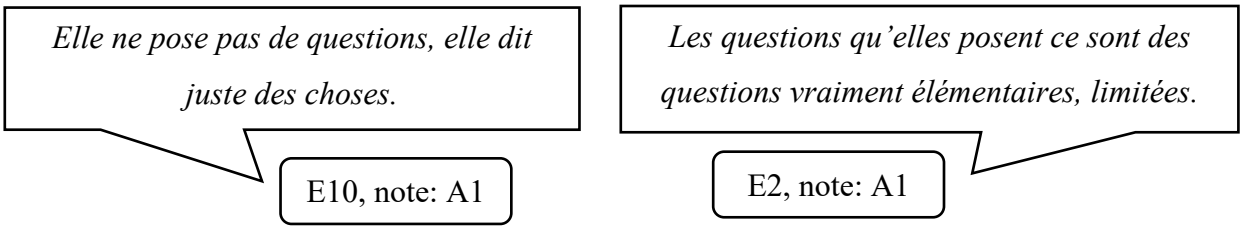
Tableau 22 - Répartition des scores du critère 1 de la candidate Nora

Scores	< A1	A1	A2-1
Nombre d'examinateurs	2	6	2

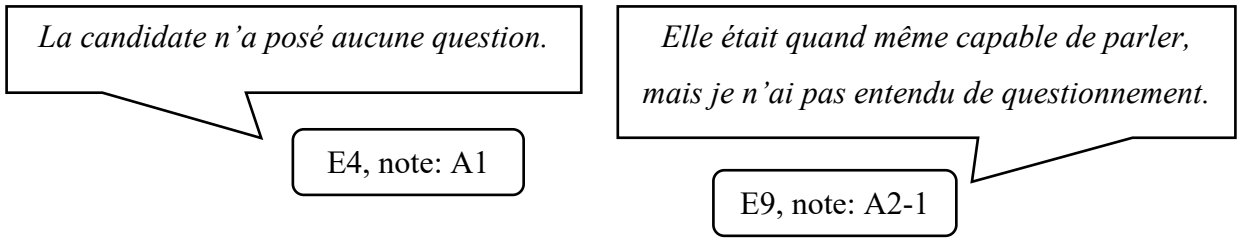
Les examinateurs ont eu une perception différente de la performance de la candidate et ne sont pas tous d'accord avec ce qu'ils ont entendu. Parmi les dix examinateurs, six affirment que la candidate n'a pas posé de questions (soit E5, E8, E1, E4, E10 et E9), les notes sont respectivement < A1, <

A1, A1, A1, A1, A2-1. Les quatre autres (E7, E2, E3 et E6) reconnaissent qu'elle en a posé, les notes sont A1, A1, A1, A2-1.

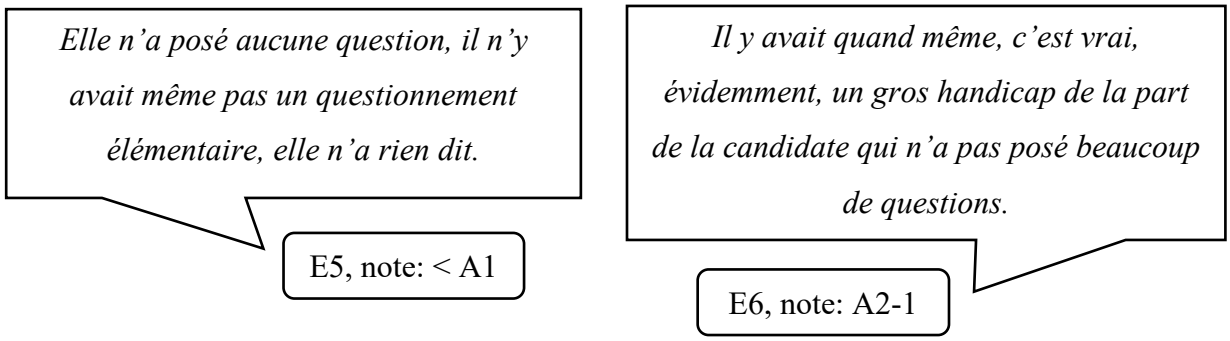
Les notes similaires et les commentaires divergents



Les notes divergentes et les commentaires similaires



Le degré de sévérité des commentaires différents



4.1.1.2. Les propos en lien avec les descripteurs de la grille d'évaluation

On observe dans nos analyses que certaines notes d'examinateurs ne concordent pas avec les descripteurs de la grille d'évaluation. Par exemple, parmi ceux qui affirment que la candidate n'a posé aucune question, trois d'entre eux (E1, E4, E10) lui ont attribué le score A1, et l'un d'entre eux (E9) le score A2-1. Cependant, sur la grille d'évaluation, on mentionne dans les descripteurs de ces niveaux-ci qu'il existe un questionnement, car les descripteurs pour les niveaux A1 et A2 sont respectivement nommés : « Questionnement élémentaire et limité. La conversation dépend entièrement de l'examinateur » et « Questionnement simple et général. Quelques demandes de clarification ou de reformulation. Les échanges sont peu suivis ». Ce qui a été observé par les

examinateurs ne correspond donc pas en grande partie aux descripteurs du niveau qu'ils ont choisis. D'ailleurs trois examinateurs (E1, E4 et E10) soulignent cet aspect dans leur commentaire. Ces derniers n'ont entendu aucune question de la part de la candidate, or ils ont opté pour le niveau A1 dont une partie du descripteur mentionne : « Questionnement élémentaire et limité ». E4 admet par ailleurs avoir été « *forcé* », selon ses propres termes, à choisir ce niveau, étant donné que la grille n'offre pas d'autres alternatives. Toutefois, ces trois examinateurs s'entendent pour dire que le deuxième énoncé du descripteur A1 : « La conversation dépend entièrement de l'examinateur » s'applique à ce qu'ils ont entendu.

Un autre cas relevé concerne le descripteur du score inférieur à A1 (< A1) « Absence d'observables ». On observe ici dans les commentaires un manque de consensus dans la compréhension de ce descripteur. Selon deux examinateurs (E5 et E8) une absence d'observables signifie que le candidat ne pose pas de questions, alors que selon trois autres examinateurs (E1, E3 et E10) cela signifie que le candidat ne parle pas du tout.

4.1.1.3. L'attitude de l'animatrice

Des commentaires envers l'attitude de l'animatrice ont été exprimés. Ils portent sur trois aspects différents : 1) le fait qu'elle ait décidé d'écourter la durée de la conversation; 2) le fait qu'elle aurait ou n'aurait pas expliqué la tâche à la candidate; 3) sa manière d'animer de façon globale.

La décision d'écourter la durée de la conversation

Six examinateurs (E3, E4, E6, E7, E8 et E9) ont commenté le fait que l'animatrice ait décidé d'écourter la durée de la section A de l'épreuve qui a été de 3 minutes 40 au lieu de 5 minutes. Parmi eux, E3 et E8 ont considéré que cette durée était suffisante, tandis que pour E4, E6, E7 et E9, il n'aurait pas fallu écourter la conversation. Ici, les deux types de commentaires soulèvent la question de la durée de la section A face à des candidats rencontrant de grandes difficultés à effectuer la tâche demandée. La question est de savoir s'il faut mettre fin à la conversation avant la fin en présumant que cela n'apporterait pas de changement à leur performance, ou s'il faut coûte que coûte tenter de poursuivre l'échange afin de respecter le temps imparti, malgré de nombreuses relances infructueuses.

L'explication de la tâche à la candidate

Face aux grandes difficultés de la candidate à bien comprendre l'exercice, les trois examinateurs (E9, E4 et E8) ont relevé le fait que l'animatrice aurait ou n'aurait pas expliqué clairement la tâche à effectuer au préalable. D'un côté, E9 et E4 ont supposé que l'animatrice aurait négligé l'explication et qu'elle aurait débuté la conversation sans s'être assurée que son interlocutrice ait totalement saisi ce qu'elle devait faire. Et d'un autre côté, d'après E8, l'animatrice aurait expliqué à la candidate son objectif en insistant sur les mots difficiles du document de support, mais celle-ci n'aurait de toute façon rien compris. Les a priori au sujet de la façon dont l'exercice a été annoncé sont alors différents.

L'animation globale

L'attitude générale de l'animatrice a été perçue différemment par les deux examinateurs E6 et E8. D'après E6, l'animatrice n'a pas bien donné suite à l'intervention de la candidate dès le départ, et ses premiers propos ont été abrupts. Par la suite, elle n'a pas compris ce que la candidate exprimait, c'est-à-dire son besoin d'apprendre le français pour répondre à des situations précises de sa vie quotidienne. Elle a alors coupé court à la conversation au lieu de revenir sur les éléments de contexte fournis par son interlocutrice. Par conséquent, l'attitude de l'animatrice a eu une influence néfaste sur le nombre de questions devant être posées. En revanche, d'après E8, l'animatrice a fait tout son possible pour aider la candidate en essayant par tous les moyens de lui faire poser des questions. Elle a adapté son niveau de langue en parlant clairement et lentement tout au long de l'échange, et a globalement agi de « *façon extraordinaire* » selon ses propres termes.

4.1.1.4. Les références extérieures à la grille d'évaluation

Deux types de commentaires n'ayant pas de lien avec la grille d'évaluation ont été émis : des commentaires faisant référence au niveau de scolarisation de la candidate et des commentaires faisant référence à sa culture.

Les références au niveau de scolarisation de la candidate

Les examinateurs E5 et E8 ont fait référence au niveau de scolarisation de la candidate. Pour E5, elle proviendrait d'un milieu peu instruit, et pour E8, elle serait même analphabète. Selon eux, ces aspects justifieraient la grande difficulté de la candidate à comprendre la tâche demandée.

Les références à la culture de la candidate

Les quatre examinateurs E2, E6, E7, E9 ont évoqué des aspects liés à l'origine culturelle de la candidate Nora. Par exemple, pour E6, E7 et E9, la candidate pourrait être originaire du continent africain ou d'Haïti. E7 constate à ce sujet que les candidats provenant d'Afrique du Nord ou des Caraïbes ont souvent de la difficulté à se projeter dans les jeux de rôle de l'épreuve. Il ajoute par ailleurs que souvent pour les candidats originaires de ces régions, jouer le rôle de l'ami de leur interlocuteur lors de la section B est embarrassant lorsque les deux sexes (du candidat et de l'animateur) sont opposés. Selon lui, comme l'amitié entre hommes et femmes serait une chose peu commune, il serait difficilement naturel de simuler cet état pour ces candidats.

D'après E2, dans la culture d'origine de la candidate Nora, le type de questionnement qui est exigé, c'est-à-dire le questionnement direct avec des questions et des réponses, n'existerait pas. Dans ladite culture, le questionnement serait plutôt indirect, autrement dit avec des phrases non interrogatives, mais déclaratives et sans intonation montante, comme la structure suivante : « J'ai besoin de savoir si... », au lieu de « Savez-vous si...? ». E2 explique que la candidate parle en continu, qu'il comprend ce qu'elle dit, mais que les choses qu'elle dit ne sont pas des questions au sens propre du terme. Cette différence culturelle justifierait la difficulté de la candidate à se faire comprendre. En outre, cet aspect soulevé fait écho avec un autre point qui a été relevé par E6. Ce dernier cite une réplique de la candidate, qu'il a modifiée, qui est la suivante : « *Moi je veux apprendre le français pour acheter mes légumes, acheter mes fruits, pour appeler les pompiers* ». Selon lui, la candidate aurait exprimé sa volonté en apportant un contexte précis, sa demande de renseignement aurait été alors sous-entendue, mais sans formulations de phrases interrogatives explicites.

4.1.2. L'évaluation du critère 2 : section B « Capacité à présenter et débattre »

Dans la section B, la consigne demandée à la candidate Nora est de présenter à son amie un document concernant des meubles fabriqués en carton puis de la convaincre d'en acheter en ligne.

4.1.2.1. Les notes et les commentaires

Les scores pour le critère 2 de la candidate Nora vont de < A1 (score inférieur à A1) à B1-1. Un examinateur a accordé le < A1 (le score inférieur à A1), sept le score A1, un le score A2-1 et un le score B1-1 (Tableau 23).

Tableau 23 - Répartition des scores du critère 2 de la candidate Nora

Scores	< A1	A1	A2-1	A2-2	B1-1
Nombre d'examinateurs	1	7	1	0	1

Dans l'ensemble, tous les examinateurs s'entendent sur les qualificatifs de la performance de la candidate Nora. Pour eux, la présentation était soit inexistante, soit très rudimentaire, et le débat soit inexistant, soit très pauvre.

Comme l'objectif de la section B est de présenter et de débattre, les examinateurs doivent observer ces deux objets, pourtant, il n'y a pas d'unanimité concernant leurs observations chez la candidate. Dans tableau ci-dessous (Tableau 24), on remarque que d'après E9, il y a eu une présentation et un débat, sa note est B1-1; d'après E6, E4, E10 et E2, il y a eu une présentation, mais pas de débat, leur note est A1, A1, A1, A2-1; d'après E8, il n'y a pas eu de présentation, mais un débat, sa note est A1; puis d'après E5, E1, E3 et E7, il n'y a eu ni présentation ni débat, leur note est < A1, A1, A1, A1.

Tableau 24 - Les éléments observés et les scores attribués pour le critère 2 de la candidate Nora

Éléments observés	Scores attribués	Identités des examinateurs	Nombre d'examinateurs
Une présentation et un débat	B1-1	E9	1
Une présentation, pas de débat	A1, A1, A1, A2-1	E6, E4, E10, E2	4
Pas de présentation, un débat	A1	E8	1
Pas de présentation, pas de débat	< A1, A1, A1, A1	E5, E1, E3, E7	4

Les notes similaires et les commentaires divergents

Pour essayer de convaincre, on est loin du compte, parce qu'elle-même dit « Ah non, il ne faut pas acheter les meubles en carton », c'était comme une très mauvaise vendeuse si elle devait convaincre son amie, la mission n'est pas accomplie, elle fait tout le contraire de convaincre, elle la dissuade.

E3, note: A1

De temps en temps elle donne son avis, mais de façon très simple c'est vrai. Elle va dire son avis sur les meubles, sur le bois, le carton. Et puis des fois elle dit que ça peut être dangereux, donc elle va donner son avis, donc il y a quelques bribes de communication qui rejoindraient une partie de débat, donc elle a un peu donné son avis.

E8, note: A1

Les notes divergentes et les commentaires similaires

La présentation c'était très sommaire, elle n'a pas dit grand-chose en fait.

E4, note: A1

Ça a limité énormément la présentation initiale, mais quand même elle a lancé une présentation, elle a présenté un autre type de meubles, des meubles en bois, mais quand même elle a présenté. quelque chose

E2, note: A2-1

Le degré de sévérité des commentaires différents

Elle n'a même pas essayé de convaincre, il n'y avait pas de présentation, il n'y avait pas la moindre intention de convaincre (...), la candidate elle réagissait, mais ce n'était pas un débat.

E5, note: < A1

Elle a commencé avec des choses qui n'étaient pas correctes, donc avec le temps, elle a réussi quand même à reprendre ce qu'elle devait dire, mais elle n'avait pas beaucoup d'arguments.

E9, note: B1-1

4.1.2.2. Les propos en lien avec les descripteurs de la grille d'évaluation

Les éléments observés chez certains examinateurs ne reflètent pas toujours ce qui est mentionné dans les descripteurs de la grille d'évaluation. Par exemple, les trois examinateurs E1, E3, E7 ont attribué le score A1, bien qu'ils n'aient pas entendu de présentation, ni le moindre échange de la part de la candidate, alors que le descripteur du niveau A1 mentionne ceci : « Présentation sommaire, simple lecture. Donne son avis de façon très simple. Les échanges sont difficiles et très limités. ». Concernant la présentation du document, la mention « simple lecture » du descripteur gêne les examinateurs E2, E8 et E10, car selon eux, la candidate n'a pas lu le document qu'elle devait présenter. Cette mention pose surtout un problème à E2 qui déclare que la candidate a produit des bribes de présentation, mais que le fait qu'elle n'ait pas lu le document la situerait au niveau B1 selon la grille d'évaluation. Or l'examinateur estime que la candidate n'a pas le niveau B1. À titre d'information, on nomme une présentation du document par une lecture dans les niveaux A1 et A2 (A1 : « Présentation sommaire, simple lecture », A2 : « Présentation très simple, lecture et paraphrase. »). Par ailleurs, E2 avoue que, de manière générale, lorsqu'il évalue des candidats, il est souvent compliqué pour lui de se fier aux descripteurs de la grille pour prendre une décision. Il utilise alors ses propres connaissances des niveaux du CECRL ainsi que son expérience en tant qu'enseignant-évaluateur pour attribuer des résultats qui s'approchent le plus possible des performances des candidats. Il ajoute que la grille d'évaluation n'est pas suffisante pour l'aider à porter son jugement, et que par conséquent, les descripteurs poussent parfois les examinateurs à se tromper.

Comme lors de la section A, on observe un manque de consensus dans la compréhension du descripteur « Absence d'observables » du niveau inférieur à A1 (< A1) dans la section B. Selon E5, une absence d'observables s'applique lorsqu'aucun des deux objectifs de la section B n'a été atteint (présenter et débattre), même si la candidate est capable d'avoir une conversation. À l'inverse, pour les trois examinateurs E1, E3 et E10, une absence d'observables ne s'applique pas lorsque la candidate produit juste des phrases.

4.1.2.3. L'attitude de l'animatrice

Parmi les commentaires sur l'attitude de l'animatrice dans cette section, E3, E7 et E8 ont mentionné que son travail d'animation était bon. Selon eux, elle a fait tout son possible afin de déclencher un débat : elle s'est ajustée au niveau de la candidate, elle l'a relancée en permanence

avec de nombreuses questions, et a veillé à ne pas dominer la discussion. Elle a été perçue dans l'ensemble comme encourageante et empathique.

En revanche, E4 et E9 ont vu que la candidate était dans une position de hors sujet et ont ressenti de la négligence de la part de l'animatrice. Les deux examinateurs pensent que dès le tout début de la conversation, cette dernière aurait dû rediriger explicitement la candidate vers le sujet devant être abordé, en l'occurrence celui de l'achat de meubles en carton. Les examinateurs présument que l'animatrice n'a pas pris la peine d'expliquer clairement l'objectif de l'exercice à la candidate avant le début de la conversation, et d'après eux, elle n'a alors pas correctement rempli sa fonction d'animation.

4.1.2.4. Les références extérieures à la grille d'évaluation

Les références à diverses compétences de la candidate

Compte tenu de la grande difficulté de la candidate, E8 et E10 ont conclu qu'elle ne savait pas lire ou qu'elle avait un manque de compétence en lecture. E10 précise que bien qu'il soit écrit sur le document « meubles en carton » et « Karton design », et qu'il y ait la présence d'une photo de divers meubles, la candidate n'a pas pu être en mesure de comprendre la nature du produit à présenter, et sa difficulté à lire en serait donc la cause.

Les références à l'âge de la candidate

E7, E8 et E9 ont fait allusion à l'âge de la candidate et ont déduit d'après sa voix qu'elle était assez âgée. Pour eux, l'âge de la candidate serait un facteur qui expliquerait sa difficulté à comprendre la consigne de l'épreuve (susciter un débat à partir de la présentation d'un produit) ou le concept du sujet (acheter des meubles en carton en ligne) qui est une pratique moins courante chez les personnes d'un certain âge. E7 précise que si le concept du sujet avait été plus conventionnel (comme acheter des meubles fabriqués avec des matériaux classiques dans un magasin) ou plus en lien avec le quotidien de la candidate, celle-ci aurait pu mieux réussir.

4.1.3. L'évaluation du critère 3 : « Syntaxe »

L'évaluation de la syntaxe se réalise au travers de la section A et de la section B.

4.1.3.1. Les notes et les commentaires

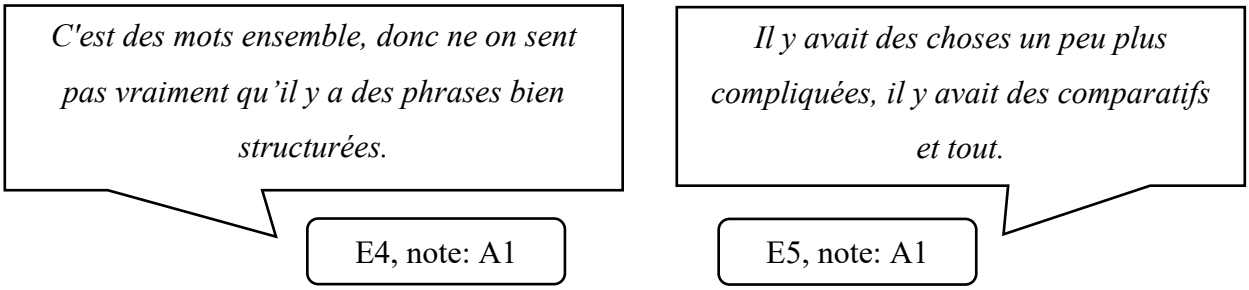
Concernant la répartition des scores pour le critère 3 de la candidate Nora, six examinateurs ont accordé le score A1, deux le score A2-1, un le score A2-2 et un le score B1-1. Les scores vont donc de A1 à B1-1 (Tableau 25).

Tableau 25 - Répartition des scores du critère 3 de la candidate Nora

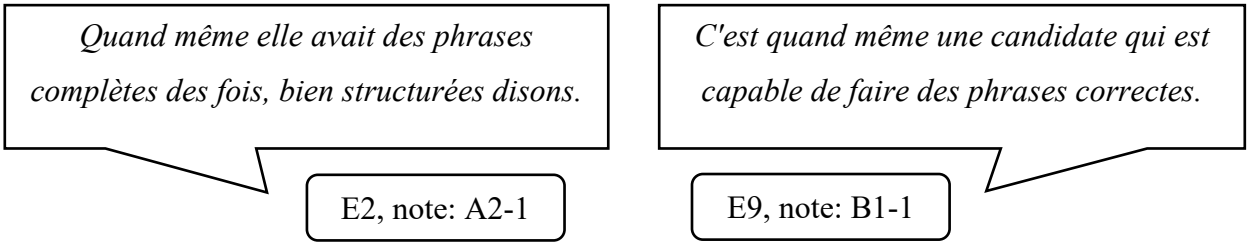
Scores	A1	A2-1	A2-2	B1-1
Nombre d'examinateurs	6	2	1	1

La moitié des examinateurs (E1, E3, E4, E8 et E10) ont relevé que la syntaxe de la candidate était très limitée et lacunaire et ont attribué A1. L'autre moitié, E5, E2, E7, E6 et E9, ont également observé une syntaxe très limitée et lacunaire, mais ont reconnu que la candidate était tout de même capable d'utiliser occasionnellement des phrases correctes, les notes sont respectivement A1, A2-1, A2-1, A2-2, B1-1.

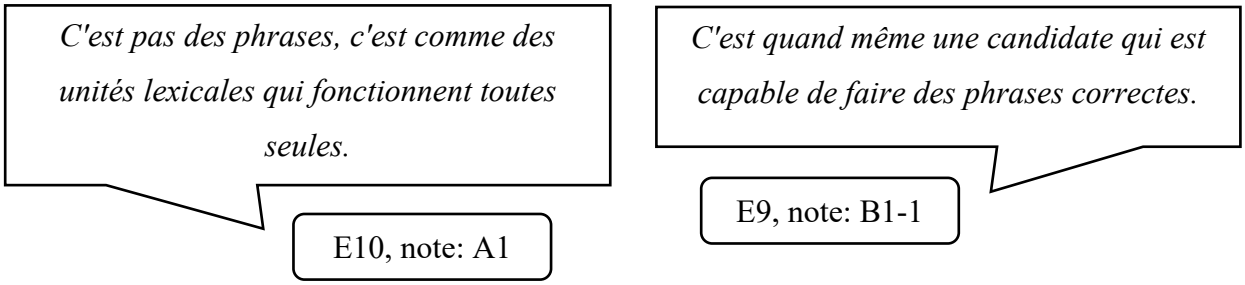
Les notes similaires et les commentaires divergents



Les notes divergentes et les commentaires similaires



Le degré de sévérité des commentaires différents



4.1.3.2. Les propos en lien avec les descripteurs de la grille d'évaluation

Les trois examinateurs E10, E2 et E5 ont relevé des aspects imprécis ou non applicables dans les descripteurs de la grille d'évaluation pour le critère de la syntaxe.

Selon E10, le juste niveau de la syntaxe de la candidate se situerait dans une zone inexistante entre < A1 et A1, étant donné qu'il n'y a pas eu d'absence d'observables (car la candidate s'est exprimée), que sa syntaxe était pratiquement inexistante, et donc que ses phrases n'étaient même pas d'un niveau élémentaire. À titre informatif, le descripteur du niveau < A1 s'intitule : « Absence d'observables » et celui du niveau A1 : « Structures et phrases élémentaires, souvent mémorisées ». L'absence de case entre ces deux niveaux contrarie l'examinateur et l'amène à faire un choix final par défaut, à savoir le niveau A1.

E2, quant à lui, pointe du doigt le descripteur du niveau A2 qui s'intitule « Phrases simples et stéréotypées. Erreurs systématiques, mais l'ensemble est compréhensible », et plus particulièrement le terme « l'ensemble » qu'il trouve ambigu. En effet, il ne sait pas à quoi ce terme fait référence et se demande si cela renvoie à l'ensemble du discours ou à l'ensemble des phrases simples et stéréotypées. Ce doute le fait hésiter dans ses choix.

Enfin, E5 trouve que le terme « mémorisées » dans le descripteur du niveau A1 qui se nomme : « Structures et phrases élémentaires, souvent mémorisées » ne s'applique pas à ce qu'il a observé chez la candidate étant donné que celle-ci était tout à fait spontanée, malgré son très faible niveau de syntaxe. Il affirme par ailleurs qu'il ne sait pas comment interpréter correctement ce terme, car il est normal selon lui d'utiliser des structures de phrases que l'on a mémorisées lorsqu'on apprend une langue. Il ajoute qu'il lui est souvent difficile de voir la limite entre une tournure de phrase dite de mémorisation et une tournure de phrase plus naturelle, surtout avec les candidats situés entre les niveaux A1 et B1.

4.1.4. L'évaluation du critère 4 : « Lexique »

L'évaluation du lexique se réalise au travers de la section A et de la section B.

4.1.4.1. Les notes et les commentaires

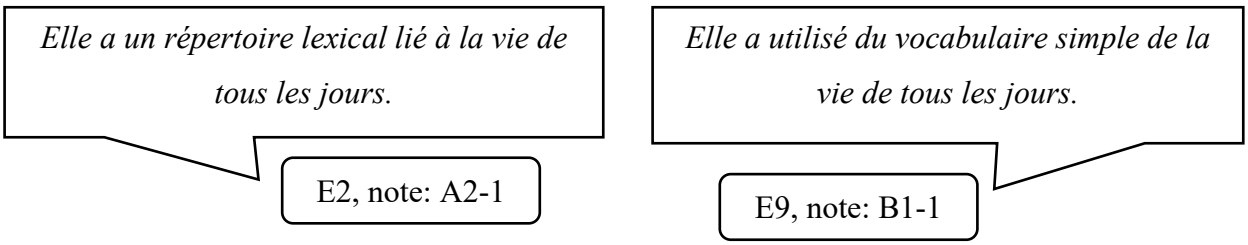
La distribution des scores pour le critère 4 de la candidate Nora s'échelonne ainsi : trois examinateurs ont accordé le score A1, quatre le score A2-1, deux le score A2-2 et un le score B1-1 (Tableau 26).

Tableau 26 - Répartition des scores du critère 4 de la candidate Nora

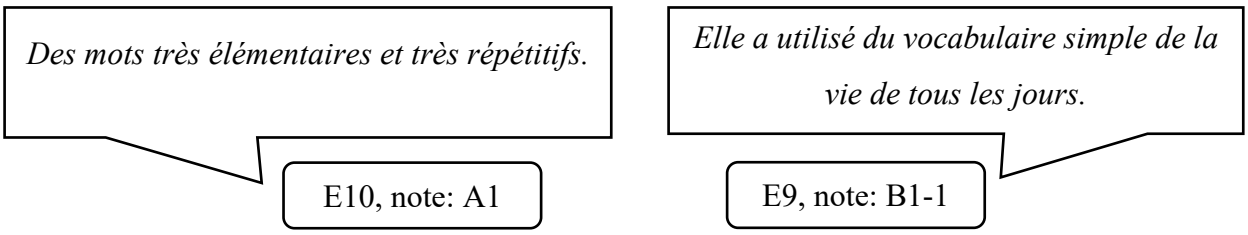
Scores	A1	A2-1	A2-2	B1-1
Nombre d'examinateurs	3	4	2	1

Tous les examinateurs s'entendent sur le fait que la candidate possède un lexique plus ou moins restreint. Cependant, on observe deux grandes catégories dans les notes et commentaires : ceux ayant juste perçu un lexique très élémentaire et ceux ayant constaté l'utilisation d'un vocabulaire un peu plus développé et lié à la vie quotidienne. Dans la première catégorie, trois examinateurs (E8, E3 et E10) ont octroyé la note A1, et dans la seconde catégorie, les sept autres (E1, E2, E4, E7, E5, E6 et E9) ont respectivement accordé A2-1, A2-1, A2-1, A2-1, A2-2, A2-1, B1-1.

Les notes divergentes et les commentaires similaires



Le degré de sévérité des commentaires différents



Lorsque l'on examine le commentaire de E9 qui a accordé B1-1, on remarque qu'il n'est pas en conformité avec le descripteur du niveau B1 de la grille qui se qualifie ainsi : « Répertoire lexical plus large permettant de s'adapter à la situation, à l'aide de périphrases ou d'emprunts à d'autres langues ». Le commentaire de E9 s'associerait plutôt au niveau A2 : « Répertoire lexical restreint aux situations quotidiennes et familières ».

4.1.5. L'évaluation du critère 5 : « Aisance à l'oral, élocution »

L'évaluation de l'aisance à l'oral et de l'élocution se réalisent au travers de la section A et de la section B.

4.1.5.1. Les notes et les commentaires

Les scores pour le critère 5 de la candidate Nora sont les suivants : six examinateurs ont accordé le score A1, un le score A2-1, un le score A2-2 et un le score B1-1 (Tableau 27).

Tableau 27 - Répartition des scores du critère 5 de la candidate Nora

Scores	A1	A2-1	A2-2	B1-1
Nombre d'examinateurs	6	2	1	1

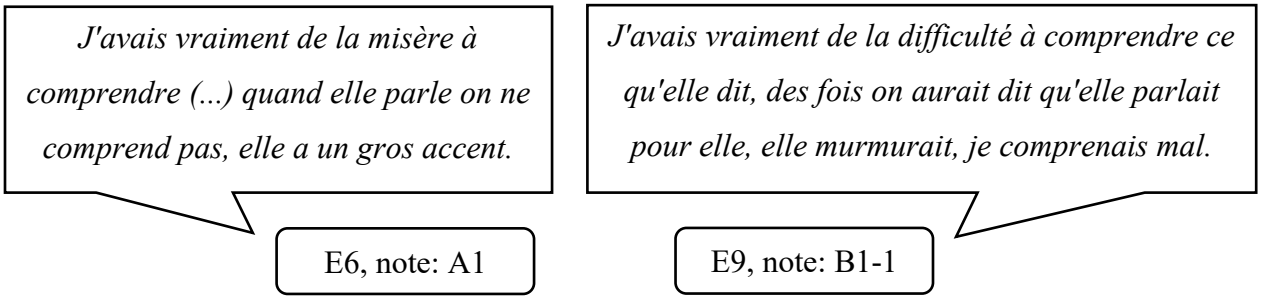
Au regard des descripteurs du critère 5, celui-ci englobe quatre composantes : la prononciation, le débit, l'intonation et l'accent. Parmi ces quatre composantes, la prononciation et le débit ont surtout été commentés. Huit examinateurs (E2, E3, E4, E6, E7, E8, E10 et E9) pensent que le discours de la candidate était très peu intelligible et que son manque d'articulation rendait la compréhension fort difficile, les notes sont respectivement A1, A1, A1, A1, A1, A1, A1, B1-1. Pour les deux autres, E1 et E5, le discours était plus ou moins compréhensible malgré de nombreuses faiblesses, leur note est A2-1.

Lorsque l'on regarde de près le commentaire de E9 qui donné la note la plus élevée, soit B1-1 : « *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* », on remarque qu'il ne coïncide pas avec le descripteur de la note B1-1 qui mentionne : « Hésitations fréquentes, mais le discours est clair. Des erreurs de prononciation peuvent parfois gêner la compréhension ». Le commentaire de E9 est plutôt en lien avec la note A1 dont le descripteur indique : « Discours peu intelligible, la compréhension est difficile et demande beaucoup d'efforts ».

D’autre part, E1 et E5, qui ont tous deux attribué la note A2-1, font remarquer qu’ils auraient pu donner une note plus haute à la candidate. Par exemple, E1 aurait pu accorder B2, car il trouve que celle-ci a une élocution continue et correcte en dépit de toutes ses difficultés. De plus, l’examinateur reconnait qu’il est familier avec la prosodie de la candidate étant donné qu’il affirme avoir les mêmes origines géographiques qu’elle, à savoir le Maghreb. Il dit bien comprendre la candidate et connaitre beaucoup de femmes provenant de cette région qui parlent français comme elle. Cependant, il avoue que cette familiarité représente une part de subjectivité.

E5 qui a accordé A2-1 admet qu’il aurait pu donner B1 à la candidate, car malgré ses grandes faiblesses de prononciation, il a trouvé son intonation naturelle et spontanée. Selon lui, la candidate a l’habitude de parler français et semble être à l’aise, car elle n’a pas « *la voix robotique d’une étudiante débutante qui cherche ses mots dans la tête* » selon ses propres termes. Par ailleurs, il suppose que si la candidate avait eu un sujet qui lui était plus familier, c’est-à-dire traitant d’une situation ordinaire de la vie quotidienne comme une conversation avec un marchand de fruits, celle-ci aurait fait mieux.

Les notes divergentes et les commentaires similaires



4.1.5.2. Les propos en lien avec les descripteurs de la grille d’évaluation

E10 a attribué A1 et reconnait que son choix de note ne correspond pas à ce qu’il aurait véritablement voulu donner. Il déclare que le discours de la candidate n’était aucunement intelligible, qu’il ne comprenait absolument rien, et que sa vraie note se situerait dans une zone fictive entre les niveaux < A1 et A1 étant donné qu’aucun des descripteurs de ces deux niveaux ne s’applique à ce qu’il a observé. Pour lui, une mention « *Discours non intelligible, la compréhension est impossible* » devrait exister dans la grille d’évaluation. Pour rappel, le descripteur complet du niveau A1 indique : « Pauses très nombreuses. Discours peu intelligible, la compréhension est difficile et demande beaucoup d’efforts », et celui du niveau < A1 : « Absence d’observables ».

4.1.6. Les compléments de l'évaluation de la candidate Nora

Après avoir terminé l'exercice d'évaluation, cinq examinateurs (E1, E5, E8, E9 et E10) ont fait la synthèse de la performance de la candidate et sont revenus sur divers éléments comme ses difficultés, son encadrement durant le test, ses caractéristiques personnelles et notamment sur son niveau de scolarisation.

E9 a accordé le score général le plus élevé, soit A2, et explique que la candidate était capable de parler français, mais qu'elle a éprouvé de la difficulté à comprendre la tâche qu'elle devait réaliser, c'est-à-dire jouer un jeu de rôle en recueillant des informations pertinentes dans la première section, puis en présentant une situation et en développant des idées dans la deuxième section. Selon lui, la candidate n'aurait pas suffisamment obtenu d'encadrement pour être dans des conditions optimales pour passer le test, car il pense que l'animatrice ne lui aurait pas expliqué de manière claire ce qu'elle devait faire**.** De plus, l'examinateur affirme que le fait que la candidate soit âgée et issue d'une culture différente constitue un obstacle pouvant entraver sa compréhension des sujets du test. Il remet alors en question certains sujets conçus par l'organisme concepteur du test, qui selon lui, peuvent être difficilement compréhensibles chez des candidats d'un certain âge et d'une certaine culture.

