Évaluation des compétences linguistiques des RAPEC

Phase II : Résultats de la mise à l'essai

RAPPORT FINAL

directions
Evidence + Policy Research Group

cmec
Conseil des ministres de l'Éducation (Canada)
Council of Ministers of Education, Canada

Fondé en 1967, le Conseil des ministres de l'Éducation (Canada) [CMEC] donne aux ministres de l'Éducation au Canada une voix collective et leur permet d'assumer leur leadership en éducation aux échelons pancanadien et international. L'organisme aide les provinces et les territoires à exercer leur compétence exclusive en éducation.

•

Les Registraires de l'agrément du personnel enseignant Canada (RAPEC), un comité établi en 1999 à la demande du CMEC, échangent des informations sur la réglementation de la profession enseignante partout au Canada. De plus, ils coordonnent la mise en œuvre de l'Accord de libre-échange canadien (ALEC) pour la profession enseignante. Le comité regroupe les registraires de l'agrément du personnel enseignant de l'ensemble des provinces et des territoires.

•

Le Groupe de travail sur les compétences linguistiques (GTCL) est un groupe de travail de durée limitée chargé d'appuyer le projet « Centre d'évaluation et intégration à l'échelle pancanadienne des enseignantes et enseignants formés à l'étranger », sous la direction des RAPEC.

Auteur

Le Groupe *Directions* : Recherche et analyse des politiques est l'auteur du présent rapport, dont le financement s'inscrit dans le projet « Centre d'évaluation et intégration à l'échelle pancanadienne des enseignantes et enseignants formés à l'étranger » mené dans le cadre du Programme de reconnaissance des titres de compétences étrangers (PRTCE) d'Emploi et Développement social Canada (EDSC) et des Registraires de l'agrément du personnel enseignant Canada (RAPEC).

Avertissement

Les avis, les interprétations, les conclusions et les recommandations formulés dans ce rapport sont ceux de l'auteur. Ils ne représentent pas forcément la politique, les positions ou les points de vue officiels du Conseil des ministres de l'Éducation (Canada) [CMEC], des gouvernements provinciaux et territoriaux du Canada, ni des organismes de réglementation en enseignement des provinces et des territoires du Canada.

Remerciements

L'analyse exposée dans le présent rapport a bénéficié de la participation des entités suivantes :

- les Registraires de l'agrément du personnel enseignant Canada (RAPEC);
- le Groupe de travail sur les compétences linguistiques (GTCL);
- Wired Solutions;
- Eunice Eunhee Jang, Ph. D.

This document is also available in English with this title:

RTCC Language-competency Assessment – Phase II: Results of the Pilot-Testing Process

Table des matières

Sommaire

Les enseignantes et enseignants formés à l'étranger (EEFE) souhaitant obtenir l'agrément au Canada ont l'obligation, quand leur formation à l'enseignement ne s'est faite ni en anglais ni en français, de prouver qu'ils sont capables de communiquer dans l'une des langues officielles du Canada. Les provinces et les territoires, qui ont, au Canada, la compétence exclusive en matière d'éducation, n'ont pas de méthode permettant de déterminer si les EEFE ont les compétences linguistiques répondant tout particulièrement aux exigences de l'enseignement. Les Registraires de l'agrément du personnel enseignant Canada (RAPEC, sous les auspices du Conseil des ministres de l'Éducation [Canada] — CMEC) ont cherché à se doter d'un outil d'évaluation des compétences linguistiques qui serait commun à l'ensemble des provinces et des territoires et qui les aiderait à respecter leurs obligations en matière de mobilité de la main-d'œuvre au Canada. L'examen des évaluations existantes a permis de montrer qu'aucune d'entre elles ne se concentrait suffisamment sur les compétences linguistiques exigées par l'enseignement. Les RAPEC ont par conséquent choisi de mettre au point une évaluation propre à la profession enseignante.

Le projet d'évaluation des compétences linguistiques des RAPEC comprend trois phases. La phase I (2010-2013) a comporté une analyse des travaux de recherche sur les compétences linguistiques exigées pour faire un bon travail d'enseignement, la définition d'un cadre de compétences linguistiques pour la profession enseignante et la mise au point d'outils d'évaluation des compétences linguistiques en français et en anglais pour tester ces compétences. La phase II (2019-2022) a comporté la mise à l'essai de ces outils d'évaluation, en vue d'en confirmer la validité sur le plan psychométrique et de permettre la création d'un modèle pour le déploiement du test. La phase III sera le déploiement du test en tant que tel.

