Canadian Journal of Applied Linguistics
Revue canadienne de linguistique appliquée

Adossement des épreuves d'expression orale et écrite du Test de connaissance du français (TCF) sur les Niveaux de compétences linguistiques canadiens (NCLC) et correspondance avec les niveaux du Cadre européen commun de référence pour les langues (CECRL)

Vincent Folny

Volume 23, numéro 2, automne 2020

Special Issue: The Canadian National Frameworks for English and French Language Proficiency: Application, Implication, and Impact
Numéro spécial : Niveaux de compétence linguistique canadiens pour la compétence langagière en français et en anglais : impact, application et implication

URI : https://id.erudit.org/iderudit/1072969ar
DOI : https://doi.org/10.37213/cjal.2020.30437

Aller au sommaire du numéro

Éditeur(s)

University of New Brunswick

ISSN

1920-1818 (numérique)

Découvrir la revue

Citer cet article

Folny, V. (2020). Adossement des épreuves d'expression orale et écrite du Test de connaissance du français (TCF) sur les Niveaux de compétences linguistiques canadiens (NCLC) et correspondance avec les niveaux du Cadre européen commun de référence pour les langues (CECRL). *Canadian Journal of Applied Linguistics / Revue canadienne de linguistique appliquée*, *23*(2), 20–72. https://doi.org/10.37213/cjal.2020.30437

Résumé de l'article

En 2015, le CIEP a mené une étude afin de pouvoir adosser les productions écrites et orales du Test de connaissance de français (TCF) aux NCLC et d'établir une correspondance avec les niveaux du CECRL. In fine, Il s'agissait d'assurer aux candidats à l'immigration au Canada une bonne interprétation de leur niveau de compétence et de pouvoir expliquer les procédures mise en place pour l'interprétation des scores. Pour assurer l'adossement des épreuves d'expression écrites et orales du TCF aux niveaux NCLC, plusieurs procédures et études ont été mises en place : utilisation des niveaux CECR attribués à une sélection de productions au cours des corrections du TCF, séminaire organisé avec des panélistes pour attribuer des niveaux NCLC à ces mêmes productions, analyses psychométriques pour le calibrage de la sélection de productions, évaluation du lien entre les niveaux attribués avec les deux échelles (NCLC et CECR) afin de vérifier la convergence des résultats, analyses qualitatives des descripteurs NCLC et CECR et mise en correspondance.

Adossement des épreuves d'expression orale et écrite du Test de connaissance du français sur les Niveaux de compétences linguistiques canadiens et correspondance avec les niveaux du Cadre européen commun de référence pour les langues

Vincent Folny
France Education international

Résumé

En 2015, France Education international (FEI) a mené une étude afin de pouvoir adosser les productions écrites et orales du Test de connaissance du français (TCF) aux Niveaux de compétences linguistiques canadiens (NCLC) et d'établir une correspondance avec les niveaux du Cadre européen commun de référence pour les langues (CECRL). *In fine*, il s'agissait d'assurer aux candidats à l'immigration au Canada une bonne interprétation de leur niveau de compétence et de pouvoir expliquer les procédures mises en place pour l'interprétation des scores. Pour assurer l'adossement des épreuves d'expression écrites et orales du TCF aux niveaux NCLC, plusieurs procédures et études ont été mises en place : utilisation des niveaux CECRL attribués à une sélection de productions au cours des corrections du TCF, séminaire organisé avec des panélistes pour attribuer des niveaux NCLC à ces mêmes productions, analyses psychométriques pour le calibrage de la sélection de productions, évaluation du lien entre les niveaux attribués avec les deux échelles (NCLC et CECRL) afin de vérifier la convergence des résultats, analyses qualitatives des descripteurs NCLC et CECRL et mise en correspondance.

Abstract

In 2015, France Education international (FEI) conducted a study in order to be able to align the written and oral productions of the Test de connaissance du français (TCF) with the NCLC and to establish a correspondence with the levels of the CEFR. Ultimately, it was a question of ensuring candidates for immigration to Canada a good interpretation of their skill level and of being able to explain the procedures put in place for the interpretation of scores. To ensure the alignment of the written and oral expression tests of the TCF at the NCLC levels, several procedures and studies have been put in place: use of the CEFR levels attributed to a selection of productions during the TCF corrections, a seminar organized with panellists to assign NCLC levels to these same productions, psychometric analyses for the calibration of the selection of productions, evaluation of the link between the levels assigned with the two scales (NCLC and CECRL) in order to check the convergence of results, qualitative analyses of descriptors NCLC and CEFR and mapping.

Adossement des épreuves d'expression orale et écrite du Test de connaissance du français sur les Niveaux de compétences linguistiques canadiens et correspondance avec les niveaux du Cadre européen commun de référence pour les langues

Contexte d'émergence de l'étude

Au début des années 2000, FEI, un établissement public sous la tutelle du ministère de l'Éducation nationale en France, a développé un test de français, le TCF évaluant la langue française en compréhensions orale et écrite et expressions orale et écrite sur les 6 niveaux du CECRL (Conseil de l'Europe, 2001). En 2009, une étude a permis d'établir rigoureusement l'adossement du test aux niveaux du CECRL. Une vingtaine de panélistes avait été réunie et la méthode des marque-pages utilisée (Cizek & Bunch, 2007). Afin d'assurer l'utilisation idoine du TCF dans le cadre de l'immigration canadienne, il était nécessaire d'adosser les résultats aux NCLC (Centre des niveaux de compétence linguistique canadiens, 2012). Si une première étude a été menée pour les épreuves de compréhension, le présent article fait état de la seconde étude menée pour les épreuves d'expressions écrite et orale. Deux milles quinze est l'année durant laquelle FEI a mené cette étude.

Méthodologie

Pour assurer l'adossement des épreuves d'expression du TCF aux niveaux NCLC, plusieurs procédures et études ont été mises en place. Les experts de FEI ont procédé à une sélection de productions écrites et orales dont le niveau avait été évalué au cours des séances de corrections du TCF avec les niveaux du CECRL. Un séminaire a été organisé avec des panélistes pour attribuer des niveaux NCLC à ces mêmes productions. Des analyses psychométriques ont fait suite à ce calibrage de la sélection de productions afin d'obtenir des informations sur le niveau de fidélité des panélistes. Des analyses ont permis d'évaluer la force du lien entre les niveaux attribués indépendamment avec les deux référentiels (NCLC et CECRL) et de vérifier la convergence des résultats. Enfin, des analyses qualitatives des descripteurs NCLC et CECRL sont venus compléter l'alignement des productions avec les deux référentiels.

L'évaluation des épreuves d'expressions écrite et orale du TCF se faisant dans l'opérationnel à partir des niveaux du Cadre européen commun de compétence en langue (CECRL) (Conseil de l'Europe, 2009), il convenait de proposer une correspondance entre les échelles NCLC et CECRL. Pour l'expression écrite (TCF EE) et l'expression orale (TCF EO), il a été décidé de mettre en correspondance les scores exprimés sur une échelle de 20 points avec les niveaux NCLC. Le score exprimé sur une échelle de 20 points est la traduction en 20 catégories de la fréquence des niveaux CECRL obtenus pour chacune des 3 tâches TCF EO ou TCF EE à l'issue des doubles corrections (soit 6 observations par candidat). Ces catégories traduisent des niveaux CECRL clairement identifiés et des niveaux intermédiaires qui constituent un découpage plus fin de la compétence en langue que celui des 6 niveaux du CECRL. Cette distribution progressive des niveaux permet d'utiliser des performances qui ne sont pas toutes « typiques » des niveaux du CECRL et donne l'occasion de positionner les niveaux NCLC avec plus de précision et de pertinence. Cela permet encore d'éviter que les panélistes cherchent à établir en priorité une

correspondance entre les 6 niveaux du CECRL et les 12 niveaux NCLC alors que l'activité qui leur était demandée consistait à évaluer le niveau de performance des candidats en utilisant les descripteurs NCLC.

Sélection des productions orales et écrites évaluées avec les niveaux du CECRL

Pour le TCF EO, 21 productions orales ont été sélectionnées parmi 15573 productions ayant bénéficié d'une double correction indépendante (recueil de données du premier septembre deux mille quatorze au cind février deux mille quinze). Pour le TCF EE, 21 productions écrites ont été sélectionnées parmi 5795 productions ayant reçu une double correction indépendante (recueil de données du premier septembre deux mille quatorze au cind février deux mille quinze).
Les productions ont été sélectionnées selon les principes suivants :

1. Productions issues de passations réelles ;
2. Productions corrigées par les correcteurs les plus consensuels (taux d'accords exacts les plus élevés) ;
3. Productions présentant des profils de scores similaires entre le correcteur 1 et le correcteur 2 ;
4. Les 21 productions correspondent aux 21 catégories de scores attribuables pour le TCF (de 0 à 20).

Certaines productions sont clairement positionnées sur l'un des 6 niveaux du CECRL ; d'autres se situent à la frontière de deux niveaux adjacents (ex. : A1/A2). Les productions apparaissent telles qu'elles ont été collectées à l'issue des passations sans aucune modification. Les panélistes ignorent le niveau final déjà attribué aux candidats.

Cadre de référence utilisé pour adosser les productions du TCF aux niveaux NCLC

Les procédures utilisées pour adosser les productions du TCF aux NCLC et établir la correspondance avec les niveaux du CECRL sont pour l'essentiel empruntées aux méthodes référencées dans les ouvrages traitant de la définition des scores de césure et de l'étalonnage (Cizek & Bunch, 2007; Figueras et coll., 2009; Hambleton & Pitoniak, 2006; Zieky et coll., 2008). Ces mêmes ouvrages et l'article de Hambleton & Pitoniak ont servi à cerner les enjeux de l'adossement des examens aux référentiels de compétence et à repérer les points critiques à considérer du point de vue méthodologique. Le *Manuel pour relier les examens au CECRL* (Conseil de l'Europe, 2009) a été consulté pour la phase de familiarisation avec l'échelle NCLC et ses descripteurs. L'ouvrage de Figueras et coll. a été consulté pour la somme d'expériences significatives qu'il représente. L'ouvrage *Niveaux de compétence linguistique canadiens Français langue seconde pour adultes* (Centre des niveaux de compétence linguistique canadiens, 2012) a servi de référence pour l'appropriation des échelles et descripteurs NCLC durant toute la procédure d'adossement et de mise en correspondance. La méthodologie utilisée pour assurer l'alignement des productions du TCF sur les niveaux NCLC traduit les choix faits parmi les options théoriques proposées dans la littérature scientifique et professionnelle. L'auteur de cet

article a préféré alléger la lecture en centrant l'explicitation des choix théoriques dans la méthodologie.

Panélistes pour le TCF EE et TCF EO

Douze panélistes ont été sélectionnés parmi les experts du département Évaluation et Certifications (DEC) à FEI. En moyenne, les panélistes ont 42 ans. Plus des 2/3 sont des femmes. Ils ont en moyenne 11 ans d'ancienneté dans le domaine du français comme langue étrangère / seconde. Un tiers des panélistes a déjà participé à des procédures de définition des scores de césure. Ils sont titulaires d'un diplôme de niveau master 2 en didactique des langues ou bien en sciences du langage. Tous utilisent quotidiennement des échelles du CECRL. Certains sont familiarisés à l'utilisation de plusieurs référentiels de compétence en langue.

Familiarisation avec les niveaux NCLC, procédure suivie

Pour la familiarisation, la formation des panélistes et la procédure d'alignement (Tableau 1), les actions mises en place sont inspirées de celles qui sont recensées et décrites dans le « Manuel pour relier les examens de langues au CECRL » (Conseil de l'Europe, 2009) et par Hambleton & Pitoniak (2006).

Tableau 1

Actions de formations et procédure d'alignement retenue

Action	Descriptif	Attendus
Formation initiale (une journée)	Définition des scores de césure et méthodologie retenue Appropriation des descripteurs NCLC	Montée en compétence des participants
Tour 1	Travail individuel : classement et attribution d'un niveau NCLC aux productions	Jugements indépendants
Tour 2	Analyses des résultats du tour 1 et discussions par paire de panélistes avant un nouveau classement et évaluation des productions	Confronter le jugement des panélistes pour stimuler le consensus.
Tour 3	Analyse des résultats du tour 2 et des discussions en grand groupe avant un ultime classement et une dernière évaluation des productions. Prise en compte possible de l'impact des résultats pour les candidats.	Favoriser un large consensus et prendre en considération un éventuel impact des résultats pour les candidats

Analyses effectuées après chacun des tours et graphiques proposés aux panélistes

Pour les analyses (Tableau 2), les « meilleures pratiques » recensées dans la littérature scientifique pour le traitement de la fidélité inter et intracorrecteur ont été privilégiées (Linacre, 2013; Linacre & Wright, 1994; Stemler & Tsai, 2008).

Tableau 2

Indices et valeurs retenus pour l'analyse de la fidélité inter et intrapanélistes

Type d'analyses	Indices	Valeurs visées
Estimation du consensus	Accords exacts	>70%
	Kappa de Cohen	>0,5
Indices de cohérence interne	Corrélation de Pearson/ W de kendall	>0,7
	Alpha de Cronbach	>0,7
Estimation de la qualité de la mesure	Variance expliquée	>70%
	Indices d'adéquation des données au modèle (indices infit)	>0,5 <1,5

À l'issue du premier tour, les accords exacts et adjacents (+ ou - 1 niveau NCLC) ont été calculés. Ces calculs d'accords exacts ont été établis à partir de la médiane du classement des productions et du niveau médian attribué par les 12 panélistes à chacune des 21 performances orales et écrites. Le W de Kendall a été utilisé ainsi qu'une analyse en composantes principales.

Après chaque tour, des représentations graphiques présentant la distribution des jugements, selon que les productions étaient classées ou évaluées avec les NCLC, ont été présentées aux panélistes (classement final des productions et niveaux attribués à chacune des productions par les 12 panélistes).