Trois examinateurs (E1, E8 et E10) sont revenus sur le niveau de scolarisation de la candidate. Par exemple, E1 a conclu que la candidate représentait un certain profil de femmes, originaires du Maghreb, ayant acquis des bases suffisantes en français pour être fonctionnelles dans un environnement proche de la vie courante, mais n'ayant jamais progressé du fait de leur absence de scolarisation. Un commentaire de E8 rejoint les propos de E1. Selon lui, la candidate aurait un faible niveau de scolarisation, car elle aurait probablement quitté l'école étant jeune pour se lancer sur le marché du travail ou pour s'occuper de sa famille. Elle n'aurait jamais fait d'études de sa vie, n'aurait pas l'expérience de l'apprentissage théorique de manière générale, et cela expliquerait son faible niveau de français. Par ailleurs, E8 s'interroge sur l'épreuve d'expression orale du TEF en faisant remarquer que cette dernière n'est pas adaptée pour la candidate dans la mesure où elle ne sait pas lire, et donc dans la mesure où elle ne peut pas exploiter d'informations écrites.

E10, quant à lui, dit n'avoir jamais eu de candidates comme Nora durant son expérience en tant qu'examinateur TEF, c'est-à-dire des femmes originaires d'Afrique, ne sachant pas lire et ayant un niveau grand débutant (A1) en français. Lorsqu'il évoque son expérience en tant

qu'examinateur TEF, il déclare que les candidates qu'il a déjà eues provenant de la même zone géographique que Nora et ne sachant pas lire étaient toutes assez compétentes à l'oral, contrairement à Nora. L'examinateur évoque par ailleurs qu'à cause de leur défaillance en lecture et dans leur façon d'argumenter, ces candidates-là perdaient beaucoup de points au test TEF.

Enfin, E5 sort du cadre de l'évaluation du test et examine les choses sous une perspective plus large. Il évoque l'enjeu élevé du test qui a pour finalité l'intégration des immigrants dans un pays donné, et mentionne que l'épreuve d'expression orale devrait mieux refléter la capacité des personnes à faire face à des situations concrètes du quotidien. L'examinateur affirme que lorsqu'il est face à des candidats de faible niveau, il se questionne souvent sur leur capacité à se débrouiller dans la vraie vie, mais avoue toutefois que cela représente un biais pouvant fortement influencer ses choix de notes. En parlant de la candidate Nora, il affirme que celle-ci aurait été capable d'assez bien se débrouiller au quotidien dans un territoire francophone, bien qu'elle ait de grandes difficultés sous tous les angles en français. Cependant, la note globale qu'il a attribuée à la candidate Nora est A1, car il reconnait que le cadre de la grille d'évaluation l'a dirigé vers ce résultat.

4.1.7. Le bilan de l'évaluation de la candidate Nora

Ainsi, beaucoup de divergences ont été relevées avec la candidate Nora, en particulier au sujet de l'identification des objectifs visés de la section A et de la section B, ainsi que sur la notion d'absence d'observables. Concernant l'utilisation de la grille d'évaluation, certains examinateurs ont été amenés à faire des choix par défaut en raison de la non-application de certains descripteurs. Par ailleurs, les grandes difficultés de la candidate ont soulevé beaucoup de questions périphériques assez récurrentes, et l'animatrice n'a pas été perçue de façon identique.

Le niveau global estimé de la candidate Nora par l'équipe du Français des affaires[38] (ÉFA) est de A2, mais nous constatons que seulement deux examinateurs sur dix lui ont attribué ce niveau (E6 et E9). Les huit autres examinateurs ont accordé le niveau A1. La moyenne et la médiane sont A1. (Tableau 28).

Tableau 28 - Niveau global de la candidate Nora par les dix examinateurs et par l'équipe du Français des affaires (ÉFA)

[38] Étant donné que nous n'avions pas accès aux données quantitatives de la grille d'évaluation, l'équipe du Français des affaires a effectué la conversion des échelons de la grille en valeurs chiffrées pour chaque candidat.

Examinateurs	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	Moyenne	Médiane	ÉFA
Niveau global	A1	A1	A1	A1	A1	A2	A1	A1	A2	A1	A1	A1	A2

Dans la partie suivante, nous ferons l'analyse de l'évaluation de la deuxième candidate, Jane.

4.2. L'évaluation de la candidate Jane

Dans cette section, nous présenterons l'évaluation des 5 critères concernant la candidate Jane.

4.2.1. L'évaluation du critère 1 : section A « Capacité à obtenir des informations »

La consigne demandée à la candidate Jane dans la section A est d'obtenir des informations à propos de l'achat d'un appartement.

4.2.1.1. Les notes et les commentaires

Les notes pour le critère 1 de la candidate se répartissent de A2-2 à C1-2, ce qui est une répartition étendue. Un examinateur a accordé la note A2-2, un la note B1-1, cinq la note B2-1, deux la note B2-2, et un la note C1-2 (Tableau 29).

Tableau 29 - Répartition des scores du critère 1 de la candidate Jane

Scores	A2-2	B1-1	B1-2	B2-1	B2-2	C1-1	C1-2
Nombre d'examinateurs	1	1	0	5	2	0	1

Les avis des examinateurs sont un peu mitigés. Par exemple, pour E2, les questions de la candidate étaient très élémentaires et l'échange peu suivi, sa note est A2-2. Pour E1, les questions étaient pertinentes, mais pas suffisantes, sa note est B2-1. Puis pour les huit autres examinateurs (E7, E9, E3, E4, E5, E6, E10 et E8), la candidate a posé beaucoup de questions adéquates et l'échange avec l'animatrice était suivi, les notes sont respectivement B1-1, B1-2, B2-1, B2-1, B2-1, B2-2, B2-2, C1-1.

Les notes similaires et les commentaires divergents

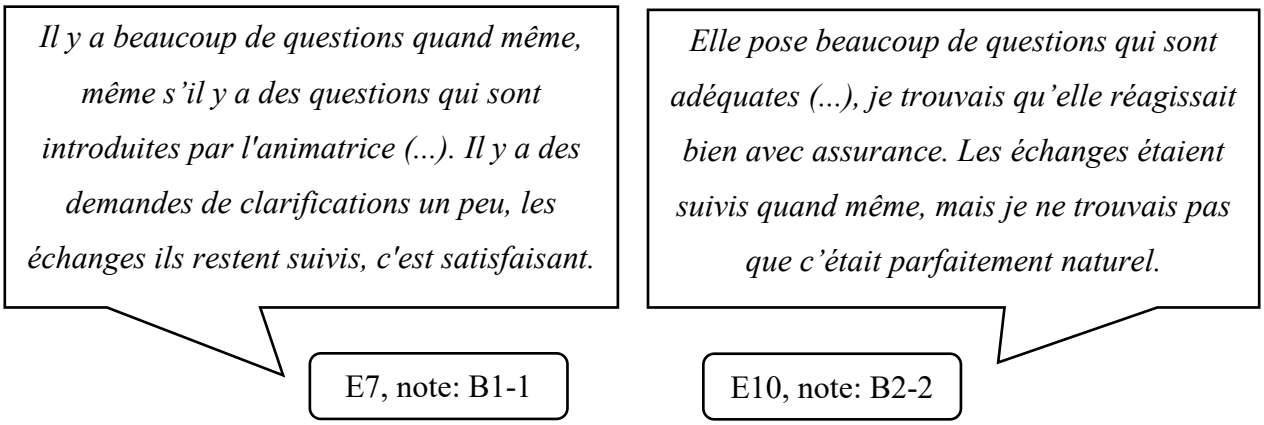
Elle a posé beaucoup de questions, vraiment beaucoup de questions, y compris des questions après les interventions de l'animatrice (...) concernant l'achat d'un appartement c'est bon.

E5, note: B2-1

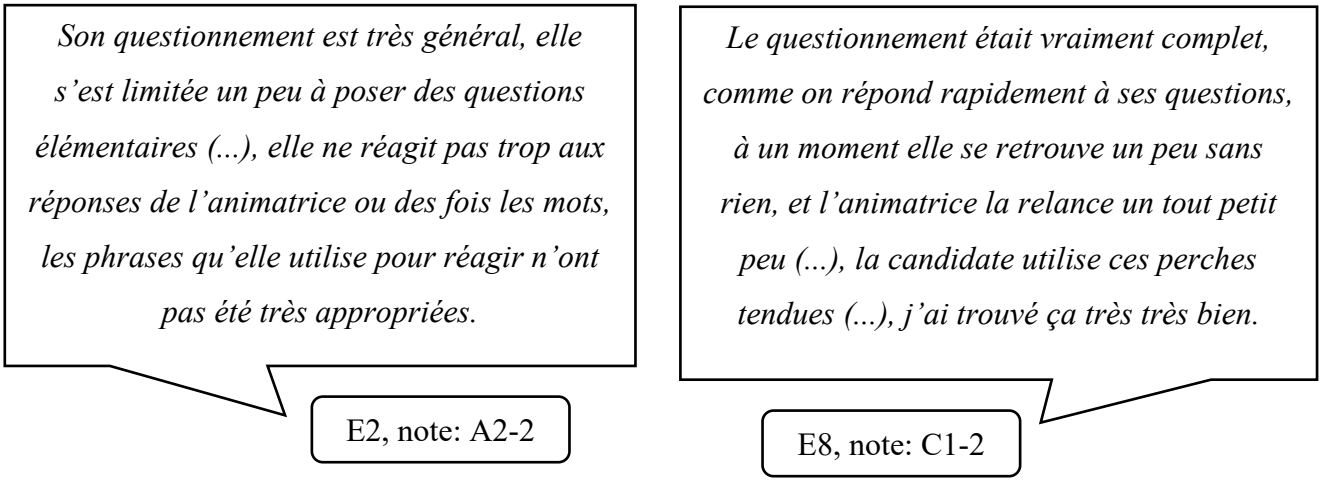
Il y a eu quand même pas mal de bonnes questions, mais pas suffisamment (...) à un moment elle a bloqué, elle n'avait plus envie de poser des questions.

E1, note: B2-1

Les notes divergentes et les commentaires similaires



Le degré de sévérité des commentaires différents



4.2.1.2. Les références extérieures à la grille d'évaluation

Les examinateurs ont émis quelques commentaires sur des éléments linguistiques au sein des critères communicatifs.

L'enchevêtrement des critères linguistiques et des critères communicatifs

Dans la section A, les cinq examinateurs E2, E3, E4, E7 et E10 ont relevé des éléments linguistiques comme des lacunes syntaxiques et lexicales. Ils mentionnent par exemple que souvent les structures des questions de la candidate n'étaient pas libellées parfaitement et qu'elle utilisait quelques mots en anglais dans la conversation. Bien entendu, les éléments linguistiques ne relèvent pas du critère 1 qui porte sur la section A, car cette dernière cible la capacité à obtenir

des informations. Les éléments relevés font référence aux critères 3 et 4 qui se rapportent à la syntaxe et au lexique.

Les extraits de dialogues interprétés différemment

Dans la section A, où le sujet de la conversation porte sur la vente d'un appartement, nous avons observé dans les commentaires que deux extraits de dialogue ont été interprétés de façon différente. Nous avons transcrit les deux extraits du dialogue entre la candidate et l'animatrice.

Extrait 1 :

Minute : 1:27
La candidate : − *Quel haut est l'appartement ?*
L'animatrice : − *À quel étage ?*
La candidate : − *Dans l'appartement quel haut ?*
L'animatrice : − *La hauteur des plafonds?*
La candidate : − *Oui, oui.*
L'animatrice : − *Ils sont très hauts, ils font 3 mètres 50.*

Dans ce premier extrait de dialogue, E8, E9 et E10 ont commenté l'utilisation du mot « *haut* » dans « *quel haut est l'appartement ?* ». D'après E8, la candidate parlait de la hauteur de l'appartement, alors que d'après E9 et E10, celle-ci souhaitait parler de l'étage de l'appartement. E8 pense que la question est bonne, car il est pertinent d'obtenir des informations sur la hauteur des plafonds. En revanche, E9 et E10 sont certains que la candidate a utilisé le mot « *haut* » parce qu'elle ne connaissait pas le mot « *étage* », bien que l'animatrice lui ait fourni le mot. E9 rajoute que de peur d'être pénalisée, la candidate s'est finalement accommodée de la question de l'animatrice : « *la hauteur des plafonds ?* » en répondant oui afin de ne pas briser le fil de la conversation.

Extrait 2 :

Minute : 0:41
La candidate : − *Combien de chambres il y a dans l'appartement?*
L'animatrice : − *Actuellement, il y a deux chambres, mais il y a possibilité de faire une troisième chambre.*
Minute : 2:04
La candidate : − *Alors on peut construire une autre chambre?*
L'animatrice : − *Oui absolument, vous pouvez agrandir.*

Dans ce deuxième extrait de dialogue, à partir de la minute 0:41, la candidate demande le nombre de chambres dans l'appartement, l'animatrice lui répond deux chambres, puis ajoute qu'il y a une possibilité d'en faire une troisième. Plus tard, à partir de la minute 2:04, la candidate revient sur ce point et demande si on peut construire une autre chambre. Les deux examinateurs E6 et E9 ont commenté cette seconde partie du dialogue (celle à partir de la minute 2:04). Pour E9, la candidate a juste repris la réponse de l'animatrice pour en faire une nouvelle question, mais qui est plutôt une confirmation. E6, quant à lui, ne semble pas avoir remarqué cela, pour lui, la question : « *alors on peut construire une autre chambre?* » a été très pertinente, car il affirme ne jamais avoir entendu de telle question avec ce type de sujet. Il a d'ailleurs trouvé quelques-unes de ses questions assez originales, et grâce à cette originalité, la candidate mériterait une note plus élevée selon lui.

Ainsi, il peut exister différentes interprétations du sens dans de courtes séquences de dialogue avec un vocabulaire et une syntaxe relativement simple.

4.2.1.3. Les éléments extérieurs à la grille d'évaluation

Des commentaires extérieurs à la grille d'évaluation portant sur la durée de l'épreuve ont été émis. En effet, deux examinateurs (E6 et E7) ont signalé que la section A avait duré environ une minute de plus, et que cela n'est normalement pas permis. E7 ajoute que l'animatrice a autorisé beaucoup de questions vers la fin du temps imparti, alors qu'elle aurait dû clore l'échange.

4.2.2. L'évaluation du critère 2 : section B « Capacité à présenter et débattre »

La consigne demandée à la candidate Jane dans la section B est de présenter à son amie une activité où il s'agit de rencontres dites express de sept minutes pour trouver un emploi, puis d'essayer de la convaincre de participer à cet événement.

4.2.2.1. Les notes et les commentaires

La répartition des notes pour le critère 2 de la candidate est assez uniforme, puisque six examinateurs ont accordé la note B1-1 et quatre la note B1-2 (Tableau 30).

Tableau 30 - Répartition des scores du critère 2 de la candidate Jane

Scores	B1-1	B1-2
Nombre d'examinateurs	6	4

L'objectif de la section B étant à la fois de présenter le contenu d'un document et de débattre, seulement trois examinateurs (E1, E4 et E7) ont fait référence à la capacité à présenter, et cela, de façon très succincte. Les propos de la totalité des examinateurs se sont donc principalement rapportés à la capacité à débattre de la candidate.

Les commentaires des trois examinateurs (E1, E4 et E7) au sujet de la présentation de l'activité sont analogues. Il a été dit par exemple que la présentation était simple, mais assez claire, et que la candidate n'avait pas dépassé le cadre du contenu du document. Concernant sa capacité à débattre, les commentaires sont également homogènes, car tous les examinateurs s'entendent sur le fait que ses arguments étaient assez bons de façon globale, mais qu'ils manquaient trop de développement et qu'ils étaient parfois peu clairs. Cet accord général explique la distribution des notes : B1-1 et B1-2. Pour E8, par exemple, les arguments étaient souvent des relectures des parties du document de support, sa note est B1-1. E3, quant à lui, a accordé B1-2 et trouve que la candidate n'était pas très convaincante, car elle se laissait facilement contester, elle utilisait à maintes reprises des formules non adaptées comme « *oui, peut-être que tu as raison* ».

4.2.2.2. Les révisions de notes

À la fin de l'écoute de la section A, E6 a mis la note C1-1 à la candidate pour le critère 1, mais après avoir terminé d'écouter la section B, il a décidé de revenir sur la note du critère 1 et de la baisser en mettant B2-2, car il a constaté qu'il y avait un trop grand écart entre la note du critère 1 (de la section A) qui est C1-1, et la note du critère 2 (de la section B) qui est B1-2. Selon lui, les notes des deux sections devraient logiquement être homogènes, car dans sa pratique, il remarque que lorsque l'écart est très vaste entre les notes des deux sections (lorsque la section A est meilleure que la section B), souvent les candidats sont extrêmement préparés, ils apprennent par cœur les questions et même les éventuelles réponses des sujets de la section A grâce à des stratégies particulières afin de dissimuler leur manque de compétence en français. À titre de rappel, la tâche de la section A est d'un niveau de complexité inférieur à la tache de la section B. Néanmoins, E6 admet qu'il n'a pas décelé de discours appris par cœur chez la candidate qu'il a évaluée dans la section A, mais il reconnait qu'il s'est laissé impressionner par certaines questions qu'il qualifie de très intelligentes et d'originales, comme la hauteur des plafonds de l'appartement, l'équipement de la cuisine et la construction d'une future chambre.

On observe également chez E8 un écart de notes assez important entre la section A et la section B, soit C1-2 pour la section A, et B1-1 pour la section B. Pourtant, l'examinateur n'a pas décidé de revenir sur sa note de la section A afin de l'homogénéiser avec celle de la section B, contrairement à E6. E8 explique le décalage de ses deux notes par le fait que le sujet de la section A (l'achat d'un appartement) soit plus facile en comparaison avec le sujet de la section B (des rencontres express pour trouver un emploi) qui est plus complexe. Il affirme de plus que le thème principal du sujet de la section A, à savoir le thème de la maison, est étudié assez tôt dans l'apprentissage d'une langue, et que de plus, la mission de la section A qui consiste à recueillir des informations est beaucoup plus sommaire que celle de la section B qui consiste à s'approprier une situation et à convaincre son interlocuteur de faire une activité à l'aide d'arguments persuasifs. Pour E8, la section B est plus révélatrice du véritable niveau des candidats que la section A, et dès lors, le décalage de ses deux notes, C1-2 et B1-1, démontre juste que les compétences en français de la candidate Jane étaient limitées, car celle-ci n'a pas été capable de s'exprimer sur des questions assez profondes.

4.2.2.3. Les références extérieures à la grille d'évaluation

Trois types de références extérieures à la grille d'évaluation ont été relevés : des critères linguistiques à la place des critères communicatifs, des remarques quant à la culture de la candidate, puis quant à l'attitude de l'animatrice.

L'enchevêtrement des critères linguistiques et des critères communicatifs

Dans la section B, les quatre examinateurs E4, E5, E6 et E8 ont relevé des éléments linguistiques dans leur commentaire comme des erreurs syntaxiques et des lacunes lexicales. Comme nous l'avons vu précédemment, ces éléments linguistiques doivent être pris en compte dans les critères 3 et 4, puisque le critère 2 ne vise que des éléments communicatifs.

Les références à la culture de la candidate

Les trois examinateurs, E2, E6 et E8, pensent que les difficultés de la candidate à trouver des arguments pour débattre peuvent ne pas résulter d'un manque de compétence en français, mais d'une non-connaissance ou d'une mauvaise interprétation du concept de son sujet, c'est-à-dire des entretiens d'embauche express d'une durée de sept minutes dans le cadre d'un salon de l'emploi. E2 précise que cette situation s'applique à beaucoup de candidats qui passent le test de manière

générale, et que cet aspect demeure culturel. Selon lui, tout apprenant d'une langue étrangère devrait en principe en savoir davantage sur les pratiques propres à la culture cible, car la langue et la culture entretiennent des rapports très étroits. Néanmoins, l'examinateur reconnait que la culture de la langue française est multiple étant donné la diversité géographique des pays francophones.

Dans la même veine, E8 fait allusion à d'autres sujets du test qui ne font pas partie des pratiques courantes des candidats de certaines cultures et également d'un certain âge. Il cite l'exemple du concept de l'échange de maisons entre particuliers dans le cadre des voyages à l'étranger. Selon lui, l'idée qu'une personne soit hébergée dans un autre pays que le sien chez des inconnus sans leur présence pendant que ces derniers sont chez elle pourrait paraitre inconcevable pour certains candidats, notamment ceux qui sont originaires de pays non occidentaux.

Ainsi, les examinateurs sont conscients que des disparités culturelles dans les sujets proposés peuvent entraver la performance des candidats, et par conséquent rendre leur évaluation difficile.

4.2.2.4. L'attitude de l'animatrice

Sept examinateurs ont commenté l'attitude de l'animatrice, mais suivant deux points de vue bien différents. D'après quatre examinateurs, celle-ci a parlé de façon excessive, alors que d'après trois autres, elle s'est montrée très encourageante.

D'après E1, E6, E7 et E9, l'animatrice a accaparé la conversation avec toute une série de contre-arguments, et de ce fait, la candidate n'a pas eu suffisamment de temps pour exprimer son opinion et développer ses idées lors du débat. E6 déclare que cette attitude a déstabilisé la candidate, et E1, très étonné, met en évidence le fait que cela ait eu un impact considérable sur le résultat final de la candidate. Par ailleurs, E9 pense que la candidate aurait pu obtenir un score plus élevé que le score qu'elle a obtenu, c'est-à-dire B2-1 ou B2-2 au lieu de B1-1, si celle-ci avait été dans de meilleures conditions. Selon lui, la candidate s'est trouvée face à une circonstance non favorable et aurait dû interagir avec un animateur bienséant.

En revanche, d'après E3, E4 et E8, le travail de l'animatrice a été très bénéfique : elle a favorisé la discussion en maintenant l'interaction avec la candidate, elle s'est mise à sa place, elle l'a judicieusement aidée à reformuler ses idées, et lui a donné l'occasion de mettre en lumière ses capacités. E3 en conclut que l'attitude bienveillante de l'animatrice résulte de son expérience professionnelle.

Les techniques d'animation ont également été commentées. Par exemple, les trois examinateurs E2, E7 et E9 mentionnent que les interventions de l'animatrice se présentaient surtout sous forme de questions et que cette technique n'est pas adéquate. E7 et E9 signalent que les interventions pour introduire les contre-arguments doivent principalement se faire par des phrases déclaratives et non interrogatives, par exemple : « *moi, je me demande si c'est une bonne idée parce que...* » au lieu de « *il n'y a pas d'autres possibilités?* ».

4.2.3. L'évaluation du critère 3 : « Syntaxe »

L'évaluation de la syntaxe se réalise au travers de la section A et de la section B.

4.2.3.1. Les notes et les commentaires

Les scores pour le critère 3 de la candidate Jane vont de A2-2 à B2-1. Deux examinateurs ont accordé le score A2-2, trois le score B1-1, quatre le score B1-2 et un le score B2-1 (Tableau 31).

Tableau 31 - Répartition des scores du critère 3 de la candidate Jane

Scores	A2-2	B1-1	B1-2	B2-1
Nombre d'examinateurs	2	3	4	1

Tous les examinateurs s'entendent pour dire que dans l'ensemble, la candidate Jane faisait fréquemment des erreurs de syntaxe dans son discours. E3 évoque « *quelques erreurs* », sa note est B1-2, tandis que E6 parle d' « *énormément d'erreurs* » et sa note est B1-1. Les huit autres examinateurs ont commenté d'autres points que les erreurs. Par exemple, pour les cinq examinateurs, E5, E7, E9, E4 et E8, le discours était par-dessus tout marqué par des phrases simples et manquait de structures complexes, les notes respectives sont A2-1, B1-1, B1-1, B1-2, B1-2. E2 a relevé l'incapacité de la candidate à s'autocorriger malgré les corrections de l'animatrice, il a attribué A2-2. E1, quant à lui, a donné la note B1-2 et justifie son choix en mentionnant que la candidate n'a pas réussi à atteindre les exigences du niveau B2. Pour lui, les fautes qu'il a relevées sont des fautes que normalement un candidat de niveau B2 ne commettrait pas. Enfin, E10 qui a attribué la note la plus haute pour le critère 3, soit B2-1, justifie son choix en expliquant que ce qui caractérise le niveau B2 du niveau B1 sont les mots clés de la grille d'évaluation : « Les erreurs syntaxiques ne conduisent pas à des malentendus ». Comme il n'a pas constaté de malentendus, il situe alors la candidate dans le niveau B2.

Les notes similaires et les commentaires divergents

Il y a de petits détails au niveau de la syntaxe, manque de prépositions, d'articles, questions d'accords aussi peut-être.

E9, note: B1-1

Il y avait énormément d'erreurs que j'ai notées, dans la section A, elle a dit « je peux équiper de la cuisine de mon goût », « la prix ».

E6, note: B1-1

Le degré de sévérité des commentaires différents

C'est quelqu'un qui a pris pas mal d'heures, mais qui ne se soucie pas trop de la syntaxe, pour elle, le principal, c'est que le message passe, le reste ce n'est pas grave.

E5, note: A2-2

Il n'y a jamais rien qui conduit à des malentendus, c'est vraiment ce mot auquel je me fie souvent lorsque je fais des évaluations, c'est « est-ce qu'il y a des malentendus? » et puis s'il n'y en a pas, c'est vraiment plus un B2 que la catégorie d'avant.

E10, note: B2-1

4.2.4. L'évaluation du critère 4 : « Lexique »

L'évaluation du lexique se réalise au travers de la section A et de la section B.

4.2.4.1. Les notes et les commentaires

La distribution des scores pour le critère 4 de la candidate Jane s'échelonne ainsi : un examinateur a accordé le score A2-2, cinq le score B1-1, trois le score B1-2 et un le score B2-1 (Tableau 32).

Tableau 32 - Répartition des scores du critère 4 de la candidate Jane

Scores	A2-2	B1-1	B1-2	B2-1
Nombre d'examinateurs	1	5	3	1

Neuf examinateurs s'entendent pour dire que la candidate Jane commettait souvent des erreurs de lexique, qu'elle possédait un répertoire lexical assez limité, qu'elle ne prenait pas de risque, et qu'elle faisait des emprunts d'une autre langue (de l'anglais). Les notes sont A2-2, B1-1, B1-1, B1-1, B1-1, B1-1, B1-2, B1-2, B1-2. E10 se différencie des neuf autres examinateurs dans son commentaire, il a accordé la note la plus haute, soit B2-1, car il juge le lexique satisfaisant. Son choix se justifie par les termes des descripteurs de la grille d'évaluation du niveau B2 : « Les confusions ou approximations ne gênent pas la communication ». Ainsi, selon lui, le lexique de la candidate était satisfaisant dans la mesure où il n'entravait pas la communication.

Le degré de sévérité des commentaires différents

Un répertoire limité, disons à des mots qui étaient sur l'annonce, mais elle n'a pas utilisé d'autre vocabulaire pour enrichir la présentation, pour démontrer qu'elle connaissait assez bien le sujet.

E2, note: A2-2

C'est quand même assez bien (...) ce qui fait mon choix dans la case c'est « ne gênent pas la communication » », mais le lexique en tant que tel ça va.

E10, note: B2-1

4.2.4.2. L'attitude de l'animatrice

Les trois examinateurs E4, E7, E9 ont ajouté que parfois l'animatrice traduisait les mots anglais de la candidate en français, elle lui plaçait « les mots dans la bouche » et reformulait correctement les phrases à sa place.

4.2.5. L'évaluation du critère 5 : « Aisance à l'oral, élocution »

L'évaluation de l'aisance à l'oral et de l'élocution se réalisent au travers de la section A et de la section B.

4.2.5.1. Les notes et les commentaires

Les scores pour le critère 5 de la candidate Jane sont les suivants : un examinateur a accordé le score A2-1, deux le score B1-1, quatre le score B1-2 et trois le score B2-1 (Tableau 33).

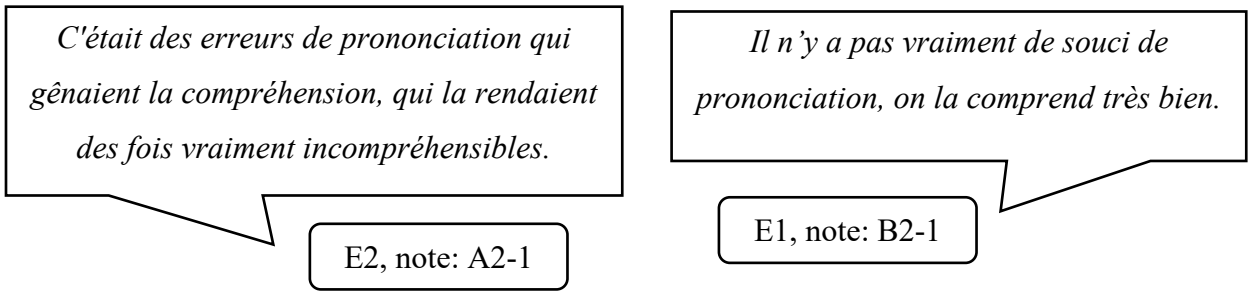
Tableau 33 - Répartition des scores du critère 5 de la candidate Jane

Scores	A2-1	A2-2	B1-1	B1-2	B2-1
Nombre d'examinateurs	1	0	2	4	3

Les commentaires des examinateurs ont surtout porté sur la prononciation et le débit, rarement sur l'accent, et aucunement sur l'intonation. Ils ont été regroupés en trois grandes catégories : les appréciations positives, les appréciations mitigées et les appréciations négatives.

Premièrement, pour E1, E5, E10 qui ont accordé la note B2-1, les erreurs de prononciation étaient plutôt mineures et sporadiques, elles n'entravaient pas le sens de l'ensemble du discours. Le débit était assez régulier et continu, ce qui ne posait pas de grand problème. Deuxièmement, E4, E7, E8 et E9 ont donné les notes B1-1 et B1-2. Pour eux, les erreurs de prononciation gênaient également la compréhension, mais de façon occasionnelle. Le débit des phrases était lent, c'est-à-dire marqué par beaucoup d'hésitations, d'interruptions et de pauses. Troisièmement, E2, E3 et E6 ont attribué les notes A2-1, B1-1 et B1-2. Pour eux, les erreurs de prononciation étaient assez bien marquées et plutôt fréquentes, elles rendaient le discours difficile à comprendre. Quant au débit, il était lent, les commentaires à ce propos étaient semblables à ceux du groupe d'examinateurs précédent. Enfin, concernant l'accent de la candidate, celui-ci a été commenté de façon contradictoire par seulement deux examinateurs: pour E6, il était très marqué, alors que pour E5, il n'était pas prononcé.

Le degré de sévérité des commentaires différents



4.2.6. Le bilan de l'évaluation de la candidate Jane

En conclusion, des divergences ont été observées dans la grande majorité des critères, dans la perception de l'attitude l'animatrice, ainsi que dans de courts extraits de dialogues. Il a également été intéressant de constater la manière dont ont été envisagés les écarts élevés de notes dans les deux premières sections, puis de découvrir une concordance entre les notes et les commentaires dans la section B.

Le niveau global estimé de la candidate Jane par l'équipe du Français des affaires est de B1. D'après le tableau ci-dessous (Tableau 34), un examinateur, E2, a attribué un niveau inférieur, soit A2, et un examinateur a attribué un niveau supérieur, soit B2. Les niveaux des huit autres examinateurs correspondent au niveau estimé de départ. La moyenne ainsi que la médiane sont B1.

Tableau 34 - Niveau global de la candidate Jane par les dix examinateurs et par l'équipe du Français des affaires (ÉFA)

Examinateurs	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**E9**	**E10**	**Moyenne**	**Médiane**	**ÉFA**
Niveau global	B1	A2	B1	B1	B1	B1	B1	B1	B1	B2	B1	B1	B1

Dans la section suivante, nous ferons l'analyse de l'évaluation de la troisième candidate, Mina.

4.3. L'évaluation de la candidate Mina

Dans cette section, nous présenterons l'évaluation des 5 critères concernant la candidate Mina.

4.3.1. L'évaluation du critère 1 : section A « Capacité à obtenir des informations »

La tâche demandée à la candidate Mina dans la section A est de se renseigner à propos des services d'un photographe pour une fête de famille.

4.3.1.1. Les notes et les commentaires

Concernant la répartition des notes pour le critère 1, nous observons qu'elles s'échelonnent de B1-2 à C1-2. Trois examinateurs ont accordé la note B1-2, trois la note B2-1, un la note B2-2, deux la note C1-1 et un la note C1-2 (Tableau 35).

Tableau 35 - Répartition des scores du critère 1 de la candidate Mina

Scores	B1-2	B2-1	B2-2	C1-1	C1-2
Nombre d'examinateurs	3	3	1	2	1

Dans l'ensemble, tous les examinateurs ont constaté que les questions de la candidate étaient pertinentes durant la première partie de la conversation, mais qu'elles étaient plus limitées durant la seconde partie.

Pour six d'entre eux (E4, E9, E2, E8, E6 et E10), le comportement de l'animatrice a eu un impact sur cette baisse de performance (nous reviendrons sur ce point dans la section suivante), les notes sont respectivement B1-2, B1-2, B2-1, B2-1, B2-2, C1-2. Pour trois d'entre eux (E7, E1 et E3), les questions de la candidate étaient limitées et étaient principalement en rapport avec les prix. La candidate manquait d'imagination et elle n'a pas su rebondir adéquatement aux relances proposées. Les notes sont B2-1, C1-1, C1-1. Pour E5, la candidate était vraisemblablement stressée et n'a juste pas pu improviser au moment voulu afin de trouver de nouvelles idées de questions. L'examinateur a donné le score B1-2, mais reconnait que la candidate mériterait plutôt B2-1, car il affirme avec certitude que dans une vraie situation, celle-ci aurait pu mieux faire, qu'elle aurait été tout à fait capable d'entretenir une conversation, de comprendre et de réagir convenablement. De plus, il avoue que ce qui influencerait ce choix de note, soit B2-1, est également le fait que la

candidate ait besoin d'obtenir des points pour sa demande d'immigration. Cet aspect représente un biais dont est conscient l'examinateur.

Les notes divergentes et les commentaires similaires

Au début je l'ai trouvée assez indépendante, mais après j'ai remarqué qu'elle avait quand même oublié de poser des questions sur la durée, la date, la réservation, donc il y avait des choses qui manquaient.

E9, note: B1-2

Elle mène bien son questionnaire, mais il y avait quand même certaines petites lacunes, elle aurait peut-être pu poser plus de questions, elle s'est limitée à l'essentiel.

E3, note: C1-1

Le degré de sévérité des commentaires différents

C'était un questionnement satisfaisant (...), mais ça m'a laissé quand même un petit sentiment qu'elle aurait pu poser plus de questions (...), je n'ai pas senti d'assurance.

E5, note: B1-2

Le questionnement est complet et précis (...) il y a certaines hésitations, des fois on entend des « euh », mais pour moi ce n'est pas une difficulté de langue, c'est parce que la candidate est en train de chercher ses idées.

E10, note: C1-2

4.3.1.2. L'attitude de l'animatrice

Les six examinateurs E2, E4, E6, E8, E9 et E10 ont commenté l'attitude de l'animatrice en précisant que dans la seconde partie de la conversation, les rôles entre les deux interlocutrices ont commencé à s'inverser. D'après eux, les réponses de l'animatrice étaient d'emblée trop détaillées, puis voyant que la candidate ne suivait plus le rythme, elle est intervenue en faisant plusieurs relances, et a commencé à poser de plus en plus de questions afin de combler le temps restant. Les examinateurs affirment que cette attitude a eu un impact négatif sur la performance de la candidate, et par conséquent sur sa note. Certains citent qu'à cause de cela, les questions de la candidate n'ont

pas pu être intéressantes et que sa créativité a été gâchée. D'autres déclarent qu'ils auraient pu accorder une note supérieure par rapport à la note qu'ils ont réellement accordée. Par exemple, E4 et E9 ont accordé B1-2 et déclarent qu'ils auraient pu attribuer la note B2 si l'animatrice s'était conformée à son rôle. E8, ayant donné B2-1, dit que s'il avait lui-même été à la place de l'animatrice, il aurait animé la conversation différemment, et donc la candidate aurait peut-être pu obtenir le score C1. Enfin, E10, qui a donné C1-2, mentionne que la candidate aurait pu obtenir C2 si l'animatrice n'avait pas posé les questions auxquelles l'on s'attendait et si celle-ci n'avait pas trop apporté d'informations.

4.3.1.3. Les références extérieures à la grille d'évaluation

Des éléments extérieurs à la grille d'évaluation ont été commentés sous trois formes différentes : des références à la culture de la candidate, des références à des éléments linguistiques dans les critères communicatifs, et des références à la durée de l'épreuve.

Les références à la culture de la candidate

E9 déclare que le fait que la candidate ait oublié de poser plusieurs questions sur le sujet de son document serait entre autres lié à sa culture qu'il identifie comme étant asiatique. Comme le sujet de la section A porte sur les services d'un photographe pour une fête familiale, selon l'examinateur (qui par ailleurs déclare enseigner à de nombreux apprenants provenant d'Asie), il existerait dans la tradition asiatique certaines pratiques de prise de photos bien particulières qui sont différentes de celles qui ont été suggérées par l'animatrice, comme le fait de prendre des photos à l'extérieur, d'écrire sur les photos, ou l'absence de thématiques spéciales.