C'est l'organisme appelé Groupe *Directions* : Recherche et analyse des politiques qui a été embauché par la Corporation du Conseil des ministres de l'Éducation, Canada (CCMEC) comme consultant principal pour la mise à l'essai de la phase II. Ses principales activités ont été de faire passer les tests en français et en anglais à une population sélectionnée pour la mise à l'essai, de faire corriger les réponses aux tests (par des enseignantes et enseignants expérimentés au Canada), d'effectuer des analyses psychométriques pour vérifier la fiabilité des évaluations et l'absence de tout biais et d'établir les normes que chaque EEFE a pour obligation de respecter pour pouvoir recevoir l'agrément. Les normes ont été établies par un registraire ou par une personne nommée par le registraire qui avait de l'expérience en enseignement et qui était familière des tests normalisés.

Le groupe *Directions* s'est servi d'un cadre pragmatique applicable aux examens d'agrément pour effectuer une étude sur les données issues de la mise à l'essai des tests, afin de veiller à ce que les évaluations des compétences linguistiques des RAPEC mesurent bien les compétences linguistiques appropriées, que les données obtenues soient fiables et dépourvues de tout parti pris et que les normes soient établies de façon appropriée.

Le groupe *Directions* recommande vivement l'adoption des évaluations des compétences linguistiques des RAPEC (en français et en anglais), parce que l'utilisation réfléchie de ces évaluations fournit une méthode équitable et défendable pour repérer, parmi les EEFE, celles et ceux qui répondent aux normes fixées par les RAPEC pour les compétences linguistiques nécessaires en vue de pouvoir exercer la profession enseignante au Canada. Après l'adoption des évaluations, il faudra que les améliorations apportées au test pour le perfectionner continuent d'être guidées par le cadre de compétences linguistiques des RAPEC. Il est également important de créer un guide de correction et il faudra effectuer régulièrement un travail de rassemblement et d'analyse des données produites par le test.

Contexte

La profession enseignante est la profession réglementée qui compte le plus de membres au Canada et elle est aussi l'une des 14 professions ciblées par le *Cadre pancanadien d'évaluation et de reconnaissance des qualifications professionnelles acquises à l'étranger*[1] du Forum des ministres du marché du travail. Les Registraires de l'agrément du personnel enseignant Canada (RAPEC)[2] ont réalisé des progrès importants en vue de renforcer l'équité, la transparence, la cohérence et la rapidité des procédures d'évaluation et de reconnaissance des qualifications.

Les RAPEC reçoivent plus de 5 000 demandes d'agrément par an de la part d'enseignantes et enseignants formés à l'étranger (EEFE), qui font face à des obstacles importants à se lancer sur le marché du travail au Canada ou dans leur mobilité au sein de ce marché du travail. L'un des obstacles qui perdurent pour les EEFE est l'évaluation de leurs compétences linguistiques. Près de la moitié des provinces et des territoires du Canada n'ont aucune exigence linguistique pour l'agrément dans l'enseignement et les instances n'ont pas, pour la majorité d'entre elles, de test uniforme en français pour évaluer les compétences linguistiques des enseignantes et enseignants des écoles de langue française. Les registraires qui exigent la maîtrise de la langue avant d'accorder l'agrément font appel à divers tests de compétences linguistiques, comme l'International English Language Testing System – IELTS (système international d'évaluation de l'anglais), le Test d'évaluation de français (TEF) pour le Canada et le Programme canadien d'évaluation des compétences linguistiques en anglais, qui ne sont pas conçus pour la profession enseignante et qui, par conséquent, ne ciblent pas les compétences qu'exige cette profession.

Le Conseil des ministres de l'Éducation (Canada) (CMEC) travaille, sous la direction des RAPEC, à la mise au point d'une évaluation des compétences linguistiques pour les EEFE souhaitant exercer au Canada, à laquelle ils participeront lors de l'évaluation initiale de leurs diplômes d'études et de leurs qualifications professionnelles[3]. Le test linguistique propre à la profession enseignante doit servir à évaluer les aptitudes linguistiques des EEFE qui n'ont pas suivi de programme acceptable de formation à l'enseignement en français ou en anglais. Le but de ce test est de veiller à ce que les candidates et candidats possèdent les compétences linguistiques exigées pour pouvoir prodiguer un enseignement à la majorité anglophone ou francophone et aussi en contexte minoritaire.