Analyse des jugements des panélistes avec un modèle multifacette de Rasch

Depuis quelques années, il est d'usage d'analyser le jugement des panélistes pour vérifier leur cohérence et leur fidélité. En Europe, le *Manuel pour Relier les examens de langues au CECRL* (Conseil de l'Europe, 2009) propose d'entreprendre cette démarche en faisant appel à un modèle multifacette de Rasch. Ailleurs, de nombreux auteurs ou chercheurs invitent à entreprendre cette démarche pour évaluer la qualité des données (Engelhard, 2009, 2011, 2013 ; Engelhard & Gordon, 2000 ; Kaliski et coll., 2013 ; Kecker & Eckes, 2010 ; Lunz, 2000). Nous avons pris la décision d'analyser séparément les jugements des panélistes émis pour l'expression orale et écrite. Il s'agit de pouvoir utiliser l'estimation du niveau des productions après ajustement pour le niveau de sévérité / générosité des panélistes. Pour faire ces analyses, le logiciel FACETS a été utilisé dans sa version 3.71.3 (Linacre, 2013). Les ouvrages de Eckes (2011) et Engelhard (2013) ont servi de référence pour la méthodologie. Pour l'analyse, les recommandations de Linacre

(Linacre, 2013, p. 267) concernant l'utilisation des indices d'adéquation des données au modèle ont été retenues :

Interprétation des carrés moyens *infit* / *outfit* :

- >2 dégrade la qualité du système de mesure
- 1,5 – 2,0 inutile pour la construction de l'échelle de mesure, mais ne la dégrade pas
- 0,5- 1,5 utile pour la construction de l'échelle de mesure
- < 0,5 moins utile, mais ne dégrade pas la qualité de l'échelle de mesure. Peut produire de bons indices de fidélité qui induisent en erreur.

Toutefois, les données n'étant pas indépendantes (les panélistes ont interagi après le premier tour et les productions sont restées les mêmes durant toute la procédure), pour vérifier l'adéquation des données au modèle (indices *infit* et *outfit*), les valeurs au-delà de 1,5 ont été retenues comme indiquant un problème. Ces valeurs positives signalent plus de variance qu'attendu. Les valeurs inférieures à 0,5 ne sont pas analysées comme problématiques (Engelhard & Gordon, 2000). L'indépendance des données n'ayant pas été respectée, il est normal d'avoir des indices d'adéquation au modèle signalant moins de variance qu'attendu.

Pour l'interprétation des indices concernant la fidélité, il convient de faire attention à l'orientation de la facette (Eckes, 2011) :

- Pour la facette « candidat » (expression écrite ou orale), un indice de fidélité proche de 1 signale que le niveau de chaque candidat est clairement identifié et distingué de celui des autres candidats. Ici, c'est la sensibilité de l'instrument qu'on cherche à apprécier. Une valeur proche de zéro signifierait une absence de distinction de niveau entre les candidats ce qui est peu désirable.
- Pour les facettes « tour » et « panélistes », l'orientation est inverse. Une valeur proche de zéro signale qu'il n'y a pas de différence entre les tours et les panélistes. Dans l'idéal, il ne faudrait pas observer de changements de niveau drastiques entre les tours. Pour les panélistes, une fidélité proche de zéro signale que les panélistes sont interchangeables. Cette absence d'impact de la sévérité / générosité du panéliste sur les données est précisément ce qu'on cherche à observer.

Mise en correspondance des descripteurs du CECRL et NCLC

En parallèle de la procédure de définition des scores de césure, une analyse qualitative des descripteurs des deux échelles a été menée. Quatre experts de FEI ont mis en correspondance les descripteurs du CECRL et des NCLC pour les productions orales et écrites. Les experts ont été invités à sélectionner les éléments clés dans les descripteurs du CECRL et des NCLC permettant d'établir une correspondance entre les deux échelles. Chacun des experts a consigné dans un tableur la correspondance proposée à partir de l'analyse des descripteurs ainsi que les extraits des descripteurs permettant de justifier cette correspondance. L'objectif consistait à orienter les experts vers une démarche argumentative. Il s'agissait aussi de garder des traces des décisions prises et de pouvoir juger de la pertinence de l'alignement. Pour le CECRL, les échelles générales de

productions écrites et orales ont été utilisées. Pour les NCLC, non seulement les descripteurs de l'écrit et de l'oral ont été utilisés, mais également les descripteurs de compétences clés et les exemples de tâches.

Mécanisme de la validation définitive

À l'issue du troisième tour, les résultats ont été validés par une commission composée d'experts à FEI. Geisinger & McCormick (2010) expliquent qu'après une procédure de définition des scores de césure, il est normal de procéder à des ajustements de nature plus « politiques ». Nous pensons qu'une telle procédure est également nécessaire lorsqu'on adosse des examens à une échelle de niveau. Cette nature plus « politique » de l'ajustement est d'autant plus nécessaire qu'un autre test de français, le Test d'évaluation de français (TEF) est déjà adossé à cette échelle (Crendal, 2005; Demeuse et coll., 2004). Pour les candidats, pour les autorités canadiennes, il serait malaisé d'avoir deux tests avec des correspondances divergentes pour les deux échelles de niveau. Si tel devait être le cas, il conviendrait alors que les responsables de FEI en soient informés. Les prises de décision impliqueraient alors de forts impacts. Si des écarts « faibles » étaient constatés (ce qui est attendu), il s'agirait de trouver un équilibre entre :

- le besoin de respecter la validité scientifique de l'adossement,
- le besoin d'harmonisation de l'adossement des deux tests aux NCLC.
- Avant de commencer l'étude, FEI a fait le choix de prendre une décision définitive pour l'adossement :
- en fonction des résultats de l'évaluation des productions avec les descripteurs NCLC ;
- à la lumière de la validité convergente entre les niveaux attribués par les correcteurs (CECRL) et les panélistes (NCLC);
- en considérant les résultats de l'étude qualitative de la mise en correspondance des descripteurs du CECRL et des NCLC ;
- et enfin, en dernier lieu, en comparant les résultats avec ceux établis pour le TEF.

Il est attendu que les différences de classement et de niveaux les plus importantes sont liées à trois facteurs :

1. Le nombre de niveaux à distinguer sur les échelles (plus ce nombre est grand, moins le taux d'accords exacts attendus est élevé) ;
2. Le positionnement du niveau de la production sur l'échelle (ex. : une production à la frontière de deux niveaux est plus difficile à classer) ;
3. À la connaissance, maîtrise et compréhension des niveaux NCLC chez les panélistes.

Résultats

Analyses de la fidélité après chacun des 3 tours

Comme renseigné dans la littérature scientifique (Cizek & Bunch, 2007), le premier tour, est celui qui présente le nombre de jugements discordants le plus important.

Tour 1

Expression orale tour 1

Les moyennes des accords exacts lorsque les productions sont classées par ordre croissant de compétence (Annexe A) ou évaluées à partir des niveaux NCLC sont respectivement de 45% (Annexe B) et 54% (Annexe C). L'examen attentif du classement des productions (Annexe C) et de leur évaluation à partir des niveaux NCLC (Annexe D) laisse apparaître que les plus grosses difficultés sont concentrées sur les productions au milieu de l'échelle de compétence.

L'examen des principaux indices statistiques (Tableau 3) montre, dès le premier tour, une bonne cohérence interne des jugements des panélistes. Les accords exacts ont une valeur de 54% ce qui n'est pas encore suffisant pour considérer le consensus entre les panélistes comme fort (la valeur attendue est de 70% [Stemler & Tsai, 2008]). En revanche, l'alpha de Cronbach et le pourcentage de variance expliquée ont des valeurs relativement élevées. Il est raisonnable de considérer que les jugements des panélistes sont homogènes (Stemler & Tsai, 2008). À titre de comparaison, la société Pearson rapporte un pourcentage d'accords adjacents à l'issue du premier tour d'un séminaire d'adossement des scores du PTE (Pearson test of English) aux niveaux du CECRL pour l'expression écrite de 86% (De Jong, 2013). Toutefois, il convient de préciser que les participants faisaient connaître leurs jugements en levant une petite ardoise alors que, pour le TCF, le vote était totalement secret et les observations indépendantes.

Tableau 3
Indices statistiques du taux de fidélité des panélistes (expression orale, tour 1)

Consensus	Cohérence interne
Accords exacts (médiane) : 54%	Alpha de Cronbach : 0,76
Accords exacts et adjacents : 91%	% variance expliquée (ACP) : 95,93%
W de Kendall : 0,96 (sig. À 0,01)	

À l'issue du premier tour, pour l'expression orale, on peut considérer que les jugements sont cohérents, mais qu'ils manquent de précision. Les panélistes partagent une définition des niveaux de compétence NCLC uniforme, mais ils ont encore besoin de discuter de leur interprétation de l'échelle. L'examen des niveaux NCLC attribués aux productions orales (Annexe D) montre que certaines d'entre elles ont posé des difficultés aux panélistes. En effet, parfois, jusqu'à 6 niveaux différents ont été attribués à une même

production. Cela étant, lorsqu'on observe une telle étendue dans l'attribution des niveaux, la majorité de panélistes attribuent un niveau identique.

Concernant la fidélité des panélistes, dans l'ensemble (Annexe C, Annexe D), on peut déduire des analyses que les panélistes partagent une vision relativement homogène. Toutefois, on constate que l'étendue des niveaux d'accords exacts (Annexe A, Annexe B) varie entre 29% et 76%. Sans doute que les panélistes les moins « consensuels » bénéficieraient d'une discussion avec un de leurs homologues pour mieux interpréter les niveaux NCLC.

Mutatis mutandis, on peut comparer ces résultats avec ceux obtenus pour le TestDaf en Allemagne (Kecker & Eckes, 2010). Huit panélistes ont évalué 9 productions sur 9 niveaux. Alors que dans les phases d'entrainement le taux d'accords exacts était de 75%, il était de 32% au moment d'évaluer les productions ayant servi à définir les scores de césures (les jugements sont indépendants). On peut remarquer que les résultats obtenus lors du premier tour de l'expression orale (54% d'accords exacts) avec 12 panélistes, 21 productions et 12 niveaux sont conformes aux niveaux de fidélité renseignés dans la littérature scientifique et les retours d'expérience.

À l'issue du premier tour, chaque panéliste a été informé des taux moyens d'accords exacts et de leur propre taux d'accords exacts. Ceux qui avaient des jugements divergents ont ainsi pu prendre conscience de leur idiosyncrasie et ont pu les réviser avant le tour suivant. Ils ont eu ainsi l'occasion d'ajuster leur interprétation des niveaux NCLC.

Expression écrite tour 1

Lorsque les productions sont classées par ordre croissant de compétence (Annexe E) ou évaluées à partir des niveaux NCLC (Annexe F), les accords exacts sont respectivement de 35% (Annexe G) et 49% (Annexe H). L'examen attentif du classement des productions (Annexe E) et de leur évaluation avec les niveaux NCLC (Annexe F) laisse apparaître que les difficultés significatives sont situées sur les productions en dehors des extrémités de l'échelle de compétence. L'examen des principaux indices statistiques (Tableau 4) montre que dès le premier tour le taux de consensus et le niveau de cohérence interne est substantiel.

Tableau 4

Indices statistiques fidélité des panélistes (expression écrite, tour 1)

Consensus	Cohérence interne
Accords exacts (médiane) : 49%	Alpha de Cronbach : 0,88
Accords exacts et adjacents : 83%	% de variance expliquée (ACP) : 92,54%
W de Kendall : 0,92 (sig. à 0,01)	

Si on constate que les accords exacts ont une valeur de 49% (le consensus n'est pas encore suffisant), l'alpha de Cronbach et le pourcentage de variance expliquée ont des valeurs adéquates pour considérer le jugement des panélistes comme consistant (Stemler & Tsai, 2008).

À l'issue du premier tour pour l'expression écrite, on peut considérer que, dans l'ensemble, les jugements des panélistes sont cohérents, mais qu'ils manquent encore de précision. Autrement dit, les panélistes partagent la même définition des niveaux de compétence NCLC, mais ils doivent encore discuter de leur interprétation. L'examen des niveaux attribués aux productions (Annexe F) montre que certaines productions posent des difficultés d'appréciation. Parfois, 5 à 6 niveaux différents ont été attribués à une même production. Contrairement à l'expression orale, il est difficile de dégager une tendance majoritaire dans l'attribution des niveaux de certaines performances. Pour cette compétence, il est indispensable de disposer d'un deuxième tour pour pouvoir définir l'adossement.

Concernant la fidélité, les panélistes partagent une vision plus homogène de l'estimation du niveau des productions avec les descripteurs NCLC que de leur simple classement (Annexe G, Annexe H). Pour l'estimation du niveau des productions avec les niveaux NCLC, l'étendue des accords exacts (Annexe H) fluctue de 10% à 76%. Les panélistes les moins « consensuels » auront besoin de la discussion avec un autre panéliste pour confronter leur interprétation des niveaux NCLC.

Comme pour l'expression orale, il est possible de comparer ces résultats avec ceux du TestDaF (Kecker & Eckes, 2010). On peut alors remarquer que les résultats obtenus lors du premier tour de l'expression écrite (49% d'accords exacts) avec 12 panélistes, 21 productions et 12 niveaux sont conformes aux niveaux de fidélité renseignés dans les écrits scientifiques. À l'issue du premier tour, chaque panéliste a été informé des taux d'accords exacts. Certains panélistes ont ainsi pu prendre conscience de l'idiosyncrasie de leurs jugements.

Tour 2

Pour le tour 2 (et également le tour 3), l'analyse a été faite à partir des accords exacts et adjacents. Les autres indices de fidélité ne sont pas utilisés, car ils supposent une indépendance des jugements. Le taux d'accords exacts et adjacents permettra d'apprécier la qualité générale du processus, le niveau de consensus et de porter un diagnostic sur le comportement des panélistes. Pour le deuxième tour, il a été rappelé aux panélistes qu'ils devaient toujours faire leurs choix en toute indépendance.