L'enchevêtrement des critères linguistiques et des critères communicatifs

Cinq examinateurs (E3, E4, E6, E8 et E10) ont fait des commentaires concernant les éléments linguistiques lors de la section A, or ces éléments n'ont pas leur place à cet endroit, car ils s'appliquent plutôt aux critères 3, 4 et 5. Comme nous l'avons mentionné précédemment, le critère 1 a trait à la dimension communicative de la langue puisqu'il vise uniquement la capacité à obtenir des informations. Parmi les éléments linguistiques cités, les examinateurs ont mentionné que la candidate avait un très bon niveau de français, qu'elle se débrouillait très bien syntaxiquement, qu'elle maitrisait bien certaines expressions idiomatiques, et même qu'elle serait francophone étant

donné la qualité de sa langue. E8 avoue d'ailleurs que son excellent niveau de langue a un peu pris le dessus dans l'évaluation de sa performance de la section A.

Les références à la durée de l'épreuve

E2, E7 et E10 ont remarqué que le temps de l'épreuve de la section A avait été plus long que le temps alloué, soit 6 minutes 05 au lieu de 5 minutes. E10 précise que l'animatrice a étiré l'entrevue et n'a pas clôt la discussion avant la fin. Il ajoute que dans sa pratique d'évaluation-animation du TEF, il s'assure de toujours respecter le temps imparti à l'aide d'un chronomètre et de clore naturellement la discussion de façon précise vers 4 minutes 40.

Sous une autre perspective, E10 a émis un commentaire à propos de la durée imposée de la section A. L'examinateur explique que la candidate n'a eu besoin que de 3 minutes pour poser l'essentiel des questions dont elle avait besoin, étant donné que cette dernière est francophone selon lui. Il est d'avis que la durée totale de la section A, soit 5 minutes, est trop longue pour une personne de niveau avancé (C1-2, C2), car les informations demandées sont relativement simples. Il pense que lorsque les examinateurs sont face à des utilisateurs expérimentés du français, ils peuvent très rapidement se rendre compte de leur compétence à communiquer. Dès lors, la tâche demandée de la section A, qui est de poser une dizaine de questions en 5 minutes, peut très bien se réaliser en 2 ou 3 minutes avec ce profil de candidats. Par ailleurs, l'examinateur ajoute que même dans un contexte de vie réelle, un locuteur francophone aurait rarement une conversation téléphonique avec un interlocuteur pendant une durée d'environ 5 minutes dans le but d'obtenir des renseignements à propos d'un service ou d'un produit. Selon lui, la tâche demandée n'est pas transférable dans la réalité d'un locuteur natif.

4.3.2. L'évaluation du critère 2 : section B « Capacité à présenter et débattre »

Dans la section B, la tâche demandée à la candidate Mina est de proposer à son amie d'héberger un étudiant étranger à la maison.

4.3.2.1. Les notes et les commentaires

Concernant les notes pour le critère 2, nous remarquons qu'elles vont de B1-2 à C2. Trois examinateurs ont accordé la note B1-2, deux la note B2-1, un la note B2-2, deux la note C1-2, et deux la note C2 (Tableau 36).

Tableau 36 - Répartition des scores du critère 2 de la candidate Mina

Scores	B1-2	B2-1	B2-2	C1-1	C1-2	C2
Nombre d'examinateurs	3	2	1	0	2	2

Sept examinateurs ont brièvement fait référence à la capacité à présenter le document. Parmi eux, E2, E4, E5 et E6 ont globalement perçu une présentation claire avec des reformulations d'informations, mais simple et peu détaillée. En revanche, pour E1, E3 et E8, la présentation était bien développée, structurée et même excellente.

En ce qui concerne la capacité à débattre, les commentaires se divisent en trois catégories : ceux qui ont trouvé les arguments peu développés, ceux qui ont trouvé les arguments bons au début puis faibles vers la fin, puis ceux qui trouvé les arguments très bons. Les deux examinateurs E2 et E6 ont trouvé que les arguments étaient très peu développés, ils ont attribué la note B1-2. E2, par exemple, déclare que les propos de la candidate n'étaient pas assez forts pour mener un vrai débat, et que ses arguments étaient surtout des descriptions sommaires. Pour E6, l'argumentation était quasiment absente, la candidate était à court d'idées et ses propos étaient trop ordinaires.

Les six examinateurs E4, E9, E5, E7, E3 et E8 et ont constaté que la conversation avait très bien commencé, la candidate possédait de bons moyens de débattre, les arguments apportés étaient bien développés et intéressants. Mais progressivement, les arguments se sont affaiblis à force de beaucoup de « *je ne sais pas* », selon les termes de E4, et par conséquent, la candidate n'a pas pu répliquer de manière adéquate. Les notes sont respectivement B1-2, B2-1, B2-1, B2-2, C1-2, C1-2.

E1 et E10, quant à eux, ont jugé les arguments de la candidate excellents, nuancés et complexes, et ont conclu que celle-ci avait largement réussi à convaincre son interlocutrice. E1 ajoute qu'elle avait beaucoup de répondant, et E10 ajoute que la discussion était naturelle, telle une discussion entre deux personnes qui se connaissent réellement. Les notes sont C2, c'est-à-dire la note la plus élevée.

Les notes divergentes et les commentaires similaires

Les échanges ont été suivis, elle a argumenté jusqu'à un moment donné, elle a bien argumenté sur toutes les choses et après c'est un peu moins bien, mais c'est suivi.

E5, note: B2-1

Elle mène bien son discours, son argumentaire (...) au niveau de l'argumentation, la forme est bonne, à un moment donné il y a une faiblesse.

E3, note: C1-2

Le degré de sévérité des commentaires différents

Elle a expliqué vraiment dans les grandes lignes ce que c'était d'avoir un étudiant dans une famille d'accueil (...), elle comprenait très bien toutes les interventions, mais après je me suis bien rendu compte qu'elle n'avait pas de suite dans les idées (...) il n'y a aucune argumentation pratiquement.

E6, note: B1-2

C'était complet, c'était très pertinent, la présentation et les arguments, c'était super, elle a très bien réussi à convaincre, elle réagit de manière adéquate et elle a beaucoup de répondant.

E1, note: C2

4.3.2.2. Les propos en lien avec les descripteurs de la grille d'évaluation

E2 a rappelé l'incohérence du critère 2 de la grille d'évaluation dont l'objet est « Capacité à présenter et débattre », ce même examinateur a d'ailleurs exprimé les mêmes propos précédemment lors de l'évaluation de la première candidate Nora. D'après lui, présenter et débattre sont deux actions indépendantes l'une de l'autre, et ce problème le gêne beaucoup dans sa pratique d'évaluation. Il soulève que très souvent, il est confronté à des candidats qui présentent très bien un document en se l'appropriant, mais qui ont des lacunes lorsque vient le temps d'échanger des points de vue dans un débat, et inversement. Il ajoute que cela produit des performances différentes, et il suggère que ces deux éléments sont distincts dans l'évaluation.

4.3.2.3. L'attitude de l'animatrice

Six examinateurs (E2, E4, E5, E8, E9 et E10) ont fait référence à l'attitude de l'animatrice dans leur commentaire, et plus précisément au fait qu'elle dominait la deuxième partie de la conversation avec de nombreux contre-arguments. Pour l'un d'eux, E10, l'attitude de l'animatrice ne posait pas de problème, car il considère que contre-argumenter fait partie du travail normal de l'animation. De plus, l'examinateur a trouvé l'animatrice très expérimentée, car elle maitrisait aisément ses fonctions.

En revanche, pour les cinq autres examinateurs (E2, E4, E5, E8 et E9), l'attitude de l'animatrice a été perçue de façon très négative, car ils pensent que cette dernière a intimidé la candidate, et par conséquent, cela a directement conduit à l'amenuisement de sa capacité à débattre. E5 reconnait que l'intention de l'animatrice était bonne, qu'elle voulait apporter son soutien à la candidate en essayant de lui « *tendre des perches* », selon ses propres termes, mais que finalement le résultat souhaité a été inversé. E9 ajoute par ailleurs que l'animatrice est sans doute novice et que son manque d'expérience professionnelle explique un tel comportement. D'autre part, on constate que le vocabulaire employé par les examinateurs pour évoquer cette situation possède une connotation assez « dramatique » : « *l'animatrice a été un peu agressive* » (E2), « *j'ai un peu souffert avec la candidate* » (E2), « *à cause de la force émotive de l'animatrice* » (E4), « *elle l'a assommée* (E5) », « *elle était bloquée* » (E9), « *la pauvre* » (E8).

Après réflexion, deux de ces cinq examinateurs, E4 et E5, tempèrent leurs propos et prennent un certain recul par rapport à l'attitude de l'animatrice. Par exemple E5 se remet lui-même en question, car il reconnait que dans sa pratique, il lui arrive parfois dans la section B de prendre l'ascendant dans la conversation lorsque le candidat est à court d'idées. L'examinateur explique qu'il agit ainsi de peur de laisser trop de vide dans les échanges. Il admet également que donner plus de matière aux candidats en imaginant les aider n'est pas toujours la solution assurée pour alimenter le débat, car il n'y a aucune garantie que cela fonctionne, et cela peut même créer l'effet inverse.

E4, quant à lui, comprend que les examinateurs peuvent être tentés de beaucoup parler lors de leur travail d'animation, et que cela peut arriver à tous lors d'une période d'égarement. Il souligne alors l'importance d'être en tout temps vigilant et de se remémorer les bonnes techniques d'animation

vues lors des formations, comme le fait de constamment s'assurer de ne pas dominer la conversation et de faire en sorte que l'on entende davantage les candidats s'exprimer.

4.3.2.4. Les références extérieures à la grille d'évaluation

Parmi les références extérieures à la grille d'évaluation qui ont été relevées, nous retrouvons des allusions à la culture de la candidate.

Les références à la culture de la candidate

E2 est d'avis qu'un effet culturel a joué un rôle important dans la performance de la candidate où il est question de débattre. Dans le jeu de rôle de la section B, la candidate dit être d'origine japonaise, dès lors, l'examinateur associe cette information avec le fait qu'elle vienne d'une culture où les rapports humains sont plutôt consensuels et où la confrontation n'est pas courante. D'après lui, un animateur qui se montre trop « agressif » face à un candidat ne garantit pas toujours une stimulation du débat, car cela dépend des caractéristiques de la société d'origine de ce dernier. Il cite l'exemple d'un candidat d'origine latino-américaine qui se montrerait plus réactif qu'un candidat d'origine japonaise face à un animateur « *combatif* », selon ses propres termes. Ainsi, l'examinateur estime que le niveau de langue de la candidate Mina n'est pas la cause de sa difficulté à convaincre, car il atteste que son niveau de français est bon. La raison est possiblement culturelle, c'est pourquoi il est difficile pour lui d'évaluer cette section B.

4.3.3. L'évaluation du critère 3 : « Syntaxe »

L'évaluation de la syntaxe se réalise au travers de la section A et de la section B.

4.3.3.1. Les notes et les commentaires

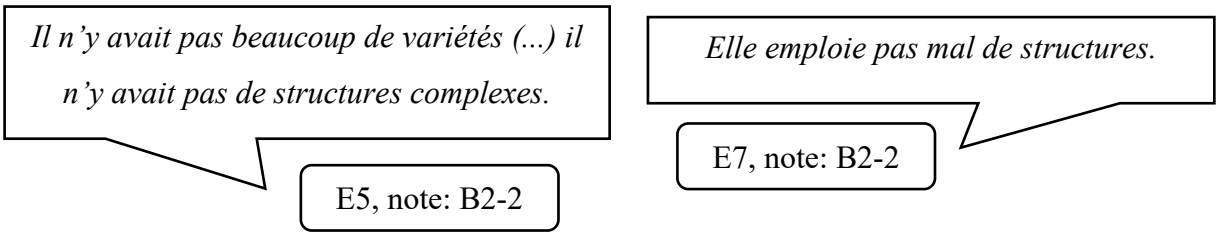
Les notes pour le critère 3 de la candidate Mina vont de B2-1 à C2. Deux examinateurs ont accordé la note B2-1, trois la note B2-2, un la note C1-1, deux la note C1-2 et deux la note C2 (Tableau 37).

Tableau 37 - Répartition des scores du critère 3 de la candidate Mina

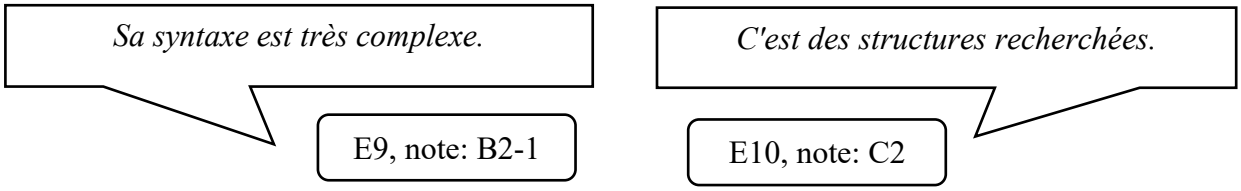
Scores	B2-1	B2-2	C1-1	C1-2	C2
Nombre d'examinateurs	2	3	1	2	2

Les avis concernant le critère 3 sont disparates. Pour deux examinateurs, E8 et E10, la candidate n'a commis aucune erreur syntaxique, ses structures de phrases étaient très complexes et variées, les notes sont C2. Pour trois autres, E9, E7 et E1, la syntaxe était élaborée et diversifiée, mais il y avait tout de même des erreurs mineures, les notes sont B2-1, B2-2, C1-2. Quant aux cinq autres, E2, E5, E4, E6 et E3, ils ont observé que les erreurs étaient peu nombreuses et qu'elles n'affectaient pas l'ensemble du discours, cependant, les structures de phrases n'étaient pas très recherchées ni d'une grande variété, les notes sont B2-1, B2-1, B2-2, B2-2, C1-2.

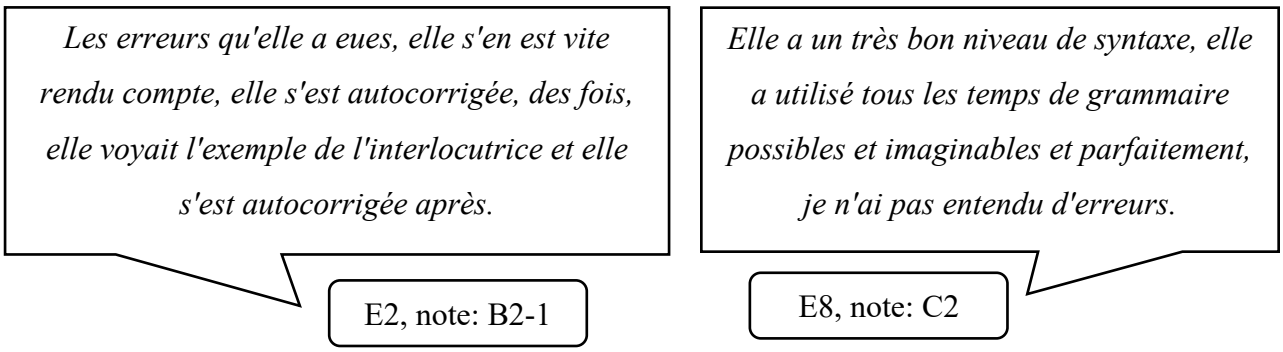
Les notes similaires et les commentaires divergents



Les notes divergentes et les commentaires similaires



Le degré de sévérité des commentaires différents



4.3.4. L'évaluation du critère 4 : « Lexique »

L'évaluation du lexique se réalise au travers de la section A et de la section B.

4.3.4.1. Les notes et les commentaires

Les scores pour le critère 4 de la candidate Mina varient de B2-1 à C2. Trois examinateurs ont donné le score B2-1, trois le score B2-2, un le score C1-2 et trois le score C2 (Tableau 38).

Tableau 38 - Répartition des scores du critère 4 de la candidate Mina

Scores	B2-1	B2-2	C1-1	C1-2	C2
Nombre d'examinateurs	3	3	0	1	3

Tous les examinateurs estiment que dans l'ensemble, le lexique était convenable ou excellent. D'après cinq d'entre eux (E2, E5, E9, E6 et E7), le lexique de la candidate était bon, mais pas assez large, les notes sont respectivement B2-1, B2-1, B2-1, B2-2, B2-2. Certains précisent par exemple qu'il n'y avait pas de grande variété d'expressions, que les mots étaient en lien avec certains domaines particuliers (vocabulaire du quotidien et des échanges relationnels et interculturels), qu'elle utilisait très souvent les mêmes mots ou que les situations rencontrées ne lui avaient pas permis de montrer davantage ce dont elle était capable. Pour deux examinateurs (E3 et E4), le lexique était approprié et assez vaste, les notes sont B2-2, C1-1. E3 cite dans son commentaire quelques exemples d'erreurs de syntaxe en pensant qu'il s'agit d'erreurs de lexique, il relève par exemple des erreurs de genre d'article et de pronom complément, comme « *le famille, il peut le changer* ».

Pour les trois autres examinateurs (E1, E8 et E10), le lexique était très riche, nuancé et parfaitement maitrisé, les notes sont C2. D'ailleurs, E8 et E10 signalent que le niveau de lexique de la candidate était du même niveau que celui d'un locuteur natif.

Les notes similaires et les commentaires divergents

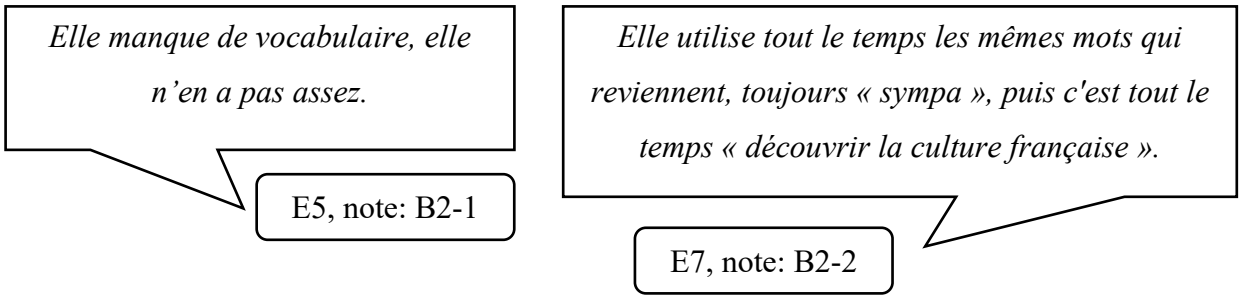
Le lexique c'est assez large vraiment (...), on trouve du vocabulaire assez riche, que ce soit dans la première ou dans la deuxième section.

E4, note: B2-2

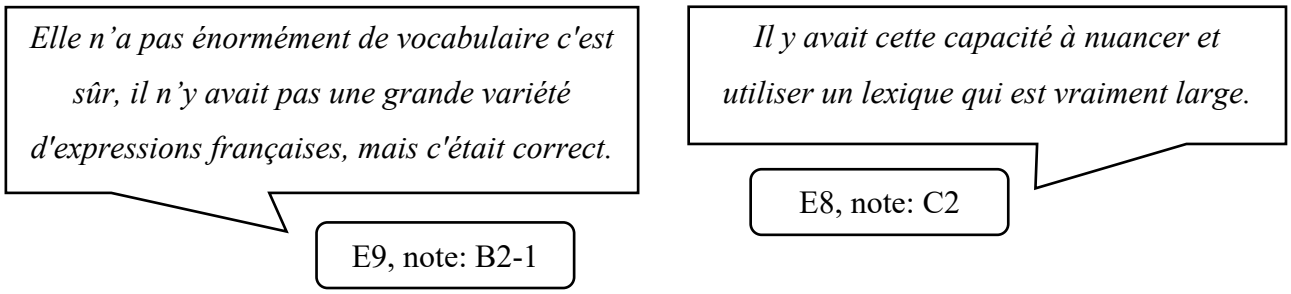
Le lexique est bien employé, adapté à la situation, mais ce n'était pas si varié que ça.

E6, note: B2-2

Les notes divergentes et les commentaires similaires



Le degré de sévérité des commentaires différents



4.3.5. L'évaluation du critère 5 : « Aisance à l'oral, élocution »

L'évaluation de l'aisance à l'oral et de l'élocution se réalisent au travers de la section A et de la section B.

4.3.5.1. Les notes et les commentaires

Les scores pour le critère 5 de la candidate Mina vont de B2-2 à C2. Trois examinateurs ont donné le score B2-2, un le score C1-1, deux le score C1-2 et quatre le score C2 (Tableau 39).

Tableau 39 - Répartition des scores du critère 5 de la candidate Mina

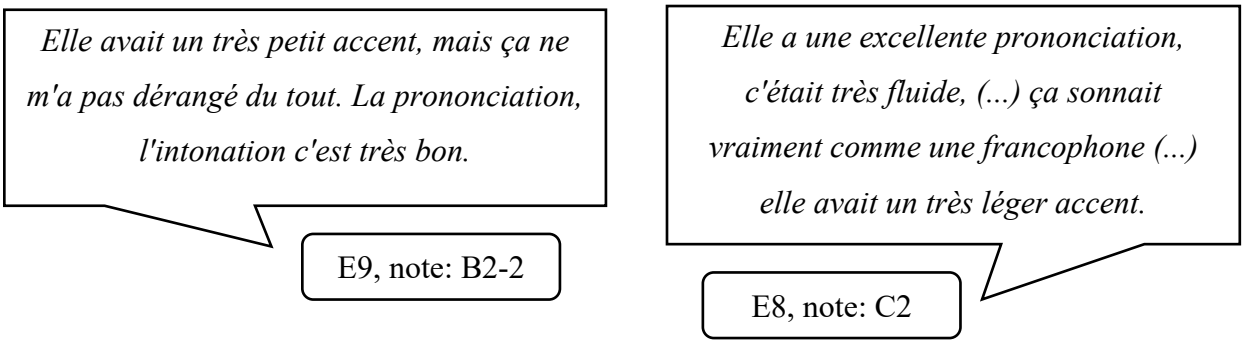
Scores	B2-2	C1-1	C1-2	C2
Nombre d'examinateurs	3	1	2	4

La totalité des examinateurs considère que l'aisance à l'oral et l'élocution de la candidate étaient bonnes ou très bonnes, et aucun commentaire concernant une quelconque difficulté liée aux éléments prosodiques n'a été émis. Dans tous les commentaires, on évoque une prononciation juste, un accent léger qui ne gêne nullement la compréhension, ainsi qu'une intonation et une

fluidité continuellement adaptées au contexte. Par conséquent, une certaine similitude apparait dans les propos des examinateurs malgré la distribution des scores. Lorsque l'on observe de près les propos des quatre examinateurs ayant accordé la note C2, c'est-à-dire E1, E5, E8 et E10, ils se distinguent des autres examinateurs par leur utilisation de termes mélioratifs comme « *excellent* », « *parfait* », de plus, ils assimilent l'aisance à l'oral et l'élocution de la candidate à celles d'un locuteur natif. E5 ajoute par exemple que la candidate a sans doute longtemps été en totale immersion dans un milieu francophone étant donné le caractère naturel et sans effort de son discours. E8 affirme par ailleurs qu'il fait abstraction du léger accent de la candidate dans son évaluation, car d'après lui, comme de nombreuses personnes de diverses origines cohabitent au Québec, le fait d'avoir un tel accent reflète la réalité sociale multiculturelle de la province canadienne.

Pour E6 qui a attribué C1-2, son statut d'enseignant de français langue seconde représente un biais pour évaluer l'aisance à l'oral et l'élocution de la candidate. Comme il déclare très bien connaitre les difficultés prosodiques que rencontrent généralement les apprenants asiatiques (principalement chinois, japonais et coréens) en français, il réalise que la candidate, qui est japonaise selon lui, s'exprime extrêmement bien en comparaison avec les apprenants à qui il a l'habitude d'enseigner le français. Il dit par exemple que la candidate Mina accentue beaucoup le son [r] qui est un son problématique pour les apprenants asiatiques. Il dit également que le ton avec lequel celle-ci exprime certaines locutions courantes, comme « *Oh ça va, ça va!* », est authentique et qu'il est rare d'observer cela chez les apprenants asiatiques. De ce fait, l'examinateur est conscient que cette caractéristique l'influence de façon positive et avoue que cela l'incite à donner une note plus élevée pour le critère de l'aisance à l'oral et l'élocution.

Les notes divergentes et les commentaires similaires



4.3.6. Les compléments de l'évaluation de la candidate Mina

Après avoir terminé leur évaluation de la candidate Mina, les quatre examinateurs E3, E4, E7 et E9 ont dressé un bilan succinct à propos de la « réussite » de la candidate au test de français. Par réussite, ils évoquent le fait d'obtenir au moins le niveau B2, c'est-à-dire le seuil minimal à compter duquel des points sont attribués pour une demande d'immigration au Québec. À titre de rappel, le TEF n'a pas de caractère de réussite ou d'échec, car il situe juste les candidats dans un niveau de langue. E3 affirme par exemple que la candidate a « *réussi l'entretien haut la main* » selon ses propres termes, car toutes les notes qu'il a données pour l'ensemble des cinq critères se situent dans le niveau C1. En revanche, E4 n'est pas certain de sa note globale finale, car ses notes pour les cinq critères oscillent entre B1 et B2. Il ne sait donc pas si la candidate pourrait « *réussir* » le test et se questionne sur le total des points. Il ajoute que si la candidate ne réussissait pas à obtenir un niveau B2, l'animatrice en serait quelque peu responsable étant donné que celle-ci ne l'a pas mise dans de bonnes conditions d'évaluation. E9 évoque cette même question au sujet de l'animatrice, il est conscient que cette dernière a affecté la performance de la candidate et il souhaiterait sincèrement que la candidate puisse obtenir le niveau B2, car il considère qu'elle en a les capacités. Ses notes pour les cinq critères se situent entre B1 et B2 et il mentionne que l'obtention d'un niveau B2 risque d'être trop juste. Enfin, E7 affirme ne pas inquiet à propos de la « *réussite* » de la candidate puisqu'il ne lui a accordé aucune note inférieure à B2 pour les cinq critères.

4.3.7. Le bilan de l'évaluation de la candidate Mina

En conclusion, de nombreuses divergences ont été reportées dans tous les critères, mais pour celui de l'aisance à l'oral et l'élocution, une concordance dans les propos est apparue malgré la distribution des scores. Concernant la capacité à débattre de la section B, il a été relevé que le profil culturel de la candidate avait sans doute nui à sa performance. Par ailleurs, beaucoup d'examinateurs ont fait part de l'attitude de l'animatrice, et plus particulièrement de son impact négatif sur la performance de la candidate, qui a d'ailleurs suscité diverses réflexions. Enfin, avec certains, l'évaluation s'est terminée sur un questionnement des points de la grille d'évaluation afin d'atteindre le niveau B2.

Le niveau global estimé de la candidate Mina par l'équipe du Français des affaires est de B2. D'après le tableau ci-dessous (Tableau 40), dix examinateurs ont attribué le niveau B2, et quatre

autres ont attribué un niveau supérieur, à savoir deux examinateurs le niveau C1, E3 et E8, et deux examinateurs le niveau C2, E1 et E10. La moyenne est de C1 et la médiane B2.

Tableau 40 - Niveau global de la candidate Mina par les dix examinateurs et par l'équipe du Français des affaires (ÉFA)

Examinateurs	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**E9**	**E10**	**Moyenne**	**Médiane**	**ÉFA**
Niveau global	C2	B2	C1	B2	B2	B2	B2	C1	B2	C2	C1	B2	B2

Dans la partie qui suit, nous ferons l'analyse de l'évaluation du quatrième candidat, Rayan.

4.4. L'évaluation du candidat Rayan

Dans cette section, nous présenterons l'évaluation des 5 critères concernant le candidat Rayan.

4.4.1. L'évaluation du critère 1 : section A « Capacité à obtenir des informations »

La consigne demandée au candidat Rayan dans la section A est de poser des questions sur une école de langues.

4.4.1.1. Les notes et les commentaires

Concernant la répartition des notes pour le critère 1 du candidat, nous observons qu'elles s'échelonnent de B2-1 à C2. Deux examinateurs ont accordé la note B2-1, deux la note B2-2, trois la note C1-1, deux la note C1-2 et un la note C2 (Tableau 41).

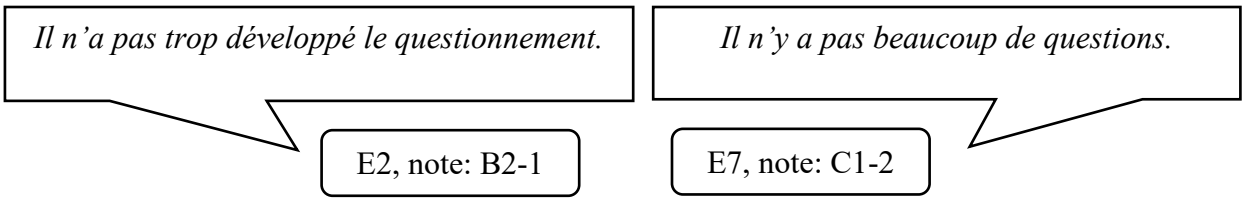
Tableau 41 - Répartition des scores du critère 1 du candidat Rayan

Scores	B2-1	B2-2	C1-1	C1-2	C2
Nombre d'examinateurs	2	2	3	2	1

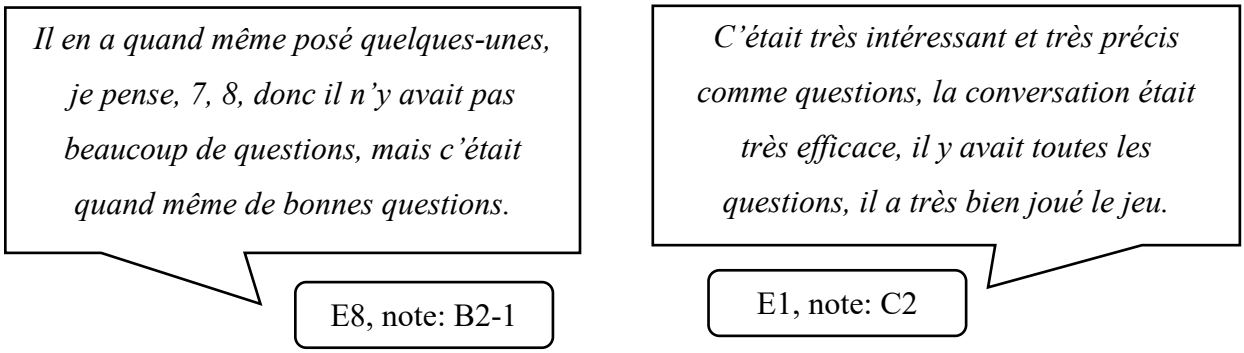
Tous les dix examinateurs sont d'avis que le candidat peut facilement suivre l'échange avec son interlocuteur et intervenir de façon appropriée et naturelle face à l'imprévu. Concernant le questionnement, un seul examinateur, E1, a remarqué qu'il était exhaustif et très précis, sa note est C2. En revanche, les neuf autres ont constaté qu'il n'y avait pas suffisamment de questions, à savoir cinq, six ou sept d'après certains commentaires, au lieu d'une dizaine comme l'exige la consigne. Les neuf autres examinateurs ayant émis la même réflexion concernant le nombre de questions ont attribué des notes variées : B2-1, B2-1, B2-2, B2-2, C1-1, C1-1, C1-1, C1-2, C1-2. À titre d'information, les descripteurs du niveau B2 spécifient « Questionnement approprié » et ceux du niveau C1 « Questionnement complet et précis ». Cinq examinateurs, E4, E6, E7, E9 et E10, ont accordé C1-1 ou C1-2 bien qu'ils aient observé que le questionnement du candidat n'était ni complet ni précis. Les justifications du choix de leur note sont variées. Par exemple, E9 dévoile que lors de la formation des examinateurs dans son centre de passation du TEF, il a été demandé à ces derniers d'accorder automatiquement un score minimum de C1 à un candidat francophone pour tous les critères de la grille d'évaluation. Pourtant, E9, qui reconnait que le candidat Rayan est francophone, n'approuve pas la décision imposée par son centre et mentionne que son véritable choix de note pour la section A serait plutôt B2. Pour E6, le fait que le candidat intervienne avec

justesse dans la conversation avec l'animateur et que celle-ci soit naturelle, telle une conversation entre deux francophones, est une caractéristique du niveau C1. E4 et E7 ont exprimé des propos similaires, leur choix de note s'explique par le bon niveau de français du candidat associé à sa capacité à répliquer aux interrogations de l'animateur. Enfin, E10 mentionne que le candidat a les capacités linguistiques d'un niveau C1, car il peut converser de façon efficace, mais que ce dernier n'a pas pu prendre le dessus et s'affirmer davantage à cause de sa personnalité peu affirmée. Ces cinq examinateurs ayant accordé C1-1 ou C1-2 n'ont alors pas tenu compte de la mention du descripteur de la grille d'évaluation « Questionnement complet et précis » pour leur note de la section A.

Les notes divergentes et les commentaires similaires



Le degré de sévérité des commentaires différents



4.4.1.2. L'attitude de l'animateur

La totalité des examinateurs ont commenté l'attitude de l'animateur, et notamment le fait qu'il s'exprimait trop. Ils ont observé par exemple que celui-ci avait une grande emprise sur le début de la conversation, c'est-à-dire qu'il posait beaucoup trop de questions (cinq d'après certains commentaires), qu'il faisait de nombreuses remarques assez longues sur les propos du candidat, et qu'il parlait trop rapidement. De ce fait, le temps consacré au candidat a été limité, et ce n'est que vers la fin de la conversation que ce dernier a pu se sentir plus libre pour pouvoir développer son

questionnement. Beaucoup d'examinateurs ont déclaré que cet incident les avait gênés dans leur prise de décision.

Concernant la conduite de l'animateur, E5 et E7 réalisent que poser des questions lors de la tâche d'animation correspond tout à fait à un contexte de vraie vie, car il est très naturel en tant qu'interlocuteur au téléphone, qui répond à des questions à propos d'un service ou d'un produit, de demander des informations à son tour. E5 précise que dans le jeu de rôle de la section A, il est demandé aux candidats de « *faire comme si c'était dans la vraie vie* », selon ses propres termes, mais que cela n'est pas cohérent avec les consignes de la formation des examinateurs qui exigent que l'examinateur-animateur évite le plus possible de poser lui-même des questions et qu'il fournisse intentionnellement des renseignements incomplets dans le but de faire rebondir son interlocuteur. E5 avoue que cette posture dans laquelle il doit contrôler ses propos le bloque quelquefois. Il avoue également que lorsque la fatigue s'installe chez lui, généralement au bout du quatrième ou du cinquième candidat qu'il évalue, il lui arrive de poser quelques questions lors de la section A, puis lors de la section B, d'accommoder davantage un candidat stressé ou ayant un tempérament naturellement peu enjoué qui mériterait d'obtenir le niveau B2. L'effet de fatigue l'amène alors à s'écarter du cadre de l'animation défini issu de la formation des examinateurs.

4.4.1.3. Les références extérieures à la grille d'évaluation

Des références extérieures à la grille d'évaluation ayant trait à la durée de l'épreuve ont été relevées.

Les références à la durée de l'épreuve

E3 et E4 ont signalé que la durée de la conversation avait dépassé le temps alloué, soit précisément 5 minutes 50 au lieu de 5 minutes. E3 rappelle qu'il est important de conclure la conversation trente secondes avant la fin afin de respecter le temps. E4, quant à lui, ajoute que tous les candidats doivent disposer de la même durée pour l'épreuve par souci d'équité.

4.4.2. L'évaluation du critère 2 : section B « Capacité à présenter et débattre »

La consigne demandée au candidat Rayan dans la section B est de présenter à son ami une publicité au sujet de meubles en carton puis de le convaincre d'en acheter en ligne.