Le projet d'évaluation des compétences linguistiques des RAPEC comprend trois phases :

- Phase I (2010-2013)
  - analyse des travaux de recherche sur les compétences linguistiques exigées pour faire un bon travail d'enseignement;
  - tour d'horizon des provinces et des territoires du Canada pour voir s'il existe une évaluation des compétences linguistiques convenant aux professionnels de l'enseignement formés à l'étranger qui souhaitent obtenir l'agrément au Canada;
  - définition d'un cadre de compétences linguistiques pour la profession enseignante;
  - mise au point d'items pour l'évaluation des compétences linguistiques en français et en anglais et de versions de l'évaluation pour tester ces compétences;
- Phase II (2019-2022)
  - mise à l'essai de ces outils d'évaluation, en vue d'en confirmer la fiabilité et la validité et de vérifier que les décisions reposant sur ces outils sont défendables;
  - création d'un modèle pour le déploiement du test;
- Phase III (2023)
  - déploiement en bonne et due forme de l'évaluation des compétences linguistiques des RAPEC.

Dans ce rapport, nous nous concentrons sur la mise à l'essai de la phase II, effectuée par le Groupe *Directions* : Recherche et analyse des politiques, consultant principal responsable de cette phase.

[1] *Cadre pancanadien d'évaluation et de reconnaissance des qualifications professionnelles acquises à l'étranger*, Forum des ministres du marché du travail, 2009; sur Internet : https://www.canada.ca/content/dam/esdc-edsc/documents/programs/foreign-credential-recognition/CA-561-11-09-FR.pdf

[2] Les Registraires de l'agrément du personnel enseignant Canada (RAPEC) est un comité établi en 1999 à la demande du CMEC. Les registraires membres de ce comité échangent des informations sur la réglementation de la profession enseignante au Canada. Ils assurent également la coordination de la mise en œuvre de l'Accord de libre-échange canadien (ALEC) pour la profession enseignante. Le comité regroupe les registraires de l'agrément du personnel enseignant de l'ensemble des provinces et des territoires.

[3] Les RAPEC font référence à une *évaluation* des compétences linguistiques dans le contexte de l'enseignement. Dans la terminologie psychométrique, la définition du mot *test* est très générale et peut inclure des tâches d'évaluation qui ne se présentent pas sous la forme d'épreuves conventionnelles avec un crayon et du papier. Dans le présent rapport, nous utilisons les termes *test* et *évaluation* de façon interchangeable, mais il est entendu que nous faisons référence à l'*évaluation* des compétences linguistiques relevant du mandat des RAPEC.

Phase I : Mise au point de l'évaluation des compétences linguistiques des RAPEC

Analyse des travaux de recherche et tour d'horizon des provinces et des territoires

La première étape, dans la mise au point de l'évaluation, était d'analyser les travaux de recherche sur la question, afin de déterminer les compétences linguistiques exigées de la part des enseignantes et enseignants du primaire-secondaire, dans les écoles d'anglais langue maternelle et de français langue maternelle du Canada, pour qu'ils puissent bien faire leur métier. Le groupe *Directions* a examiné les travaux de recherche dans le domaine de l'enseignement de la langue, de l'apprentissage de la langue et des compétences pour l'enseignement (soit 761 communications en anglais et 394 communications en français). Cet examen indique que l'enseignement exige des enseignantes et enseignants qu'ils possèdent un large éventail de compétences linguistiques variées pour qu'ils puissent connaître la réussite dans l'exercice de la profession[4]. La panoplie de compétences est la même pour les enseignantes et enseignants en contexte d'anglais langue maternelle et en contexte de français langue maternelle, mais la réalité des différences dans les contextes linguistiques (langue maternelle, langue minoritaire, etc.) peut faire que les exigences soient différentes pour les enseignantes et enseignants en ce qui concerne leurs compétences et connaissances linguistiques.