Expression orale tour 2

À l'issue du deuxième tour, on observe une forte homogénéité entre le classement des productions orales aux tours 1 et 2 (Annexe C, Annexe I). Dix-neuf productions sur 21 sont classées dans le même ordre de difficulté. La corrélation entre le classement des productions au tour 1 et tour 2 à partir de l'utilisation des descripteurs NCLC a une valeur de 0,998. Le deuxième tour a été une occasion pour les panélistes d'affiner leurs jugements (Annexe D, Annexe J). Cet ajustement est observable dans les écarts types et erreurs standards de mesure associés à chacune des productions. L'erreur standard de mesure moyenne est passée de 0,34 à 0,13 au deuxième tour. Toujours au deuxième tour (Annexe J), les différences d'appréciation pour chacune des productions ne dépassent pas une étendue de 3 niveaux. Les niveaux de 6 productions sur 21 font l'unanimité (100% d'accords). Le niveau moyen des productions est stable puisqu'il est de 6,80 au tour 1 et de

6,79 au tour 2. L'écart type associé à cette moyenne au tour 1 est de 3,6 et de 3,7 au tour 2. Les ajustements effectués à l'issue du premier tour ont permis aux panélistes d'être plus précis sans modifier radicalement les résultats.

Concernant la fidélité des panélistes, on constate une augmentation significative des accords exacts et adjacents (Annexe K, Annexe L), ce, aussi bien pour le classement des productions que pour l'estimation de leur niveau NCLC. Alors que la moyenne des accords exacts des productions évaluées avec les NCLC était de 54% au premier tour (Annexe A), elle est de 86% au deuxième tour (Annexe L). Dès le deuxième tour, le taux d'accord est supérieur à celui requis pour valider un niveau de fidélité adéquat (Stemler & Tsai, 2008). Dès le deuxième tour, un consensus s'est formé quant aux « niveaux NCLC » des productions.

Expression écrite tour 2

Après le tour 2, on observe une bonne homogénéité entre le classement des productions écrites du tour 1 et 2 (Annexe E, Annexe M). 11 productions occupent un rang identique, 7 occupent un rang adjacent, 2 deux rangs en moins et 1 trois rangs en moins. Lorsque les descripteurs NCLC sont utilisés, la corrélation entre le classement des productions au tour 1 et tour 2 a une valeur de 0,991. Comme pour l'oral, le deuxième tour a constitué une occasion pour les panélistes d'affiner leurs jugements portant sur le « niveaux NCLC » des productions (Annexe F et Annexe N). Cet ajustement est observable dans les écarts types et erreurs standards de mesure associés à chacune des productions évaluées avec les descripteurs NCLC. Alors que l'écart type moyen des 21 productions est de 1 au tour 1, il est de 0,44 au tour 2. L'erreur standard de mesure moyenne est passée de 0,30 à 0,13. Le niveau moyen des productions est relativement stable puisqu'il est de 7,1 au tour 1 et de 7,4 au tour 2. L'écart type associé à cette moyenne au tour 1 est de 2,9 aux tours 1 et 2. Enfin, l'erreur standard de mesure qui est de 0,65 au tour 1 a une valeur de 0,66 au tour 2. Les ajustements effectués après le premier tour ont permis aux panélistes d'être plus précis sans avoir à modifier radicalement les jugements portés au premier tour.

Concernant la fidélité des panélistes, on constate une augmentation significative des accords exacts et adjacents (Annexe O, Annexe P) pour le classement des productions ou leur positionnement effectué avec les descripteurs NCLC. Alors que les accords exacts sont de 49% au premier tour (Annexe H), ils sont de 77% au deuxième tour (Annexe P). Dès le deuxième tour, le taux d'accord est supérieur à celui requis pour valider un niveau de fidélité suffisant des panélistes (Stemler & Tsai, 2008).

Tour 3

Après le deuxième tour, les panélistes ont été réunis pour une présentation des résultats. Ils ont été invités à réexaminer leurs jugements les moins consensuels.

Expression orale tour 3

Après le troisième tour, on observe une homogénéité parfaite entre le classement des productions orales aux tours 2 et 3 (Annexe I et Annexe Q). Pour les panélistes, le troisième tour a essentiellement été une occasion d'affiner les niveaux NCLC attribués aux

productions (Annexe J, Annexe R). Cet ajustement est observable dans les écarts types et erreurs standards de mesure associés à chacune des productions. L'erreur standard de mesure est ainsi passée de 0,13 à 0,06 au tour 3. Au troisième tour (Annexe R), on note que les différences d'appréciation pour chacune des productions ne dépassent pas 1 niveau et que les niveaux de 13 productions sont unanimement reconnus. Le niveau moyen des productions a légèrement augmenté puisqu'il était de 6,79 au tour 2 et qu'il est de 6,85 au tour 3. L'écart type associé à cette moyenne aux tours 2 et 3 est de 3,7. Pour l'oral, les trois tours n'ont eu aucun impact significatif sur l'estimation du niveau moyen des 21 productions orales. Les ajustements qui ont eu lieu à l'issue du troisième tour ont permis aux panélistes d'être plus précis sans modifier radicalement les résultats issus du premier et deuxième tour.

À l'issue du troisième tour, le consensus entre les panélistes est fort (Annexe S, Annexe T). Les accords exacts pour le classement des productions et l'estimation de leurs niveaux NCLC ont une valeur de 89% et 91%. Les accords exacts et adjacents sont de 99% et de 100%. De nombreux panélistes ont un taux d'accords exacts ou exacts et adjacents de 100%.

Expression écrite tour 3

Après le troisième tour, on observe une forte homogénéité entre le classement des productions écrites aux tours 2 et 3 (Annexe L, Annexe U). La corrélation entre le positionnement des productions (niveaux NCLC) au tour 2 et 3 est de 0,998. À l'instar de l'expression orale, le troisième tour a permis aux panélistes d'affiner l'attribution des niveaux NCLC aux productions (Annexe N et Annexe V). Cet ajustement est observable dans les écarts types et les erreurs standards de mesure associés à chacune des productions évaluées avec les descripteurs NCLC. Alors que l'écart type moyen est de 0,44 au tour 2, il est de 0,14 au tour 3. L'erreur standard de mesure moyenne est passée de à 0,13 au tour 2 à 0,04 au tour 3. Au troisième tour (Annexe V), on note que les différences d'appréciation pour chacune des productions ne dépassent pas un niveau. Les niveaux de 13 productions sur 21 sont unanimement reconnus (100% d'accords exacts). Le niveau moyen des productions est stabilisé à 7,4 au tour 2 et au tour 3. L'écart type associé à cette moyenne au tour 2 est de 2,95 et de 2,97 au tour 3. Enfin, l'erreur standard de mesure a conservé la même valeur à 0,66. Pour l'écrit comme pour l'oral, les trois tours n'ont pas eu d'impact sur l'estimation du niveau moyen des 21 productions.

Les ajustements à l'issue du troisième tour ont permis aux panélistes d'être plus précis sans modifier radicalement les résultats issus du premier et deuxième tour. À l'issue du troisième tour, le consensus entre les panélistes est fort (Annexe W et Annexe X). En effet, les accords exacts pour le classement des productions en fonction de leur niveau et le classement des productions en utilisant les descripteurs NCLC ont respectivement une valeur de 82% et 92%. Les accords adjacents sont de 98% et 100%.

Expression orale : analyse du jugement des panélistes avec un modèle multifacette de Rasch

Pour l'expression orale (Annexe A), deux panélistes (8 et 9) ont dû être retirés de l'échantillon leurs indices d'adéquation des données au modèle (*infit* et *outfit*) ayant des

valeurs égales ou supérieures à 1,5. Après retrait de ces deux panélistes, une augmentation de la qualité des résultats est observable. Le taux de variance unique expliquée est passé de 95% à 98%. Les deux panélistes 5 et 1 présentent des indices de l'*infit* à la limite inférieure (<0,5) (Tableau 3). Il a décidé de les garder dans l'échantillon, car leur présence ne dégrade pas la qualité de la mesure. À l'issue du calibrage, pour les panélistes, la valeur moyenne de l'*infit* est de 0,81 et pour l'*outfit* de 0,89. Ces valeurs sont satisfaisantes.

Tableau 5

Adéquation des données au modèle pour les panélistes (calibrage final expression orale)

Panéliste	Score total	N observations	Mesure moyenne	Mesure ajustée	Mesure en logit	S.E.	Infit Carrés moyens	Infit standardisé	Outfit Carrés moyens	Outfit standardisé
2	414	63	6,57	6,12	0,9	0,23	1,45	2,07	1,36	1,1
10	437	63	6,94	6,61	-0,34	0,24	1,44	1,92	1,18	0,65
11	440	63	6,98	6,69	-0,5	0,23	0,85	-0,67	0,7	-0,87
3	432	63	6,86	6,47	-0,06	0,23	0,75	-1,25	1,13	0,49
4	442	63	7,02	6,75	-0,61	0,23	0,75	-1,24	0,74	-0,69
7	426	63	6,76	6,34	0,27	0,23	0,7	-1,51	0,59	-1,43
6	430	63	6,83	6,43	0,05	0,23	0,66	-1,75	0,56	-1,66
12	431	63	6,84	6,45	-0,01	0,23	0,55	-2,47	1	0,11
5	419	63	6,65	6,2	0,64	0,23	0,48	-3,1	0,42	-2,27
1	437	63	6,94	6,61	-0,34	0,24	0,47	-3,1	1,19	0,69

Pour l'expression orale, le « niveau » moyen de chacun des 3 tours est strictement identique (indice de fidélité à 0) (Annexe Y). L'indice de séparation « *strata* » (qui permet de savoir combien de groupes statistiquement différents sont détectés) indiqué par le logiciel FACETS a une valeur de 0,33. Les 21 candidats ont des niveaux de compétences clairement différents (indice *strata* =29,62 ; fidélité =1). Pour ce qui est des panélistes, leur niveau de sévérité/ générosité est fortement semblable (indice *strata* = 2,63). Toutefois, entre panélistes, il existe des différences d'appréciation notables. L'indice de fidélité a une valeur de 0,75. Si les panélistes ne sont pas « interchangeables », leur niveau de cohérence interne est important. En effet, on ne peut scinder le groupe de panélistes qu'en trois groupes de niveau de sévérité différents. Par ailleurs, comme en atteste la carte des tours, candidats, panélistes et niveaux (Annexe Y), la différence de sévérité n'est pas importante entre les panélistes. Pour avoir une idée exacte du niveau de précision des panélistes, l'étendue entre le panéliste le plus sévère et le plus généreux est de 1,51 *logit*. Cette étendue représente uniquement 3,7% de l'étendue totale de l'échelle de mesure (Annexe Y).

L'examen des seuils et catégories des « niveaux NCLC », proposé par FACETS, nous apprend que pour l'oral, les niveaux NCLC sont significativement différents les uns des autres (Tableau 6). Les seuils (endroits sur l'échelle où se produit un changement de niveau) sont correctement ordonnés (cf. *Expectation measure at category* / *Rasch-Thurstone Thresholds*).

Tableau 6

Fonctionnement des NCLC pour les 21 productions d'expression orale

DATA				QUALITY CONTROL		RASCH-ANDRICH	EXPECTATION	MOST	RASCH-	Cat	Response
Category	Counts		Cum.	Avge	Exp. OUTFIT	Thresholds	Measure at	PROBABLE	THURSTONE	PEAK	Category
Score	Total	Used	%	%	Meas	MnSq	Measure	S.E.	Category -0.5	from	Thresholds	Prob	Name
0	29	29	5%	5%	-21.99	-21.93	.3		(-19.59)	low	low	100%	
1	27	27	4%	9%	-13.73	-13.69	1.0	-18.53	.92	-15.44	-18.51	-18.53	-18.53	91%	NCLC 1
2	35	35	6%	14%	-10.80	-10.77	.8	-12.54	.35	-10.94	-12.61	-12.54	-12.57	71%	NCLC 2
3	33	33	5%	20%	-8.23	-8.02	.9	-9.34	.31	-8.27	-9.47	-9.34	-9.40	59%	NCLC 3
4	58	58	9%	29%	-5.50	-5.42	.6	-7.24	.28	-5.12	-7.02	-7.24	-7.15	80%	NCLC 4
5	53	53	8%	37%	-1.08	-1.00	.6	-3.08	.33	-1.72	-3.18	-3.08	-3.13	66%	NCLC 5
6	77	77	12%	50%	1.09	.94	.8	-.39	.21	1.32	-.29	-.39	-.35	74%	NCLC 6
7	29	29	5%	54%	3.44	3.55	.8	3.13	.27	3.69	2.74	3.13	2.92	46%	NCLC 7
8	52	52	8%	62%	5.81	5.93	.7	4.28	.28	5.61	4.56	4.28	4.43	63%	NCLC 8
9	56	56	9%	71%	8.17	7.96	1.0	6.82	.23	8.06	6.82	6.82	6.81	63%	NCLC 9
10	45	45	7%	78%	10.30	10.23	.6	9.31	.24	10.38	9.25	9.31	9.27	58%	NCLC 10
11	52	52	8%	87%	13.29	13.34	1.4	11.41	.26	13.75	11.63	11.41	11.49	84%	NCLC 11
12	84	84	13%	100%	18.94	18.92	.9	16.16	.34	(17.24)	16.17	16.16	16.16	100%	NCLC 12
										(Mean)	(Modal)	(Median)			

Expression écrite : analyse du jugement des panélistes avec un modèle multifacette de Rasch

Pour l'expression écrite (Annexe Z), un seul panéliste a un indice *outfit* (carrés moyens) supérieur à 1,77 (panéliste 3). Ce panéliste a été retiré de l'échantillon, mais, après retrait, nous n'avons observé aucune amélioration notable (la variance unique explique toujours 95,5% de la variance totale).

Le « niveau » de chacun des tours n'est pas identique (indice de fidélité d'une valeur de 0,92). L'indice « *strata* » à une valeur de 4,71. Alors qu'au premier tour le niveau moyen des productions était de 7,07, au deuxième tour, il est de 7,37 et 7,39 au dernier tour. Si la différence de niveau entre les tours est significative, en revanche, ces écarts ne sont pas importants (écart maximal de 0,39 soit moins de la moitié d'un niveau NCLC). L'ajustement a eu lieu uniquement après le premier tour. Aucune différence n'est observable entre les tours 2 et 3 (Annexe Z).