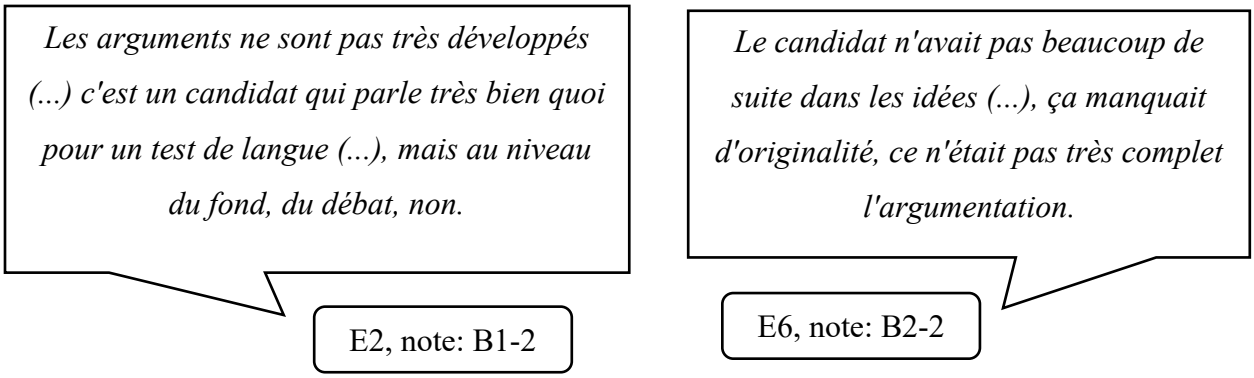
4.4.2.1. Les notes et les commentaires

Les notes du candidat pour le critère 2 vont de B2-1 à C1-2. Un examinateur a accordé la note B1-2, deux la note B2-2, un la note C1-1 et six la note C1-2 (Tableau 42).

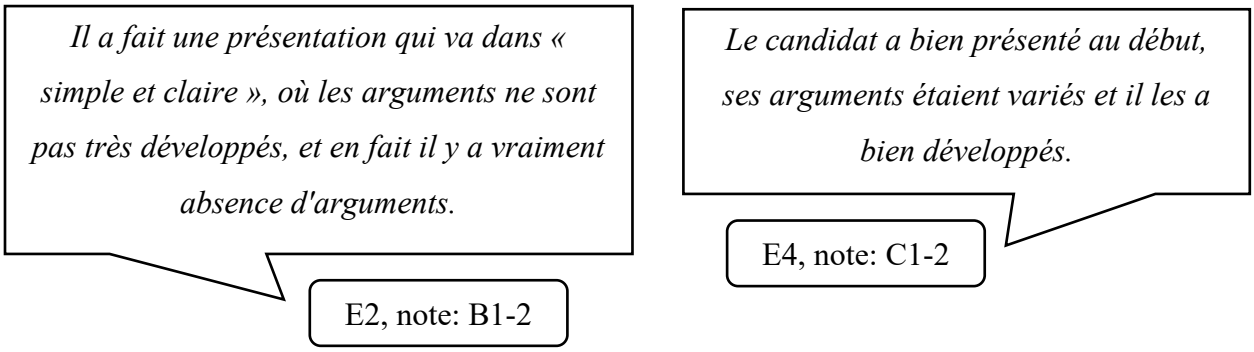
Tableau 42 - Répartition des scores du critère 2 du candidat Rayan

Scores	B1-2	B2-1	B2-2	C1-1	C1-2
Nombre d'examinateurs	1	0	2	1	6

Les notes divergentes et les commentaires similaires



Le degré de sévérité des commentaires différents



Seulement quatre examinateurs ont fait référence à la capacité du candidat à présenter le document, mais de façon très succincte. Deux d'entre eux, E5 et E7, ont précisé que sa capacité à présenter était d'un niveau plus faible que sa capacité à débattre, et que leur note de la section B était uniquement représentative de la capacité à débattre. Étant donné que l'action de présenter et que

l'action de débattre ne peuvent pas être indissociables, les examinateurs ont fait le choix de faire fi de l'action de présenter.

En ce qui concerne la capacité à débattre, quatre examinateurs (E2, E3, E6 et E9) ont déclaré que les arguments du candidat n'étaient pas assez bons ni assez développés, notamment à cause d'une panne d'idées. Ils ont mentionné de plus que les arguments étaient parfois même contradictoires, et de ce fait, le candidat était passablement convaincant dans l'ensemble. Les notes accordées sont respectivement B1-2, B2-2, B2-2, C1-2. En ce qui concerne l'examinateur ayant accordé C1-2, E9, nous rappelons que celui-ci a mentionné précédemment (lors de l'évaluation de la section A du même candidat) qu'il était d'usage, dans son centre de passation du test, d'attribuer un score minimum de C1 à un candidat francophone.

Quant aux six autres examinateurs (E5, E1, E4, E7, E8 et E10), ils sont tous d'avis que les arguments apportés par le candidat étaient variés, intéressants, profonds et bien développés, et que celui-ci avait bien réussi à mener le débat. Les notes sont C1-1 pour E5 et C1-2 pour les autres.

4.4.2.2. L'attitude de l'animateur

Comme lors de la section A, l'attitude peu collaborative de l'animateur a été commentée dans la section B, non pas par la totalité des examinateurs, mais par trois d'entre eux (E1, E2 et E8). Par exemple, E1 explique que plusieurs fois, lorsque le candidat essayait de développer un argument, l'animateur intervenait au mauvais moment, et par conséquent, cela coupait son élan. De manière générale, E1 pense que certains animateurs ne sont pas pleinement attentifs à ce que disent les candidats, car ils doivent continuellement anticiper en enchainant avec les contre-arguments qu'on leur propose de sorte que le flot de la conversation soit maintenu.

D'après E2, le comportement peu accommodant de l'animateur a surtout affecté la capacité du candidat à présenter le document. L'examinateur a remarqué que dès le tout début de la conversation, l'animateur est intervenu en réagissant à la première phrase que le candidat a prononcée, dès lors, ce dernier n'a pas suffisamment eu de temps de faire une présentation complète et intéressante du document. La conversation n'a alors pas pu se dérouler dans l'ordre demandé, car l'animateur l'a immédiatement dirigé vers la voie du débat avec une multitude de contestations. Pour E2, cet accroc a eu un impact négatif sur le résultat de la partie consacrée à la capacité à présenter du critère 2.

Dans la même lignée, E8 a trouvé que l'animateur était extrêmement familier, que ses propos étaient parfois déplacés, qu'il parlait trop vite, et qu'il passait brusquement d'un sujet à un autre dans la discussion. Cependant, malgré de tels agissements de la part de l'animateur, l'examinateur a constaté que le candidat était tout de même capable d'apporter des réponses et qu'il s'était bien débrouillé dans l'ensemble.

4.4.3. L'évaluation du critère 3 : « Syntaxe »

L'évaluation de la syntaxe se réalise au travers de la section A et de la section B.

4.4.3.1. Les notes et les commentaires

Les notes du candidat Rayan pour le critère 3 vont de B2-2 à C2. Un examinateur a accordé la note B2-2, quatre la note C1-1, deux la note C1-2 et trois la note C2 (Tableau 43).

Tableau 43 - Répartition des scores du critère 3 du candidat Rayan

Scores	B2-2	C1-1	C1-2	C2
Nombre d'examinateurs	1	4	2	3

Les deux examinateurs E8 et E10 ont constaté que la syntaxe du candidat était entièrement correcte et qu'elle ne contenait aucune erreur. Ils ont attribué la note C2. Les huit autres (E3, E2, E5, E6, E9, E4, E7 et E1) ont constaté qu'il y avait des erreurs de syntaxe mineures et relativement peu fréquentes, les notes sont respectivement B2-2, C1-1, C1-1, C1-1, C1-1, C1-2, C1-2, C2. Parmi eux, E6 a accordé la note C1-1, mais il réalise après réflexion qu'il a peut-être été trop indulgent et qu'il aurait pu donner la note B2-2, car selon lui, de manière générale, les candidats de niveau B2-2 utilisent une meilleure syntaxe avec des tournures de phrases plus complexes en comparaison avec le candidat Rayan. E6 a alors douté de sa note, mais ne l'a pas modifiée, car il a finalement déclaré que cela n'était pas si grave. Un autre examinateur, E1, a accordé C2, c'est-à-dire la note maximale, mais est conscient que sa note est trop élevée et qu'il aurait pu donner C1-2. Toutefois, il a décidé de ne pas revenir sur sa note et se justifie par le fait qu'il considère que les erreurs mineures et occasionnelles du candidat peuvent être admises dans l'absolu, car elles sont le reflet de la syntaxe orale du français québécois qui est quelquefois approximative et erronée. Concernant E9, l'examinateur qui a précédemment déclaré qu'il devait octroyer au minimum C1 à un candidat

francophone selon les instructions de son centre de passation TEF, celui-ci a donné la note C1-1 et avoue que son véritable choix de note serait plutôt B2-2. Il justifie cela par le manque de structures de phrases complexes et la non-maitrise totale de certaines notions de grammaire comme le subjonctif et les pronoms compléments d'objet direct et indirect.

Les notes similaires et les commentaires divergents

Il n'y a pas de soucis particuliers à avoir (...), certaines erreurs de syntaxe. Parfois, il utilise le conditionnel au lieu de l'imparfait: « si je pourrais », il le dit plusieurs fois.

E1, note: C2

C'était très bien, pour moi c'est une personne francophone, c'est très naturel, c'est excellent.

E10, note: C2

Les notes divergentes et les commentaires similaires

Il fait des phrases tout à fait correctes grammaticalement, mais par-ci par-là, il y a de petites erreurs, ce n'est pas impeccable partout.

E3, note: B2-2

Il n'y a pas de soucis particuliers à avoir (...), certaines erreurs de syntaxe. Parfois, il utilise le conditionnel au lieu de l'imparfait: « si je pourrais », il le dit plusieurs fois.

E1, note: C2

Le degré de sévérité des commentaires différents

Il fait des phrases tout à fait correctes grammaticalement, mais par-ci par-là, il y a de petites erreurs, c'est pas impeccable partout.

E3, note: B2-2

Tout ce qu'il a utilisé c'était bien, toute sa syntaxe était correcte, tout ce qu'il a utilisé comme temps de grammaire c'était correct.

E8, note: C2

4.4.3.2. Les références extérieures à la grille d'évaluation

Des références à l'utilisation des mauvais pronoms d'adresse ont été relevées, c'est-à-dire au vouvoiement au lieu du tutoiement dans une situation de communication informelle.

Les références au vouvoiement

Trois examinateurs, E6, E7, E8, ont évoqué le fait que le candidat vouvoyait l'animateur tout le long de la conversation de la section B, et sont étonnés de voir que l'animateur ne l'a pas repris. Pour rappel, le dialogue doit être informel dans la section B, le candidat doit jouer le rôle de l'ami de l'animateur, et par conséquent, les deux doivent se tutoyer. E7 ne sait pas s'il peut pénaliser cela, mais il fait remarquer que d'après les directives de la formation qu'il a suivies en présentiel, il est important d'insister sur le tutoiement. Toutefois, il reconnait que pour des candidats, le pronom *tu* peut être délicat à utiliser face à un interlocuteur inconnu dans un contexte officiel de passation de test, et que cet inconfort peut être lié à leur éducation ou à leur culture. Il ajoute que dans sa pratique d'animation, il commence à tutoyer les candidats dès l'annonce de la consigne de la section B afin de faciliter leur entrée dans le jeu de rôle. Puis lorsque les candidats dévient fortuitement vers le vouvoiement, il fait toujours en sorte de les rediriger vers le tutoiement avec humour.

4.4.4. L'évaluation du critère 4 : « Lexique »

L'évaluation du lexique se réalise au travers de la section A et de la section B.

4.4.4.1. Les notes et les commentaires

Les notes du candidat Rayan pour le critère 4 vont de B2-2 à C2. Deux examinateurs ont accordé la note B2-2, trois la note C1-1, deux la note C1-2 et trois la note C2 (Tableau 44).

Tableau 44 - Répartition des scores du critère 4 du candidat Rayan

Scores	B2-2	C1-1	C1-2	C2
Nombre d'examinateurs	2	3	2	3

Pour quatre examinateurs (E3, E5, E10 et E9), le lexique du candidat était bien maitrisé dans l'ensemble, mais avec tout de même quelques lacunes comme des confusions, des imperfections et un manque de variété. Les notes sont respectivement B2-2, B2-2, C1-1, C1-2. Pour les six autres,

(E2, E6, E4, E1, E7 et E8), aucune lacune n'a été relevée, les expressions courantes étaient tout à fait maitrisées, la gamme lexicale était riche et même excellente. Les notes sont respectivement C1-1, C1-1, C1-2, C2, C2, C2.

Les notes similaires et les commentaires divergents

Il utilise une gamme lexicale étendue de ce qu'on attend pour ce type de discussion, comme des mots associés au monde de l'environnement, du recyclage, je pense que c'est très bien.

E2, note: C1-1

Il a commencé à parler des rejets de CO2, il a montré qu'il avait quand même du vocabulaire, mais ce n'était pas approprié du tout dans la conversation.

E6, note: C1-1

Les notes divergentes et les commentaires similaires

Quand même il y avait un peu d'erreurs et des confusions.

E5, note: B2-2

Son lexique n'est pas adapté (...) on comprend le sens global, c'est juste pas les bonnes formulations.

E10, note: C1-1

Le degré de sévérité des commentaires différents

Ce n'est pas une maitrise absolue du lexique, il a une bonne maitrise, mais pas hors de tout doute, il y a de petites imperfections, de petites imprécisions.

E3, note: B2-2

Il a vraiment montré un lexique qui était vraiment riche, il montrait une maitrise parfaite du français, le lexique est bien développé, bien riche.

E8, note: C2

4.4.5. L'évaluation du critère 5 : « Aisance à l'oral, élocution »

L'évaluation de l'aisance à l'oral et de l'élocution se réalisent au travers de la section A et de la section B.

4.4.5.1. Les notes et les commentaires

Les notes du candidat Rayan pour le critère 5 vont de B2-2 à C2. Un examinateur a accordé la note B2-2, deux la note C1-2 et sept la note C2 (Tableau 45).

Tableau 45 - Répartition des scores du critère 5 du candidat Rayan

Scores	B2-2	C1-1	C1-2	C2
Nombre d'examinateurs	1	0	2	7

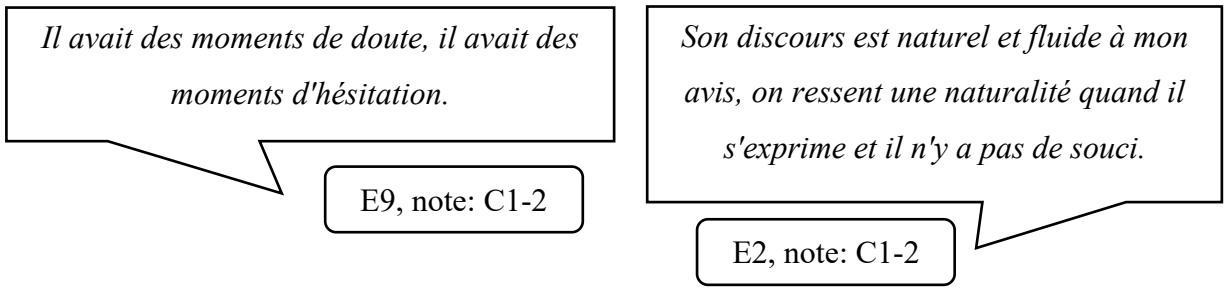
Tous les examinateurs ont constaté dans l'ensemble que le candidat avait une élocution et une aisance à l'oral naturelles, qu'il s'exprimait sans effort et que son intonation était adaptée à la situation, les notes sont B2-2, C1-2 et majoritairement C2. Les quatre examinateurs E6, E8, E9 et E10 ont ajouté que les traits prosodiques du candidat étaient proches ou similaires à ceux d'un locuteur natif, et en particulier d'un francophone de la région du Maghreb en raison de son accent. À ce sujet, E6 mentionne par exemple que l'une des particularités de cet accent est le fait de prononcer le son [ɑ̃] comme le son [ɔ̃].

Un des examinateurs ayant accordé C1-2, E9, affirme qu'il aurait pu accorder la note C2 s'il n'y avait pas eu de doutes et d'hésitations dans le discours du candidat. E8 a également relevé cet aspect dans son commentaire, mais déclare que cela n'a pas affecté sa note, à savoir C2. Pour lui, les silences et les hésitations ne sont pas considérés comme des lacunes, mais plutôt comme des moments de réflexion dont le candidat a besoin pour exprimer sa pensée le mieux possible. Il reconnait en outre qu'il n'était pas facile de rétorquer face aux nombreuses objections de l'animateur qui par ailleurs était très loquace.

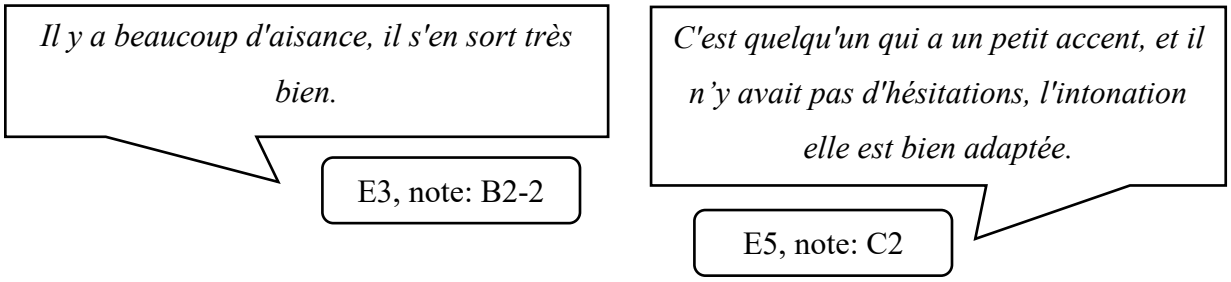
Concernant l'examinateur ayant attribué B2-2, E3, celui-ci souligne l'aisance du discours du candidat et le fait que ce dernier s'en sorte très bien. Il ne relève aucune difficulté particulière et justifie sa note en affirmant qu'il est assez convaincu que le candidat se situe dans le niveau B2-2 de façon globale, c'est-à-dire pour l'ensemble des critères de la grille d'évaluation. D'ailleurs

toutes ses notes pour les cinq critères sont centralisées sur l'échelon B2-2. E3 ajoute par ailleurs qu'il serait sceptique à l'idée de le classer au-delà.

Les notes similaires et les commentaires divergents



Les notes divergentes et les commentaires similaires



4.4.6. Le bilan de l'évaluation du candidat Rayan

En conclusion, des divergences ont été notées dans tous les critères. Certains examinateurs n'ont pas tenu compte de certains énoncés dans les descripteurs de la grille d'évaluation pour leur note de la section A. Par ailleurs, il est apparu à travers les propos que certaines notes pour la syntaxe ont délibérément été trop élevées pour diverses raisons, puis il s'est dégagé des perceptions différentes à l'égard des moments de silence et d'hésitation pour le critère de l'aisance à l'oral. Un autre point récurrent qui a été souligné dans les commentaires est l'attitude de l'animateur ayant été accaparante, mais ayant tout compte fait peu affecté le candidat dans l'ensemble. Enfin, les références à l'utilisation du vouvoiement dans le cadre d'une discussion informelle ont soulevé la question de la prise en compte de cet aspect dans la notation.

Le niveau global estimé du candidat Rayan par l'équipe du Français des affaires est de C1. D'après le tableau ci-dessous (Tableau 46), deux examinateurs, E2 et E3, ont accordé un niveau inférieur, soit B2, et un autre examinateur, E1, a accordé un niveau supérieur, soit C2. Les sept autres ont accordé le niveau estimé de départ, soit C1. La moyenne et la médiane sont C1.

Tableau 46 - Niveau global du candidat Rayan par les dix examinateurs et par l'équipe du Français des affaires (ÉFA)

Examinateurs	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**E9**	**E10**	**Moyenne**	**Médiane**	**ÉFA**
Niveau global	C2	B2	B2	C1	C1	C1	C1	C1	C1	C1	C1	C1	C1

Dans la partie suivante, nous présenterons les réponses des examinateurs concernant nos cinq questions de l'entrevue qui portent sur leur appropriation et sur leur appréciation de la grille d'évaluation.

4.5. Les perceptions des examinateurs lors de l'entrevue

Dans cette section, nous présenterons les perceptions des examinateurs concernant nos cinq questions de l'entrevue qui sont les suivantes :

Q1. « Quel(s) critère(s) est(sont) plus facile(s) à évaluer selon vous ? Justifiez. »

Q2. « Quel(s) critère(s) est(sont) plus difficile(s) à évaluer selon vous ? »

Q3. « En quoi les descripteurs sont-ils appropriés selon vous ? »

Q4. « Quel(s) échelon(s) vous posent le plus de problèmes ? Justifiez. »

Q5. « Avez-vous déjà utilisé l'ancienne grille d'évaluation ? Si oui, entre l'ancienne grille d'évaluation et la nouvelle, laquelle considérez-vous la plus facile utiliser ? Justifiez. »

4.5.1. Perceptions des examinateurs concernant les critères les plus faciles évaluer, Q1

À la question « Quel(s) critère(s) est(sont) plus facile(s) à évaluer selon vous ? Justifiez. », les critères les plus cités par ordre d'importance sont le critère 5 : élocution, aisance à l'oral (cité par huit examinateurs), le critère 1 : capacité à obtenir des informations dans la section A (cité par six examinateurs) et le critère 3 : syntaxe (cité par cinq examinateurs).

Le critère qui s'avère être le plus facile à évaluer est le critère 5 qui porte sur l'élocution et l'aisance à l'oral. Selon huit examinateurs (E1, E2, E5, E6, E7, E8, E9 et E10), ce critère se rapporte à une balise concrète et bien définie sur laquelle ils peuvent se baser pour guider leur évaluation. De plus, il est sans ambiguïté, il permet une certaine objectivité, et de ce fait, il peut être identifié et mesuré assez facilement et rapidement. E6 et E10 mentionnent que leur profession d'enseignant en FLS les amène davantage à porter leur attention sur le critère 5, qui renvoie à la langue en elle-même. Les deux examinateurs expliquent par exemple que lorsqu'ils sont en classe avec leurs étudiants, ils sont quotidiennement confrontés à une large variété de langues et d'accents de tous les niveaux. Placer le niveau d'élocution et d'aisance à l'oral d'un candidat sur la grille d'évaluation est alors « *comme un jeu* », selon les propos de E6, car l'exercice lui est très intuitif. E10 ajoute que le critère 5 est celui qu'il observe en premier lorsqu'il évalue un candidat, car ce critère tient lieu de repère. D'après lui, très souvent, lorsqu'un candidat a un bon débit, celui-ci a également un bon lexique ainsi qu'une bonne syntaxe, alors que cela n'est pas le cas dans une situation inverse, c'est-à-dire qu'un candidat qui possède un lexique très vaste et une syntaxe

correcte peut parler avec un débit très lent. Ainsi, selon l'examinateur, parler avec un débit adéquat et naturel englobe de fait une bonne maitrise des éléments lexicaux et syntaxiques. Par ailleurs, pour E5, qui précise entre parenthèses que le français n'est pas sa langue maternelle, le choix du critère 5 s'explique par le fait qu'il perçoive la langue comme une mélodie qu'il connait bien et qu'il aime entendre. Selon ses propres termes, lorsque « *ça chante faux, ça irrite* », c'est-à-dire qu'il peut instantanément percevoir « *les fausses notes* » d'un candidat. Pour lui, lorsque les éléments prosodiques d'un discours sont lacunaires, cela renvoie à la fausseté d'une mélodie devenue méconnaissable.

Le deuxième critère le plus cité est le critère 1, soit la capacité à obtenir des informations, qui fait uniquement référence à la section A de l'épreuve. Pour six examinateurs (E2, E3, E4, E6, E7, E8), ce critère est facile à évaluer, car il correspond à un exercice simple. En effet, comme la tâche du candidat consiste uniquement à poser des questions à propos d'une annonce de journal, le champ est très restreint et prévisible. Dans la mesure où les questions sont très souvent de même ordre, les examinateurs savent à quoi s'attendre, et selon le sujet de l'annonce, ils peuvent préalablement connaitre les questions allant être posées. D'ailleurs, il a été ajouté par E8 que très souvent lorsque les examinateurs sont en interaction avec le candidat dans la section A, il ne leur est presque pas nécessaire de prendre des notes, car les questions allant être posées sont si prévisibles qu'elles n'entravent pas la mémoire à court terme. Pour E2, E3, E4, E7 et E8, il est simple de se focaliser sur le contenu de la section A, car la durée est relativement courte (cinq minutes) et l'exercice consiste à faire le calcul des questions pertinentes. De ce fait, ils peuvent aisément et rapidement avoir une vue d'ensemble de l'éventail des propos du candidat.

Pour cinq examinateurs (E1, E5, E8, E9 et E10), le troisième critère considéré comme le plus facile à évaluer est le critère 3, qui fait référence à la syntaxe. Cette facilité s'explique par le fait que les structures de phrases soient inhérentes à la langue, elles sont alors plus tangibles et univoques. Étant donné que les examinateurs sont majoritairement des enseignants en français langue seconde, l'enseignement et l'évaluation de la syntaxe et de la grammaire représentent des points centraux dans leur profession. Les erreurs, tout comme les constructions complexes et variées, chez les candidats sont plus faciles et rapides à repérer. Par ailleurs, E5 et E9 ajoutent qu'ils se sentent plus confiants et qu'ils ont rarement des doutes concernant la note qu'ils attribuent au critère 3. Enfin, E1 et E8 déclarent que sur la grille d'évaluation, les frontières entre les échelons du critère 3 sont

plus visibles et distinctes que celles des autres critères, ils peuvent ainsi mieux saisir la variation graduelle au fil des différents descripteurs.

4.5.2. Perceptions des examinateurs concernant les critères les plus difficiles évaluer, Q2

À la question « Quel(s) critère(s) est(sont) plus difficile(s) à évaluer selon vous ? », les critères les plus cités par ordre d'importance sont le critère 2 : capacité à présenter et débattre (par neuf examinateurs), le critère 4 : lexique (par deux examinateurs), et le critère 1 : capacité à obtenir des informations (par deux examinateurs).

Justifications du choix du critère 2

Pour neuf examinateurs (E1, E2, E3, E5, E6, E7, E8, E9 et E10), le critère qui s'avère être le plus difficile à évaluer est le critère 2 qui porte sur la capacité à présenter et débattre dans la section B de l'épreuve. La difficulté du critère 2 réside principalement dans son libellé, à savoir la capacité à présenter et la capacité à débattre, et concurremment dans ses différents descripteurs qui tiennent compte d'une progression parallèle entre ces deux capacités. Pour les examinateurs, ces deux capacités correspondent à deux actions bien distinctes, car d'après leur expérience sur le terrain, les candidats n'accomplissent pas toujours conjointement ces deux actions de façon égale. Ils expliquent qu'ils sont parfois confrontés à des candidats qui font des présentations très sommaires du document de support, mais qui débattent avec des arguments variés et bien détaillés, et inversement, c'est-à-dire des candidats qui présentent le document de support de façon structurée et très bien développée, mais qui apportent des arguments simples et peu illustrés. Entre présenter et débattre, les examinateurs ont des interrogations, car ils ne savent pas lequel de ces deux éléments doit avoir le plus de poids. Cette dualité paralyse leur prise de décision, et par conséquent, il leur est très difficile d'opter pour le bon échelon sur la grille d'évaluation. E7 cite qu'il est difficile de « rentrer dans les cases » de la grille d'évaluation, car il peut accorder par exemple B1 pour la présentation du document, puis une note plus élevée comme C1 pour les arguments. Cependant, lorsqu'il fait la moyenne de B1 et de C1, le résultat, qui serait logiquement B2, ne correspond pas au descripteur B2 de la grille d'évaluation. Par ailleurs, il ajoute qu'il est parfois difficile de bien faire comprendre aux candidats, même francophones, que la section B de l'épreuve d'expression orale comporte deux objectifs (présenter et débattre) qui ont une importance équivalente, car il observe que les candidats délaissent quelquefois la présentation au profit du débat. Enfin, E1 et E10 mentionnent qu'ils auraient souhaité que le critère 2 soit divisé en deux,

comme il l'était dans l'ancienne version de la grille d'évaluation (datant d'avant octobre 2018), car la division leur permettrait de faciliter la tâche d'évaluation.

Pour E3, E9 et E10, le critère 2 est le plus difficile à évaluer, car les candidats peuvent parfois manquer d'idées, et le manque d'idées d'un candidat peut quelquefois être dû à sa personnalité, plutôt qu'à sa réelle compétence en français. Selon eux, être capable de bien débattre dans la section B dépend entre autres du trait de caractère naturel des personnes, car même les candidats francophones ou ceux ayant un très bon niveau de français peuvent être incapables de défendre une opinion, de produire une argumentation riche et variée, ainsi que de rebondir aux contre-arguments de leur interlocuteur de façon spontanée. Si les candidats ont des difficultés à convaincre naturellement dans leur propre langue maternelle, il y a de fortes chances qu'ils aient ces mêmes difficultés dans la section B de l'épreuve. Compte tenu de cet aspect, E10 ajoute que le débat ne devrait pas avoir sa place dans un test ayant pour objectif d'évaluer une langue.

Par ailleurs, E5 relève que les candidats ont souvent les moyens linguistiques de débattre, qu'ils sont capables de réagir en ayant des réactions appropriées, mais qu'ils n'ont parfois pas les bonnes idées d'arguments qui arrivent au bon moment lors de l'épreuve de la section B. Par conséquent, les candidats ne parviennent pas à accomplir la tâche demandée de l'épreuve de façon optimale. E5 constate cela avec les candidats avec qui il communique de façon parallèle en échangeant des informations à des stades différents de la passation du test, c'est-à-dire lorsqu'il les accueille dans les salles d'examen, lorsqu'il leur explique les consignes, puis lorsqu'il met fin à la passation du test. Quand il est face à ce type de candidats, l'examinateur déclare qu'il est pénible de devoir les pénaliser juste parce qu'ils n'ont eu pas les bonnes idées à point nommé. Il admet même être parfois un peu plus généreux en termes de notes pour le critère 2 lorsqu'il se rend compte d'un décalage entre la performance observée d'un candidat dans le cadre du test et « *comment il est en vrai en dehors du test* », selon ses propres termes.

D'autre part, le manque d'idées des candidats peut parfois provenir des sujets des documents proposés aux candidats dans la mesure où certains sujets sont moins inspirants que d'autres. D'après E3 et E9, si par exemple un candidat n'a pas déjà vécu une expérience personnelle similaire au sujet de son document ou s'il n'est pas familier avec, il sera plus compliqué pour lui d'imaginer des arguments afin d'enrichir le débat. Il a également été évoqué que les sujets des documents de la section B ont des degrés de difficulté variables, c'est-à-dire que certains sont

ordinaires tandis que d'autres sont plus abstraits ou complexes, et cet aspect constitue un biais. E3 pense que lorsque les sujets sont proches de la réalité des candidats, ces derniers sont chanceux.

De plus, la difficulté du critère 2 peut également être liée à la différence culturelle de certains candidats. Par exemple, E5 et E10 observent que très souvent, les candidats d'origine asiatique, principalement de Chine et de Corée du Sud, sont réservés et hésitants dans la manière dont ils mènent le débat, puis qu'ils concèdent facilement même si leur niveau de compétence langagière en français est bon. D'après leur expérience, les examinateurs constatent que ces candidats n'insistent généralement pas beaucoup pour convaincre leur interlocuteur (qui joue le rôle d'un ami réticent à faire une activité) et qu'ils répondent rarement avec persévérance aux propos opposés. E5 pense que les candidats considèrent le fait d'être insistant comme une forme de pression et même une forme d'impolitesse, d'autant plus qu'ils interagissent avec une personne inconnue ayant une position hiérarchique supérieure. Lors du débat de la section B, malgré les explications données dans la consigne et les efforts déployés afin de mettre à l'aise les candidats dans le jeu de rôle, E5 est convaincu que ces derniers perçoivent toujours les examinateurs comme des professeurs qu'ils doivent respecter sans enfreindre les codes de bonne conduite issus de leur culture.

Il a été mentionné en outre par E10 que même dans la culture québécoise, le débat est une pratique relativement peu courante par rapport à la culture française. Bien que le Québec et la France partagent la même langue, le sens du débat n'est pas le même dans les deux cultures : il est beaucoup plus aiguisé chez les Français alors qu'il est peu valorisé chez les Québécois où il est vu comme une « *chicane argumentée* » selon les termes de l'examinateur. Celui-ci rappelle d'ailleurs que comme le test TEF a été conçu en France, ce dernier relève des pratiques culturelles de ce pays.

Il apparait également que le critère 2 est le plus difficile à évaluer avec les candidats de niveau intermédiaire, et plus précisément avec ceux situés entre le niveau A2-2 et le niveau B2-2. Selon E6, il est plus difficile pour lui de savoir réellement où placer les candidats de ces niveaux-là de façon précise sur la grille d'évaluation, surtout lorsque ces derniers cherchent leurs mots et s'expriment avec beaucoup d'hésitation. Même si les candidats réalisent les objectifs de la section B, il se trouve qu'il lui est compliqué de faire le tri à la fin de l'échange oral entre ce qui fait partie d'un simple discours sans profondeur, et ce qui fait partie d'un discours où l'on retrouve une vraie

intention de convaincre l'interlocuteur. De plus, l'examinateur ajoute qu'avec ce profil de candidats (de niveau intermédiaire), il est facile de perdre sa concentration, surtout lorsque l'on évalue plusieurs candidats d'affilé sans interruption, et lorsque l'on effectue au préalable d'autres tâches liées au test comme des tâches administratives et des surveillances d'épreuves de compréhension orale.

Pour E8 et E9, le critère 2 est le plus difficile à évaluer entre autres en raison de sa durée (qui est de dix minutes) et de la complexité de la section B qui exigent davantage d'écoute et de concentration comparativement à la section A. Durant une période de dix minutes, les aspects auxquels les examinateurs doivent prêter attention sont multiples et simultanés : ils doivent être en mesure de repérer chez le candidat sa manière de présenter un document écrit (c'est-à-dire distinguer une façon de faire originale et étoffée d'une relecture sans reformulation), de se focaliser sur la qualité et la pertinence des arguments apportés, de mener un travail d'animation en improvisant le rôle d'un ami ayant un comportement désapprobateur, et enfin d'être attentif aux aspects linguistiques.

Justifications du choix du critère 4

Le critère 4, qui se rapporte au lexique, a également été cité comme étant l'un des plus difficiles à évaluer par E4, E5 et E7. D'après les constatations de E5, le problème vient des candidats francophones de niveau maîtrise (C2), et plus particulièrement des candidats originaires de France, car certains possèdent un vocabulaire qui est excellent et d'autres un vocabulaire assez bon, sans être excellent. Lorsque le niveau de vocabulaire de ces candidats est assez bon, l'examinateur réalise que lui-même ainsi que la majorité de ses collègues leur accordent le bénéfice du doute en attribuant d'office la note C2, la note la plus avancée. Toutefois, il est conscient que cela n'est pas cohérent au vu de l'énoncé du descripteur du niveau C2 de la grille d'évaluation qui indique que le répertoire lexical doit être riche et nuancé, adapté et très bien maîtrisé. Cette situation préoccupe l'examinateur, car il lui est difficile d'affirmer si ces candidats francophones ont de réelles lacunes lexicales ou s'ils se contentent délibérément du minimum.

Par ailleurs, il apparait que le critère 4 est l'un des plus difficiles à évaluer en raison de certains descripteurs du lexique de la grille d'évaluation. E7 déclare ne pas s'y retrouver facilement entre les termes : « Répertoire lexical assez large », « Répertoire lexical vaste et bien maitrisé » et « Répertoire lexical riche et nuancé ». Ces derniers manquent de nuances et le font beaucoup

hésiter dans son évaluation. De plus, l'examinateur s'interroge sur le descripteur du niveau C1 où il est mentionné : « Les lacunes sont compensées sans effort apparent », il se demande alors si l'on peut vraiment parler de lacunes si celles-ci sont compensées sans effort apparent. Enfin, il avoue d'autre part qu'il enlève souvent des points aux candidats lorsqu'ils ne comprennent pas ce qu'il dit, car selon lui, une faiblesse dans la compréhension orale témoigne d'un manque de lexique.

Justifications du choix du critère 1

Le critère 1 qui porte sur la capacité à obtenir des informations dans la section A a également été cité comme étant l'un des plus difficiles à évaluer par E9. L'examinateur spécifie que la difficulté se produit plus spécifiquement avec les candidats francophones qui ne posent pas suffisamment de questions, qui posent le bon nombre de questions, mais qui stoppent la conversation au bout de deux ou trois minutes (au lieu de cinq minutes), qui parlent plus qu'ils ne posent de questions ou qui ne rebondissent pas aux réponses incomplètes ou ambigües de l'interlocuteur. Avec ce profil de candidats, l'examinateur n'est alors pas certain de l'aspect dont il doit davantage tenir compte dans sa note, à savoir le nombre de questions posées, le respect de la durée de cinq minutes, ou le fait de rebondir en demandant des précisions aux informations partielles de l'animateur.