L'examen des évaluations existantes portant sur les compétences linguistiques révèle qu'elles présentent plusieurs limites et qu'il n'existe aucune évaluation qui évalue les compétences linguistiques en compréhension écrite, en expression écrite, en compréhension orale et en expression orale telles qu'elles s'appliquent à l'enseignement au primaire-secondaire au Canada, dans le contexte anglophone et dans le contexte francophone. Les évaluations générales de langue seconde en milieu de travail n'évaluent pas les compétences linguistiques telles qu'elles s'appliquent à l'enseignement (enseignement et évaluation, gestion de la classe et du comportement des élèves et communications avec les parents et avec d'autres spécialistes professionnels). Les évaluations linguistiques portant tout particulièrement sur l'enseignement n'évaluent pas l'ensemble des quatre modalités linguistiques et n'évaluent pas tous les domaines linguistiques propres à l'enseignement (enseignement et évaluation, gestion de la classe et du comportement des élèves et communications avec les parents et avec d'autres spécialistes professionnels). Il s'agit aussi d'évaluations conçues pour l'enseignement dans des contextes extérieurs au Canada ou pour lesquelles les données relatives à la validité et à la fiabilité ne sont pas claires.

Cadre des compétences linguistiques

L'analyse des travaux de recherche a servi à éclairer la mise au point du cadre des compétences linguistiques et des niveaux de compétence linguistique pour le personnel enseignant[5]. Les compétences linguistiques sont une série d'énoncés décrivant les aptitudes linguistiques de l'individu en anglais ou en français selon chacune des quatre modalités suivantes : expression orale, compréhension orale, expression écrite et compréhension écrite. Ces modalités se retrouvent couramment dans les cadres relatifs à la maîtrise de la langue. Pour chaque domaine de compétence, le cadre définit des résultats visés dans trois domaines d'activité : (A) enseignement et évaluation; (B) gestion de la classe et du comportement des élèves; et (C) communication avec les parents et d'autres professionnels.

À titre d'exemple, la première compétence dans le domaine de l'expression écrite est la suivante : « écrire des textes cohérents, dans le langage formel ou informel, qui résument et évaluent des renseignements et

[4] *Parlons d'excellence : Compétences linguistiques pour un enseignement efficace*, Conseil des ministres de l'Éducation (Canada), 2013; sur Internet : https://www.cmec.ca/Publications/Lists/Publications/Attachments/320/Parlons_dexcellence.pdf.

[5] *Ibid.*

des idées complexes tirés de sources diverses ». Les résultats correspondants dans les différents domaines d'activité sont les suivants :

- **enseignement et évaluation** — rédiger des plans de leçons, des plans de cours, des descriptions de cours, de la documentation destinée aux élèves ou du matériel pédagogique;
- **gestion de la classe et du comportement des élèves** — rédiger des résumés des attentes et des objectifs relatifs à la classe;
- **communication avec les parents et d'autres professionnels** — rédiger des courriers électroniques, des lettres ou des rapports destinés à d'autres professionnels du milieu scolaire en utilisant un langage à caractère technique ou non.

Mise au point des items et construction du test

En 2012, le groupe *Directions* a mis au point environ 1 600 items d'évaluation en français et 1 600 items d'évaluation en anglais, à partir du cadre de compétences linguistiques. Les items en expression écrite dans le domaine de l'enseignement et de l'évaluation, par exemple, comprenaient la rédaction de plans de leçons, de descriptions de cours et d'autres ressources pédagogiques. Les items ont été mis au point de façon indépendante en anglais et en français par un groupe composé de divers éducateurs et éducatrices (enseignantes et enseignants en exercice, anciens enseignants et enseignantes travaillant toujours dans l'éducation et spécialistes des évaluations linguistiques). À partir de cette banque d'items, le groupe *Directions* a mis au point cinq versions du test en français et cinq versions du test en anglais, qui ont été examinées par des spécialistes externes (spécialistes reconnus sur la scène internationale dans le domaine de l'évaluation dans l'enseignement, du perfectionnement professionnel du personnel enseignant, de la langue dans l'éducation et de l'élaboration de programmes d'études) en anglais et en français, pour en vérifier l'authenticité, la validité apparente et la validité du contenu et pour recueillir des suggestions d'améliorations des items. L'équipe a révisé et remplacé ses items en incorporant les commentaires et suggestions à la fois de ces spécialistes externes et des membres du Sous-comité des RAPEC chargé des compétences linguistiques.