Les 21 candidats ont des niveaux de compétences clairement différents (indice *strata* =21,05 ; fidélité =1). Pour ce qui est des panélistes (Tableau 5), leur niveau de sévérité/ générosité est fortement semblable (indice *strata* = 3,14). Toutefois, il existe des différences d'appréciation comme le signale l'indice de fidélité (valeur de 0,82). Comme pour l'oral, si les panélistes ne sont pas entièrement interchangeables, leur niveau de cohérence interne est important.

À l'issue du calibrage final, tous les panélistes ont des indices d'adéquation des données au modèle satisfaisant (Tableau 5). La moyenne des *infit* est de 0,85 et celles de *outfit* de 0,89. Ces valeurs sont satisfaisantes pour pouvoir dire que le groupe de panélistes jouit d'un niveau de fidélité suffisant à l'issue du calibrage des productions.

Tableau 7

Adéquation des données au modèle pour les panélistes (calibrage final expression écrite)

Panéliste	Score total	N observations	Mesure moyenne	Mesure ajustée	Mesure en logit	S.E.	Infit Carrés moyens	Infit standardisé	Outfit Carrés moyens	Outfit standardisé
11	447	63	7,1	6,72	0,4	0,18	1,31	1,51	1,09	0,47
4	457	63	7,25	6,87	0,06	0,19	1,21	1	1,18	0,8
8	471	63	7,48	7,1	-0,44	0,19	1,2	0,97	1,07	0,36
9	448	63	7,11	6,74	0,37	0,18	1,03	0,19	0,89	-0,47
1	435	63	6,9	6,56	0,79	0,18	0,91	-0,45	0,92	-0,38
3	473	63	7,51	7,13	-0,51	0,19	0,9	-0,42	1,77	2,53
12	442	63	7,02	6,65	0,57	0,18	0,7	-1,65	0,63	-2,01
10	472	63	7,49	7,11	-0,47	0,19	0,69	-1,57	0,94	-0,15
7	469	63	7,44	7,06	-0,37	0,19	0,64	-1,85	0,54	-2,17
5	454	63	7,21	6,83	0,16	0,19	0,63	-2,03	0,57	-2,21
2	470	63	7,46	7,08	-0,4	0,19	0,55	-2,45	0,69	-1,29
6	463	63	7,35	6,97	-0,15	0,19	0,45	-3,26	0,38	-3,37

L'examen des seuils et catégories de niveaux NCLC montre que les niveaux NCLC identifiés pour l'expression écrite sont significativement différents les uns des autres (Tableau 8). La valeur du niveau médian de chacun de ces seuils montre que les niveaux NCLC utilisés par les panélistes sont correctement ordonnés (cf. expectation measure at category / Rasch-Thurstone Thresholds). Toutefois, pour l'écrit, la distinction entre les niveaux NCLC 1 et 2 et NCLC 11 et 12 a été plus difficile à opérer que pour l'oral. Le niveau 1 n'a été attribué que deux fois alors que des productions de faibles niveaux avaient été sélectionnées pour constituer l'échantillon.

Tableau 8

Seuils pour le passage, examens du fonctionnement des NCLC pour les 21 productions écrites

DATA					QUALITY CONTROL		RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat	Response
Category Counts			Cum.		Avge	Exp. OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK	Category
Score	Total	Used	%	%	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob	Name
1	2	2	0%	0%	-9.67	-9.43	.7			(-12.31)		low	low	100% NCLC 1
2	9	9	1%	1%	-9.55	-9.14	.6	-10.79	.72	-10.83	-11.65		-11.29	32% NCLC 2
3	74	74	10%	11%	-7.87	-7.82	.9	-10.87	.34	-8.03	-10.02	-10.83	-10.37	88% NCLC 3
4	95	95	13%	24%	-3.49	-3.42	.6	-5.43	.26	-3.92	-5.50	-5.43	-5.46	70% NCLC 4
5	58	58	8%	31%	-1.87	-2.04	.9	-2.26	.18	-1.97	-2.76	-2.26	-2.55	40% NCLC 5
6	95	95	13%	44%	-.37	-.20	1.0	-1.64	.18	-.25	-1.21	-1.64	-1.38	63% NCLC 6
7	88	88	12%	56%	1.93	1.89	.5	.94	.18	2.05	.90	.94	.91	60% NCLC 7
8	54	54	7%	63%	3.82	3.67	.8	3.27	.19	3.80	3.04	3.27	3.14	46% NCLC 8
9	48	48	6%	69%	5.45	5.34	.7	4.67	.21	4.98	4.41	4.67	4.48	35% NCLC 9
10	108	108	14%	83%	6.39	6.37	1.8	5.06	.17	6.65	5.66	5.06	5.41	68% NCLC 10
11	41	41	5%	89%	8.30	8.43	1.0	8.17	.22	8.56	7.72	8.17	7.90	41% NCLC 11
12	84	84	11%	100%	10.41	10.36	.7	8.88	.22	(10.22)	9.48	8.88	9.18	100% NCLC 12
									(Mean)		(Modal)	(Median)		

Le niveau ajusté calculé pour les productions écrites ne diffère pas beaucoup du score moyen calculé pour chacune des productions. On peut interpréter ce résultat comme étant une preuve de robustesse de l'ensemble de la procédure.

Comparaison de l'estimation du niveau des productions orales et écrites avec les descripteurs du CECRL et des NCLC

Après le calibrage des performances avec le modèle multifacette de Rasch, nous pouvons comparer le classement des candidats selon qu'ils ont été évalués dans le processus opérationnel du TCF avec les niveaux CECRL par les correcteurs ou qu'ils ont été évalués avec les niveaux NCLC par les panélistes. Pour mesurer la force du lien entre l'évaluation des candidats avec les descripteurs du CECRL et des NCLC, il a été décidé d'utiliser les corrélations de Pearson, Spearman et le tau de Kendall. Cela permet d'obtenir des indices pour la cohérence interne des jugements (Stemler & Tsai, 2008).

Pour l'expression orale (Tableau 9), les corrélations entre le classement des candidats après les doubles corrections (utilisation des descripteurs CECRL) et les évaluations faites par les 12 panélistes (utilisation des descripteurs NCLC) sont fortes. Si on observe une légère différence entre la corrélation de Spearman et le tau de Kendall, on constate que le lien est fort.

Tableau 9
Coefficients de corrélation entre le niveau CECRL et le niveau médian NCLC (expression orale)

	Tour 1	Tour 2	Tour 3
Pearson	0,97	0,98	0,98
Spearman	0,99	0,99	0,99
tau de Kendall	0,96	0,97	0,97

Note. toutes les corrélations sont significatives à p<0,001

Pour l'expression écrite (Tableau 8), les corrélations entre le classement des candidats après les doubles -corrections (utilisation des descripteurs CECRL) et les évaluations faites par les 12 panélistes (utilisation des descripteurs NCLC) sont fortes, et ce, aussi bien pour la corrélation de Pearson, de Spearman que celle du tau de Kendall.

Tableau 10
Coefficients de corrélation entre le niveau CECRL et le niveau médian NCLC (expression écrite)

	Tour 1	Tour 2	Tour 3
Pearson	0,98	0,98	0,98
Spearman	0,98	0,98	0,98
tau de Kendall	0,94	0,93	0,94

Note. toutes les corrélations sont significatives à p<0,001

La lecture de ces résultats indique que la validité des épreuves du TCF est forte. Alors que des correcteurs et des panélistes issus de deux ensembles indépendants ont évalué les candidats avec deux échelles de compétences différentes (CECRL et NCLC), les classements des candidats se trouvent peu affectés et sont fortement similaires. Cette donnée valide la capacité des panélistes à discriminer des niveaux de maîtrise en langue française avec les échelles NCLC. On peut déduire que l'utilisation d'instruments de mesure différents n'a pas altéré la qualité de l'estimation du niveau des performances orales et écrites. Les résultats de ce calibrage de performances identiques avec deux échelles différentes et deux ensembles d'évaluateurs différents permettent d'envisager que non seulement le taux d'accord entre les panélistes est satisfaisant, mais qu'en plus leur interprétation des niveaux des échelles NCLC est correcte. On peut croire raisonnablement que les panélistes sont à la fois exacts et précis dans leur interprétation des niveaux NCLC. Toutefois, la forte corrélation entre les résultats obtenus avec les NCLC et le CECRL n'est pas « surprenante » puisque la définition du construit de la compétence langagière utilisée par les deux référentiels est issue des travaux de Lyle Bachman (Bachman, 1990; Bachman & Palmer, 1996, 2010; Centre des niveaux de compétence linguistique canadiens, 2012; Conseil de l'Europe, 2001).

Synthèse des analyses qualitatives des descripteurs

Afin d'atteindre une plus grande certitude sur la correspondance entre les niveaux CECRL et NCLC, il a été décidé de procéder à une analyse qualitative des contenus des descripteurs issus des deux référentiels. Si la correspondance entre les contenus des descripteurs est suffisamment similaire à celle obtenue à l'issue du calibrage par les évaluateurs, ce sera un faisceau de preuve complémentaire. Pour cette étude, 4 experts en contenu ont mis en relation les descripteurs NCLC et CECRL. Chacun devait proposer une correspondance. Ils devaient citer les éléments leur permettant de rapprocher et d'attribuer les niveaux.

Pour l'expression orale, les jugements indépendants des experts sont consensuels sauf pour le NCLC 5 (Tableau 11). Pour l'expression écrite, le consensus est atteint pour tous les niveaux (Tableau 12). À l'issue de ce travail d'analyses qualitatives, on constate que la correspondance établie pour l'expression orale et écrite est presque identique.

Tableau 11

Correspondance entre les niveaux NCLC et CECRL (expression orale)

NIVEAU NCLC	A1 non-atteint	A1	A2	B1	B2	C1	C2
1	**3**	**1**					
2		**4**					
3			**4**				
4			**4**				
5			**2**	**2**			
6				**4**			
7				**1**	**3**		
8					**4**		
9						**4**	
10						**4**	
11							**4**
12							**4**

Tableau 12

Correspondance entre les niveaux NCLC et CECRL (expression écrite)

Niveau NCLC	A1 non-atteint	A1	A2	B1	B2	C1	C2
1	**4**						
2		**4**					
3		**1**	**3**				
4			**4**				
5				**4**			
6				**4**			
7					**4**		
8					**4**		
9						**4**	
10						**4**	
11							**4**
12							**4**

Synthèse de la mise en correspondance des descripteurs du CECRL et des NCLC

Afin d'assurer l'adossement, plusieurs procédures quantitatives et qualitatives ont été mises en place. Le résultat de la prise en compte de ces analyses est consigné dans les Tableaux 13 et 14. Le niveau de français étant également évalué pour une utilisation au Canada par le Test d'évaluation de français (TEF) de la Chambre de commerce et d'industrie de Paris (CCIP, 2015; Crendal, 2005), l'adossement établi par la CCIP complète la triangulation.

Tableau 13

Correspondance CECRL/NCLC (expression orale)

NCLC	CCIP	CIEP (Évaluation des productions)	CIEP (Analyse qualitative des descripteurs)	Niveau final après arbitrage
0	A1	A1 non atteint	-	A1 non atteint
1	A1	A1	A1 non atteint	A1
2	A2	A2	A1	A2
3	A2	A2/B1	A2	A2
4	B1	A2/B1	A2	A2
5	B1	B1	A2/B1	B1
6	B2	B1	B1	B1
7	B2	B2	B2	B2
8	C1	B2	B2	B2
9	C1	C1	C1	C1
10	C1	C1	C1	C1
11	C2	C2	C2	C2
12	C2	C2	C2	C2

Tableau 14

Correspondance CECRL/NCLC (expression écrite)

NCLC	CCIP	CIEP (Évaluation des productions)	CIEP (Analyse qualitative des descripteurs)	Niveau final après arbitrage
0	A1	A1 non atteint / A1	-	A1 non atteint
1	A1	A1	A1 non atteint	A1
2	A2	A2	A1	A2
3	A2	A2	A2	A2
4	B1	A2	A2	A2
5	B1	B1	B1	B1
6	B2	B1/B2	B1	B1
7	B2	B2	B2	B2
8	C1	B2	B2	B2
9	C1	B2/C1	C1	C1
10	C1	C1	C1	C1
11	C2	C2	C2	C2
12	C2	C2	C2	C2

Correspondance finale entre les niveaux NCLC et CECRL

Pour parvenir à l'adossement définitif, une séance de travail FEI a été organisée. Les éléments pris en compte ont été :

1) Les résultats des analyses quantitatives et qualitatives ;
2) La cohérence de l'adossement de l'écrit et de l'oral ;
3) La cohérence avec l'adossement du TEF.

Les décisions collégiales finales restent proches des résultats obtenus au moment des analyses. Toutefois, elles présentent l'avantage d'une plus grande facilité d'utilisation du fait d'une forte homogénéité (Tableau 15, Tableau 16).