4.5.3. Perceptions des examinateurs concernant les descripteurs, Q3

Concernant la troisième question de l'entrevue : « En quoi les descripteurs sont-ils appropriés selon vous ? », E1, E4 et E6 ont mentionné que dans l'ensemble, les descripteurs étaient appropriés dans la mesure où l'on pouvait assez bien saisir les variations graduelles au fil des différents niveaux grâce à certains mots ou phrases qui marquent la différence. Par exemple, E1 cite le critère 2 (capacité à présenter et débattre) où l'on retrouve pour le niveau B1 « Présentation simple et claire », puis pour le niveau B2 « Présentation claire et détaillée ». Ici, le fait qu'il y ait le terme « détaillé » lui permet de mieux se positionner. E4 donne l'exemple du critère 1 pour le niveau B2 où l'on a « Questionnement approprié [...]. Les échanges sont suivis », puis pour le niveau C1 « Questionnement complet et précis [...]. La conversation est soutenue ». Lorsque l'examinateur trouve le questionnement du candidat bon, mais qu'il a un doute entre les niveaux B2 et C1, les derniers énoncés des descripteurs lui permettent de statuer.

Pour E2, E3, E6, E7 et E10, certains descripteurs ne leur permettent pas de faire de distinction entre les différents échelons, car ils sont trop similaires dans leur formulation. Par exemple, E3 dit ne pas percevoir de différence dans la manière dont sont énoncées les premières parties des

descripteurs des niveaux C1 et C2 du critère 1 (portant sur la capacité à obtenir des informations) : « Questionnement complet et précis » et « Questionnement exhaustif et pertinent », puis des niveaux B2 et C1 du critère 2 (portant sur la capacité à présenter un document et à débattre) : « Présentation claire et détaillée. Arguments illustrés ou détaillés. » et « Présentation développée et structurée. Arguments variés et bien développés. ». Dans la même optique, E10 ne distingue pas clairement les deuxièmes parties des descripteurs des niveaux et B1 et B2 du critère 3 (portant sur la syntaxe) : « Les erreurs sont fréquentes, mais le sens général est clair. » et « Les erreurs ne conduisent pas à des malentendus. ». Ainsi, une trop grande présence de synonymes dans plusieurs descripteurs juxtaposés rend la distinction entre les différents échelons de la grille d'évaluation difficile. De façon à mieux les guider dans leur choix de notations, ces examinateurs souhaiteraient que les descripteurs vagues contiennent davantage d'informations précises, concrètes et univoques.

Pour E5 et E8, certains descripteurs regroupent parfois plusieurs actions assez distinctes et rendent leur prise de décision difficile. Par exemple, E8 cite le descripteur du niveau C1 du critère 1 qui indique : « Questionnement complet et précis. Intervient avec justesse. La conversation est soutenue. ». D'après lui, il peut arriver quelquefois chez un candidat que les deux premiers éléments, c'est-à-dire le questionnement et l'intervention, soient adaptés au descripteur du niveau C1, mais que le troisième élément, c'est-à-dire la conversation, ne le soit pas. Dans ce cas de figure, l'examinateur admet qu'il fait le choix de ne pas tenir compte de l'une des parties du descripteur, car toutes les actions du candidat peuvent parfois ne pas s'accomplir de façon harmonieuse.

Il a par ailleurs été soulevé par E7 que les descripteurs sont appropriés dans l'ensemble, mais que certains manquent amplement de précision, par exemple les termes des descripteurs du critère 3 (portant sur la syntaxe) comme « phrases simples », « phrases complexes », « grande variété de structures », puis du critère 4 (portant sur le lexique) comme « répertoire lexical plus large », « répertoire lexical assez large », « répertoire lexical vaste », « répertoire lexical riche ».

E7 est conscient que la grille d'évaluation seule n'est pas suffisante pour exercer son jugement et qu'il doit avoir en amont une très bonne connaissance du référentiel du CECRL, dont il juge en outre sa lecture assez laborieuse. Afin de trouver un bon compromis entre le manque de précision de certains descripteurs et le document du CECRL, l'examinateur suggère, à titre d'illustration, de placer au verso de la grille d'évaluation des exemples concrets de ce qui est attendu pour les

critères, tels que des modèles de phrases simples et de phrases complexes courantes (faisant référence au critère 3 : syntaxe), ainsi que des modèles de présentations limpides suscitant l'intérêt (faisant référence au critère 2 : capacité à présenter et débattre).

4.5.4. Perceptions des examinateurs concernant les échelons, Q4

À la question : « Quel(s) échelon(s) vous posent le plus de problèmes ? Justifiez. », les échelons B1, B2, A1 et C2 ont été cités.

Les échelons B1 et B2 sont majoritairement les échelons qui posent le plus de problèmes pour neuf examinateurs (E1, E2, E3, E4, E5, E6, E7, E8, E10). Parmi les justifications, l'enjeu très élevé que représente le test a été évoqué. En effet, comme nous l'avons déjà mentionné à plusieurs reprises, le niveau B2 est le niveau à partir duquel des points sont attribués aux candidats ayant un projet d'émigration durable vers la province du Québec. Étant donné que les examinateurs incarnent une instance et qu'ils représentent une des parties prenantes du réseau d'émigration en tant que personnes chargées d'une mission pour le compte de l'administrateur du test TEF, le fait que les examinateurs optent pour B1 ou B2 sur la grille d'évaluation a donc un impact considérable sur le cheminement de la vie des candidats. Cette importante responsabilité pèse sur leur jugement, car ne pas avoir « le droit à l'erreur » afin d'être le plus juste possible provoque chez certains une lutte intérieure. En effet, donner la « vraie note » s'avère très difficile pour les examinateurs, car ils sont conscients qu'ils ne doivent pas favoriser des profils particuliers de candidats ni être trop drastiques au risque de compromettre leurs projets d'avenir. De plus, certains gardent à l'esprit que le jour de la passation du test correspond à une grande situation de stress où les personnes ne sont généralement pas à cent pour cent de leurs capacités. Lorsque la performance orale d'un candidat oscille entre l'échelon B1 et l'échelon B2, cela génère chez les examinateurs davantage de doute et d'hésitation, et par conséquent, le temps de réflexion et de concentration est plus important. E7 ajoute que bien que ses horaires de passation du test soient limités (en mode présentiel face aux candidats), il prend souvent l'initiative de réécouter une à deux fois la conversation afin de s'assurer de fournir la notation la plus juste possible. D'autres examinateurs, comme E3 et E6, qualifient les niveaux B1 et B2 de zone grise, et la frontière entre les niveaux B1-2 et B2-1 de ligne sensible, car il leur est toujours très difficile de faire leur choix. Ainsi, « *le seuil de réussite B2* » (d'après les termes de E4) pose un problème, étant donné que chaque point attribué autour de cette aire a une importance majeure et qu'il n'est pas possible de revenir en arrière.

Pour E1 et E5, les échelons B1 et B2 sont ceux qui leur posent le plus de problèmes et plus particulièrement avec les candidats qui se situent partiellement dans un de ces deux niveaux. Les candidats de niveau B2 sont caractérisés, de façon générale, comme étant des utilisateurs indépendants de la langue, capables d'aborder des sujets abstraits tout en ayant une bonne maitrise des éléments linguistiques. Or, il a été observé chez ces examinateurs qu'il y a parfois un décalage chez ces candidats entre leurs compétences communicatives (questionner, débattre) qui sont plus élevées et leurs compétences linguistiques qui sont plus fragiles, surtout en ce qui concerne la syntaxe et le lexique. Par exemple, lorsque les compétences communicatives d'un candidat correspondent à l'échelon B2, mais que ses compétences syntaxiques et lexicales correspondent à l'échelon B1, certains examinateurs admettent qu'ils accordent parfois le niveau B2 à ces deux compétences malgré tout. Ils ferment alors les yeux sur les lacunes linguistiques en vue d'unifier le niveau global sur la grille d'évaluation.

D'autre part, d'après E5, le fait que le discours d'un candidat de niveau B2 ne contienne pas de vocabulaire nuancé, d'expressions imagées, ni de phrases complexes comme il est prescrit dans la grille, lui pose un problème dans sa notation. Selon lui, la norme syntaxique et lexicale du niveau B2 de la grille d'évaluation convient davantage à des cadres formels ou académiques, par exemple lorsqu'il s'agit de faire des discours devant un auditoire, de rédiger des comptes-rendus, des lettres officielles, ou des dissertations. Or, pour des candidats n'étant pas en lien avec cette réalité, mais se destinant à exercer au Québec une activité professionnelle dans un secteur très différent, par exemple dans un secteur technique comme celui du bâtiment ou de la restauration, l'examinateur estime que les standards de la syntaxe et du lexique du niveau B2 ne sont pas indispensables. Enfin, E6 rejoint ce propos et affirme qu'il est conscient que les modèles que l'on applique pour représenter l'expérience de la communication ne peuvent pas rendre compte de la propre expérience et des propres besoins de chacun des candidats ni de leur difficulté.

Pour E9, les échelons A1 et C2 posent le plus de problèmes, car ils ne contiennent qu'un sous échelon, contrairement aux autres échelons de la grille d'évaluation qui en contiennent deux. L'examinateur n'en connait pas la raison et ne comprend pas la logique de ce découpage. Lorsqu'il est face à des candidats de niveau C2, qui sont majoritairement des francophones, il constate que bien que ces derniers maitrisent entièrement la langue française, ils ne répondent pas pleinement aux exigences de la norme, car ils n'emploient pas constamment de grandes variétés de

constructions de phrases ni d'expressions idiomatiques, et n'ont pas systématiquement un répertoire lexical très riche et très nuancé. Selon l'examinateur, il en est de même en ce qui concerne les compétences communicatives (faisant référence au critère 1 : « Capacité à obtenir des informations » et au critère 2 : « Capacité à présenter et débattre »), car il est tout à fait envisageable que les candidats avancés et débutants se situent dans un « C2 faible » ou un « C2 fort », tout comme dans un « A1 faible » ou un « A1 fort ». L'examinateur juge qu'il serait pertinent de subdiviser les échelons A1 et C2, car cela offrirait plus de marge de manœuvre et donnerait lieu à une meilleure précision dans son évaluation. D'ailleurs E10 ajoute à ce propos qu'il serait souhaitable d'avoir trois sous échelons à tous les niveaux de la grille, car cela permettrait d'avoir une option neutre, et ainsi de mieux nuancer leur jugement.

4.5.5. Perceptions des examinateurs concernant l'ancienne et la nouvelle grille d'évaluation, Q5

À la question : « Avez-vous déjà utilisé l'ancienne grille d'évaluation ? Si oui, entre l'ancienne grille d'évaluation et la nouvelle, laquelle considérez-vous la plus facile utiliser ? Justifiez. », tous les examinateurs ont déclaré avoir déjà utilisé l'ancienne grille d'évaluation. Quant aux durées d'utilisation, elles sont variables, elles vont de un à dix ans, mais sont majoritairement d'environ cinq ans.

Tous les examinateurs considèrent que la nouvelle grille d'évaluation est celle qui est la plus facile à utiliser, et parmi les justifications, la rapidité de son utilisation a été évoquée de façon unanime. À titre de rappel, elle a été adoptée en octobre 2018 et est passée de douze critères à cinq critères, le nombre d'échelons est resté inchangé (sept), mais le nombre de sous échelons a été amplement réduit, passant de vingt-et-un à onze. Pour les examinateurs, la grille actuelle leur simplifie le travail, car les nombreux regroupements leur donnent une vue d'ensemble plus holistique.

Par ailleurs, il a été relevé par E3, E5 et E8 que la nouvelle grille convient mieux à la nouvelle réforme du fonctionnement de l'évaluation, datant également d'octobre 2018. Avant cette date, chaque candidat était face à un jury constitué de deux examinateurs, et chaque examinateur était à tour de rôle animateur et observateur. L'examinateur-animateur jouait le rôle de l'interlocuteur dans l'une des deux sections de l'épreuve, pendant que l'examinateur-observateur contrôlait le temps de passation et gérait l'enregistrement sonore. Les examinateurs intervertissaient leur rôle pour la deuxième section de l'épreuve. Quant à la notation, les deux examinateurs remplissaient

d'abord la grille d'évaluation de façon individuelle, puis comparaient leurs notes et négociaient pour arriver à un consensus. Depuis la mise en place de la nouvelle réforme, l'évaluation se réalise toujours en binôme, mais de façon non interactive et isolée, c'est-à-dire qu'un examinateur est en contexte face au candidat et l'autre est hors contexte et a recours aux enregistrements audio. Cette nouvelle procédure du fonctionnement de l'évaluation accompagnée de sa nouvelle grille convient aux examinateurs, car il a été mentionné que grâce au gain de temps procuré, ils peuvent davantage se consacrer à l'interaction avec les candidats et ainsi mieux apprécier la performance de ces derniers. Avec l'ancienne grille d'évaluation qui était plus étoffée, les examinateurs étaient parfois plus préoccupés par les nombreux choix qu'ils allaient devoir effectuer et ne se focalisaient pas pleinement sur l'écoute des candidats.

Lorsque les examinateurs évoquent l'ancienne procédure d'évaluation, certains, comme E6 et E10, abordent spontanément le fait qu'ils préféraient être face à un pair, et d'autres, comme E4 et E8, disent qu'ils préfèrent travailler en autonomie en étant seuls. Ceux qui avaient une préférence pour l'interaction entre pairs déclarent qu'ils pouvaient obtenir des rétroactions en cas d'hésitations et mettre en parallèle les deux points de vue. Négocier à deux pour arriver à un consensus sur le résultat final du candidat nécessitait que chacun explicite sa démarche en étayant ses arguments à haute voix, et cela leur permettait de développer une expertise du raisonnement évaluatif. Selon eux, la nouvelle procédure a progressivement conduit à un amenuisement de cette expertise. Ceux qui préfèrent évaluer de façon isolée mentionnent qu'avec l'ancienne méthode de travail et l'ancienne grille d'évaluation qui était plus consistante et plus analytique, ils passaient beaucoup trop de temps à justifier leur point de vue et à se mettre d'accord avec leur coéquipier. Par conséquent, cela leur faisait ressentir une lourdeur et lorsque les avis entre pairs étaient trop divergents, et cela engendrait même parfois des conflits latents. E8 ajoute par ailleurs qu'avant la nouvelle réforme, il comparait souvent les passations de l'épreuve d'expression orale du TEF avec celles de tests d'anglais standardisés équivalents dans le centre de langues où il travaillait. Il observait que pour les tests d'anglais, un seul examinateur était présent face à un candidat et il trouvait alors que cette configuration d'évaluation était beaucoup plus limpide et qu'elle paraissait tout aussi efficace. Pour lui, la disposition d'un jury en dyade pour le TEF représentait une double « gestion » : celle du candidat, mais aussi celle de son collègue.

Comme il a déjà été souligné dans la question 2 de l'entrevue, le regroupement des trois critères communicationnels de la section B (de la grille précédente) en un seul critère (dans l'actuelle grille) embarrasse un grand nombre d'examinateurs et quelques-uns (E1, E2, E3 et E7) ont à nouveau insisté sur ce point lors de la question 5 de l'entrevue. Nous rappelons que les trois critères communicationnels de la section B de l'ancienne grille ont fusionné en un seul critère dans la nouvelle grille, c'est-à-dire que les critères 4 (présentation des faits), 5 (qualité de l'argumentation) et 6 (qualité des échanges) sont devenus le critère 2 (section B - capacité à présenter et débattre). Le fait que le critère 2 regroupe deux actions distinctes gêne les examinateurs dans leur prise de décision, car ils observent dans leur pratique que les candidats n'accomplissent pas toujours conjointement les deux actions exigées (présenter le document de support et mener un débat) de façon égale.

Un autre aspect qui a été relevé dans les réponses des examinateurs est le manque de précision de la nouvelle grille en raison de la réduction du nombre de critères et du choix des sous échelons. Selon E3, E4, E5, E6, E7, E8 et E10, comme le contenu est plus allégé, il y a moins de nuance et de flexibilité, et cela les amène à faire des choix plus tranchés, surtout avec les critères 1 et 2 (critère 1 : « Capacité à obtenir des informations » dans la section A; critère 2 : « Capacité à présenter et débattre » dans la section B). De ce fait, leur choix est parfois marqué par de l'incertitude malgré l'apparence commode de l'outil. Avec l'ancienne version de la grille, les examinateurs ont déclaré qu'ils avaient le sentiment que leur jugement était plus juste dans la mesure où les résultats étaient plus subtils, et par conséquent, cela était à l'avantage des candidats.

Bien que ces examinateurs considèrent que l'ancienne grille permettait d'obtenir des résultats plus précis, ils éprouvaient toutefois des blocages avec quelques critères. Par exemple à l'égard du critère 8 qui ciblait la cohésion du discours (la manière dont les idées s'enchaînent), E9 n'était pas sûr s'il devait systématiquement prendre en compte l'utilisation de connecteurs logiques comme il était prescrit dans quelques descripteurs (quatre descripteurs sur sept). L'examinateur constatait très souvent que les candidats de niveau avancé n'en utilisaient jamais alors qu'il leur accordait la note la plus élevée pour ce critère.

Le critère 9 qui visait l'étendue du lexique (le lexique dont le candidat est en possession) et le critère 10 qui visait la maitrise du lexique (la capacité du candidat à maîtriser l'usage du lexique) causaient également des difficultés à E7, car ils étaient perçus comme trop similaires, malgré la

présence d'informations dans les descripteurs. Même si la distinction était assez compréhensible d'un point de vue théorique, dans la pratique, l'examinateur rencontrait fréquemment des difficultés à dissocier les deux critères, et de ce fait, il les combinait en attribuant la même note.

En outre, il a été évoqué par E10 que les deux contextes d'évaluation (en mode présentiel et à distance en mode asynchrone) pouvaient donner des perspectives différentes. Par exemple, l'examinateur qui est en mode présentiel est soumis à des contraintes temporelles : il doit respecter un horaire précis, gérer le matériel d'enregistrement, co-construire l'interaction avec le candidat en contrôlant la durée, puis émettre son jugement en temps réel immédiatement à la fin de l'échange. Par ailleurs, il développe d'une certaine façon une forme de lien socioaffectif éphémère avec le candidat en l'accueillant dans la salle d'examen, en le mettant à l'aise avec des attentions subtiles (sourires, ton bienveillant de la voix), en le rassurant s'il est stressé, et en jouant le rôle de son ami avec l'utilisation du tutoiement dans la section B de l'épreuve. Ainsi, il a été relevé que les exigences cognitives sont plus importantes pour l'examinateur en mode présentiel que pour l'examinateur à distance, car il gère et reçoit beaucoup plus d'informations. De ce fait, les caractéristiques évoquées peuvent exercer une influence indirecte sur le jugement et jouer sur l'objectivité de la note.

Enfin, des suggestions de mises à jour concernant la grille d'évaluation actuelle ont été faites par E2 et E7, comme le fait de pouvoir obtenir des éclaircissements sur la pondération des points attribués pour chaque échelon et sous échelon. Lors des formations, comme aucune indication n'est donnée à ce sujet, les examinateurs pensent qu'il est important pour eux, par souci de transparence et d'équité, d'avoir une connaissance de base du système de calculs des points. Par exemple, concernant les deux sous échelons des niveaux A2, B1, B2 et C1, les examinateurs ont mentionné vouloir connaitre leur pourcentage, c'est-à-dire savoir s'ils représentent chacun 50% ou plutôt 70% et 30%. Ils se sont également questionnés sur le poids des deux catégories des critères (communicatifs et linguistiques) de la grille d'évaluation et se demandent si le volet linguistique domine par rapport au volet communicatif étant donné qu'il y a trois critères linguistiques et deux critères communicatifs.

Ce quatrième chapitre a permis de présenter les méthodes d'analyse utilisées pour répondre à nos questions de recherche et les résultats obtenus à l'aide de ces méthodes. Dans le chapitre suivant,

nous poursuivons notre démarche de recherche par la discussion des résultats présentés afin de formuler des réponses claires à nos deux questions de recherche.

CHAPITRE 5: DISCUSSION

Introduction

Ce dernier chapitre s'attarde à discuter des résultats présentés au quatrième chapitre afin de répondre à nos questions de recherche. Pour chaque partie, nous commenterons les points saillants et les interpréterons conceptuellement en les comparant aux études réalisées dans le domaine et dont nous avons fait état précédemment. Dans un premier temps, nous présenterons un résumé des résultats en ce qui concerne les diverses formes de divergences (ce qui correspond à la première question de la recherche), puis nous dresserons l'état des lieux de l'appropriation et de l'appréciation de la grille d'évaluation (ce qui correspond à la deuxième question de la recherche). Pour finir, nous proposerons nos suggestions.

5.1. Les divergences observées chez les examinateurs

Les divergences que l'on a observées portent sur le raisonnement évaluatif et la note, sur la familiarité avec l'accent des candidats, sur les inférences non pertinentes, sur la perception de l'attitude de l'animateur ainsi que sur d'autres aspects.

5.1.1. Les divergences dans le raisonnement évaluatif et la note

Des études empiriques ont révélé que le raisonnement évaluatif et la note que les examinateurs attribuaient pouvaient différer. En effet, 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 manière similaire et attribuer des notes différentes. Des notes quantitativement identiques n'excluent donc pas des différences qualitatives dans la prise de décision de l'examinateur ou dans l'interprétation du construit (Ang-Aw et Goh, 2011; Douglas, 1994; Douglas et Selinker, 1992, 1993; Orr, 2002).

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

Les commentaires nous permettent de mieux saisir le sens donné à la note sous une perspective plus « interne ». Cela laisse alors supposer que s'il y avait absence de grille d'évaluation, les résultats au test pourraient être différents, les commentaires permettraient de savoir à quel point les candidats répondent aux attentes que l'on a d'eux.

Parallèlement, nous avons très souvent observé une large dispersion des notes et des commentaires pour un même candidat à un même critère. Par exemple, concernant le critère de la capacité à présenter et débattre d'une même candidate, nous avons eu ces deux notes : B1-2 et C2, avec des commentaires très opposés qui sont respectivement les suivants : « *Elle a expliqué vraiment dans les grandes lignes ce que c'était d'avoir un étudiant dans une famille d'accueil (...), mais après je me suis bien rendu compte qu'elle n'avait pas de suite dans les idées (...) il n'y a aucune argumentation pratiquement* », puis : « *C'était complet, c'était très pertinent, la présentation et les arguments, c'était super, elle a très bien réussi à convaincre, elle réagit de manière adéquate et elle a beaucoup de répondant* ». Tout au long de l'exercice de verbalisation, les disparités dans les notes étaient récurrentes, les dix examinateurs n'ont à aucun moment tous attribué la même note à un même critère pour l'ensemble des quatre candidats. Ces dispersions signifient que les conséquences sont très différentes en termes d'obtention de points au TEF.

Malgré les mêmes formations suivies et l'utilisation d'un même outil basé sur une norme commune, les examinateurs n'attribuent pas la même signification aux performances des

candidats. Ils valorisent des aspects différents et leurs sensibilités par rapport aux différents éléments de la grille d'évaluation sont très variables. En conséquence, ces disparités nécessitent une harmonisation et ce point crucial devrait être envisagé dans le contenu des formations des examinateurs.

5.1.2. Les divergences dans la familiarité avec l'accent des candidats

Dans d'autres études empiriques sur les effets des examinateurs, nous avons constaté que la familiarité des examinateurs avec l'accent des candidats avait des répercussions positives sur le score de la prononciation (Carey *et al.*, 2011; Huang *et al.*, 2016; Huang et Jun, 2014; Hsieh, 2011; Winke *et al.*, 2011, 2012). Il est admis d'ailleurs en linguistique et en sciences cognitives que les accents familiers de langue étrangère sont plus faciles à comprendre que les accents inconnus. De plus, la perception que les auditeurs attribuent à certaines caractéristiques de la prononciation change avec l'expérience linguistique, c'est-à-dire que plus nous sommes en contact avec une langue étrangère, plus nous devenons familiers avec celle-ci (Nittrouer *et al.*, 1993; Zhang *et al*., 2005).

Dans notre étude, quelques examinateurs ont pris conscience qu'ils avaient eu une perception différente des traits phoniques de quelques 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 à celle-ci, et par conséquent, il a affirmé qu'il la comprenait très bien, en comparaison des 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 rehausser 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 FLS) issus de la même région. 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 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 que cette familiarité l'avait influencé de façon positive.

Ainsi, la familiarité de l'examinateur avec l'accent du candidat peut compromettre l'exactitude et l'interprétation d'une évaluation. Par conséquent, elle ne doit pas être sous-estimée et les études à son sujet doivent être abordées pendant les formations des examinateurs.

5.1.3. Les divergences dans les inférences

D'autres études sur les effets des examinateurs ont montré que ces derniers avaient très souvent tendance à faire des inférences sur les candidats, et que celles-ci étaient typiquement utilisées pour excuser ou expliquer certains modèles de comportement des candidats ou pour justifier l'attribution de certaines notes. Par exemple, il a été constaté que les difficultés prosodiques de candidats pouvaient être expliquées par les examinateurs comme étant liées à leur nervosité, à leur manque d'intérêt pour le sujet de discussion, ou simplement par le fait de chercher leurs mots. D'autres difficultés comme celles liées à la grammaire, au lexique et aux idées des candidats peuvent être expliquées par leur manque de culture générale ou leur personnalité, par exemple leur immaturité d'esprit ou leur volonté ou réticence à dialoguer (Brown, 2000, 2006; Pollitt et Murray, 1996).

Selon les chercheurs, les examinateurs se retrouvent souvent incertains face aux vraies causes des problèmes rencontrées par les candidats et ont de la difficulté à émettre leurs jugements. Ces inférences représentent un problème majeur, car elles ne forment pas une base adéquate pour la formulation d'un jugement (Brown, 2000, 2006; Orr, 2002). Elles peuvent être vues comme des « dangers », car elles creusent l'écart entre l'information à l'état brut, c'est-à-dire l'information telle qu'elle a été recueillie, et sa traduction par l'examinateur (Roegiers, 2004).

Dans notre 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 de l'épreuve ou le sujet, 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 les jeux de rôle, et par ailleurs, elle aurait utilisé des styles de phrases typiques de sa zone géographique qui n'étaient pas celles demandées dans le test.

Concernant les difficultés d'une autre candidate à trouver des arguments dans la section B de l'épreuve, quelques examinateurs ont associé cela à un manque de connaissance d'un concept lié à un aspect culturel. Enfin concernant une autre candidate, il a été dit que ses difficultés à trouver des questions dans la section A et à se confronter à son interlocutrice dans la section B étaient dues à une différence culturelle. À ce propos, la difficulté à débattre dans la section B pour les candidats d'origine asiatique (plus particulièrement de Chine, de Corée du Sud et du Japon) a souvent été évoquée par les examinateurs. D'après leurs observations sur le terrain, ces candidats sont plutôt réservés dans la manière dont ils mènent le débat et concèdent facilement, même si leur niveau de compétence langagière en français est bon. Un examinateur a mentionné que ces candidats considèrent le fait d'être insistant comme une forme de pression et d'impolitesse, et qu'ils perçoivent les examinateurs comme des professeurs qu'ils doivent respecter sans enfreindre les codes de bonne conduite de leur culture. Par ailleurs, comme l'a souligné un examinateur, même dans la culture québécoise, le débat dans la société est une pratique relativement peu courante par rapport à la culture française. Bien que le Québec et la France partagent la même langue, le sens du débat n'est pas le même dans les deux cultures : il est beaucoup plus aiguisé chez les Français alors qu'il est peu valorisé chez les Québécois. La section B de l'épreuve orale du TEF est alors davantage imprégnée des pratiques culturelles françaises, et cela peut être perçu comme un biais culturel chez certains candidats ayant une bonne connaissance des pratiques de la société québécoise.

En somme, les types d'inférence émis par les examinateurs leur permettent d'interpréter les difficultés rencontrées par les candidats, or, comme les vraies raisons de ces difficultés sont totalement inconnues, elles représentent des visions très personnelles et sont inhérentes à la subjectivité des examinateurs. Comme ces inférences sont en réalité des commentaires périphériques sans réelle production de sens, elles doivent être évitées. Ce point devrait alors être évoqué lors des formations.

5.1.4. Les divergences dans la perception de l'attitude de l'animateur/animatrice

Dans les situations de test où l'examinateur n'interagit pas avec le candidat, il a été observé que les examinateurs tiennent compte du comportement de l'animateur (l'intervieweur) dans leur évaluation, notamment de leur empathie (Orr, 2002), de leur encouragement à l'égard des candidats (Pollitt et Murray, 1996), du degré de difficulté de leurs questions, du nombre de

questions fermées qu'ils posent, de leur attitude de dénigrement, du caractère inapproprié des sujets qu'ils abordent (sujets délicats, ennuyeux), de leurs interruptions du discours, du temps qu'ils accordent aux candidats pour répondre, et de leurs difficultés à comprendre les arguments des candidats (Brown, 2000).

Dans notre étude, l'attitude de l'animateur a largement été commentée avec les quatre candidats, mais selon des angles de vue différents.

Avec la première candidate, l'attitude globale de l'animatrice a été perçue différemment dans la section A et dans la section B de l'épreuve. Certains examinateurs pensent qu'elle a été négligente, qu'elle n'a pas bien donné suite à l'intervention de la candidate, qu'elle a mal guidé cette dernière, et que ses propos ont parfois été abrupts. Par conséquent, les examinateurs ont reconnu que le comportement de l'animatrice avait exercé une influence néfaste sur la performance de la candidate. En revanche, d'autres examinateurs pensent que l'animatrice a très bien agi, qu'elle a fait tout son possible pour s'adapter à la candidate sans jamais dominer la discussion. Elle a été perçue comme encourageante et empathique.

Avec la deuxième candidate, dans la section B de l'épreuve, l'attitude de l'animatrice a également été perçue différemment. Certains examinateurs pensent qu'elle n'a pas bien respecté les techniques d'animation, qu'elle a accaparé la conversation avec toute une série de contre-arguments, et que cela a eu un impact négatif sur le résultat de la candidate. À l'opposé, d'autres examinateurs ont trouvé la contribution de l'animatrice très bénéfique, car celle-ci a donné l'occasion à la candidate de mettre en avant toutes ses compétences.

Avec la troisième candidate, dans la section A de l'épreuve, certains examinateurs ont constaté que les rôles entre les deux interlocutrices s'étaient inversés dans la seconde partie de la conversation, c'est-à-dire que c'est l'animatrice qui posait les questions. Les examinateurs ont avoué que si cette faute n'avait pas eu lieu, ils auraient pu accorder une note plus élevée à la candidate. Par ailleurs, dans la section B, il a été dit que l'animatrice dominait trop la deuxième partie de la conversation, et que cela avait intimidé la candidate qui n'a pas pu correctement manifester sa capacité à débattre.

Avec le quatrième candidat, tous les examinateurs ont observé dans la section A que l'animateur s'exprimait beaucoup trop et que ses propos n'étaient parfois pas pertinents. De ce fait, le temps

consacré au candidat a été limité, et ce dernier n'a pas pu suffisamment développer son questionnement. De plus, dans la section B, beaucoup d'examinateurs ont remarqué que la conversation n'a pas pu se dérouler dans le bon ordre à cause de l'attitude peu collaborative de l'animateur qui apportait une multitude de contestations.

Dans l'ensemble, beaucoup d'examinateurs ont souvent mentionné que les façons de faire des animateurs avaient eu un impact négatif sur le résultat du candidat et que cela les avait gênés dans leur prise de décision. Le fait que l'attitude des animateurs affecte les performances des candidats fait écho avec les résultats des recherches empiriques sur les effets des examinateurs traitant de l'interaction entre les examinateurs et les candidats.

Dans ces études, il est mentionné que la manière dont les animateurs-examinateurs interagissent avec leurs candidats peut constituer une source de variabilité, et par conséquent, ces deniers sont traités de façon inégale, car ils risquent soit d'être avantagés soit d'être désavantagés (Brown et Hill, 1998; Lazaraton, 1996a). En effet, en fonction de la posture de l'animateur-examinateur, un même candidat peut produire des performances différentes et recevoir des notes différentes. Selon son propre style interactionnel, c'est-à-dire sa manière de structurer les séquences de conversations, de poser des questions ou de fournir des répliques, le comportement et le langage de l'animateur-examinateur peut affecter la performance des candidats lors des tests oraux de L2 (Cafarella, 1994; Filipi, 1994; Lazaraton, 1996a, 1996b; Lazaraton et Saville, 1994; Morton *et al.*, 1997). Les chercheurs ont alors identifié deux grands types d'animateurs : ceux dits « faciles » et ceux dits « difficiles » en ce qui a trait aux accommodements offerts aux candidats. Ceux interagissant avec des animateurs « faciles » ont en majorité des notes plus élevées que ceux interagissant avec des animateurs « difficiles » (Brown, 2003; Brown, 2005; Reed et Halleck, 1997). Il est également précisé dans les études que malgré le fait que les accommodements tendent vers un climat propice à une communication efficace, trop accommoder un candidat risque d'être contreproductifs, car cela ne lui donne pas l'opportunité d'exprimer son plein potentiel et son manque de compétence peut parfois être dissimulé (Lazaraton, 1996b; Ross et Berwick, 1992).

Ainsi, bien que les techniques d'animation soient déjà présentées dans le cadre des formations des examinateurs, les étudier de manière plus approfondie s'avère essentiel. Les modèles de comportements étant appropriés et ceux ne l'étant pas devraient être bien discernés. L'objectif principal étant de faire en sorte que tous les candidats soient traités de façon égale en termes de

soutien et de défis offerts par l'animateur, et non d'uniformiser le comportement de ce dernier à tel point que tous les candidats reçoivent des répliques similaires.

5.1.5. Les autres aspects de divergences

D'autres formes de divergences ont été observées dans notre étude. Par exemple, les examinateurs pouvaient souvent assister à la même scène et la décrire de façon très différente. Lors d'une évaluation de la section B de l'épreuve, nous avons constaté différentes interprétations d'une réplique d'une candidate. Par exemple, dans un contexte où il s'agit de s'informer sur un logement, le mot « *haut* » dans la question « *Quel haut est l'appartement ?* » a été interprété comme étant la hauteur d'un plafond selon un examinateur, et comme étant l'étage de l'appartement selon deux autres examinateurs. Ainsi, même dans de courtes séquences de dialogue contenant un vocabulaire et une syntaxe simples, les interprétations peuvent différer.

D'autres formes de divergence ont été relevées au travers des pratiques évaluatives comme l'attitude face à un trop grand écart entre la note de la section A et la note de la section B. Par exemple, selon un examinateur, les notes des deux sections doivent en principe être proches sur la grille d'évaluation, car d'après sa pratique sur le terrain, il remarque que lorsque la section A est beaucoup mieux réussie que la section B, souvent les candidats ont un manque de compétence en français, mais sont paradoxalement extrêmement bien préparés. Autrement dit, ils apprennent par cœur les formules que l'on peut utiliser en toutes circonstances grâce à des stratégies particulières afin de dissimuler leurs lacunes. À l'opposé, selon un autre examinateur, le trop grand décalage des notes entre les deux sections ne pose pas de problème, car la tâche de la section A est beaucoup plus sommaire que celle de la section B, et de ce fait, la section B est plus révélatrice du véritable niveau des candidats que la section A. En somme, les conduites au sujet des écarts de notes dans les deux sections ne semblent pas claires et soulèvent des interrogations.

Concernant les deux contextes d'évaluation, à savoir le contexte en mode présentiel synchrone et celui en mode à distance asynchrone, ils ont été perçus de différente façon. Le mode présentiel soumet les examinateurs à diverses contraintes temporelles comme le respect d'un horaire précis, la co-construction de l'interaction avec le candidat, la gestion du matériel d'enregistrement, le contrôle de la durée de l'interaction, puis la production du jugement en temps réel immédiatement à la fin de l'échange. De plus, une forme de lien socioaffectif éphémère se développe avec le candidat lors de son accueil dans la salle d'examen, car l'examinateur entre en communication

avec lui, le met à l'aise s'il est stressé, et adopte une posture familière lorsqu'il joue le rôle de son ami en utilisant le tutoiement dans une partie de l'épreuve. Dans le mode à distance asynchrone, le travail de l'examinateur est monotâche (évaluation du candidat), il est moins soumis à des contraintes temporelles, matérielles et n'est pas soumis à des contraintes humaines.

De ce fait, les deux contextes d'évaluation peuvent donner des perspectives différentes. La note de l'examinateur en mode présentiel peut être moins objective que celle de l'examinateur à distance, car il fait face à beaucoup plus de contraintes cognitives et gère beaucoup plus d'informations. Ces caractéristiques devraient être abordées lors des formations, car elles peuvent exercer une influence indirecte sur le jugement et jouer sur l'objectivité de la note.