En 2013, deux versions de l'évaluation, l'une française et l'autre anglaise, ont été testées en ligne dans le cadre d'une procédure de test alpha, afin de veiller à ce que l'interface soit utilisable et que les exigences de l'évaluation soient claires pour les participantes et participants visés[6]. L'interface ou plateforme et les versions des tests ont été révisées en fonction des résultats de ce test alpha, ce qui a marqué l'achèvement de la phase I du projet.

[6] Les participantes et participants au test alpha étaient 38 volontaires, pour la plupart des enseignantes et enseignants formés à l'étranger.

Phase II : Mise à l'essai de l'évaluation des compétences linguistiques

Buts

Le groupe *Directions* a été embauché comme consultant principal pour effectuer la mise à l'essai des deux versions (anglaise et française) de l'évaluation et chargé de diriger un processus pancanadien de validation de la fiabilité de l'évaluation linguistique en ligne des RAPEC pour la profession enseignante au Canada. Le plan initial pour la phase II comprenait l'établissement de centres régionaux pour la mise à l'essai, le recrutement par le Groupe de travail sur les compétences linguistiques (GTCL) de jusqu'à 2 400 participantes et participants pour les tests (dont 1 200 pour le test en français et 1 200 pour le test en anglais), la préparation de jusqu'à trois évaluations de langue anglaise et trois évaluations de langue française à l'aide des items mis au point à la phase I, la coordination du processus de formation et de correction en personne, l'exécution d'analyses psychométriques portant sur jusqu'à 120 000 réponses aux items de la mise à l'essai (dont 60 000 en français et 60 000 en anglais), la révision des items en fonction des conclusions et l'appui à la définition de normes minimales de compétence linguistique pour l'agrément au niveau pancanadien.

Planification et mise en œuvre : répercussions de la pandémie de COVID-19

Les préparatifs des sessions en personne pour les tests eux-mêmes et pour la correction des tests, prévues à l'été 2020, étaient en cours quand la pandémie de COVID-19 a été déclarée, en mars 2020. Le tableau 1 fournit un résumé des répercussions de la pandémie pour le projet et son calendrier.

Tableau 1. Répercussions de la pandémie de COVID-19 pour la mise à l'essai et le calendrier de la phase II

Activités prévues	Activités effectuées en réalité	Calendrier
Déroulement du test : quatre centres régionaux pour les tests (C.-B., Prairies, Ontario, provinces de l'Atlantique) avec la plateforme de test XpressLab	Déroulement du test au domicile des participantes et participants partout au Canada, à l'aide de la plateforme de test XpressLab et des services de surveillance en ligne Proctortrack	• janvier – mars 2021 • septembre – décembre 2021
Recrutement des participantes et participants : recrutement dans l'ensemble des provinces et des territoires de participantes et participants au test parmi les personnes ayant nouvellement intégré la profession enseignante (avec une formation à l'enseignement en anglais ou en français) et parmi les enseignantes et enseignants formés à l'étranger (avec une formation à l'enseignement ni en anglais ni en français)[7], grâce à un recrutement ciblé auprès des personnes suivantes :	Recrutement dans l'ensemble des provinces et des territoires de participantes et participants au test dans les catégories initialement prévues. La faiblesse des effectifs au début 2021 et les difficultés de recrutement au niveau provincial et territorial ont exigé l'élargissement du recrutement des participantes et participants aux individus : • se préparant à enseigner; • exerçant à l'heure actuelle la profession enseignante; • ayant pris leur retraite de l'enseignement.	• novembre 2020 – mars 2021 • août – décembre 2021

[7] Les données sur les enseignantes et enseignants débutants en anglais ou en français fournissent des indications pour l'établissement des normes, l'objectif étant que le test pour les EEFE ne soit pas plus exigeant que la norme pour les candidates et candidats ayant suivi une formation en anglais ou en français.