Tableau 15

Correspondance finale (TCF expression orale)

Notes	Niveau CECRL	Niveau NCLC
0	A1 non atteint (4 à 6 jugements)	Niveau 1 non atteint
1	A1 (4 jugements)	1
1	A1 (5 à 6 jugements)	2
2	A2 (4 jugements)	3
3	A2 (5 jugements)	3
4	A2 (6 jugements)	4
5	A2 (4 à 6 jugements + 0 à 2 jugements >A2)	4
6	B1 (4 jugements)	5
7	B1 (5 jugements)	6
8	B1 (6 jugements)	6
9	B1 (4 à 6 jugements + 0 à 2 jugements >B1)	6
10	B2 (4 jugements)	7
11	B2 (5 jugements)	7
12	B2 (6 jugements)	8
13	B2 (4 à 6 jugements + 0 à 2 jugements >B2)	8
14	C1 (4 jugements)	9
15	C1 (5 jugements)	9
16	C1 (6 jugements)	10
17	C1 (4 à 6 jugements + 0 à 2 jugements >C1)	10
18	C2 (4 jugements)	11
19	C2 (5 jugements)	12
20	C2 (6 jugements)	12

Tableau 16
Correspondance finale (TCF expression écrite)

Notes	Niveau CECRL	Niveau NCLC
0	A1 non atteint (4 à 6 jugements)	Niveau 1 non atteint
1	A1 (4 à 6 jugements)	1
2	A2 (4 jugements)	2
3	A2 (5 jugements)	3
4	A2 (6 jugements)	4
5	A2 (4 à 6 jugements + 0 à 2 jugements >A2)	4
6	B1 (4 jugements)	5
7	B1 (5 jugements)	6
8	B1 (6 jugements)	6
9	B1 (4 à 6 jugements + 0 à 2 jugements >B1)	6
10	B2 (4 jugements)	7
11	B2 (5 jugements)	7
12	B2 (6 jugements)	8
13	B2 (4 à 6 jugements + 0 à 2 jugements >B2)	8
14	C1 (4 jugements)	9
15	C1 (5 jugements)	9
16	C1 (6 jugements)	10
17	C1 (4 à 6 jugements + 0 à 2 jugements >C1)	10
18	C2 (4 jugements)	11
19	C2 (5 jugements)	12
20	C2 (6 jugements)	12

Limites de cette étude

Les procédures standardisées de définition des scores de césure évoluent constamment. Ces évolutions sont autant liées aux avancées de la recherche qu'aux retours d'expériences. Si l'évaluation en langue répond à des standards de qualité (par exemple, en Europe, la *« Q-mark »* d'ALTE), il est faux d'affirmer qu'elle correspond à des normes. L'évaluation de la qualité d'un adossement à un référentiel dépend de consensus savants et professionnels (état de l'art), de l'utilisation d'un argumentaire (pour expliquer ses choix théoriques et méthodologiques), de la nécessaire prise en compte du contexte d'utilisation des examens, autant que des valeurs à atteindre pour des indices statistiques ou de tailles d'échantillon minimales.

Plutôt que de s'attarder sur ces débats irrésolus, il nous semble plus important de pointer les limites liées à la généralisabilité des résultats. Si l'adossement du TCF est multiréférentiel, si les productions utilisées pour procéder à l'adossement ont dûment été sélectionnées, si les panélistes ont été choisis parmi un ensemble d'experts attestés, si des faisceaux de données qualitatives et quantitatives ont été triangulés, si l'impact de l'alignement a été analysé, il n'en reste pas moins qu'un certain nombre d'opportunités d'amélioration subsistent.

Tout d'abord, il y a des limites connues des praticiens. Un échantillon d'experts plus fourni et ayant une pratique professionnelle régulière des NCLC aurait été utile, travailler avec plusieurs échantillons d'experts et différents lots de productions du TCF aurait pu corriger l'erreur liée à une sélection particulière de productions et un panel particulier d'experts. Si ces limites sont importantes, il faut toutefois les pondérer, car, bien souvent, elles sont liées à des problèmes de ressources financières, temporelles, de disponibilité de l'expertise, mais aussi de l'intérêt des experts pour ce type de projet et, enfin, de conflits d'intérêts.

Ensuite, la nécessité d'utiliser des procédures connues et validées par la communauté des chercheurs et des praticiens peut amener à un certain conservatisme et manque d'innovation. L'approche privilégiée dans cette étude est une approche essentiellement centrée sur l'évaluateur, choix largement privilégié par la communauté professionnelle. Il existe pourtant des méthodes d'adossement centrées sur le contenu des productions ou des corpus. Pour le TCF, si la triangulation des données quantitatives et qualitatives a sérieusement été considérée, peu a été fait pour relier le contenu des performances aux descripteurs NCLC.

Étant donné la complexité de la composition et de l'analyse des corpus langagiers, l'utilisation des procédures de l'intelligence artificielle pourrait permettre d'enrichir la compréhension que nous avons des contenus correspondant aux descripteurs des échelles des référentiels que ce soit le CECRL ou bien les NCLC. Cette voie, si elle est exploratoire, est prometteuse, car elle viendra combler l'angle négligé de l'analyse des contenus dans les procédures d'adossement.

Enfin, cet adossement du TCF aux NCLC est le premier mené par FEI. Une phase itérative sera nécessaire, car elle permettra, à la fois, de mieux vérifier la reproductibilité des résultats, de considérer les nouveaux résultats obtenus par les autres tests et d'actualiser les pratiques professionnelles au regard des retours d'expériences consignés par les associations d'évaluation en langue.[1]

Conclusion

Le travail d'adossement de productions orales et écrites sur des échelles de niveau est un travail qui demande de la patience et de l'humilité. Pour arriver à un alignement valide et utile, il convient de faire preuve de rigueur dans les analyses, mais également d'une absence de dogmatisme. S'il est utopique de croire que l'adossement de productions orales et écrites est le fruit d'une « pure » approche scientifique, il est tout aussi absurde de croire que les résultats obtenus ne sont pas reproductibles du fait d'un manque de rigueur scientifique et qu'il ne s'agit que d'une « construction sociale ». Les organismes certificateurs en langue ont le devoir d'utiliser des procédures rigoureuses, des approches non dogmatiques et plurielles. L'adossement s'enrichit de la triangulation méthodologique. Il convient de mettre en garde contre une vision trop puriste de ce que sont les niveaux de compétences en langue et d'appeler à la responsabilité sur l'utilisation des résultats à un test. Si l'objectivation est grande, elle n'est et ne sera jamais absolue et, ce, d'autant plus que les niveaux de langue ne sont pas des entités réelles, mais bien le fruit d'un consensus, d'une expertise, d'une reconnaissance. Le processus d'adossement à un cadre de référence est un processus itératif triangulé et doit se construire et être amélioré dans le temps en

privilégiant une utilisation raisonnée des ressources.

La correspondance devrait être adressée à Vincent Folny.
Courriel : folny@france-education-international.fr

Notes

[1] ALTE (Association of Language Testers in Europe), EALTA (European Association for language Testing and Assessment), ILTA (International Language testing Association), ACEL (Association Canadienne pour l'évaluation des langues), …

Références

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice : Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice : Developing language assessments and justifying their use in the real world*. Oxford University Press.

Chambre de commerce et d'industrie de Paris (2015). *Brochure TEF*. http://www.francais.cci-paris-idf.fr/wp-content/uploads/downloads/2015/02/Brochure-TEF-maj-20151.pdf

Centre des niveaux de compétence linguistique canadiens. (2012). *Niveaux de compétence linguistique canadiens, français langue seconde pour adultes*. Immigration, réfugies et citoyenneté Canada. http://publications.gc.ca/collections/collection_2012/cic/Ci63-26-2012-fra.pdf

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting : A guide to establishing and evaluating performance standards on tests.* SAGE Publications Ltd.

Conseil de l'Europe. (2001). *Cadre européen commun de référence pour les langues*. Didier.

Conseil de l'Europe. (2009). *Manuel pour Relier les examens de langues au Cadre européen commun de référence pour les langues (CECR)*. http://www.coe.int/t/dg4/linguistic/Manuel1_FR.asp#Manual

Crendal, A. (2005). Vers la multiréférentialisation du TEF. *Point commun*, *24*. http://www.centredelanguefrancaise.paris/wp-content/uploads/2011/05/vers_la_multireferentialisation_du_tef.pdf

De Jong, J. H. A. L. (2013). *The Stages Of standard Setting*. BAAL TEA SIG. http://www.beds.ac.uk/__data/assets/pdf_file/0008/256787/John-De-Jong-2013-06-BAAL-TEA-SIG.pdf

Demeuse, M., Desroches, F., Crendal, A., Renaud, F., Oster, P., & Leroux, X. (2004). L'évaluation des compétences linguistiques des adultes en français langue étrangère dans une perspective de multiréférentialisation. 17e colloque international de l'Association pour le Développement des Méthodologies d'Évaluation en Éducation (ADMEE-Europe), Lisbonne. http://www.centredelanguefrancaise.paris/wp-content/uploads/downloads/2011/10/1.tef-admee2004-multireferentialisation.pdf

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

Engelhard, G. (2009). Evaluating the Judgments of Standard-Setting Panelists using Rasch Measurement Theory. In E. V. Smith & G. E. Stone (Éds.), *Criterion referenced testing : Practice analysis to score reporting using Rasch measurement models*. JAM Press.

Engelhard, G. (2011). Evaluating the Bookmark Judgments of Standard-Setting Panelists. *Educational and Psychological Measurement*, *71*(6), 909-924. https://doi.org/10.1177/0013164410395934

Engelhard, G. (2013). *Invariant measurement : Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Engelhard, G., & Gordon, B. (2000). Setting and Evaluating Performances Standards For high Stakes Writting Asessments. In M. Wilson & G. Engelhard (Éds.), *Objective measurement : Theory into practice. Vol. 5 : [Papers presented at the Ninth International Objective Measurement Workshop (IOMW9), University of Chicago, Illinois, March 1997]* (p. 3-14). Ablex Publishing Corporation.

Figueras, N., Noijons, J. (Eds). (2009). *Linking to the CEFR levels : Research perspectives*. Council of Europe.

Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores : Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, *29*(1), 38–44.

Hambleton, R., K., & Pitoniak, M. J. (2006). Setting Performance Standards. In R. L. Brennan, National Council on Measurement in Education, & American Council on Education (Éds.), *Educational measurement* (4. ed). Praeger Publ.

Kaliski, P. K., Wind, S. A., Engelhard, G., Morgan, D. L., Plake, B. S., & Reshetar, R. A. (2013). Using the Many-Faceted Rasch Model to Evaluate Standard Setting Judgments : An Illustration With the Advanced Placement Environmental Science Exam. *Educational and Psychological Measurement*, *73*(3), 386-411. https://doi.org/10.1177/0013164412468448

Kecker, G., & Eckes, T. (2010). Putting the Manual to the test : The TestDaF-CEFR linking project. In *Aligning Tests with the CEFR : reflection on using the Council of Europe's draft Manual* (Cambridge University Press, Vol. 33, p. 50-79). UCLES.

Linacre, J. M. (2013). *A User's Guide to FACETS Rasch-Model Computer Programs. Program Manual 3.71.2*.

Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370-371.

Lunz, M. E. L. (2000). Setting Standards on Performance Examinations. In M. Wilson & G. Engelhard (Éds.), *Objective measurement : Theory into practice.* (Vol. 5, p. 181-202). Ablex Publishing Corporation.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability : Three common approaches. *Best practices in quantitative methods*, 29–49.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores : A manual for setting standards of performance on educational and occupational tests*. Educational Testing Service.

Annexe A
Fidélité des panélistes pour le classement des productions orales (tour 1)

TCF EO NCLC	Panéliste 1	Panéliste 2	Panéliste 3	Panéliste 4	Panéliste 5	Panéliste 6
Accords exacts	52%	52%	48%	67%	52%	38%
Accords exacts et adjacents	90%	86%	76%	95%	81%	90%

TCF EO NCLC	Panéliste 7	Panéliste 8	Panéliste 9	Panéliste 10	Panéliste 11	Panéliste 12
Accords exacts	52%	29%	24%	33%	57%	38%
Accords exacts et adjacents	86%	57%	97%	81%	86%	71%

TCF EO NCLC	Moyenne
Accords exacts	45%
Accords exacts et adjacents	81%

Annexe B
Fidélité des panélistes pour les productions orales avec les niveaux NCLC (tour 1)

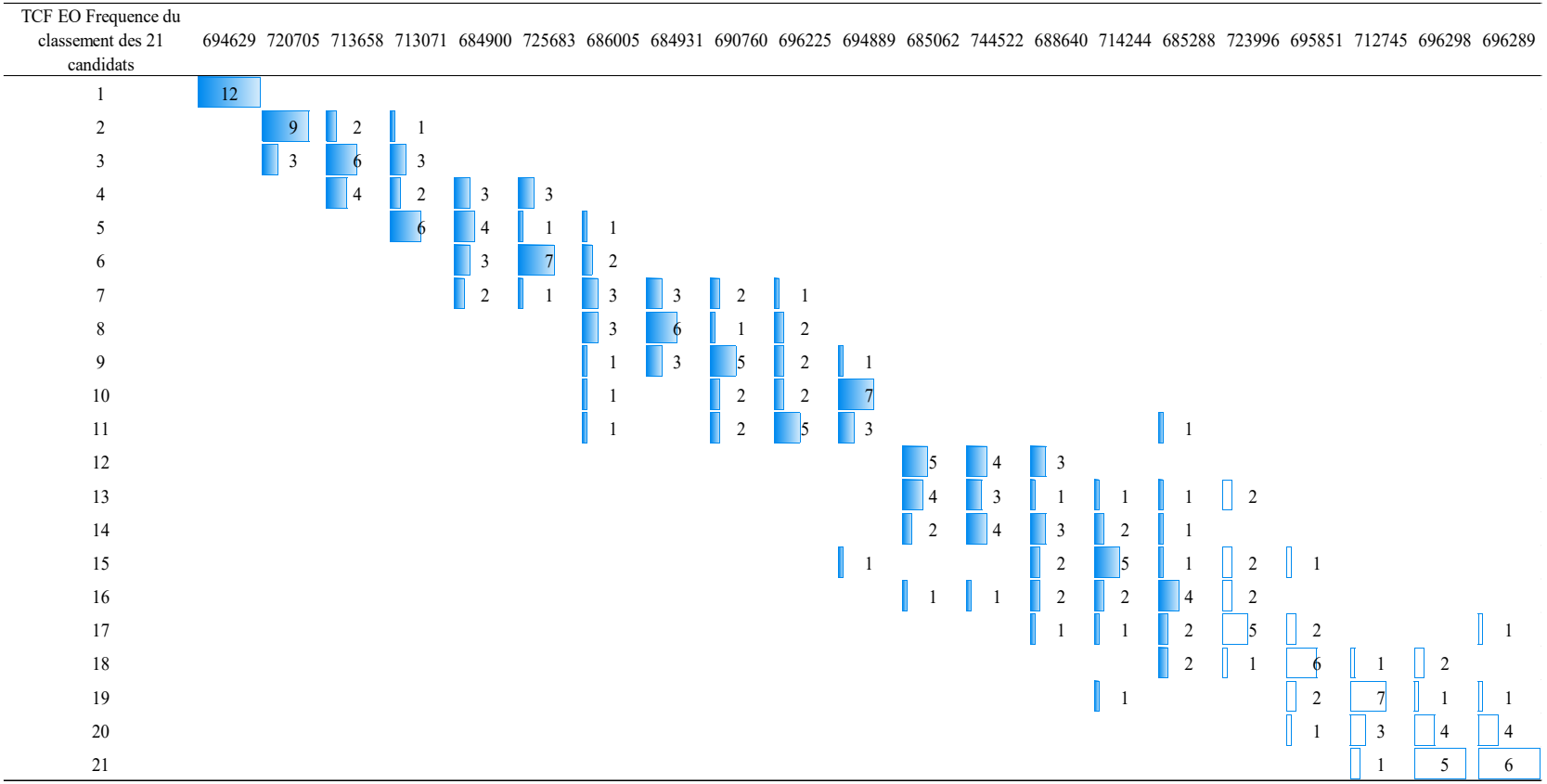
TCF EO NCLC

	Panéliste 1	Panéliste 2	Panéliste 3	Panéliste 4	Panéliste 5	Panéliste 6
Accords exacts	52%	48%	57%	52%	76%	71%
Accords exacts et adjacents	100%	81%	95%	90%	95%	95%