En somme, ces multiples formes de divergences renvoient à l'idée que le jugement dépend du parcours de formation, de l'expérience professionnelle et personnelle, ainsi que de la personnalité de la personne. Le jugement est une représentation de celui qui le pose, car il reflète la vision du monde de ce dernier et son impression d'une réalité ou d'une situation. Il fait partie d'un processus complexe composé de plusieurs opérations diverses et réunies entre elles parmi lesquelles on retrouve des questionnements, des perceptions et des intuitions (Angers, 2010; Huver et Springer, 2011; Laming, 2004; Lipman, 1992; Lumley, 2002, 2005).

5.2. Le portrait de l'appropriation et de l'appréciation par les examinateurs de la grille d'évaluation de la compétence langagière

Parmi les études empiriques sur les effets des examinateurs, on a observé que les examinateurs avaient des sensibilités très variables par rapport aux différents critères de la grille d'évaluation et qu'ils accordaient une importance différente à chaque critère (Brown, 2000; Brown *et al.*, 2005; Hamilton *et al.*, 1993; McNamara, 1996; Meiron, 1998; Orr, 2002; Pollitt et Murray, 1996). Les examinateurs doivent juger la performance des candidats à travers leur « interprétation subjective des critères de notation » (Bachman, 1990, p. 76). Plus la norme et leur propre compréhension de la norme sont éloignées, et plus il y a de l'inconstance dans la façon dont les critères sont appliqués (Bachman, 1990; Meiron et Schick, 2000; Reed et Cohen, 2001). Très souvent, malgré le caractère explicite des descripteurs d'une grille d'évaluation et malgré la standardisation mise en place au cours des sessions de formation, les différences ne peuvent pas s'éliminer. Les examinateurs ont

une perception intime de ce qui leur est acceptable et ces perceptions sont formées, dans une certaine mesure, par leur expérience antérieure et/ou leurs attributs personnels (Brown, 1993; Bejar, 2012; Eckes, 2011; Han, 2016; Lumley et McNamara, 1995; McNamara, 1993, 1996; Taguchi, 2011; Wolfe et Chiu, 1997). Chaque examinateur a donc sa propre « grille d'évaluation mentale » (Bejar, 2012).

Nous avons retrouvé ces diverses caractéristiques dans notre étude. Nous avons mis en avant les différentes interprétations ainsi que les insuffisances perçues des éléments de la grille d'évaluation. Nous avons également soulevé les préoccupations des examinateurs, puis fait un constat de leur appréciation de la grille. Ces aspects se retrouvent au sein de différentes sous-parties telles que les descripteurs, les échelons, les regroupements de critères et de descripteurs, l'actuelle version de la grille d'évaluation, ainsi que les normes de la grille d'évaluation.

5.2.1. Les descripteurs

Nous avons observé que les examinateurs avaient des compréhensions variables de mêmes termes au sein des descripteurs comme « Absence d'observables », « Phrases simples », « Phrases complexes », « Répertoire lexical vaste », « Répertoire lexical riche », etc. Beaucoup de termes ont été interprétés différemment ou jugés ambigus ou vagues, ce qui a créé des désaccords et des confusions parmi les examinateurs. Les descripteurs devraient donc contenir davantage d'informations précises, concrètes et univoques.

Dans la même lignée, certains examinateurs ont mis en avant le fait que les différents descripteurs d'un même critère s'enchainaient parfois de façon trop brute et non fluide, surtout pour les niveaux grands débutants (entre le niveau <A1 et le niveau A1). À plusieurs reprises lors de l'exercice d'évaluation, les examinateurs ont avoué qu'ils auraient situé le juste niveau d'un candidat dans une zone fictive entre deux échelons. Comme il est parfois difficile de bien saisir la variation graduelle au fil des différents niveaux, les examinateurs ont parfois de la difficulté à observer des distinctions claires. Beaucoup de termes dans les descripteurs adjacents ont été jugés trop similaires dans leur formulation, par exemple pour les niveaux B2 et C1 du critère 2 qui indiquent respectivement : « Présentation claire et détaillée. Arguments illustrés ou détaillés. » et « Présentation développée et structurée. Arguments variés et bien développés. ».

En outre, il a été mentionné que certains descripteurs de la grille regroupaient parfois plusieurs actions assez distinctes, et comme toutes les actions d'un candidat peuvent parfois ne pas s'accomplir de façon harmonieuse, cela rend le jugement difficile. Par exemple, pour le descripteur du niveau C1 du critère 1 qui indique : « Questionnement complet et précis. Intervient avec justesse. La conversation est soutenue. », il peut arriver quelquefois chez un candidat que son questionnement soit complet et précis et que son intervention soit soutenue, mais que la conversation ne soit pas soutenue. Les examinateurs doivent alors faire fi de certains énoncés dans un même descripteur, mais ils ne savent pas si cela est permis. La trop grande quantité d'énoncés indépendants au sein d'un même descripteur peut entraver le jugement donc la notation.

Lors de l'exercice de verbalisation, pour beaucoup d'examinateurs, les justifications de leurs choix de notes ne concordaient pas toujours avec les descripteurs de la grille. Certains n'en étaient pas conscients, tandis que d'autres l'étaient et l'ont signalé dans leurs propos. Ceux l'ayant signalé ont admis qu'ils avaient été amenés à faire des choix par défaut, car la grille ne fournit pas assez d'options. Par conséquent, lorsque les examinateurs estiment que certaines parties de la grille d'évaluation ne sont pas suffisantes, ils admettent qu'ils doivent tirer parti de leur propre expérience en tant qu'enseignants de FLS ainsi que de leurs propres connaissances des normes du CECRL afin d'émettre leur jugement.

Dans le but de rendre la lecture de la grille d'évaluation plus objective, il est primordial que celle-ci offre suffisamment d'indications et que ses termes soient univoques. Toutefois, il y a un équilibre à trouver entre une surdescription et une concision des éléments de la grille. Celle-ci doit rester facile à utiliser et pas trop spécifique afin qu'elle puisse être largement utilisable.

5.2.2. Les échelons

Concernant les échelons, ceux qui posent le plus de problèmes sont majoritairement les échelons B1 et B2. La principale justification est le fait que le niveau B2 soit le niveau à partir duquel des points sont attribués aux candidats ayant un projet d'émigration vers la province du Québec. À ce propos, l'obtention du niveau général B2 au test a été qualifiée de « réussite » par certains. Lorsque la performance d'un candidat oscille entre les niveaux B1 et B2, cela génère chez les examinateurs davantage de doute et d'hésitation, et par conséquent, le temps de réflexion et de concentration est plus important. Les niveaux B1 et B2, et plus particulièrement les sous-niveaux B1-2 et B2-1, ont été qualifiés de zone très sensible, car chaque point attribué autour de cette aire a un impact

considérable sur le cheminement de la vie des candidats. Les propos d'un examinateur dans une étude d'Inoue *et al.* (2021, p. 51), sur les perceptions des examinateurs à propos du test d'anglais IELTS, synthétise l'ampleur des conséquences d'un test à enjeu extrêmement élevé : « *Si on réfléchit du point de vue du candidat, la différence entre un 6,5 et un 7 peut faire la différence entre aller à l'université ou non, ou pouvoir émigrer ou non. C'est vraiment un changement de vie pour eux. Mais pour nous, c'est une différence de 0.5 sur une échelle* ». Étant donné l'enjeu très élevé que représente le TEF, il est alors important d'insister sur les niveaux B1 et B2 lors des formations dans les centres agréés du Québec.

Il a également été relevé que les échelons A1 et C2 devaient contenir chacun deux sous échelons, à l'instar des autres échelons de la grille d'évaluation, car les candidats de niveau débutant (A1) et ceux de niveau avancé (C2) peuvent très bien se situer dans un sous niveau étant plus « faible » ou plus « fort ». En outre, l'ajout d'un troisième sous échelon à tous les échelons de la grille a été suggéré, car cela permettrait d'obtenir une option plus neutre, et ainsi de mieux nuancer le jugement des examinateurs.

Un autre point suggéré est la volonté d'obtenir des éclaircissements sur la pondération des points attribués pour chaque échelon et sous échelon. Lors des formations, comme aucune indication n'est donnée à ce sujet, certains examinateurs pensent qu'il est important pour eux, par souci de transparence, d'avoir une connaissance de base du système de calculs des points. Les examinateurs se sont également questionnés sur le poids des deux catégories des critères (communicatifs et linguistiques) de la grille d'évaluation et se demandent si le volet linguistique domine par rapport au volet communicatif, étant donné qu'il y a trois critères linguistiques et deux critères communicatifs.

5.2.3. Les critères

Presque tous les examinateurs ont, à plusieurs reprises, fait part de leurs mécontentements concernant le critère 2 intitulé « Section B, Capacité à présenter et débattre ». Ces deux capacités (présenter et débattre) correspondent à deux actions bien distinctes, et d'ailleurs, sur le terrain, les candidats n'accomplissent pas toujours conjointement ces deux actions de façon égale. Ils peuvent faire des présentations très sommaires du document de support, mais débattre avec des arguments variés et bien détaillés, et inversement, c'est-à-dire présenter le document de support de façon structurée et très bien développée, mais apporter des arguments simples et peu illustrés. Nous ne

savons donc pas si ceux deux actions doivent être attestées ensemble ou si une certaine forme de compensation est permise. De plus, durant l'exercice de verbalisation, nous avons observé que la présentation du document de support avait très peu été commentée, comparativement à la capacité à débattre, cela pourrait supposer que le débat aurait plus d'importance. Il est donc essentiel de clarifier le critère 2, car cette dualité (présenter et débattre) paralyse très souvent la prise de décision des examinateurs.

5.2.4. L'actuelle version de la grille d'évaluation

Concernant l'actuelle version de la grille d'évaluation (datant d'octobre 2018), il a été relevé que la réduction du nombre de critères apportait moins de nuance et de flexibilité, et que cela amenait à faire des choix plus tranchés et moins certains, surtout pour les critères 1 et 2 (Critère 1 : capacité à obtenir des informations dans la section A; Critère 2 : capacité à présenter et débattre dans la section B). Cependant, le contenu plus allégé de l'actuelle grille en termes de nombre de critères (passant de douze à cinq critères) facilite beaucoup la tâche de tous les examinateurs, car celle-ci est plus rapide à utiliser. Grâce au gain de temps procuré, ils peuvent davantage se consacrer à l'interaction avec les candidats et mieux apprécier leur performance.

D'autre part, l'actuelle grille est mieux adaptée à la nouvelle réforme du fonctionnement de l'évaluation (datant également d'octobre 2018) où les examinateurs évaluent les candidats de façon différée. Certains préfèrent la nouvelle réforme où ils évaluent seuls les candidats, car ils se sentent plus autonomes et peuvent mieux gérer leur temps, tandis que d'autres préféraient l'ancienne méthode où ils étaient pairés pour émettre leur jugement, car ils pouvaient obtenir des rétroactions en cas d'hésitations et mettre en parallèle les différents points de vue. Ceux qui préfèrent évaluer de façon isolée mentionnent qu'avec l'ancienne méthode de travail et l'ancienne grille d'évaluation qui était plus consistante et plus analytique, ils passaient beaucoup de temps à justifier leurs opinions et à trouver un consensus avec leur coéquipier. Par conséquent, cela leur faisait ressentir une lourdeur, et lorsque les avis entre pairs étaient trop divergents, cela engendrait même des conflits latents.

Il est vrai que l'actuelle version de la grille facilite amplement le travail des examinateurs, mais sa visée plus holistique apporte moins de précisions et plus de doutes. Parallèlement, la nouvelle réforme du fonctionnement de l'évaluation ne permet plus de négocier à deux pour arriver à un consensus sur le résultat final du candidat. Cette nouvelle procédure a progressivement conduit à

un amenuisement de l'expertise du raisonnement évaluatif et a en quelque sorte confiné les examinateurs qui évaluent désormais de façon totalement isolée. Les formations des examinateurs devraient fréquemment proposer des confrontations de jugements afin d'optimiser l'uniformisation des pratiques. Chacun expliciterait à haute voix sa démarche en étayant ses arguments. Cela permettrait de mieux développer des compétences de raisonnement, et éviterait d'individualiser les points de vue.

5.2.5. Les normes

Enfin, sous une perspective plus large, on observe par ailleurs que la grille peut « enfermer » les candidats dans un cadre. Par exemple, il a été mentionné qu'une candidate à qui l'on a majoritairement accordé le niveau A1 semblait malgré tout avoir l'habitude de s'exprimer oralement en français et qu'elle aurait été capable de se débrouiller seule dans une situation ordinaire de la vie quotidienne. Cette contradiction s'explique par le fait que le profil langagier d'un candidat peut parfois ne pas adhérer au modèle unique de référence de la grille d'évaluation qui prescrit un ensemble de normes de la communication orale. Comme l'a très bien soulevé un examinateur, les modèles que l'on applique pour représenter l'expérience de la communication ne peuvent pas rendre compte de la propre expérience et des propres besoins de chacun des candidats ni de leur difficulté.

Il a également été relevé que la norme syntaxique et lexicale du niveau B2 convenait mal à des candidats se destinant à exercer une activité professionnelle dans un secteur plutôt technique au Québec (comme le bâtiment ou la restauration), car les standards linguistiques sont d'une autre réalité. En outre, un examinateur a admis qu'il faisait parfois abstraction de certaines erreurs linguistiques mineures et occasionnelles (de syntaxe et de phonologie) chez les candidats étant donné que ces erreurs sont le reflet du français oral propre des francophones de la province du Québec.

Ainsi, bien que la grille d'évaluation ait été mise à jour dernièrement, elle revêt toujours des interprétations personnelles et contient des zones de flou. Malgré les bonnes connaissances des examinateurs des normes du CECRL, nous avons observé que les éléments de la grille d'évaluation ne sont pas inscrits de façon intrinsèque chez eux. Les quelques constats que nous avons mis en évidence dévoilent des représentations divergentes et des contrariétés qui provoquent le doute chez les examinateurs et qui les amènent à faire les mauvais choix. Afin de créer une vision commune

dans l'interprétation de la grille et de la rendre plus opérationnelle, sa future mise à jour devrait se faire à l'aune de nos résultats. À travers nos suggestions dans la section suivante, nous évoquerons ce point. Au préalable, les points les plus essentiels de notre recherche seront synthétisés.

5.3. La synthèse de la discussion

Les points les plus essentiels de la discussion sont relevés dans ce tableau de synthèse (Tableau 47).

Tableau 47 - Tableau de synthèse de la discussion

Divergences chez les examinateurs du TEF
- Une large dispersion des notes et des commentaires peut être observée pour un même candidat à un même critère.
- Une même note peut être accordée à une même performance alors que les interprétations peuvent différer, et inversement.
- La familiarité avec l'accent du candidat facilite la compréhension de ce dernier. Cette familiarité peut entraîner une répercussion positive sur le score de la prononciation.
- Des inférences non pertinentes peuvent être faites pour expliquer certaines difficultés de candidats.
- Un même animateur peut être perçu différemment (encourageant et envahissant). L'attitude de l'animateur peut affecter la performance du candidat.
Portrait de l'appropriation et de l'appréciation de la grille d'évaluation
Certains descripteurs : - sont imprécis, abstraits ou ambivalents; - s'enchainaient de façon trop brute ou sont similaires au fil des différents échelons; - regroupent plusieurs actions indépendantes.
Les échelons B1 et B2 sont les plus difficiles à discriminer.
Le critère de la section B pose un problème à cause de la présence de deux actions distinctes (présenter et débattre).
L'actuelle version de la grille d'évaluation est plus facile et rapide à utiliser. Cependant, elle ne permet pas de bien nuancer le jugement.

La grille d'évaluation n'est parfois pas suffisante pour prendre en compte certains aspects de la performance d'un candidat.

5.4. Les suggestions

Comme nous l'avons constaté, de mêmes réalités peuvent aboutir à de multiples perceptions, la lecture de la grille d'évaluation peut varier d'un examinateur à l'autre, et cela conduit à des décisions différentes aux conséquences importantes. Il est donc primordial que les formations initiales et continues se fondent sur une démarche plus rigoureuse qui permette de minimiser le plus possible la subjectivité et qui atténue les différents systèmes de valeurs d'évaluation des examinateurs. Pour cela, nous suggérons de prendre en compte dans les formations l'ensemble des divers regards évaluatifs que l'on a pu identifier dans notre analyse. Nous préconisons également de porter à la connaissance les différentes études empiriques menées sur les effets des examinateurs afin de développer une meilleure prise de conscience des multiples variables pouvant représenter une menace possible à la fidélité.

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 se réaliser 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, des améliorations de l'épreuve d'expression orale pourraient être réalisées sous une forme plus sociale, c'est-à-dire en donnant la parole aux examinateurs. Étant donné que les voix des examinateurs sont considérées comme des ressources importantes qui favorisent la validité des tests d'expression orale (Ducasse et Brown, 2009; Galaczi *et al.*, 2012; Nakatsuhara *et al.*, 2017), nous pensons qu'il serait pertinent que des informations soient recueillies grâce à des sondages à grande échelle. Cela permettrait d'accorder aux examinateurs une liberté dans l'expression de leur point de vue, et éviterait qu'il n'y ait que la seule perspective d'une norme extérieure imposée dans une pratique purement technique.

Plus précisément, nous suggérons une étude à la manière de celle d'Inoue *et al.* (2021) qui a examiné les points de vue des examinateurs sur tous les aspects de l'épreuve d'expression orale du test IELTS, comme les tâches du test, les sujets, le format, les pistes de l'interlocuteur, les directives de l'examinateur, l'administration, l'évaluation, la formation et la standardisation, ainsi que l'utilisation du test. Cette étude réalisée à l'aide d'une méthode mixte a recueilli les avis de 1203 examinateurs grâce un questionnaire en ligne. De façon complémentaire, 36 examinateurs, représentatifs d'un certain nombre de régions géographiques différentes, ont été sélectionnés pour réaliser des entretiens afin d'explorer en profondeur les justifications de leurs opinions.

Dans cette étude, quelques résultats sont similaires aux nôtres concernant l'appréciation de la grille d'évaluation. Par exemple, il a été mentionné par les examinateurs que quelques énoncés étaient trop vagues dans les descripteurs, et qu'il était nécessaire d'introduire à certains endroits plus de précisions afin de mieux distinguer les niveaux. Un autre problème soulevé concerne un critère nommé « Aisance et cohérence », car celui-ci fait l'amalgame entre deux actions distinctes : l'aisance et la cohérence. Les examinateurs observent sur le terrain qu'un candidat peut avoir un rythme de parole discontinu (avoir une aisance « faible ») et qu'il peut en même temps bien structurer son discours en utilisant fréquemment des marqueurs de discours (avoir une cohérence « forte »). Somme toute, les résultats de cette étude permettent de fournir des suggestions de changements potentiels de l'épreuve et visent à améliorer sa validité.

Dans ce chapitre, nous avons apporté des éléments de réponse aux deux questions de recherche posées dans cette thèse, à savoir : (1) Quelles divergences pouvons-nous observer chez les examinateurs en considérant différents aspects du jugement évaluatif ? et (2) Quel portrait peut-on dresser de leur appropriation et de leur appréciation de la grille d'évaluation de la compétence langagière ? D'une part, ces réponses ont été comparées aux résultats des études du domaine identifiés dans la problématique et dans le cadre conceptuel de ce travail, et d'autre part, quelques nouveaux éléments ont été mis en évidence. Finalement, des suggestions visant à améliorer la validité de l'épreuve d'expression orale du TEF ont été apportées.

CONCLUSION

Cette recherche visait à comprendre le jugement des examinateurs dans le cadre de l'épreuve d'expression orale du TEF. Ce test aux enjeux très élevés permet d'attester les connaissances linguistiques en français des personnes désirant étudier ou résider de façon permanente au Canada. L'épreuve d'expression orale du TEF est constituée de jeux de rôle entre le candidat et l'examinateur, et la nature imprévisible et créative de cette structure peut poser un problème. En effet, il a été démontré que les comportements d'examinateurs humains pouvaient représenter une menace possible à la fidélité de l'épreuve malgré les nombreuses mesures prises afin de minimiser les variabilités de l'interprétation et du jugement comme l'utilisation de grilles d'évaluation critériées et l'offre de formations. Les caractéristiques des examinateurs représentent une source importante de biais et se définissent comme étant des « effets ». À l'aide d'une analyse minutieuse d'entretiens avec les examinateurs, nous avons alors tenté d'observer si ces « effets » étaient présents au sein de leurs activités d'évaluation dans le cadre de l'épreuve d'expression orale du TEF. Notre recherche a retenu deux objectifs. Le premier objectif était de découvrir l'existence de divergences chez les examinateurs, et le deuxième objectif était de brosser le portrait de l'appropriation et de l'appréciation des examinateurs de la grille d'évaluation. Compte tenu des résultats obtenus, nous avons formulé des suggestions.

En ce qui concerne le premier objectif de recherche, plusieurs divergences ont été constatées principalement dans le raisonnement évaluatif et dans l'attribution de la note, la familiarité avec l'accent des candidats, les inférences et la perception de l'attitude de l'animateur.

À propos du raisonnement évaluatif et l'attribution de la note, l'analyse des données a permis d'observer que des notes analogues pouvaient masquer des jugements divergents, et inversement, des jugements identiques (ou très proches) pouvaient reposer sur des notes différentes. De plus, des écarts dans l'attribution de notes et dans les commentaires justificatifs étaient fréquents entre les examinateurs.

Concernant la familiarité avec l'accent des candidats, certains examinateurs ont pris conscience qu'ils avaient eu une perception différente des traits phoniques de certains candidats en raison de leur familiarité avec quelques caractéristiques. Cette prise de conscience a eu des répercussions positives sur la perception de l'aisance à l'oral et de l'élocution des candidats, toutefois, la note

pour le critère correspondant n'a pas toujours été rehaussée en raison d'une volonté de l'examinateur de rester objectif.

Au sujet des inférences, elles ont été nombreuses et variées. Afin d'expliquer certaines difficultés des candidats, des raisons comme le faible niveau de scolarisation, l'âge, le manque d'encadrement et surtout les différences culturelles ont été évoquées. À ce propos, il a souvent été dit que le débat dans une des sections de l'épreuve du test posait un problème de biais culturel pour certains candidats originaires d'Extrême-Orient.

Quant à la perception de l'attitude de l'animateur, elle a beaucoup été commentée selon des angles de vue différents. Un même animateur pouvait être perçu de façon contradictoire, c'est-à-dire comme étant encourageant ou comme étant non collaboratif. De plus, le comportement de ce dernier a parfois gêné la prise de décision des examinateurs, et cela a eu une influence néfaste sur la note de la performance du candidat.

D'autres aspects divergents ont été observés. Par exemple, de trop grands écarts de notes entre la section A et la section B sur la grille d'évaluation ont soulevé des interrogations et ont conduit à des attitudes opposées. Enfin, il a été mentionné que les deux contextes d'évaluation (en présentiel et à distance) pouvaient donner des perspectives différentes, c'est-à-dire que la note de l'examinateur en mode présentiel pouvait être moins objective que celle de l'examinateur à distance en raison des contraintes cognitives et de la gestion d'une multitude d'informations.

Le deuxième objectif de recherche était de brosser le portrait de l'appropriation et de l'appréciation des examinateurs de la grille d'évaluation. Nous avons répondu à cette question à travers différentes sous-parties telles que les descripteurs, les échelons, les critères, l'actuelle version de la grille d'évaluation, ainsi que les normes de celle-ci.

Concernant les descripteurs, ils ont été interprétés différemment et ont été jugés comme étant ambigus ou vagues à plusieurs reprises. Il a aussi été relevé que quelques descripteurs adjacents ne s'enchainaient pas de façon graduelle, et notamment parce qu'ils étaient trop similaires dans leur formulation. De plus, on a observé que les justifications des choix de notes des examinateurs ne concordaient pas toujours avec les descripteurs de la grille. Certains examinateurs en étaient conscients et ont justifié cela par une insuffisance d'informations de la grille. De plus, il a été

mentionné qu'un même descripteur regroupait parfois plusieurs actions assez distinctes qui ne pouvaient pas toujours s'accomplir de façon harmonieuse.

Au sujet des échelons, ceux posant le plus de problèmes sont majoritairement les échelons B1 et B2 en raison de l'obtention de points pour les candidats ayant un projet d'émigration vers la province du Québec. Lorsque le niveau d'un candidat oscille entre ces deux échelons, cela génère chez les examinateurs davantage d'hésitation, et exige plus de concentration. En outre, quelques suggestions ont été apportées par les examinateurs, comme l'ajout de deux sous échelons pour les échelons A1 et C2, à l'instar des autres échelons, puis le partage de connaissance de la pondération des points attribués pour chaque échelon et sous échelon.

En ce qui concerne les critères, beaucoup d'examinateurs ont fait part de leurs mécontentements à l'égard du critère 2 qui contient deux actions très distinctes : présenter et débattre. Comme ces deux actions ne s'accomplissent pas toujours conjointement de façon égale, cela leur pose un problème dans la prise de décision.

À propos de l'actuelle version de la grille d'évaluation, les examinateurs la trouvent plus rapide à utiliser, mais les amène à faire des choix plus tranchés étant donné qu'elle apporte moins de nuance, surtout pour les critères 1 et 2. Il a également été soulevé que l'actuelle grille était mieux adaptée à la nouvelle réforme du fonctionnement de l'évaluation où les examinateurs ne font plus d'acccord inter-juge. À ce sujet, les points de vue étaient mitigés. Certains trouvaient que ce procédé prenait trop de temps et qu'il était difficile de trouver un consensus avec le coéquipier. Pour d'autres, cela leur permettait d'avoir des rétroactions en cas d'hésitations et de mettre en parallèle les avis.

Enfin, sous un angle plus global, il a été mentionné que les normes de la grille d'évaluation pouvaient « enfermer » les candidats dans un cadre. En effet, le profil langagier d'un candidat peut parfois ne pas adhérer à un modèle unique prescrivant les codes de la communication orale, et de plus, certains secteurs professionnels ont des standards linguistiques distincts.

Les limites de la recherche

La portée d'une recherche se mesure par ses limites (Van der Maren, 1996), et notre recherche en contient quelques-unes. Afin d'obtenir des données pour notre objet de recherche, nous avons eu recours à une méthodologie de type qualitatif. L'activité de verbalisation ainsi que l' entrevue

semi-dirigée nous a permis de recueillir beaucoup d'informations. Cependant, les données obtenues par ces moyens doivent être interprétées avec grande précaution étant donné qu'elles ne proviennent que de dix examinateurs. Des résultats plus représentatifs et, par conséquent, plus généralisables auraient pu être obtenus avec un échantillon plus vaste. Il faut donc relativiser la portée de ces recherches, car il n'est pas possible de généraliser les résultats à d'autres examinateurs et d'autres contextes. Toutefois, nous disposions de quarante évaluations différentes de candidats dans l'ensemble, et nous avons observé chacun des dix examinateurs pendant une période de presque deux heures, ce qui a donné un total d'environ vingt heures (incluant les moments d'écoute des performances des candidats). Cela permet alors accroître la validité des données.

Par ailleurs, notre recherche a été présentée au sein d'un contexte québécois et les participants étaient tous basés dans la région de Montréal, ce qui limite la représentativité des résultats obtenus à échelle mondiale. À titre de rappel, le TEF et ses déclinaisons sont diffusés grâce à un réseau de 308 centres d'examen agréés dans le monde, nous aurions donc pu orienter le contexte de notre recherche vers un axe plus international en recrutant des participants provenant de diverses régions dans le monde.

D'autre part, dans le cadre de l'élaboration de notre première question de recherche au sujet des divergences, nous n'avons pas retenu la variable du statut de la langue des examinateurs, c'est-à-dire le fait d'être locuteurs natifs et locuteurs non natifs de la langue évaluée. Cette variable a été prise en compte dans les études empiriques sur les effets des examinateurs, mais nous n'avons pas pu la prendre en compte étant donné qu'une grande majorité de nos participants avaient le français comme langue première. Nous aurions pu recruter un nombre égal d'examinateurs locuteurs natifs et d'examinateurs locuteurs non natifs et caractériser le statut de leur langue. La prise en compte de cette variable aurait pu faire émerger plus de données et enrichir nos résultats. Toutefois, si nous l'avions fait, nous aurions pris des précautions dans l'analyse de nos interprétations, car d'après nos recensions, les définitions proposées sur l'identité d'un locuteur natif sont problématiques. Des facteurs psychologiques, sociologiques, culturels et politiques entrent en jeu, et il est également difficile de déterminer un certain niveau de compétence linguistique pour qu'un locuteur non natif soit considéré comme son homologue natif.

Les retombées de la recherche

Comme nous l'avons mentionné dans le premier chapitre, les études recensées sur l'évaluation de l'épreuve d'expression orale du TEF sont laconiques et de type qualitatif, cette recherche apporte donc une contribution certaine à ce domaine. D'une part, elle apporte un « penchant » épistémologique différent, car elle porte un regard plus « psychologique » en tentant d'accéder au fonctionnement cognitif des examinateurs. D'autre part, elle enrichit la connaissance et la compréhension des gestes professionnels évaluatifs des examinateurs.

Les résultats de cette recherche peuvent être utilisés par les parties prenantes concernées comme l'équipe du Français des affaires de la Chambre de commerce et d'industrie de Paris Île-de-France et les différents centres d'examen agréés TEF dans le monde afin d'amener les examinateurs à être plus efficaces et plus constants dans leur pratique, et ainsi d'améliorer la standardisation des procédures. L'explicitation critique et raisonnée de la subjectivité des examinateurs constitue un bon moyen pour aller vers davantage d'objectivité. Les aspects sensibles et variables du comportement des examinateurs que l'on a identifiés dans nos résultats peuvent fournir de nouveaux éléments pouvant contribuer aux formations initiales et continues.

De plus, comme nous avons documenté les perceptions et les rétroactions des examinateurs sur leur utilisation concrète des grilles, les résultats obtenus peuvent également apporter de précieuses indications pour une future mise à jour de l'outil d'évaluation.

Finalement, par la même occasion, une meilleure standardisation participe au respect du principe d'égalité pour les usagers du service offert. Tous les candidats doivent obtenir les mêmes opportunités dans la possibilité de démontrer leurs habiletés, et l'assignation d'un examinateur-animateur ne doit pas être une question de chance. L'épreuve d'expression orale du TEF doit donner des résultats qui reflètent les habiletés des candidats et non leur environnement.

Les futures recherches

Les forces et les faiblesses de cette recherche pourraient servir de jalons à de futurs projets. Si nous avions la possibilité de recommencer cette recherche, la démarche serait différente et nous utiliserions une méthode mixte. Nous referions le même exercice de verbalisation avec le même nombre de candidats, mais avec plus d'examinateurs, soit vingt au lieu de dix. De plus, nous élaborerions un sondage à grande échelle en ligne. Les questions du sondage permettraient de

recueillir les points de vue des examinateurs sur les aspects de l'épreuve d'expression orale du test TEF comme la grille d'évaluation et les points problématiques de l'évaluation. Le sondage serait constitué de questions fermées, dont des questions à choix multiples, ainsi que de questions ouvertes afin que les examinateurs justifient leurs opinions. Pour l'exercice de verbalisation, le recrutement des participants se ferait à échelle internationale, c'est-à-dire que nous recruterions des candidats issus de divers centres agréés TEF dans le monde. Les examinateurs seraient représentatifs d'un certain nombre de régions géographiques différentes. Les critères d'inclusion des participants seraient les mêmes, mais comprendraient la langue première des participants, car cette variable n'a été prise en compte dans notre première question de recherche portant sur les divergences des examinateurs. À titre de rappel, les critères d'inclusion de la présente recherche étaient les suivants : être examinateur TEF certifié depuis au moins trois ans et évaluer annuellement en moyenne au moins cent candidats de l'épreuve d'expression orale dans les centres d'examen agréés. Cet axe de recherche mériterait alors d'être exploré.

Bibliographie

Adams, M. L. (1980). Five co-occurring factors in speaking proficiency. Dans J. R. Frith (dir.), *Measuring spoken language proficiency* (p. 1-6). Georgetown University Press.

Agard, E. et Dunkel, H. (1948). *An investigation of second language teaching*. Ginn and Company.

Alderson, J. C. (dir.) (2002). *Common European Framework of Reference for Languages : Learning, Teaching, Assessment: Case Studies*. Council of Europe.

Alderson, J. C. (2005). Editorial. *Language Testing, 22*(3), 257-260. https://doi.org/10.1191/0265532205lt315ed

Alderson, J. C. et Bachman, L. F. (2004). Series editors' preface to assessing speaking. Dans J. Alderson et L. Bachman (dir.), *Assessing Speaking* (p. ix-xi). Cambridge University Press.

Alderson, J. C., Clapham, C. et Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge University Press.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S. et Tardieu, C. (2004, février). *Specifications for item development and classification within the CEF : Reading and listening (English, French and German): The Dutch CEF Construct Project* [communication orale]. Workshop on research into and with the CEFR, University of Amsterdam.

Alderson, J. C. et Wall, D. (1993). Does washback exist ? *Applied linguistics, 14*(2), 115-129. https://doi.org/10.1093/applin/14.2.115

Alvarez, G. (1981). Niveau-seuil et enseignement fonctionnel du français. *Québec français*, *42*, 33-35. https://id.erudit.org/iderudit/57148ac

American Council on the Teaching of Foreign Languages (1986). *ACTFL proficiency guidelines*. American Council on the Teaching of Foreign Languages.

American Council on the Teaching of Foreign Languages (2012). A*CTFL proficiency guidelines.* American Council on the Teaching of Foreign Languages. https://www.actfl.org/resources/actfl-proficiency-guidelines-2012

American Educational Research Association, American Psychological Association et National Council on Measurement in Education (1999). *Standards for educational and psychological testing.*

American Educational Research Association, American Psychological Association et National Council on Measurement in Education (2014). *Standards for educational and psychological testing.*

Anastasi, A. (1988). *Psychological testing* (6e éd.). Macmillan.

Anastasi, A. (1994). *Introduction à la psychométrie*. Guérin Universitaire.

Ang-Aw, H. T. et 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/10.1177/0033688210390226

Angers, P. (2010). La formation du jugement. Dans M. Schleifer (dir.), *La formation du jugement* (3e éd., p. 101-120). Presses de l'Université du Québec.

Angiolillo, E. (1947). *Armed forces foreign language teaching*. Vanni.

Arens, S. A. (2006). L'étude du raisonnement dans les pratiques évaluatives. *Mesure et évaluation en éducation, 29*(3), 45-56. https://doi.org/10.7202/1086393ar

Arter, J. A. (2010). *Scoring rubrics*. Educational Testing Service.

Arter, J. A. et Chappuis, J. (2010). *Creating and recognizing quality rubrics.* Educational Testing Service.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing, 33*(1), 99-115. https://doi.org/10.1177/0265532215582283

Aubert-Gea, C. (2005). *Quelle formation pour enseigner l'oral ?* L'Harmattan.

Bachman, L. F. (1988). Problems in examining the validity of the ACTFL oral proficiency interview. *Studies in Second Language Acquisition, 10*(2), 149-164. https://www.jstor.org/stable/44488170

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

Bachman, L. F. (2000). Modern language testing at turn of the century, assuring that what we count counts. *Language Testing, 17*(1), 1-42. https://doi.org/10.1177/026553220001700101

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly, 2*(1), 1-34. https://doi.org/10.1207/s15434311laq0201_1

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

Bachman, L. F. et Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. 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. et Anderson, M. C. (2009). *Memory*. Psychological Press. https://doi.org/10.4324/9781315749860

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86-107. https://doi.org/10.1016/j.asw.2007.07.001

Barkaoui, K. (2010). Variability in ESL essay rating processes : The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*, 54-74. https://doi.org/10.1080/15434300903464418

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating : An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51-75. https://doi.org/10.1177/0265532210376379

Barkaoui, K. (2015). Test Takers' Writing Activities During the TOEFL iBT® Writing Tasks : A Stimulated Recall Study. Dans *TOEFL iBT Research Report No. 25 and ETS Research Report Series No. RR-15-04.* Educational Testing Service. https://doi.org/10.1002/ets2.12050.

Barnwell, D. (1996). *A history of foreign language testing in the United States*. Bilingual Press.

Barrat, J. et Moisei, C. (2004). *Géopolitique de la francophonie, Un nouveau souffle ?* La documentation française. Les études de la documentation française.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2,* 49-58.

Beacco, J. C. (2002). Sur les fonctions de la notion de compétence en langue en didactique des langues. Dans V. Castellotti et B. Py (dir.), *La Notion de compétence en langue. Notions en question* (no 6, p.105-113). ENS Éditions.