Tableau 1. Répercussions de la pandémie de COVID-19 pour la mise à l'essai et le calendrier de la phase II (suite)

Activités prévues	Activités effectuées en réalité	Calendrier
• candidates et candidats à la profession enseignante aux derniers stades de leur programme de formation; • diplômées et diplômés récents des programmes de formation; • enseignantes et enseignants formés à l'étranger, y compris ceux faisant une demande d'agrément; • enseignantes et enseignants certifiés se trouvant dans leurs cinq premières années d'enseignement. Un seul cycle de recrutement.	Le nombre de participantes et participants au test était insuffisant (70 en français, 143 en anglais) entre janvier et mars 2021, ce qui a conduit les RAPEC à organiser un deuxième cycle de recrutement, d'août à décembre 2021.	
Effectif de participantes et participants au test : Pour que les erreurs de mesure soient réduites à un niveau acceptable pour un test de si grande importance, l'objectif du CMEC était de recruter au minimum 400 participantes et participants pour chaque version (versions 2, 3 et 4 du test anglais et versions 2, 3 et 4 du test français), soit au total 1 200 participantes et participants en anglais et 1 200 participantes et participants en français.	En raison de la faiblesse du nombre d'inscriptions aux premières phases du recrutement, la décision a été prise de ne tester que deux versions de l'évaluation des compétences linguistiques en anglais et deux versions de l'évaluation en français. Cette approche devait permettre de répartir un plus grand nombre de participantes et participants entre un nombre plus réduit de versions. Le nombre final de participantes et participants a été : • 589 participantes et participants pour les versions 2 et 3 du test en anglais; • 349 participantes et participants pour les versions 2 et 3 du test en français.	• janvier – mars 2021 • septembre – décembre 2021
Correction des items en expression écrite et en expression orale : session en personne de formation des correctrices et correcteurs et de correction à l'aide de l'interface de correction de la plateforme de test XpressLab.	Sessions virtuelles de formation et de correction, deux sessions avec 91 correctrices et correcteurs, à l'aide de l'interface de correction de la plateforme de test XpressLab et de l'outil de communication en ligne Slack	• mars – avril 2021 • décembre 2021
Établissement des normes : réunions en personne	Réunions virtuelles (5 responsables de l'établissement des normes en anglais et 5 responsables de l'établissement des normes en français)	• février – mars 2022

Structure du test

Les versions de l'évaluation mises au point en 2012 et révisées après la phase de test alpha en 2013 ont été réexaminées pour voir si des changements étaient exigés. Les critères de correction pour les items en expression écrite et en expression orale ont été révisés pour la phase II dans l'optique de simplifier le travail de correction. Pour tester une plus vaste banque d'items, le groupe *Directions* a incorporé des items supplémentaires des versions 1 et 5 du test mis au point lors de la phase I dans les modules pour les versions 2, 3 et 4. Chaque version du test a également été réexaminée en vue de vérifier qu'elle couvrait bien les résultats visés définis dans le cadre des compétences linguistiques et des niveaux de compétence linguistique.

Pour la mise à l'essai de la phase II, quatre versions du test ont finalement été utilisées (version 2 [V2] et version 3 [V3] en anglais et version 2 [V2] et version 3 [V3] en français). Tant dans les versions françaises que dans les versions anglaises, il y a un ensemble d'items pour chaque modalité qui est commun aux deux versions. (À titre d'exemple, pour la compréhension orale, cinq items sont identiques dans V2 et dans V3[8].) Les autres items sont propres à chaque version.

Questions du test

Chaque item du test a été mis au point par rapport à un résultat de rendement figurant dans le cadre des compétences linguistiques et des niveaux de compétence linguistique. Chaque item du test comprend un stimulus (passage à lire, scénario, clip sonore, etc.) et une ou plusieurs questions ou tâches pour la participante ou le participant. Les items du test comprennent des questions à réponse choisie (choix multiples), des questions de type « Cloze » (phrase à compléter) et des questions à réponse construite (texte à rédiger ou à enregistrer en réponse aux instructions).

Exemples d'items

Vous trouverez ci-dessous des exemples d'items de test dans chaque modalité. Les résultats de rendement pour chaque item sont fournis ici à titre de référence, mais ne sont pas montrés aux participantes et participants au test.

Compréhension écrite

Résultats de rendement : Lire, évaluer et critiquer une variété d'histoires, de dissertations et d'autres textes rédigés par les élèves afin de leur fournir une rétroaction et d'évaluer leur rendement et leurs progrès.

Stimulus

Le texte suivant est une dictée rédigée par un de vos élèves dans votre classe de 4e année. Cette dictée vise principalement à évaluer la compétence des élèves à accorder les adjectifs, ainsi que leur maîtrise de la conjugaison aux personnes du pluriel. Lisez la dictée de l'élève en vue de déterminer la rétroaction appropriée sur son travail.