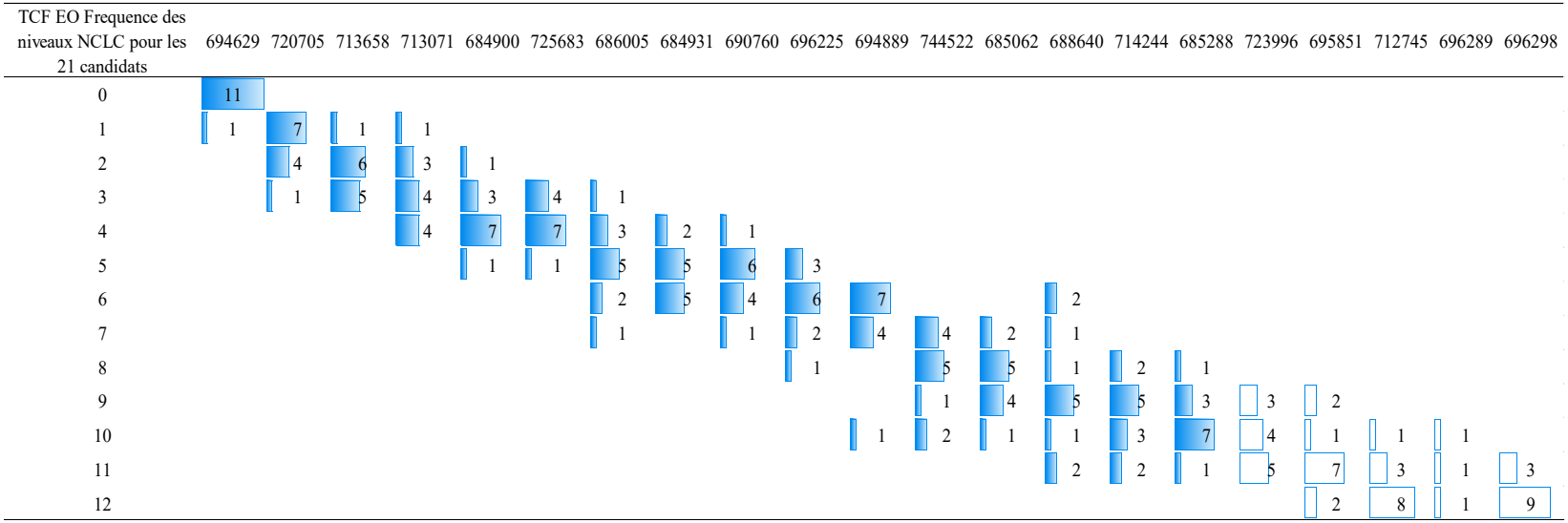
	Panéliste 7	Panéliste 8	Panéliste 9	Panéliste 10	Panéliste 11	Panéliste 12
Accords exacts	62%	33%	29%	57%	48%	62%
Accords exacts et adjacents	100%	76%	81%	86%	90%	100%

	Moyenne
Accords exacts	54%
Accords exacts et adjacents	91%

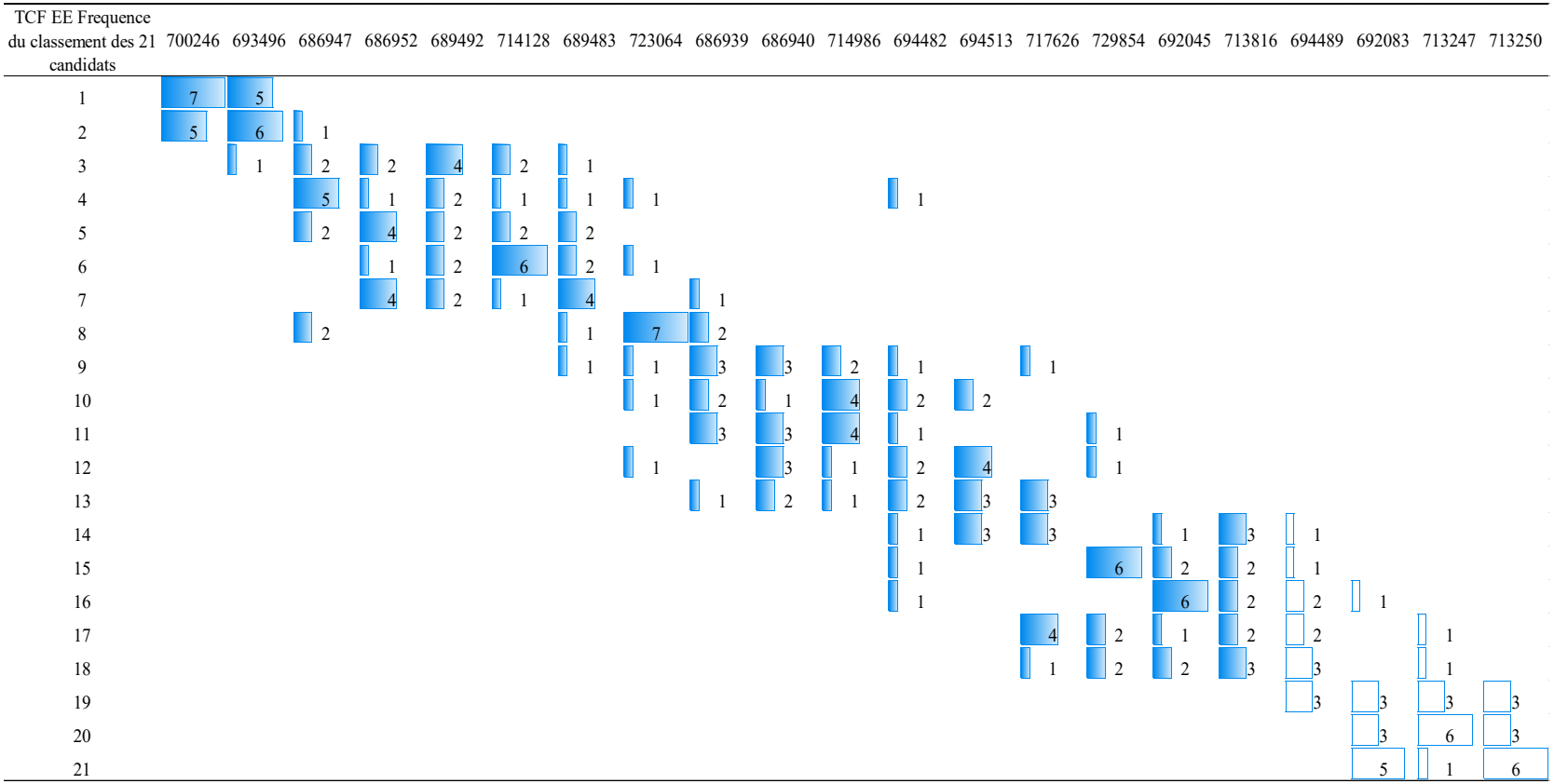
Annexe C
Classement des productions orales (tour 1)



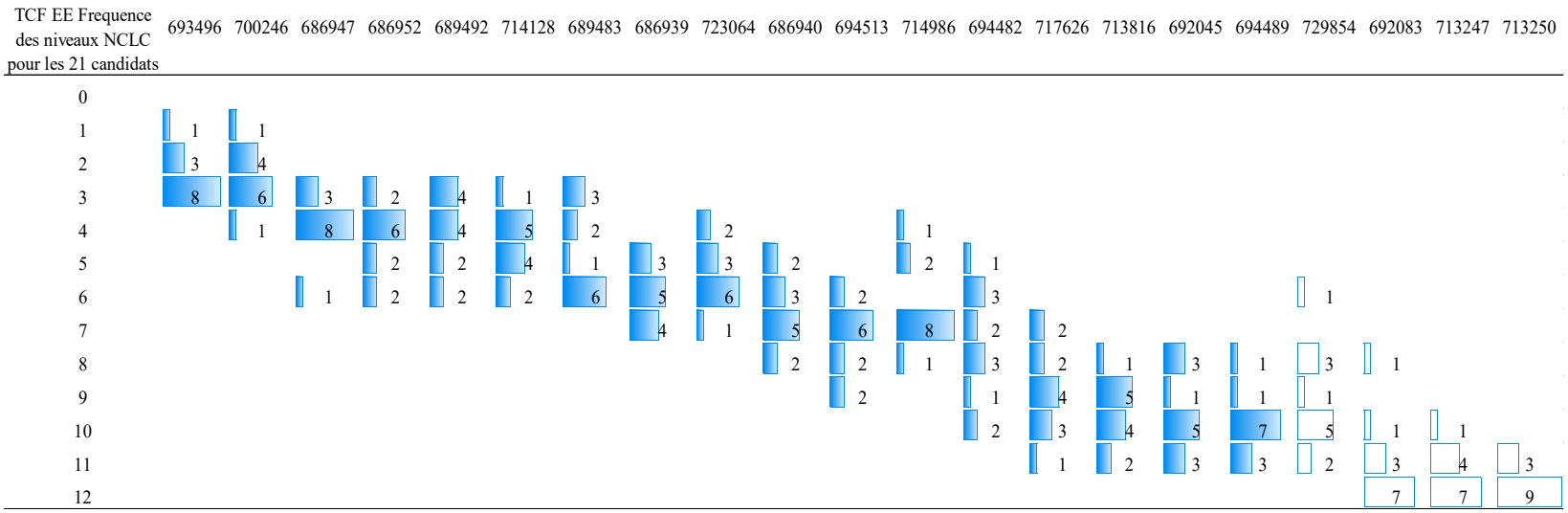
Annexe D
Positionnement des productions orales avec les descripteurs NCLC (tour 1)



Annexe E
Classement des productions écrites (tour 1)



Annexe F
Positionnement des productions écrites avec les descripteurs NCLC (tour 1)



Annexe G
Fidélité des panélistes pour le classement des productions écrites (tour 1)

TCF EE NCLC

	Panéliste 1	Panéliste 2	Panéliste 3	Panéliste 4	Panéliste 5	Panéliste 6
Accords exacts	43%	19%	33%	24%	52%	62%
Accords exacts et adjacents	81%	67%	71%	67%	81%	81%
	Panéliste 7	Panéliste 8	Panéliste 9	Panéliste 10	Panéliste 11	Panéliste 12
Accords exacts	29%	29%	24%	38%	19%	43%
Accords exacts et adjacents	86%	86%	57%	57%	62%	71%

	Moyenne
Accords exacts	35%
Accords exacts et adjacents	72%

Annexe H
Fidélité des panélistes pour l'évaluation des productions écrites avec les niveaux NCLC (tour1)

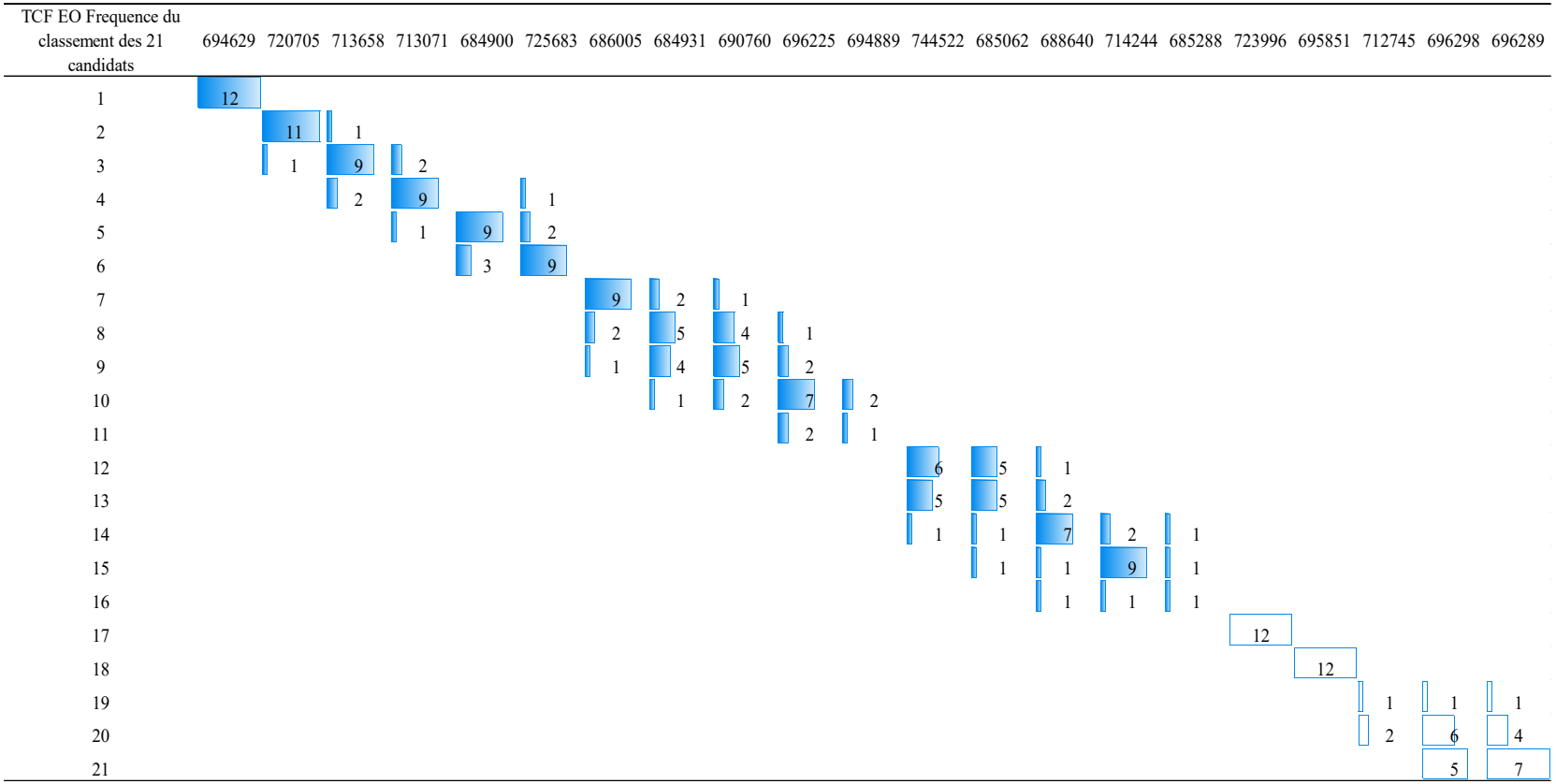
TCF EE NCLC

	Panéliste 1	Panéliste 2	Panéliste 3	Panéliste 4	Panéliste 5	Panéliste 6
Accords exacts	10%	62%	57%	48%	43%	71%
Accords exacts et adjacents	57%	95%	95%	90%	95%	95%