Beacco, J. C. (2004). *Niveau B2 pour le français, un référentiel*. Conseil de l'Europe. Didier.

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

Bélair, L. M. (2007). Défis et obstacles dans l'évaluation des compétences Professionnelles. Dans L. M. Bélair, D. Laveault et C. Lebel (dir.), *Les compétences professionnelles en enseignement et en évaluation* (p. 181-191). PUO.

Bendig, A. W. (1953). The reliability of self-ratings as a function of the amount of verbal anchoring and of the number of categories on the scale. *Journal of Applied Psychology, 37,* 38-41. https://doi.org/10.1037/h0057911

Bernaud, J. L. (2007). *Introduction à la psychométrie*. Dunod.

Berne, J. (2004). Think-Aloud Protocol and Adult Learners. Adult Basic Education: *An Interdisciplinary Journal for Adult Literacy Educational Planning,14*(3), 153-173.

Berwick, R. et Ross, S. (1996). Cross-cultural pragmatics in oral proficiency interview strategies. Dans M. Milanovic et N. Saville (dir.), *Performance testing, cognition and assessment: selected papers from the15th Language Testing Research Colloquium* (p. 34-54). Cambridge University Press.

Blais, A. (1992). Le sondage. Dans B. Gauthier (dir.), *Recherche sociale : De la problématique à la collecte des données* (p. 261-398). Presses de l'Université du Québec.

Bloom, B. S. (1953). Thoughts processes in lectures and seminars. *Journal of General Education, 7*, 160-169.

Bloom, B. S. (1954). The thought processes of students in discussion. Dans S. J. French (dir.), *Accent on teaching : Experiments in general education* (p. 23-46). Harper.

Bøhn, H. (2015). Assessing Spoken EFL Without a Common Rating Scale: Norwegian EFL Teachers' Conceptions of Construct. *SAGE Open.* https://doi.org/10.1177/2158244015621956

Bolton, S. (1987). *Évaluation de la compétence communicative en langue étrangère*. Hatier-Credif.

Bonk, W. J. et Ockey, G. (2003). A many-facet Rash analysis of second language group oral discussion talk. *Language Testing, 20*(1), 89-110. https://doi.org/10.1191/0265532203lt245oa

Boyer, J. Y. (1997). La verbalisation comme voie d'accès à la compréhension et à la production de texte. Dans J. Y. Boyer et L. Savoie-Zajc (dir.), *Didactique du français, méthodes de recherche* (p. 203-218). Éditions Logiques.

Boyer, H. (2003). *De l'autre côté du discours : Recherches sur le fonctionnement des représentations communautaires.* L'Harmattan. https://doi.org/10.4000/mots.586

Brasseur, M. (2012). L'interaction du chercheur avec son terrain en recherche-action : deux cas d'accompagnement individuel des managers. *Recherches en Sciences de Gestion, 89*, 103-118. https://doi.org/10.3917/resg.089.0101

Breiner-Sanders, K., Lowe Jr., E., Miles, J. et Swender, E. (2000). ACTFL proficiency guidelines-speaking, revised. *Foreign Language Annals, 33*, 13-18.

Bridgeman, B., Powers, D., Stone, E. et Mollaun, P. (2011). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing 29*(1), 9-108. https://doi.org/10.1177/0265532211411078

Bronckart, J. P. (1977). Théories du langage : une introduction critique. P. Mardaga.

Brown, A. (1993, 2-7 août). *The effect of rater variables in the development of an occupation-specific language performance test* [communication orale]. The meeting of the Language Testing Research Colloquium, Cambridge.

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. https://search.informit.org/doi/10.3316/informit.905752862811172

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*, 41-70. https://search.informit.org/doi/10.3316/informit.078722747791492

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

Brown, A., Iwashita, N. et McNamara, T. F. (2005). An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks. Dans *Monograph Series MS-29*. Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2005.tb01982.x

Brown, A., Iwashita, N., McNamara, T. F. et O'Hagan, S. (2002, décembre). *Getting the balance right : Criteria in integrated tasks* [communication orale]. 24th Language Testing Research Colloquium, Hong Kong Polytechnic University.

Brown, A. et Lumley, T. (1997). Interviewer variability in specific-purpose language performance tests. Dans V. Kohonen, A. Huhta, L. Kurki-Suonio et S. Luoma (dir.), *Current developments and alternatives in language assessment : proceedings of LTRC 96* (p. 137-50). University of Jyvaskyla and University of Tampere.

Brown, D. H. et Abeywickrama, P. (2010). *Language Assessment: Principles and Classroom Practices* (2e éd.). Pearson, Longman.

Brown, J. D. et Ahn, R. C. (2011). Variables that affect the dependability of L2 pragmatic tests. *Journal of Pragmatics, 43*(1), 198-217. https://doi.org/10.1016/j.pragma.2010.07.026

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

Byrnes, H. (2007). Persepctives. *The Modern Language Journal 91*(4), 641-645. https://doi.org/10.1111/j.1540-4781.2007.00627_1.x

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

Calderhead, J. (1981). Stimulated recall : A method for research on teaching. *British Journal of Educational Psychology, 51,* 211-217. https://doi.org/10.1111/j.2044-8279.1981.tb02474.x

Canale, M. et Swain, M. (1980). Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing. *Applied Linguistics, 1*(1), 1-47. https://doi.org/10.1093/applin/I.1.1

Carey, M. D., Mannel, R. H. et 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

Carr, N. T. (2011). *Designing and Analyzing Language Tests.* Oxford handbooks for language teachers. Oxford University Press.

Carroll, J. B. (1961). The Nature of Data, or How to Choose a Correlation Coefficient. *Psychometrika, 26*(4), 347-372. https://doi.org/10.1007/BF02289768

Carroll, J. B. (1983). Psychometric Theory and Language Testing. Dans J.W. Oller, Jr. (Dir.), *Issues in Language Testing research* (p. 81-107). Newbury House.

Casanova, D. et 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

Chadwick, E. (1858). On the economical, social, educational, and political influences of competitive examinations, as tests of qualifications for admission to the junior appointments in the public service. *Journal of the Statistical Society of London 21*(1), 18-51.

Chalhoub-Deville, M. (1995a). A Contextualized Approach to Describing Oral Language Proficiency. *Language Learning, 45*(2), 251-281. https://doi.org/10.1111/j.1467-1770.1995.tb00440.x

Chalhoub-Deville, M. (1995). Deriving Oral Assessment Scales Across Different Tests and Rater Group. *Language Testing, 12*, 17-33. https://doi.org/10.1177/026553229501200102

Chalhoub-Deville, M. et Fulcher, G. (2003). The Oral Proficiency Interview : A research agenda. *Foreign Language Annals, 36*(4), 498-506. https://doi.org/10.1111/j.1944-9720.2003.tb02139.x

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-272. https://doi.org/10.1017/S0267190599190135

Chapelle, C. A. (2012). Seeking Solid Theoretical Ground for the ACTFL-CEFR Crosswalk. Dans E. Tschirner (dir.), *Aligning Frameworks of Reference in Language Testing : The ACTFL Proficiency Languages* (p. 35-48). Guidelines and the Common European Framework of Reference for Tübingen : Stauffenburg Verlag.

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

Chénier, C. (2018). *Étude longitudinale de niveau de sévérité d'examinateurs en français langue étrangère* [Thèse de doctorat, Université du Québec à Montréal]. Archipel. http://archipel.uqam.ca/id/eprint/12583

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. M.LT. Press. http://www.jstor.org/stable/j.ctt17kk81z

Clark, C. M. et Peterson, P. L. (1976). *Teacher stimulated recall of interactive decisions.* Stanford Center for Research and Development in Teaching, Stanford University.

Clark, J. L. D. (1979). Direct versus semi-direct tests of speaking proficiency. Dans E. J. Briere et F. B. Hinofotis (dir.), *Concepts in language testing : Some recent studies* (p.61-74). TESOL.

Clark, J. L. D. (1988). *The proficiency-oriented testing movement in the United States and its implications for instructional program design and evaluation*. Defense Language Institute.

Comeford, R. (2009). Alerte ! L'Éducation nationale est tombée dans l'escarcelle des marchands de certifications ! *Enseigner les langues vivantes avec le Cadre européen, 18*(hors-série).

Connor-Linton, J. (1995). Looking behind the curtain : What do L2 composition ratings really mean ? *TESOL Quarterly, 29*, 762-765. https://doi.org/10.2307/3588174

Conseil de l'Europe (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer.* Didier. https://rm.coe.int/16802fc3a8

Conseil de l'Europe (2009). *Manuel pour relier les examens de langue au Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer.* https://rm.coe.int/1680667a2e

Conseil de l'Europe (2011). *Manuel pour l'élaboration et la passation de tests et d'examens de langues.* https://rm.coe.int/1680667a2c

Conseil de l'Europe (2018). *Cadre Européen Commun de Référence pour les Langues : apprendre, enseigner, évaluer.* Volume complémentaire avec de nouveaux descripteurs. http://www.coe.int/lang-cecr

Coste, D. (2007, 6-8 février). *Contextualising uses of the common European framework of reference for languages* [communication orale]. Council of Europe Policy Forum on use of the CEFR, Strasbourg.

Coste, D., Courtillon, J., Ferenczi, V., Martins-Baltar, M. et Pape, E. (1976). *Un niveau-seuil*. Hatier.

Courtillon, J. (2003). *Élaborer un cours de FLE*. Hachette, Collection F.

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

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice, 31*(3), 10-20. https://doi.org/10.1111/j.1745-3992.2012.00239.x

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31-51. https://doi.org/10.1177/026553229000700104

Cumming, A. Kantor, R. et Powers, D. (2001). Scoring TOEFL essays and TOEFL 2000 prototype writing tasks : An investigation into raters' decision making and development of a preliminary analytic framework. Dans *TOEFL Monograph Series* (vol. 22). Educational Testing Service.

Cumming, A., Kantor, R. et Powers, D. (2002). Decision making while rating ESL/EFL writing tasks : A descriptive framework. *Modern Language Journal, 86*, 67-96. https://www.jstor.org/stable/1192770

Cuq, J. P. et Davin-Chnane, F. (2007). Français langue seconde : un concept victime de son succès ? Dans M. Verdelhan-Bourgade (dir.), *Le français langue seconde. Un concept et des pratiques en évolution*. Louvain-la-Neuve, De Boeck Supérieur, « Perspectives en éducation et formation » (p. 11-28). https://hal.archives-ouvertes.fr/hal-02471528

Daunais, J. P. (1992). L'entretien non directif. Dans B. Gauthier (dir.), *Recherche sociale : De la problématique à la collecte des* données (p. 273-293). Presses de l'Université du Québec.

Davidson, F. J., Alderson, C., Douglas, D., Huhta, A., Turner, C. et Wylie, E. (1995). *Report of the Task Force on Testing Standards (TFTS) to the International Language Testing Association (ILTA).* International Language Testing Association.

Davies, A. (2008). Ethics and professionalism. Dans E. Shohamy (dir.), *Language testing and assessment* (vol. 7). Encyclopedia of language and education (p. 429-443). Springer.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. et McNamara, T. F. (1999). Dictionary of Language Testing. *Studies in language testing, 7*. Cambridge University Press.

Davis, L. E. (2012). *Rater expertise in a second language speaking assessment: The influence of training and experience* [Thèse de doctorat, Université d'Hawaï à Mānoa]. ScholarSpace. http://hdl.handle.net/10125/100897

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135. https://doi.org/10.1177%2F0265532215582282

Deaudelin, C., Lefebvre, S., Brodeur, M., Mercier, J., Dussault, M. et Richer, J. (2005). Évolution des pratiques et des conceptions de l'enseignement, de l'apprentissage et des TIC chez des enseignants du primaire en contexte de développement professionnel. *Revue des sciences de l'éducation, 31*(1), 79-110. https://doi.org/10.7202/012359ar

De Fontenay, H., (1991). Les techniques d'entrevue au service de l'évaluation de la performance orale. Actes du colloque AQEFLS 1990. *Bulletin AQEFLS, 12*(3-4), 77-97.

De Groot, D. A. (1965). *Thought and choice in chess*. Mouton

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

DeKetele, J. M. et Roegiers, X. (2009). *Méthodologie du recueil d'informations*. De Boeck Université.

De Landsheere, G. (1992). *Dictionnaire de l'évaluation et de la recherche en éducation* (2e éd.). Presses Universitaires de France.

Demeuse, M. et Artus, F. (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 du Français (TEF) de la Chambre de Commerce et d'Industrie de Paris (CCIP). Dans *Les Cahiers des Sciences de l'Éducation* (no 25 et 26, p. 131-151).

Demeuse, M., Desrochers, F., Casanova, D., Crendal, A. et Holle, A. (2010, 14-16 janvier). *Validation empirique d'un test de français langue étrangère en regard du Cadre européen commun de référence pour les langues* [communication orale]. 22e colloque international de l'ADMEE Europe. Braga, Portugal.

Demeuse, M., Desrochers, F., Crendal, A., Oster, P., Renaud, F. et Leroux, X. (2004, 18-20 novembre). *L'évaluation des compétences linguistiques des adultes en français langue étrangère dans une perspective de multi-référentialisation* [communication orale]. 17e Colloque international de l'ADMEE Europe. Lisbonne, Portugal.

Denzin, N. et Lincoln, Y. (2000) The Discipline and Practice of Qualitative Research. Dans N. K. Denzin Y. S. Lincoln (dir.), *Handbook of Qualitative Research* (p. 1-28). Sage Publications, Thousand Oaks.

De Pietro, J. F. et Wirthner, M. (1996). L'oral, bon à tout faire ?... État d'une certaine confusion dans les pratiques scolaires. *Repères, 1998(17)*, 21-40. https://doi.org/10.3406/reper.1998.2245

Deslauriers, J. P. (1991). *Recherche qualitative, guide pratique*. Thema. Chenelière/McGraw-Hill. https://doi.org/10.7202/706532ar

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

Dionne, J. P. (1996). Indices métacognitifs générés par rétrospection à partir d'épisodes de protocoles verbaux et visuels. *Revue des sciences de l'éducation, 22*(3), 539-550. https://doi.org/10.7202/031892ar

Dolz, J. (2002). L'énigme de la compétence en éducation. Des travaux en sciences de l'éducation revisités. Dans V. Castellotti et B. Py (Dir.), *La notion de compétence en langue. Notions en question, Rencontres en didactique des langues* (p. 83-104). ENS Éditions.

Dolz, J. et Schneuwly, B. (2009). *Pour un enseignement de l'oral. Initiation aux genres formels à l'école*. (4e édition). ESF éditeur, Didactique du français. https://archive-ouverte.unige.ch/unige:31461

Dörnyei, Z. (2003). *Questionnaire in second language research.* Mahway. Lawrence Erlbaum Associates.

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

Douglas, D. et Selinker, L. (1992). Analysing oral proficiency test performance in general and specific purpose contexts. *System, 20*, 317-28. https://doi.org/10.1016/0346-251X(92)90043-3

Douglas, D. et 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*. TESOL Publications, 235-256.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M. et Van Moere, A. (2008). Evaluation of the usefulness of the Versant for English test : A response. *Language Assessment Quarterly 5*(2), 160-167. https://doi.org/10.1080/15434300801934744

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

Dunn, T. G. et Lozinski, M. (2005, avril). *Tacit knowledge use in classroom management* [communication orale]. Annual meeting of the American Research Association, Montréal.

Durand, M. (2002). *Histoire du Québec*. Éditions Imago.

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

Eckes, T. (2005). Examining rater effects in TestdaF writing and speaking performance assessments : A many-facet Rasch analysis. *Language Assessment Quarterly, 2*, 197-221. https://doi.org/10.1207/s15434311laq0203_2

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society, 53*, 460-475, 644-663. https://www.jstor.org/stable/2979547

Elder, C., Knoch, U., Barkhuizen, G. et von Randow, J. (2005). Individual feedback to enhance rater training : Does it work ? *Language Assessment Quarterly, 2*, 175-196. https://doi.org/10.1207/s15434311laq0203_1

Engelhard Jr., G. et Myford, C. (2003). *Monitoring Faculty Consultant Performance in the Advanced Placement English Literature and Composition Program with a Many-Faceted Rasch Model* (publication no 2003-1 ETS RR-03-01). College Entrance Examination Board. http://dx.doi.org/10.1002/j.2333-8504.2003.tb01893.x

Ericsson, K. A. et Crutcher, R. J. (1991). Introspection and verbal reports on cognitive processes - Two approaches to the study of thinking : A response to Howe. *New Ideas in Psychology, 9*(1), 57-71. https://doi.org/10.1016/0732-118X(91)90041-J

Ericsson, K. A. et Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*, 215-251. https://doi.org/10.1037/0033-295X.87.3.215

Ericsson, K. A. et Simon, H. A. (1984). *Protocol analysis : Verbal reports as data*. MIT Press.

Ericsson, K. A. et Simon, H. A. (1987). Verbal reports on thinking. Dans C. Faerch et G. Kasper (dir.), *Introspection in second language research* (p. 24-53). Multilingual Matters.

Ericsson, K. A. et Simon, H. A. (1993). *Protocol analysis : Verbal reports as data*. MIT Press. https://doi.org/10.7551/mitpress/5657.001.0001

Extramiana, C. et Van Avermaet, P. (2010). Apprendre la langue du pays d'accueil. *Hommes et migrations*. https://doi.org/10.4000/hommesmigrations.847

Fahim, M. et Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing, 1*(1).

Farhady, H. (1983). New directions for ESL proficiency testing. Dans J. Oller (dir.), *Issues in language testing research* (p. 253-269). Newbury Publishers.

Fayer, J. M. et Krasinski, E. (1987). Native and Nonnative Judgments of Intelligibility and Irritation. *Language Learning, 37*(3), 313-326. https://doi.org/10.1111/j.1467-1770.1987.tb00573.x

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

Field, J. (2011). Cognitive validity. Dans *Examining speaking : Research and practice in assessing second language speaking* (*Studies in language testing 30*) (p. 65-111). Cambridge University Press.

Figari, G. (1994). *Évaluer : quel référentiel?* De Boeck Université. https://doi.org/10.7202/031863ar

Figueras, N. (2012). The impact of the CEFR. *ELT Journal, 66*(4). https://doi.org/10.1093/elt/ccs037

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

Fiske, S. (2008). *Psychologie sociale*. De Boeck Supérieur.

Flowerdew, J. (1994). Research of relevance to second language lecture comprehension - an overview. Dans J. Flowerdew (dir.), *Academic listening* (p. 7-29). Cambridge University Press. https://doi.org/10.1017/CBO9781139524612.004

Forel, C. et Gerber, B. (2013). L'apprentissage des langues au-delà de la linguistique : le CECR. *Cahiers Ferdinand de Saussure, 66*, 81-95. http://www.jstor.org/stable/24324191

Forget, M. H. (2013). Le développement des méthodes de verbalisation de l'action : un apport certain à la recherche qualitative. Recherches qualitatives, 32(1), 57-80.

Fortin, M. F. et Gagnon, J. (2016). *Fondements et étapes du processus de recherche : Méthodes quantitatives et qualitatives (3e éd.)*. Chenelière éducation.

Fortin, M. F. (1996). Méthodes de collecte des données. Dans M. F. Fortin (dir.), *Le processus de la recherche de la conception à la réalisation* (p. 227-263). Décarie.

Freedman, S. W. et 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.

Fulcher, G. (1987). Tests of oral performance : The need for data-based criteria. *ELT Journal 41(4)*, 287-291.

Fulcher, G. (1997). Testing speaking. Dans D. Corson et C. Clapham (dir.), *Encyclopaedia of language and education : Language testing and assessment* (vol 7, p. 75-86). Kluwer.

Fulcher, G. (2000). The 'Communicative' Legacy in Language Testing. *System, 28*, 483-497. https://doi.org/10.1016/S0346-251X(00)00033-6

Fulcher, G. (2003). *Testing second language speaking*. Pearson.

Fulcher, G. (2004). Deluded by Artifices ? The Common European Framework an Harmonization. *Language Assessment Quarterly 1*(4), 253-266. https://doi.org/10.1207/s15434311laq0104_4

Fulcher, G. (2008). Testing times ahead ? *Liaison Magazine, 1*.

Fulcher, G. (2016). Standards and frameworks. Dans D. Tsagari et J. Banerjee (dir.), *Handbook of Second Language Assessment* (p. 29-44). De Gruyter. https://doi.org/10.1515/9781614513827-005

Furneaux, C. et Rignall, M. (2007). The effect of standardization-training on rater judgements for the IELTS writing module. Dans L. Taylor et P. Falvey (dir.), *IELTS Collected Papers : Research in speaking and writing assessment* (p. 422-445). Cambridge University Press.

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

Galaczi, E. D. et ffrench, A. (2011). Context validity. Dans *Examining speaking: Research and practice in assessing second language speaking* (*Studies in language testing 30*) (p. 112-170). Cambridge University Press.

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

Garcia-Debanc, C. (2004). *Les modèles disciplinaires en acte dans les pratiques effectives d'enseignants débutants*. Actes du 9e colloque de l'AIRDF. http://www.colloqueairdf.fse.ulaval.ca/actes/

Gass, S. M. et Mackey, A. (2000). *Stimulated recall methodology in second language research*. Lawrence Erlbaum. https://doi.org/10.4324/9781410606006

Gauthier, G., Saint-Onge C. et Tavares, W. (2016). Rater cognition: review and integration of research findings. *Medical Education 50*(5), 511-522. https://doi.org/10.1111/medu.12973

Genesee, F. et Upshur, J. A. (1996). *Classroom-based Evaluation in Second Language Education*. Cambridge University Press.

Georges, S. (2013). Évaluer la production orale au travers d'une démarche scientifique. *Revue française de linguistique appliquée*, XVIII, p. 47-58. https://doi.org/10.3917/rfla.181.0047

Gephart, R. P. Jr. (1988). *Ethnostatistics : Qualitative Foundations for Quantitative Research*. Sage publications.

Gérard, F. M. et van Lint-Muguerza, S. (2000). Quel équilibre entre une appréciation globale de la compétence et le recours aux critères ? Dans C. Bosman, F. M. Gérard, X. Roegiers (dir.), *Quel avenir pour les compétences ?* (p. 135-140). De Boeck Université.

Germain, C. et Netten, J. (2002). La précision et l'aisance en FLE/FL2 : définitions, types et implications pédagogiques. Dans *Actes du colloque La didactique des langues face aux cultures linguistiques et éducatives*. Marges linguistiques.

Giles H. et Ogay, T. (2007). Communication accommodation theory. Dans B. B. Whaley et W. Samter (dir.), *Explaining communication : Contemporary theories and exemplars* (p. 325-344). Lawrence Erlbaum Associates.

Goasdoué, R. et Vantourout, M. (2017). Évaluations scolaires et étude du jugement des enseignants : pour une docimologie cognitive. Dans *L'évaluation à la lumière des contextes et des disciplines* (p. 141-168). DeBoeck Supérieur. https://doi.org/10.3917/dbu.detro.2017.01.0141

Gohier, C. (2004). De la démarcation entre critères d'ordre scientifique et d'ordre éthique en recherche interprétative. *Recherches qualitatives, 24*, 3-17. http://www.recherchequalitative.qc.ca/Textes/24gohier.pdf

Goullier, F. (2007). *Le Cadre européen commun de référence pour les langues (CECR) et l'élaboration de politiques linguistiques : défis et responsabilités*. Division des Politiques linguistiques, Conseil de l'Europe.

Goullier, F. (2008). La mise en œuvre du Cadre européen commun de référence pour les langues en Europe. Une réalité différenciée dans ses finalités et dans ses modalités. *Revue internationale d'éducation de Sèvres, 47*, 55-62. https://doi.org/10.4000/ries.367

Gouvernement du Québec (2000). *L'immigration au Québec. Un choix de développement*. Consultation 2001-2003.

Grainger, P., Purnell, K. et Kipf, R. (2008). Judging quality through substantive conversations between markers. *Assessment and Evaluation in Higher Education, 33*, 133-142. https://doi.org/10.1080/02602930601125681

Green, A. (2009). Verbal Protocol Analysis in Language Testing Research : a handbook (*Studies in language testing 5).* Cambridge University Press.

Green, A. J. K. et Gilhooly, K. J. (1990). Statistical computing: Individual differences in learning at microscopic and macroscopic levels. Dans K. J. Gilhooly, R. H. Logie, M. T. G. Kean et G. Erdos (dir.), *Lines of Thinking: Reflections on the Psychology of Thought*. Wiley.

Green, A. J. K. et Gilhooly, K. J. (1900). Individual differences and effective learning procedures : The case of statistical computing. *International Journal of Man-Machine Studies 33*, 97-119.

Green, A. J. K. et Hawkey, R. (2012). Marking Assessments : Rating Scales and Rubrics. Dans C. Coombe, P. Davidson, B. O'Sulivan et S. Stoynoff (dir.), *The Cambridge guide to second language assessment* (p. 299-306). Cambridge University Press.

Grinnell, F. (2009). *Everyday Practice of Science : Where Intuition and Passion Meet Objectivity and Logic*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195064575.001.0001

Gronlund, N. E. (1985). *Measurement and Evaluation in Teaching*. MacMillian Publishing company.

Gufoni, V. (1996). Les protocoles verbaux comme méthode d'étude de la production écrite : approche critique. *Études de linguistique appliquée, 101*, 20-32.

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

Halliday, M. (1976). *System and Function in Language*. Oxford University Press.

Hamilton, J., Lopes, M., McNamara, T. F. et Sheridan, E. (1993). Rating scales and native speaker performance on a communicatively-oriented EAP test. *Language Testing, 10*(3), 337-353. https://doi.org/10.1177/026553229301000307

Hamp-Lyons, L. (1991). *Assessing second language writing*. Ablex. https://doi.org/10.2307/328948

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

Harmegnies, B. (1997). Accent. Dans M.-L. Moreau (dir.), *Sociolinguistique - Concepts de base* (p. 9-12). Sprimont.

Hawkey, R. (2011). Consequential validity. Dans *Examining speaking : Research and practice in assessing second language speaking (Studies in language testing 30)* (p. 234-258). Cambridge University Press.

He, A. W. et Young, R. (1998). Language proficiency interviews : A discourse approach. Dans R. Young et A.W. He (dir.), *Talking and testing : Discourse approaches to the assessment of oral proficiency* (p. 1-24). John Benjamins. https://doi.org/10.1075/sibil.14.02he

Hogan, T. P. (2012). *Introduction à la psychométrie*. Chenelière-éducation.

Howe, R. et Ménard, L. (1993). *Croyances et pratiques en évaluation des apprentissages : Étude des croyances et des pratiques des enseignants des cégeps à l'égard de l'évaluation des apprentissages*. Collège Montmorency.

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. et 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. et Jun, S. A. (2014). Age Matters, And So May Raters : Rater Differences in the Assessment of Foreign Accents. *Studies in Second Language Acquisition, 37*(4), 623-650. https://doi.org/10.1017/S0272263114000576

Huberman, M. et Miles, M. B. (1991). *Analyse des données qualitatives : recueil de nouvelles méthodes*. De Boeck Université.

Hudson, T. (2012). Standards-based Testing. Dans G. Fulcher et F. Davidson (dir.), *The Routledge Handbook of Language Testing* (p. 479-494). Routledge.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. Dans M. M. Williamson et B.A. Huot (dir.), *Validating holistic scoring for writing assessment : Theoretical and empirical foundations* (p. 206-236). Hampton Press.

Hurteau, M. (2013). Aspirer à un jugement crédible dans le cadre de l'évaluation de programme. Dans R. Goasdoué, M. Romainville, M. Vantourout (dir.), *Évaluation et enseignement supérieur* (p. 145-161). Pédagogies en développement. De Boeck. https://doi.org/10.3917/dbu.romai.2013.01.0145

Hurteau, M., Houle, S. et Guillemette, F. (2012). *L'évaluation de programme axée sur le jugement crédible*. Presses de l'Université du Québec. https://doi.org/10.7202/1024550ar

Huver, E. (2014). CECR et évaluation : interprétations plurielles et logiques contradictoires. Dans *Les Cahiers du GEPE* (vol. 6/2014). Presses universitaires de Strasbourg. http://www.cahiersdugepe.fr/index.php?id=2652

Huver, E. (2017). Peut-on (encore) penser à partir du CECRL ? Perspectives critiques sur la version amplifiée, in Carette Emmanuelle : Éclectisme en didactique des langues : Hommage à Francis Carton. *Mélanges, Crapel, 38*(1), 27-42. http://www.atilf.fr/spip.php?rubrique650

Huver, E. et Springer, C. (2011). *L'évaluation en langues.* Didier.

Hymes, D. (1972). On Communicative Competence. Dans J. B. Pride et J. Holmes (dir.), *Sociolinguistics. Selected Readings* (p. 269-293). Harmondsworth : Penguin.

Hymes, D. (1984). *Vers la compétence de communication*. Col. LAL, Hatier-Credif.

Inoue, C., Khabbazbashi, N., Lam, D. et Nakatsuhara, F. (2021.) Towards new avenues for the IELTS Speaking Test : Insights from examiners' voices. *IELTS Research Reports Online Series, 2*, 1-70. http://hdl.handle.net/10547/624875

Isaacs, T. (2016). Assessing speaking. Dans D. Tsagari et J. Banerjee (dir.), *Handbook of Second Language Assessment* (p. 131-146). De Gruyter.

Isaacs, T. et Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation : Revisiting research conventions. *Language Assessment Quarterly, 10*(2), 135-159. https://doi.org/10.1080/15434303.2013.769545

Jakobson, R. (1963). *Essais de linguistique générale*. Les Éditions de Minuit.

Jodelet D. (2003). Aperçus sur les méthodologies qualitatives. Dans S. Moscovici et F. Buschini (dir.), *Les méthodes des sciences humaines* (p. 139-162), PUF Fondamental.

Joe, J. N., Harmes, J. C. et 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

Johnson, M. (2001). *The Art of Non-Conversation. A Reexamination of the Validity of the Oral Proficiency Interview*. Yale University Press.

Johnson, M. et Tyler, A. (1998). Re-analyzing the OPI : How much does it look like natural conversation ? Dans R. Young et A. He (dir.), *Talking and testing : Discourse approaches to language proficiency interviews* (p. 27-51). John Benjamins.

Jones, R. L. (1975). Testing language proficiency in the United States government. Dans R. L. Jones et B. Spolsky (dir.), *Testing language proficiency* (p. 1-9). Center for Applied Linguistics.

Jones, N. et Saville, N. (2009). European language policy assessment, learning, and the CEFR. *Annual Review of Applied Linguistics, 29*, 51-63. https://doi.org/10.1017/S0267190509090059

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

Jorro, A. (2000). *L'enseignant et l'évaluation. Des Gestes évaluatifs en question*. De Boeck.

Junker, B. H. (1960). *Field work : An introduction to the Social Sciences*. The University of Chicago Press.

Kagan, N., Krathwohl, D. R. et Miller, R. (1963). Stimulated recall in therapy using videotape - A case study. *Journal of Counselling Psychology, 10*, 237-243. https://doi.org/10.1037/h0045497

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

Kane, M. T. (2009). Validating the interpretations and uses of test scores. Dans R. W. Lissitz (dir.), *The concept of validity : Revisions, new directions and applications* (p. 39-64). IAP.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. https://doi.org/10.1111/jedm.12000

Kane, M. T., Cooks, T. et Cohen, A. (1999). Validating measures of performance. *Educational Measurement : Issues and Practice, 18*(2), 5-17. https://doi.org/10.1111/j.17453992.1999.tb00010.x

Kaplan, I. (1991). L'Évaluation des compétences fonctionnelles en langue seconde. *Bulletin de l'AQEFLS, 1*, 124-137.

Karsenti, T. et Demers, S. (2011). L'étude de cas. Dans T. Karsenti et L. Savoie-Zajc (dir.), *La recherche en éducation : Étapes et approches* (3e éd., p. 229-252). ERPI.

Kaulfers, W. V. (1944). War-time developments in modern language achievement tests. *Modern Language Journal, 70*, 366-72. https://doi.org/10.1111/J.1540-4781.1944.TB04835.X

Kenyon, D. (1992, 27 février-1 mars). *Introductory remarks at symposium on development and use of rating scale in language testing*. 14th Language Testing Research Colloquium, Vancouver.

Khalifa, H. et Weir, C. J. (2009). Examining Reading: Research and Practice in Assessing Second Language Reading (*Studies in language testing 29*). Cambridge University Press.

Khalifa, H. et Salamoura, A. (2011). Criterion-related validity. Dans *Examining speaking : Research and practice in assessing second language speaking (Studies in language testing 30*) (p. 259-292). Cambridge University Press.

Kim, H. J. (2006). Issues of rating scales in speaking performance assessment. *TESOL & Applied Linguistics, 6*(2).

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

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior - a longitudinal study. *Language Testing, 28*, 179-200. https://doi.org/10.1177/0265532210384252

Kormos, J. (1998). The use of verbal reports in L2 research : Verbal reports in L2 speech production research. *TESOL Quarterly, 32*, 353-358. https://doi.org/10.2307/3587590

Kramsch, C. (1986). From Language Proficiency to Interactional Competence. *The Modern Language Journal, 70*(4), 366-372. https://doi.org/10.2307/326815

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL quarterly, 24*, 741-746.

Lado, R. (1961). *Language testing : The construction and use of foreign language tests*. McGraw-Hill.

Lado, R. (1962). *Language testing. The construction and use of foreign language tests. A Teacher's book*. Longmans, Green.

Laming, D. (2004). *Human judgment : The eye of the beholder*. Thompson Learning.

Lantolf, J. P. et Frawley, W. (1985). Oral-Proficiency Testing. A Critical Analysis. *The Modern Language Journal, 69*(4), 337-345. https://doi.org/10.1111/j.1540-4781.1985.tb04801.x

Laplantine, F. (1996). *La description ethnographique*. Nathan.

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. (2002). *A Qualitative Approach to the Validation of Oral Language Tests*. Cambridge University Press.

Lazaraton, A. (2004). Qualitative research methods in language test development and validation. Dans *European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference July 2001* (*Studies in language testing 18*) (p. 51-72). Cambridge University Press.

Lazaraton, A. et Saville, N. (1994, 5-7 mars). *Processes and outcomes in oral assessment* [communication orale]. 16th Language Testing Research Colloquium, Washington DC.

Leclerc, J. (1989). *Qu'est-ce que la langue ?* Mondia.

Lemaire, P. (1999). *Psychologie cognitive*. De Boeck

Levelt, W. J. M. (1989). *Speaking. Cambridge*. MIT Press.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment : A longitudinal study of new and experienced raters. *Language Testing, 28*, 543-560. https://doi.org/10.1177/0265532211406422

Lim, G. S. (2018) Conceptualizing and Operationalizing Second Language Speaking Assessment: Updating the Construct for a New Century. *Language Assessment Quarterly, 15*(3), 215-218. https://doi.org/10.1080/15434303.2018.1482493

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Mesa Press.

Lincoln, Y. S. et Guba, E. G. (1985). *Naturalistic Inquiry*. Calif : Sage.

Ling, G., Mollaun, P. et Xi, X. (2014). A study on the impact of fatigue on human raters when scoring speaking response. *Language Testing, 31*(4), 479-499. https://doi.org/10.1177/0265532214530699

Lipman, M. (1992). L'éducation au jugement. Dans M. Schleifer (dir.), *La formation du jugement* (p. 99-123). Les éditions Logiques.

Liskin-Gasparro, J. E. (1984a). The ACTFL proficiency guidelines : Gateway to testing and curriculum. *Foreign Language Annals, 17*, 475-89. https://doi.org/10.1111/j.1944-9720.1984.tb01736.x

Liskin-Gasparro, J. E. (1984b). The ACTFL proficiency guidelines : A historical perspective. Dans T. V. Higgs (dir.), *Teaching for proficiency : The organizing principle* (p. 11-42). National Textbook Co.

Liskin-Gasparro, J. E. (1987). *Testing and teaching for oral proficiency*. Heinle and Heinle.

Liskin-Gasparro, J. E. (2003). The ACTFL Proficiency Guidelines and the Oral Proficiency Interview. *Foreign language annals, 36*(4).

Little, D. (2007). The Common European Framework of Reference for Languages : Perspectives on the making of supranational language education policy. *Modern Language Journal, 91*(4), 645-655. https://doi.org/10.1111/j.1540-4781.2007.00627_2.x

Ljalikova, A. (2004). La valorisation de l'évaluation certificative en Didactique de Langues-Cultures Étrangères. *Diversité de la Recherche francophone en Sciences Humaines dans l'espace baltique, 2*.

Lorenzo-Dus, N. et Meara, P. (2005). Examiner support strategies and test-taker vocabulary. International *Review of Applied Linguistics in Language Teaching, 43*(3), 239-258. https://doi.org/10.1515/iral.2005.43.3.239

Lowe, E. (1983). The ILR oral interview: Origins, applications, pitfalls, and implications. *Die Untemchtspraxis, 16*(2), 230-244. https://doi.org/10.2307/3530138

Lowe, E. (1985). The ILR proficiency scale as a synthesizing research principle : The view from the mountain. Dans C. J. James (dir.), *Foreign language proficiency in the classroom and beyond* (p. 9-53). National Textbook Co.

Lowe, E. (1987). Interagency language roundtable proficiency interview. Dans J. C. Alderson, K. J. Krahnke et C.W. Stansfield (dir.), *Reviews of English language proficiency tests* (p. 43-47). TESOL.

Lumley, T. (2002). Assessment Criteria in a large-scale writing Test : what do they really mean to the Raters ? *Language Testing, 19*(3), 247-276. https://doi.org/10.1191/0265532202lt230oa

Lumley T. (2005) *Assessing second language writing : The rater's perspective*. Peter Lang.

Lumley, T. et 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. et 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), p. 140-156. https://doi.org/10.1177/026553229701400202

Lundeberg, O. K. (1929). Recent developments in audition-speech tests. *The Modern Language Journal, 14*(3), 193-202. https://doi.org/10.1111/j.1540-4781.1929.tb01268.x

Lunz, M. E. et Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13*, 425-444. https://doi.org/10.1177/016327879001300405

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733017

Lussier, D. et Turner, C. E. (1995). *Le point sur l'évaluation en didactique des langues*. CEC.

Lyle, J. (2002). Stimulated recall : a report on its use in naturalistic research. *British Educational Research Journal, 29*(6), 861-878. http://www.jstor.org/stable/1502138

Macaire, D. (2018). Le CECRL : Quelle puissance du modèle ? Questionnements dans la recherche en didactique des langues-cultures. *Carnet des jeunes chercheurs du CREM, 7*. https://hal.archives-ouvertes.fr/hal-02573659

Mackey, A. et Gass, S. M. (2005). *Second language research : Methodology and design*. Lawrence Erlbaum Associates, Inc.

Madaus, G. F. et Stufflebeam, D. L. (2004). Program Evaluation : A Historical Overview. Dans D. L. Stufflebeam, G. F. Madaus et T. Kellagan (dir.), *Evaluation Models : Viewpoints on Educationnal and Human Services Evaluation* (p. 3-19). KIlI, ver Academie Publishers.

Magis, D., Béland, S., Tuerlinckx, F. et DeBoeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior research methods, 42*, 847-862. https://doi.org/10.3758/BRM.42.3.847

Major, R. C., Fitzmaurice, S. F., Bunta, F. et Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension : Implications for ESL assessment. *TESOL Quarterly, 36*(2). 173-190. https://doi.org/10.2307/3588329

Malone, M. E. (2003). Research on the Oral Proficiency Interview : Analysis, synthesis, and future directions. *Foreign Language Annals, 36*(4), 491-497. https://doi.org/10.1111/j.1944-9720.2003.tb02138.x

Malone, M. E. et Montee, M. J. (2010). Oral proficiency assessment : Current approaches and applications for post-secondary foreign language programs. *Language and Linguistics Compass, 4*(10), 972-986. https://doi.org/10.1111/j.1749-818X.2010.00246.x

Matell, M. S. et Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items ? *Educational and Psychological Measurement, 31*, 657-674. https://doi.org/10.1177/001316447103100307

Maurer, B. (2011). *Enseignement des langues et construction européenne - Le plurilinguisme, nouvelle idéologie dominante*. Éditions des Archives Contemporaines.

May, L. A. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing (MPLT), 11*(1), 29-51.

McIntyre, P. N. (1993). *The importance and effectiveness of moderation training on the reliability of teacher assessments of ESL writing samples* [mémoire de maîtrise, Université de Melbourne]. Minerva Access. https://minerva-access.unimelb.edu.au/handle/11343/36340

McNamara, T. F. (1993). *Second language performance assessment* [document inédit].

McNamara, T. F. (1995). Modelling performance : Opening Pandora's box. *Applied Linguistics, 16*(2), 159-179. https://doi.org/10.1093/applin/16.2.159

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

McNamara, T. F. (2000). *Language testing*. Oxford University Press.

McNamara, T. F. (2010). The use of language tests in the service of policy : issues of validity. *Revue française de linguistique appliquée, XV*, 7-23. https://doi.org/10.3917/rfla.151.0007

McNamara, T. F. et Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*, 140-56. https://doi.org/10.1177/026553229701400202

McNamara, T. F. et Shohamy, E. (2008). Language tests and human rights. *International Journal of Applied Linguistics, 18*(1), 89-95. https://doi.org/10.1111/j.1473-4192.2008.00191.x

Meiron, B. E. (1998). *Rating oral proficiency tests : A triangulated study of rater thought processes* [mémoire de maîtrise non publié, Université de Californie à Los Angeles].

Meiron, B. E. et Schick, L. (2000). Ratings, raters and test performance : An exploratory study. Dans A. J. Kunnan (dir.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando*, *Florida* (p. 153-176). Cambridge University Press.

Merrylees, B. et McDowell, C. (2007). A survey of examiner attitudes and behavior in the IELTS oral interview. Dans *IELTS Collected Papers : Research in Speaking and Writing Assessment* (*Studies in language testing 19*) (p. 142-182). Cambridge University Press.

Messick, S. (1989). Validity. Dans R. L. Linn (dir.), *Educational measurement* (3e éd., p. 13-103). American Council on Education and Macmillan.

Milanovic M. et Saville, N. (1993). The background to the first LTRC in Europe. Dans M. Milanovic et N. Saville (dir.), *Performance testing, cognition and assessment* (*Studies in language testing 3*) (p. 1-17). Cambridge University Press.

Milanovic, M., Saville, N., Pollitt, A. et Cook, A. (1996). Developing Rating Scales for CASE : Theoretical Concerns and Analyses. Dans A. Cumming et R. Berwick (dir.), *Validation in language testing* (p. 15-38). Modern Languages in Practice.

Milanovic, M., Saville, N. et Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. *Studies in language testing, 3*, 92-111.

Miller, G. (1956). The magical number seven, plus or minus two : Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97. https://doi.org/10.1037/h0043158

Ministère de l'Immigration et des Communautés Culturelles (2011). *Échelle québécoise des niveaux de compétence en français des personnes immigrantes adultes.* Gouvernement du Québec.

Moirand, S. (1982). *Enseigner à communiquer en langue étrangère*. Hachette.

Montague, M. et Applegate, B. (1993). Middle School Students' Mathematical Problem Solving: An Analysis of Think-Aloud Protocols. *Learning Disability Quarterly, 16*(1), 19-32. https://doi.org/10.2307/1511157

Morrow, K. (2004). Background on the CEF. Dans K. Morrow (dir.), *Insights from the Common European Framework* (p. 3-11). Oxford University Press.

Morton, J., Wigglesworth, G. et 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. 175-96). National Centre for English Language Teaching and Research.

Myford, C. M., Marr, D. B. et Linacre, J. M. (1996). *Reader calibration and its potential role in equating for the Test of Written English* (publication no RR-95-40, TOEFL-RR-52). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01674.x

Myford, C. M. et Wolfe, E. W. (2000). Monitoring Sources of Variability Within the Test of Spoken English Assessment System. *ETS Research Report Series, 2000(1)*, i-51.

Myford, C. M. et 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. et 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

Newton, P. E. et Shaw, S. D. (2014). *Validity in Educational & Psychological Assessment*. Thousand Oaks, Sage.

Nisbett, R. E. et Wilson, T. (1977). Telling more than we can know : verbal reports on mental processes. *Psychological Review, 84*(3), 231-259. https://doi.org/10.1037/0033-295X.84.3.231

Nittrouer, S., Manning, C. et Meyer, G. (1993). The perceptual weighting of acoustic cues changes with linguistic experience. *Journal of the Acoustical Society of America, 94*(3), 1865-1865. https://doi.org/10.1121/1.407649

Noël-Jothy, F. et Sampsonis, B. (2006). *Certifications et outils d'évaluation en FLE*. Hachette Français Langue Étrangère.

Norman, W. T. et 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

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency*. Peter Lang. https://doi.org/10.3726/978-1-4539-1059-7

North, B. (2004, 15 avril). *Europe's framework promotes language discussion, not directives.* Education Guardian. https://www.theguardian.com/education/2004/apr/15/tefl6

North, B. (2007). The CEFR Illustrative Descriptor Scales. *The Modern Language Journal 91*(4), 656-659. https://doi.org/10.1111/j.1540-4781.2007.00627_3.x

North, B. et Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15(*2), 217-262. https://doi.org/10.1191/026553298676177132

Omaggio, A. (1983). Methodology in Transition : The Focus on Proficiency. *The Modern Language Journal, 67*, 330-40. https://doi.org/10.1111/j.1540-4781.1983.tb01512.x

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. (2012). Assessing speaking. Dans C. Coombe, P. Davidson, B. O'Sullivan et S. Stoynoff (dir.), *The Cambridge guide to assessment* (p. 234-246). Cambridge University Press.

O'Sullivan, B. et Lu, Y. (2006). The impact on candidate language of examiner deviation from a set interlocutor frame in the IELTS Speaking Test. *IELTS Research Reports, 6*, 91-117.

O'Sullivan, B. et Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. Dans L. Taylor et P. Falvey (dir.), *IELTS Collected Papers : Research in speaking and writing assessment* (p. 446-478). Cambridge University Press.

O'Sullivan, B. et Weir, C. J. (2011). Test development and validation. Dans O'Sullivan, B. (dir.), *Language testing : theories and practices. Palgrave Advances in Language and Linguistics* (p. 13-32). Palgrave MacMillan.

Pae, T. I. (2004). Gender effect on reading comprehension with Korean EFL learners. *System, 32*, 265-281. https://doi.org/10.1016/j.system.2003.09.009

Pagé, M. (2011). *Politiques d'intégration et cohésion sociale.* Bibliothèque et Archives Nationales du Québec.

Pajares, M. F. (1992). Teachers' Beliefs and Educational Research : Cleaning Up a Messy Construct. *Review of Educational Research, 62*(3), 307-332. https://doi.org/10.2307/1170741

Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly, 36*, 219-233. https://doi.org/10.2307/3588333

Patton, M. (2002). *Qualitative Research and Evaluation Methods* (2e éd.). Thousand Oaks, Sage Publications.

Patz, R. J., Junker, B. W., Johnson, M. S. et Mariano, L. T. (2002). The Hierarchical Rater Model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics in Medicine, 27*, 341-384. https://doi.org/10.3102/10769986027004341

Peters, M. et Bélair, L. (2011). Caractéristiques d'activités d'évaluation de la compétence langagière à l'université. *Revue internationale de pédagogie de l'enseignement supérieur, 27*(1). https://doi.org/10.4000/ripes.439

Peterson, P. L. et Swing, S. R. (1982). Beyond time on task : Students' reports of their thoughts processes during classroom instruction. *The Elementary School Journal, 82*(5), 481-491. http://www.jstor.org/stable/1001325

Pichette, F., Raîche, G., Béland, S. et Magis, D. (2011). Évaluation d'un test de lecture en anglais par deux méthodes de détection du fonctionnement différentiel d'items. *Revue des sciences de l'éducation, 37*(3), 543-568. https://doi.org/10.7202/1014757ar

Poisson, Y. (1983). L'approche qualitative et l'approche quantitative dans les recherches en éducation. *Revue des sciences de l'éducation, 9*(3), 369-378. https://doi.org/10.7202/900420ar

Pollit, A. et 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.

Powers, D. E., Albertson, W., Florek, T., Johnson, K., Malak, J., Nemceff, B., Porzuc, M., Silvester, D., Wang, M., Weston, R., Winner, E. et Zelazny, A. (2002). Influence of irrelevant speech on standardized test performance. *ETS Research Report Series, 2002*(1). https://doi.org/10.1002/j.2333-8504.2002.tb01873.x

Préfontaine, C. et Fortier, G. (1997). Utilisation de la verbalisation dans des situations de recherche sur la production écrite. Dans J.- Y. Boyer et L. Savoie-Zajc (dir.), *Didactique du français, méthodes de recherche* (p. 219-228). Éditions Logiques.

Preston, C. C. et Colman, A. M. (2000). Optimal number of response categories in rating scales : Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104,* 1-15. https://doi.org/10.1016/S0001-6918(99)00050-5

Pula, J. J. et Huot, B. (1993). A model of background influences on holistic raters. Dans M. M. Williamson et B. Huot (dir.), *Validating holistic scoring for writing assessment : Theoretical and empirical foundations* (p. 237-265). Hampton Press.

Puren, C. (2006). De l'approche communicative à la perspective actionnelle. *Le Français dans le Monde, 347*, 37-40.

Purpura, J. E. (2012, 24-27 mars). *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.

Purpura, J. E. (2013). Cognition and language assessment. Dans A. J. Kunnan et A. J. Hoboken (dir.), *The companion to language assessment* (p. 1452-1476). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla150

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment : Affective effects on test takers. *Language Assessment Quarterly, 6*(2), 113-125. https://doi.org/10.1080/15434300902800059

Québec MCCI (1990). *Au Québec pour bâtir ensemble. Énoncé de politique en matière d'immigration et d'intégration*. Direction des affaires publiques et des communications.

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

Reed, D. J. et Halleck, G. B. (1997). Probing above the ceiling in oral interviews : what's up there ? Dans V. Kohonen, A. Huhta, 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.

Riba, P. et Mavel, M. (2005). L'harmonisation du DELF et du DALF sur les niveaux du Cadre européen commun de référence pour les langues. Dans L. Tylor et C. J. Weir (dir.), *Multilingualism and assessment (Studies in language testing 27)*. Cambridge University Press.

Rispail, M. (2005). *L'oral dans la classe : Compétences, enseignement, activités*. L'Harmattan. https://doi.org/10.4000/lidil.122

Rivière, V. (2016). Communiquer en situation d'évaluation certificative : analyse de débriefings en formation initiale d'enseignants du français langue étrangère. *Communiquer, Revue de communication sociale et publique, 18,* 103-115. https://doi.org/10.4000/communiquer.2070

Robert, M. (1988). *Fondements et étapes de la recherche scientifique en psychologie*. Edisem.

Roegiers, X. (2004). *L'école et l'évaluation. Des situations pour évaluer les compétences des élèves*. De Boeck.

Roegiers, X. (2010). *L'école et l'évaluation : Des situations complexes pour évaluer les acquis des élèves*. De Boeck Supérieur. https://doi.org/10.3917/dbu.roegis.2010.01

Roever, C. et McNamara, T. F. (2006). *Language testing : the social dimension*. Blackwell Publishing. https://doi.org/10.1111/j.1473-4192.2006.00117.x

Romainville, M. (2011). Objectivité versus subjectivité dans l'évaluation des acquis des étudiants. *Revue internationale de pédagogie de l'enseignement supérieur, 27*(2). http://ripes.revues.org/499

Rosen, E. et Reinhardt, C. (2010). *Le point sur le Cadre européen commun de référence pour les langues*. CLE International.

Rosenshine, B. V. (1986). Vers un enseignement efficace des matières structurées. Un modèle d'action inspiré par le bilan des recherches processus-produit. Dans M. Crahay et D. Lafontaine (dir.), *L'art et la science de l'enseignement* (p. 81-96). Labor.

Ross, S. (1996). Formulae and inter-interviewer variation in oral proficiency interview discourse. *Prospect, 11*(3), 3-16. https://search.informit.org/doi/10.3316/aeipt.84115

Ross, S. et Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition 14*, 159-176. http://www.jstor.org/stable/44488406

Sacks, H. (1984). Notes on methodology. Dans J. M. Atkinson et J. Heritage (dir.), *Structures of Social Action : Studies in Conversation Analysis* (p. 21-27). Cambridge University Press.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119-144. https://doi.org/10.1007/BF00117714

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment : How raters evaluate compositions. Dans A. J. Kunnan (dir.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (p. 129-152). Cambridge University Press.

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

Saville, N. (2012). The CEFR : An Evolving Framework of Reference. Dans E. Tschirner (dir.), *Aligning Frameworks of Reference in Language Testing: The ACTFL Proficiency Guidelines and the Common European Framework of Reference for Languages* (p. 57-69). Stauffenburg Verlag.

Savoie-Zajc, L. (2004). La recherche qualitative/interprétative en éducation. Dans T. Karsenti, et L. Savoie-Zajc (dir.), *Introduction à la recherche en éducation* (p. 171-198). Éditions du CRP.

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.

Savoie-Zajc, L. (2013). Le jugement professionnel. Dans S. Fontaine, L. Savoie-Zajc et A. Cadieux (dir.), *Évaluer les apprentissages - Démarche et outils d'évaluation pour le primaire et le secondaire* (p. 104-116). Les Éditions CEC, Inc.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. Renouveau pédagogique. https://doi.org/10.7202/016293ar

Schleifer, M. et Hurteau, M. (2012). Le jugement crédible : le fondement de toute démarche évaluative. Dans M. Hurteau, S. Houle et F. Guillemette, *L'évaluation de programme axée sur le jugement crédible*. Presse de l'Université du Québec.

Scriven, M. (1980). *The Logic of Evaluation*. Inverness (Calif.): Edgepress.

Shaw, S. (2002). The effect of training and standardization on rater judgement and inter-rater reliability. *Research Notes, 9*, 13-17. www.cambridgeesol.org/rs_notes/rs_nts8.pdf

Shaw, S. D. et Weir, C. J. (2007). *Examining Writing : Research and Practice in assessing second language writing. Studies in language testing 26*. Cambridge University Press. https://doi.org/10.1177/0265532209347198

Shohamy, E. (1982). Affective considerations in language testing. *Modern Language Journal, 66*(1), 3-17. https://doi.org/10.1111/j.1540-4781.1982.tb01015.x

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. (1990). Language Testing Priorities : A Different Perspective. *Foreign Language Annals, 23*(5), 385-394. https://doi.org/10.1111/j.1944-9720.1990.tb00392.x

Shohamy, E. (1990). Discourse analysis in language testing. *Annual Review of Applied Linguistics 11,* 115-131. https://doi.org/10.1017/S0267190500001999

Shohamy, E. (2001). *The power of tests : A critical perspective on the uses of language tests*. Longman/Pearson Education.

Shohamy, E. (2005). *Language policy : Hidden agendas and new approaches*. Routledge. https://doi.org/10.4324/9780203387962

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* (vol. 11, p. 521-532). Springer. https://doi.org/10.1007/978-0-387-46301-8_37

Shohamy, E., Donitsa-Schmidt, S. et Ferman, I. (1996). Test impact revisited : Washback effect over time. *Language Testing, 13*(4), 298-317. https://doi.org/10.1177/026553229601300305

Shohamy, E., Gordon, C. M. et Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal, 76*, 27-33. https://doi.org/10.2307/329895

Shohamy, E. et Walton, R. (1992). *Assessment for feedback : Testing and other strategies*. Kendall/Hunt Publishing Co., Dubuque, IO.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford University Press.

Skehan, P. (2001). Tasks and language performance assessment. Dans M. Byate, P. Skehan et E. Swain (dir.), *Researching Pedagogic Tasks* (p. 196-185). Longman.

Smagorinsky, P. (1994). Think-aloud protocol analysis : Beyond the black box. Dans P. Smagorinsky (dir.), *Speaking about writing : Reflections on research methodology* (vol. 8, p. 3-19). Thousand Oaks, Sage.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. *Studies in Immigrant English Language Assessment, 1*, 159-189.

Sollenberger, H. E. (1978). Development and current use of the FSI oral interview test. Dans J. L. Clark (dir.), *Direct testing of speaking proficiency : Theory and application* (p. 1-12). Educational Testing Service.

Someren, M. W., Barnard, Y. F. et Sandberg, J. (1994). *The think aloud method : A practical guide to modelling cognitive processes*. Academic Press.

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

Stake, R. E. et Schwandt, A. (2006). On Discerning Quality in Evaluation. Dans L. F. Shaw, J. C. Greene, et M. M. Mark (dir.), *The Sage Handbook of Evaluation* (p. 404-418). Thousand Oaks : Sage publications. https://doi.org/10.4135/9781848608078.N18

Stansfield, C. W. et Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20*, 347-64. https://doi.org/10.1016/0346-251X(92)90045-5

Tagliante, C. (2005). *L'évaluation et le cadre européen commun*. CLE International.

Tagliante, C. et Mègre, B. (2008). L'impact du CECR sur l'évaluation des compétences en FLE. *Revue japonaise de didactique du français, 3*(1), 172-178.

Taguchi, N. (2011). Rater variation in the assessment of speech acts. *Pragmatics, 21*(3), 453-471. https://doi.org/10.1075/prag.21.3.08ta

Tajeddin, Z. et Alemi, M. (2014). Pragmatic rater training : Does it affect non-native L2 teachers'rating accuracy and bias ? *Iranian Journal of Language Testing, 4*(1), 66-83.

Tanrıverdi-Köksal, F. et Ortaçtepe, D. (2017). Raters' Knowledge of Students' Proficiency Levels as a Source of Measurement Error in Oral Assessments. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H.U. Journal of Education), 32*(3), 581-599. https://doi.org/10.16986/HUJE.2017027583

Tardieu, C. (2010). Votre B1 est-il mon B1 ? L'interculturel dans les tests d'évaluation en Europe. *Recherches en Didactique des Langues et Cultures : les Cahiers de l'Acedle, 7*(7-2), 225-239. https://doi.org/10.4000/rdlc.2301

Taylor, L. (2007). The impact of the joint-funded research studies on the IELTS Speaking Module. Dans *Studies in language testing, IELTS collected papers* (vol. 19, p. 185-196). Cambridge University Press.

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

Taylor, L. (2011). Examining speaking : Research and practice in assessing second language speaking (*Studies in language testing 30*). Cambridge University Press.

Thorndyke, P. W. et Stasz, C. (1980). Individual differences in procedures for knowledge acquisition from maps. *Cognitive Psychology, 12*, 137-75. https://doi.org/10.1016/0010-0285(80)90006-7

Tochon, F. V. (1996). Rappel stimulé, objectivation clinique, réflexion partagée. Fondements méthodologiques et applications pratiques de la rétroaction vidéo en recherche et en formation. *Revue des sciences de l'éducation, 22*, 467-502. https://doi.org/10.7202/031889ar

Tochon, F. V. (1997). *Organiser des activités de communication orale*. Éditons CRP.

Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge University Press.

Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge University Press. https://doi.org/10.1017/CBO9780511840005

Trudel, P., Haughian, L. et Gilbert, W. (1996). L'utilisation de la technique du rappel stimulé pour mieux comprendre le processus d'intervention de l'entraîneur en sport. *Revue des sciences de l'éducation, 22*(3), 503-522. https://doi.org/10.7202/031890ar

Tschirner, E. (dir.) (2012). *Aligning frameworks of reference in language testing : The ACTFL Proficiency Guidelines and the Common European Framework of Reference.* Stauffenburg.

Tschirner, E. et Bärenfänger, O. (2012, 3-5 avril). *Bridging frameworks for assessment and learning : The ACTFL Guidelines and the CEFR* [communication orale]. 34th Language Testing Research Colloquium (LTRC), Princeton, NJ.

Upshur, J. A. et Turner, C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49*, 3-12. https://doi.org/10.1093/elt/49.1.3

Upshur, J. A. et 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

Vaughan, C. (1991). Holistic assessment : What goes on in the rater's mind ? Dans L. H. Lyons (dir.), *Assessing second language writing in academic contexts* (p. 111-125). Norwood.

Vidakovic, I. et Galaczi, E. D. (2013). The measurement of speaking ability 1913-2012. Dans C. J. Weir, I. Vidaković et E. D. Galaczi (dir.), *Measured constructs : A history of*

Cambridge English language examinations, 1913-2012 (Studies in language testing 37) (p. 257-346). Cambridge University Press.

Vincent, D. (2001). Les enjeux de l'analyse conversationnelle ou les enjeux de la conversation. *Revue québécoise de linguistique, 30*(1), 177-198. https://doi.org/10.7202/000517ar

Van Avermaet, P. et Rocca, L. (2013). Language testing and access. Dans E. D. Galaczi et C. J. Weir (dir.), *Exploring language frameworks : Proceedings of the ALTE Kraków conference, July 2011* (p. 11-44). Cambridge University Press.

Van den Heuvel, P. (1985). *Parole, mot, silence : Pour une poétique de l'énonciation*. Librairie José Corti.

Van der Maren, J. M. (1995). *Méthodes de recherche pour l'éducation*. Coll. Éducation et formation. Presses de l'Université de Montréal.

Van Ek, J. A. (1975). *The Threshold Level in a European Unit Credit System for Modern Language Learning by Adults*. Council for Cultural Co-operation of the Council of Europe.

Van Lier, L. (1989). Reeling, writing, drawing, stretching, and fainting coins : oral proficiency interviews as conversation. *TESOL Quarterly, 23*, 489-508. https://doi.org/10.2307/3586922

Van Someren, M., Barnard, Y. et Sandberg, J. (1994). *The think aloud method. Practical guide to modelling cognitive processes*. Academic Press.

Walters, F. S. (2004). *An Application of a conversation analysis to the development of test of second-language pragmatic competence* [Thèse de doctorat, Université d'Illinois à Urbana-Champaign]. Ideals. http://hdl.handle.net/2142/79792

Wanlin, P. (2007). L'analyse de contenu comme méthode d'analyse qualitative d'entretiens : Une comparaison entre les traitements manuels et l'utilisation de logiciels. *Recherches qualitatives, 3*, 243‑272.

Wei, J. et Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly, 12*(3), 283-304. https://doi.org/10.1080/15434303.2015.1037446

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing 11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287. https://doi.org/10.1177/026553229801500205

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment : Quantitative and qualitative approaches. *Assessing Writing, 6*, 145-178. https://doi.org/10.1016/S1075-2935(00)00010-6

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Weir, C. J. (1993). *Understanding and Developing Language Tests*. Prentice-Hall International (UK) LTD.

Weir, C. J. (2005a). *Language testing and validation, an evidence-based approach*. Palgrave Macmillan.

Weir, C. J. (2005b). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281-300. https://doi.org/10.1191/0265532205lt309oa

Wesolowski, B. C. (2016). Exploring rater cognition : A typology of raters in the context of music performance assessment. *Psychology of Music, 45*(3). https://doi.org/10.1177/0305735616665004

Widdowson, H. G. (1978). *Teaching Language as Communication*. Oxford University Press.

Wigglesworth, G. (1993). Exploring Bias Analysis as a Tool for Improving Rater Consistency in Assessing Oral Interaction. *Language Testing, 10*(3), 305-336. https://doi.org/10.1177/026553229301000306

Wilds, C. (1979). The measurement of speaking and reading proficiency in a foreign language. Dans M. L. Adams et J. R. Frith (dir.), *Testing kit : French and Spanish* (p. 1-12). Foreign Services Institute, U.S. Department of State.

Winke, P., Gass, S. et 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. https://doi.org/10.1002/j.2333-8504.2011.tb02266.x

Winke, P., Gass, S. et 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

Wolfe, E. W. et Chiu, C. W. T. (1997, mars). *Detecting rater effects with a multi-faceted rating scale model* [communication orale]. The annual meeting of the National Council on Measurement in Education, Chicago.

Wolfe E. W., Kao, C. et Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication, 15*, 465-492. https://doi.org/10.1177/0741088398015004002

Wolfe, E. W. et McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement : Issues and Practice, 31*(3), 31-37. https://doi.org/10.1111/j.1745-3992.2012.00241.x

Yee, F. P. (2008). Development of a Framework for Analysing Mathematical Problem-Solving Behaviours. *Singapore Journal of Education, 13*(1), 61-69. https://doi.org/10.1080/02188799308547704

Young, R. F. (1995). Conversational styles in language proficiency interviews. *Language Learning, 45*, 3-42. https://doi.org/10.1111/j.1467-1770.1995.tb00961.x

Young, R. F. et He, A. W. (1998). *Talking and Testing : Discourse Approaches to the Assessment of Oral Proficiency*. John Benjamins.

Young, R. F. et Milanovic, M. (1992). Discourse validation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*, 403-424. https://doi.org/10.1017/S0272263100011207

Zhang, Y. et 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

Zhang, Y., Kuhl, P. K., Imada, T., Kotani, M. et Tohkurad, Y. (2005). Effects of language experience : Neural commitment to language-specific auditory patterns. *NeuroImage, 26*, 703-720. https://doi.org/10.1016/j.neuroimage.2005.02.040

Zumbo, B. D. (2007). Three Generations of DIF Analyses : Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly, 4*(2), 223-233. https://doi.org/10.1080/15434300701375832

Annexes

Annexe 1 : Schéma de l'architecture du traitement humain de l'information des candidats

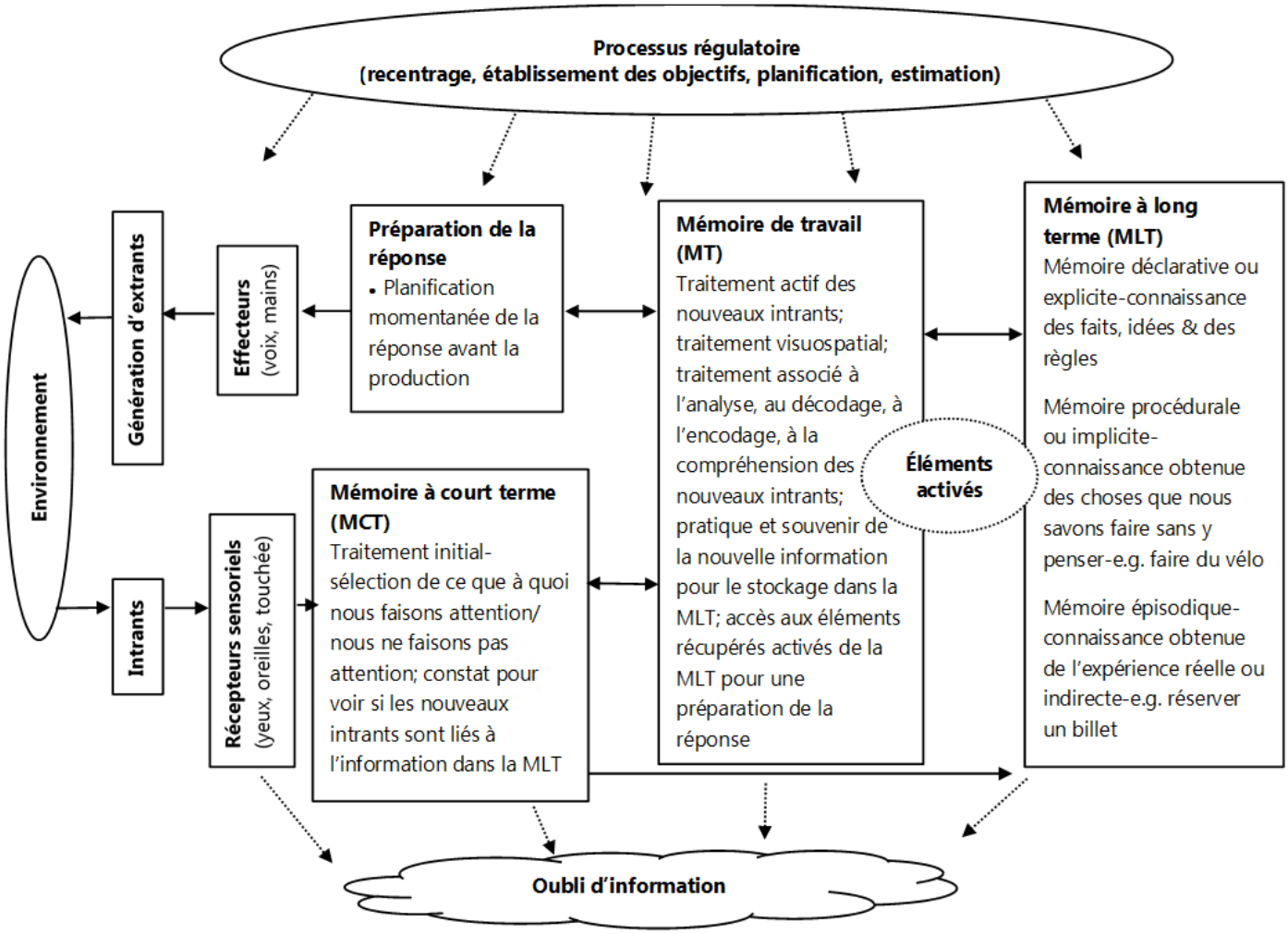


Figure 14 - L'architecture du traitement humain de l'information des candidats.

Source : Purpura, 2012 (Figure adaptée)

Annexe 2 : Schéma de l'interface de la compétence cognitive et du traitement de la L2 en évaluation

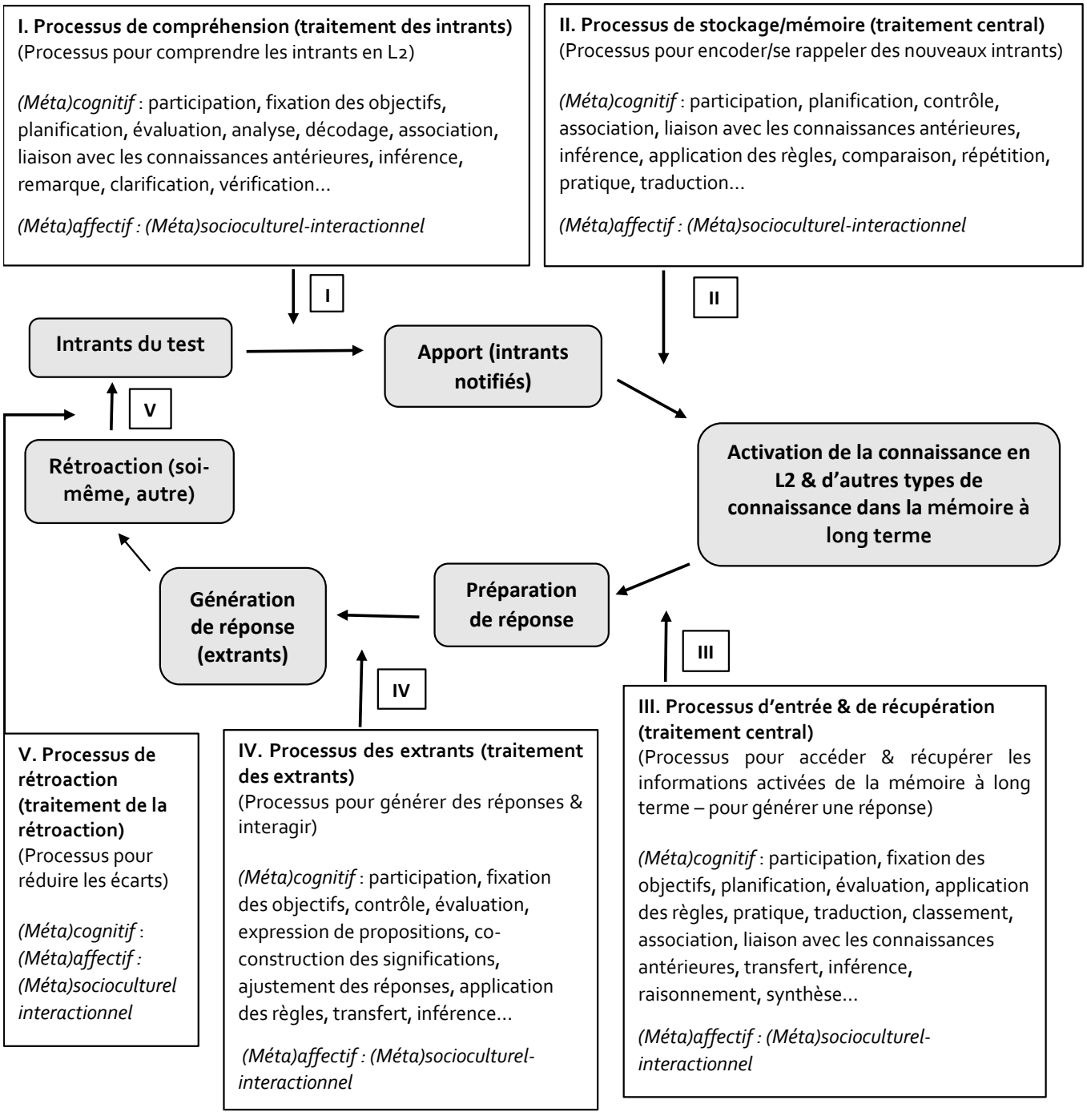


Figure 15 - L'interface de la compétence cognitive et du traitement de la L2 en évaluation.

Source : Purpura, 2012 (Figure adaptée)