	Panéliste 7	Panéliste 8	Panéliste 9	Panéliste 10	Panéliste 11	Panéliste 12
Accords exacts	76%	43%	38%	67%	33%	38%
Accords exacts et adjacents	86%	76%	71%	100%	57%	71%

	Moyenne
Accords exacts	49%
Accords exacts et adjacents	83%

Annexe I
Classement des productions orales (tour 2)



Annexe J
Positionnement des productions orales avec les descripteurs NCLC (tour 2)

TCF EO Frequence des niveaux NCLC	694629	720705	713658	713071	684900	725683	686005	690760	684931	696225	694889	685062	744522	688640	714244	685288	723996	695851	696289	696298	712745
0	12																				
1		11																			
2		1	12	3																	
3				9	3																
4					9	12															
5							12	6	5												
6								6	7	12	6										
7											6	1	3	1							
8												9	8	4							
9												2	1	7	11	3					
10															1	9	7				
11																	5	11	1		3
12																		1	11	12	9

Annexe K
Fidélité des panélistes pour le classement des productions orales (tour 2)

TCF EO
NCLC

	Panéliste 1	Panéliste 2	Panéliste 3	Panéliste 4	Panéliste 5	Panéliste 6
Accords exacts	67%	86%	71%	52%	76%	86%
Accords exacts et adjacents	100%	100%	95%	95%	100%	90%

	Panéliste 7	Panéliste 8	Panéliste 9	Panéliste 10	Panéliste 11	Panéliste 12
Accords exacts	62%	76%	57%	67%	86%	62%
Accords exacts et adjacents	95%	100%	81%	100%	100%	95%

	Moyenne
Accords exacts	71%
Accords exacts et adjacents	96%

Annexe L
Fidélité des panélistes pour l'évaluation des productions orales avec les niveaux NCLC (tour 2)

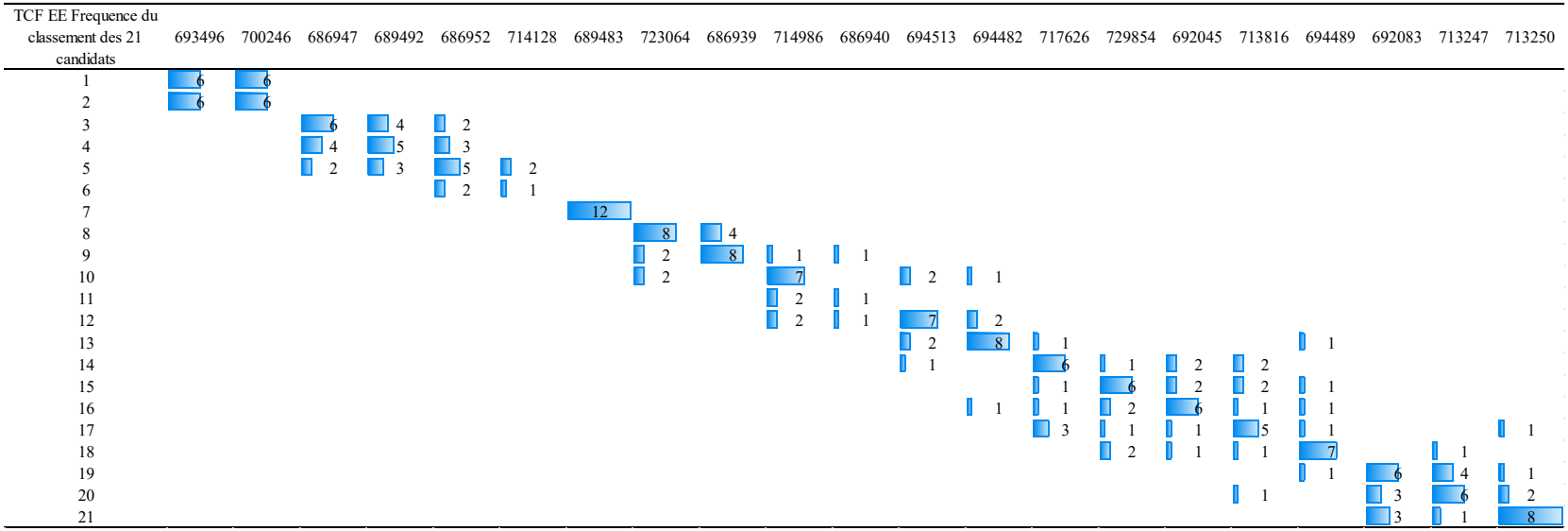
TCF EO NCL

	Panéliste 1	Panéliste 2	Panéliste 3	Panéliste 4	Panéliste 5	Panéliste 6
Accords exacts	95%	57%	95%	90%	76%	90%
Accords exacts et adjacents	100%	100%	100%	100%	100%	100%

	Panéliste 7	Panéliste 8	Panéliste 9	Panéliste 10	Panéliste 11	Panéliste 12
Accords exacts	62%	90%	67%	62%	81%	95%
Accords exacts et adjacents	100%	100%	95%	100%	100%	100%

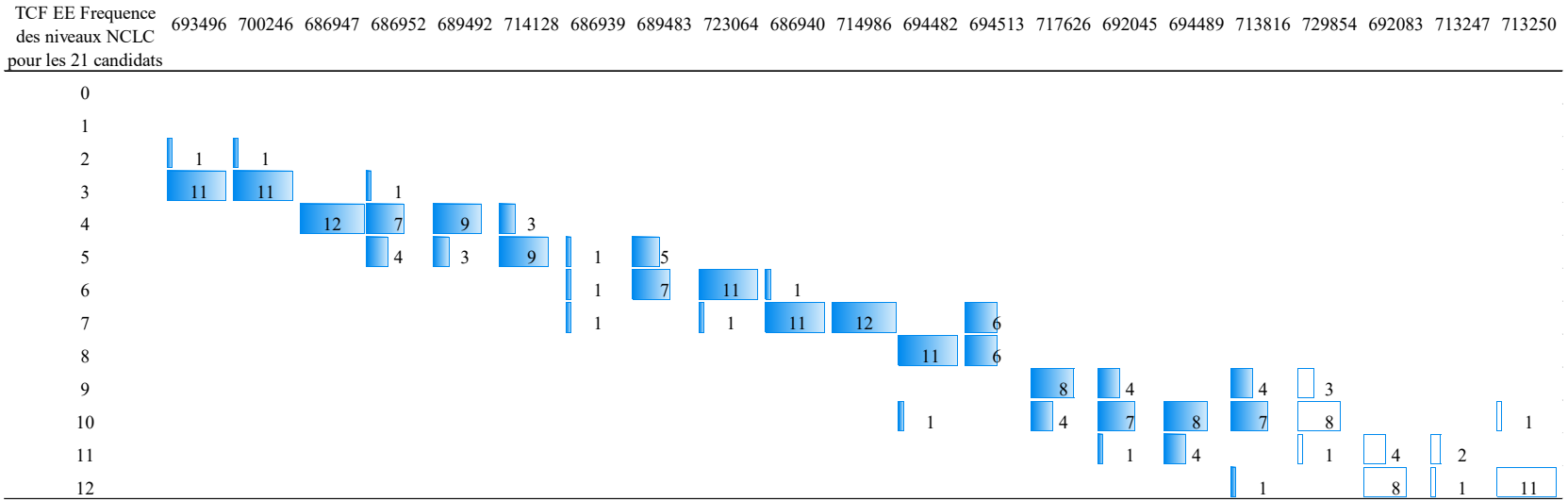
	Moyenne
Accords exacts	86%
Accords exacts et adjacents	100%

Annexe M
Classement des productions écrite (tour 2)



TCF EE Frequence du classement des 21 candidats	693496	700246	686947	689492	686952	714128	689483	723064	686939	714986	686940	694513	694482	717626	729854	692045	713816	694489	692083	713247	713250
1	6	6																			
2	6	6																			
3			6	4	2																
4			4	5	3																
5			2	3	5	2															
6					2	1															
7							12														
8								8	4												
9								2	8	1	1										
10								2		7		2	1								
11										2	1										
12										2	1	7	2								
13												2	8	1				1			
14												1		6	1	2	2				
15														1	6	2	2	1			
16													1	1	2	6	1	1			
17														3	1	1	5	1			1
18															2	1	1	7		1	
19																		1	6	4	1
20																	1		3	6	2
21																			3	1	8

Annexe N
Positionnement des productions écrites avec les descripteurs NCLC (tour 2)



Annexe O
Fidélité des panélistes pour le classement des productions écrites (tour 2)

TCF EE NCLC

	Panéliste 1	Panéliste 2	Panéliste 3	Panéliste 4	Panéliste 5	Panéliste 6
Accords exacts	67%	76%	67%	38%	52%	38%
Accords exacts et adjacents	90%	100%	86%	62%	86%	100%

	Panéliste 7	Panéliste 8	Panéliste 9	Panéliste 10	Panéliste 11	Panéliste 12
Accords exacts	76%	29%	33%	71%	86%	57%
Accords exacts et adjacents	100%	52%	76%	86%	100%	86%

	Moyenne
Accords exacts	58%
Accords exacts et adjacents	85%

Annexe P
Fidélité des panélistes pour l'évaluation des productions écrites avec les niveaux NCLC (tour 2)

TCF EE NCLC

	Panéliste 1	Panéliste 2	Panéliste 3	Panéliste 4	Panéliste 5	Panéliste 6
Accords exacts	76%	76%	86%	52%	71%	86%
Accords exacts et adjacents	100%	100%	90%	95%	100%	100%

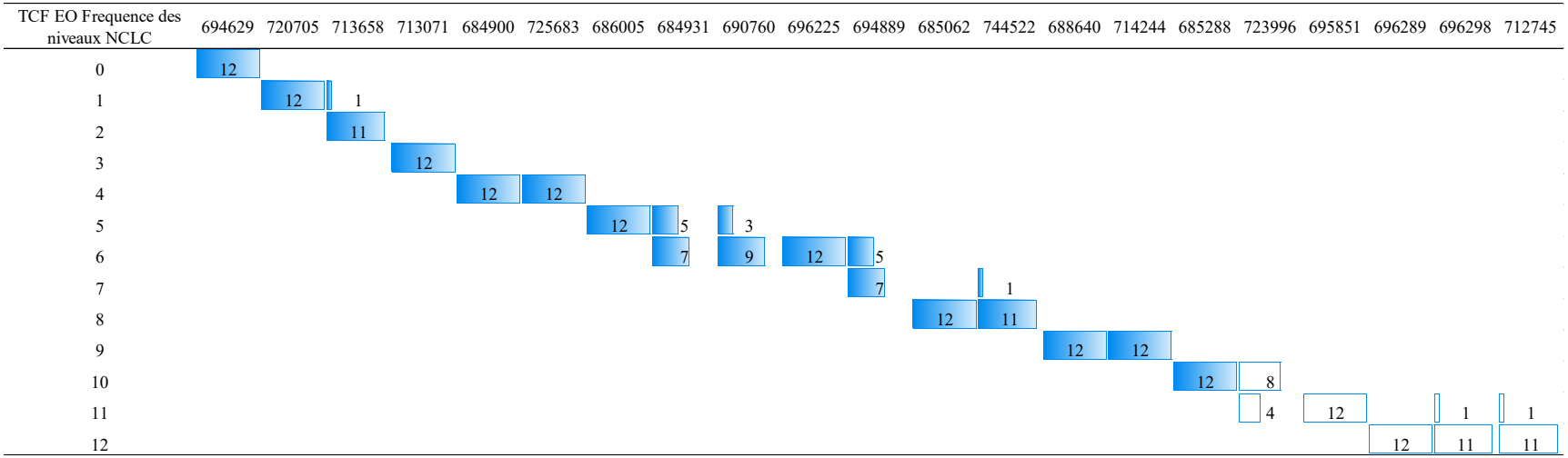
	Panéliste 7	Panéliste 8	Panéliste 9	Panéliste 10	Panéliste 11	Panéliste 12
Accords exacts	81%	86%	76%	81%	86%	67%
Accords exacts et adjacents	100%	105%	90%	105%	105%	81%

	Moyenne
Accords exacts	77%
Accords exacts et adjacents	98%

Annexe Q
Classement des productions orales (tour 3)

TCF EO Frequence du classement des 21 candidats	694629	720705	713658	713071	684900	725683	686005	684931	690760	696225	694889	685062	744522	688640	714244	685288	723996	695851	712745	696298	696289
1	12																				
2		12																			
3			12																		
4				12																	
5					11	1															
6					1	11															
7							12														
8								8	4												
9								4	8												
10										12											
11											12										
12												7	5								
13												5	7								
14														12							
15															12						
16																12					
17																	12				
18																		12			
19																			1	1	1
20																			1	9	2
21																			1	2	9

Annexe R
Positionnement des productions orales avec les descripteurs NCLC (tour 3)



Annexe S
Fidélité des panélistes pour le classement des productions orales (tour 3)

TCF EO
NCLC

	Panéliste 1	Panéliste 2	Panéliste 3	Panéliste 4	Panéliste 5	Panéliste 6
Accords exacts	100%	90%	81%	90%	100%	100%
Accords exacts et adjacents	100%	100%	100%	100%	100%	100%

	Panéliste 7	Panéliste 8	Panéliste 9	Panéliste 10	Panéliste 11	Panéliste 12
Accords exacts	81%	90%	81%	90%	81%	81%
Accords exacts et adjacents	90%	100%	100%	100%	100%	100%

	Moyenne
Accords exacts	89%
Accords exacts et adjacents	99%

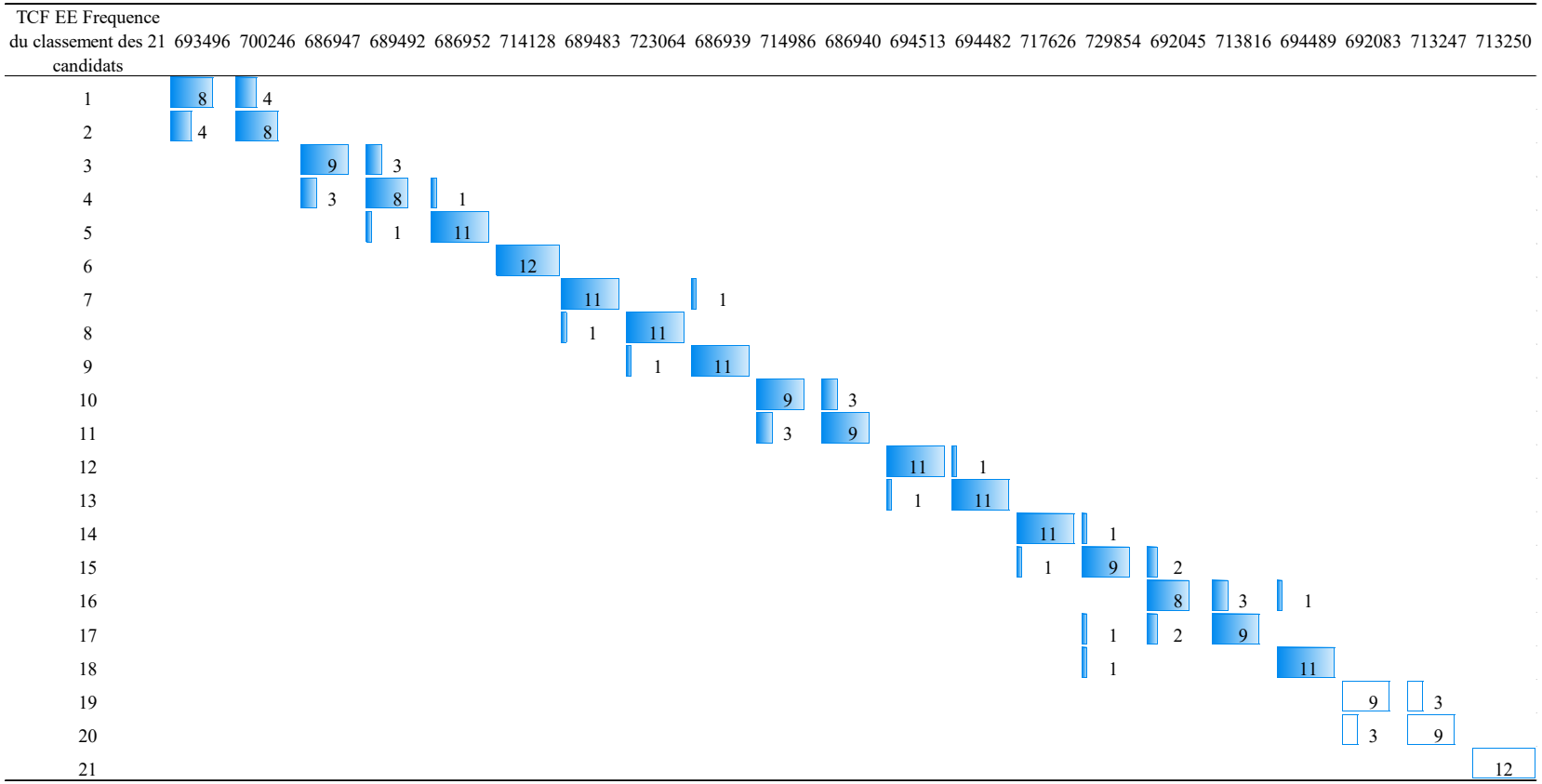
Annexe 1
Fidélité des panélistes pour l'évaluation des productions orales avec les niveaux NCLC (Tour 3)

TCF EO NCLC	Panéliste 1	Panéliste 2	Panéliste 3	Panéliste 4	Panéliste 5	Panéliste 6
Accords exacts	100%	86%	95%	81%	90%	95%
Accords exacts et adjacents	100%	100%	100%	100%	100%	100%

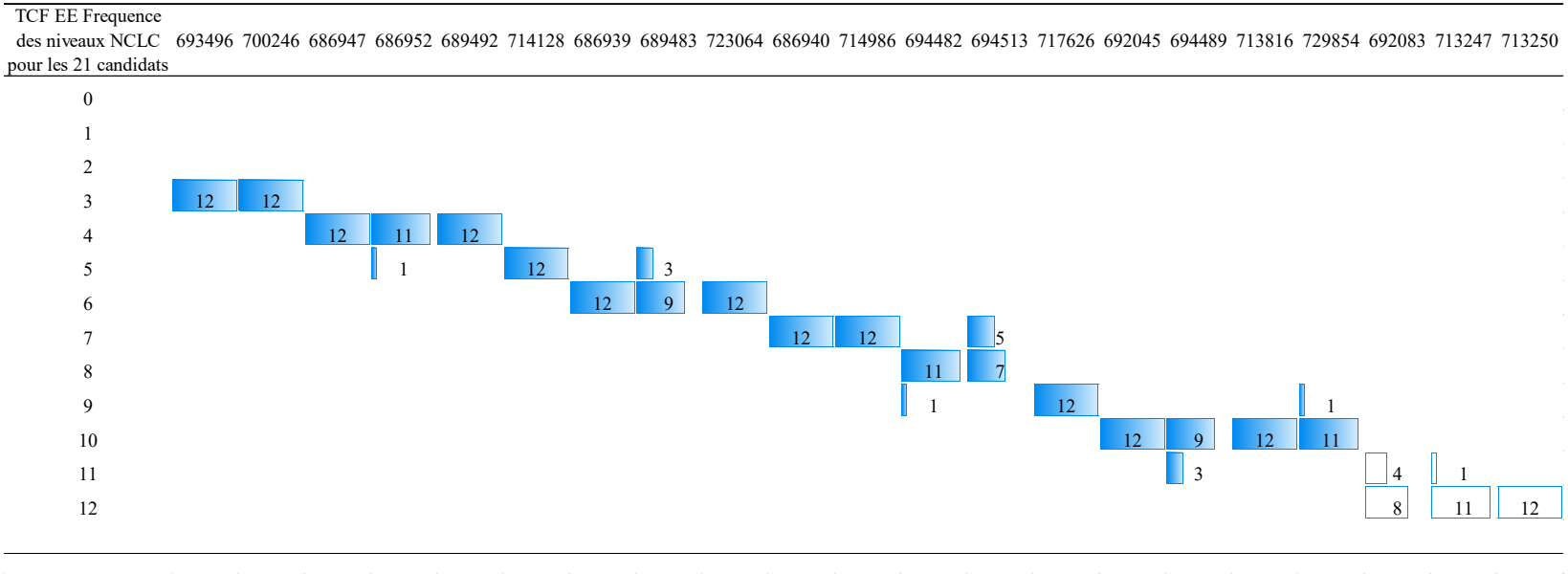
	Panéliste 7	Panéliste 8	Panéliste 9	Panéliste 10	Panéliste 11	Panéliste 12
Accords exacts	90%	95%	95%	86%	86%	100%
Accords exacts et adjacents	100%	100%	100%	100%	100%	100%

	Moyenne
Accords exacts	91%
Accords exacts et adjacents	100%

Annexe U
Classement des productions écrites (tour 3)



Annexe V
Positionnement des productions écrites avec les descripteurs NCLC (tour 3)



Annexe W
Fidélité des panélistes pour le classement des productions écrites (tour 3)

TCF EE NCLC	Panéliste 1	Panéliste 2	Panéliste 3	Panéliste 4	Panéliste 5	Panéliste 6
Accords exacts	90%	81%	100%	81%	71%	90%
Accords exacts et adjacents	100%	100%	100%	100%	100%	100%

	Panéliste 7	Panéliste 8	Panéliste 9	Panéliste 10	Panéliste 11	Panéliste 12
Accords exacts	43%	90%	90%	71%	90%	86%
Accords exacts et adjacents	86%	100%	100%	100%	100%	95%

	Moyenne
Accords exacts	82%
Accords exacts et adjacents	98%

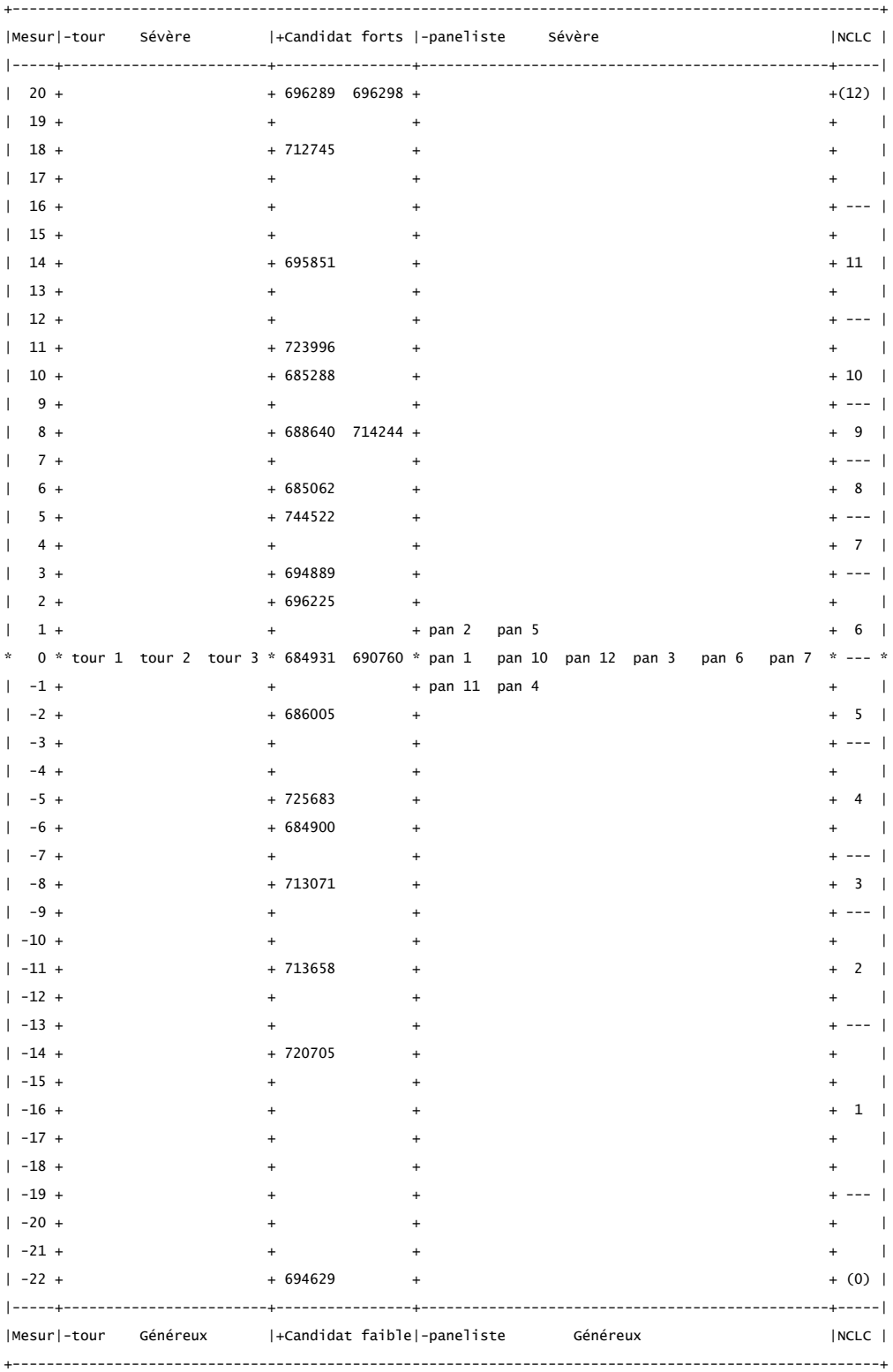
Annexe X
Fidélité des panélistes pour l'évaluation des productions écrites avec les niveaux NCLC (tour 3)

TCF EE NCLC	Panéliste 1	Panéliste 2	Panéliste 3	Panéliste 4	Panéliste 5	Panéliste 6
Accords exacts	90%	90%	90%	86%	90%	95%
Accords exacts et adjacents	100%	100%	100%	100%	100%	100%

	Panéliste 7	Panéliste 8	Panéliste 9	Panéliste 10	Panéliste 11	Panéliste 12
Accords exacts	95%	95%	95%	95%	90%	95%
Accords exacts et adjacents	100%	100%	100%	100%	100%	100%

	Moyenne
Accords exacts	92%
Accords exacts et adjacents	100%

Annexe Y
Carte des tours, candidats, panélistes, niveaux NCLC (oral)



Annexe Z
Carte des tours, candidats, panélistes, niveaux NCLC (écrit)