[8] Lors de la construction du test à la phase I, un nombre égal d'items très pertinents a été mis en évidence dans chaque modalité en anglais et en français. Ces items sont devenus les items communs aux différentes versions anglaises et aux différentes versions françaises. Les items communs en français ont été traduits en anglais et les items communs en anglais ont été traduits en français. Les révisions qui ont été apportées par la suite aux items communs en anglais ont été faites de façon indépendante des révisions apportées aux items communs en français. Il faut donc considérer que les items ne sont communs qu'aux différentes versions dans la même langue et non aux versions dans les deux langues.

Promenade en forêt.

Les éléves sont partis faire une promenade dans un petit bosquet derrière l'école. Les filles cherchent des champignons caché sous les feuilles mortes tandis que les garçons ceuillet des myrtilles odorante. Ils aimerais bien voir des écureuils mais ces petits animaux ne se montres pas volontié. De retour en classe, ils mangerons une omelette aux champignons et fabriquerons de la glace aux myrtilles.

Question

Quel énoncé fournit LA MEILLEURE ÉVALUATION de la dictée rédigée par l'élève?

A. L'élève fait principalement des erreurs d'orthographe d'usage, tandis que les erreurs de grammaire sont rares.
B. L'élève ne semble pas maitriser les règles de l'accord des adjectifs, mais son orthographe et sa conjugaison sont bien maitrisées.
C. L'élève commet l'une ou l'autre faute d'orthographe d'usage, mais se montre surtout maladroit quant aux règles de conjugaison.
D. L'élève commet quelques erreurs de conjugaison, principalement, et semble avoir manqué d'attention à l'écoute car la ponctuation et certains passages n'ont pas de sens.

Expression écrite

Résultats de rendement : Noter le rendement et les progrès des élèves sur des formulaires normalisés.

Stimulus

Vous allez enregistrer des commentaires à partir de la grille d'évaluation suivante que vous utilisez pour apprécier les productions de vos élèves dans votre cours d'arts dramatiques de 11e année.

Présentation théâtrale – 11e année : critères d'évaluation

A. Excellent	B. Bon	C. Moyen	D. Suffisant	E. Insuffisant

Les détails donnés dans cette fiche concernent le travail effectué par Mounir Prévost-Hassan, l'un de vos élèves.

Élève : Mounir Prévost-Hassan	
Présence scénique (maintien, utilisation de l'espace, attitude et énergie)	A. Mise en scène précise, juste, sans fioriture
Émotion (capacité à créer et à communiquer des états)	B. Visage, silences justes
Voix (volume, tonalité, diction)	D. Volume trop faible et donc inaudible, travail nécessaire
Scénographie (maquillage, costumes, décors, accessoires)	C. Maquillage et costumes n'apportent rien à la scène, pas nécessaire
Originalité (concept personnalisé, capacité de surprendre)	A.

Directives

En conclusion de votre fiche d'évaluation, vous réservez un espace pour résumer et interpréter le rendement de l'élève. Compte tenu des commentaires et observations présentés dans la fiche d'évaluation, achevez la rédaction du commentaire général en complétant les quatre phrases qui suivent par des informations pertinentes et adaptées. La fiche d'évaluation est destinée à votre élève de 16 ans. Certaines réponses peuvent être résumées en un seul mot, mais il est parfois nécessaire de proposer plus qu'un mot.

- _____ pour ta présentation très originale.
- Tu es parvenu à transmettre beaucoup d'émotions grâce à _____.
- Repense peut-être à ton utilisation des costumes et du maquillage, afin qu'ils _____ ta mise en scène.
- Tu dois _____ afin de mieux utiliser ta voix, que l'on n'entendait parfois pas du tout.

Critères d'évaluation : (9 points)

1. Le texte est cohérent et répond aux tâches exigées. (3 points)
2. Le texte utilise un vocabulaire approprié à l'audience. (3 points)
3. Le texte est dépourvu de fautes d'orthographe, de ponctuation et de grammaire. (3 points)

Compréhension orale

Résultats de rendement : Écouter les élèves lors d'une leçon de lecture à voix haute afin d'évaluer leurs compétences en lecture.

Stimulus

Les élèves de votre classe sont revenus de la récréation en se plaignant que la partie de soccer était injuste. Vous décidez de leur enseigner les règles de « l'esprit sportif ». Vous avez développé différentes mises en scène pour illustrer ces règles. Vous demandez à vos élèves de lire à haute voix ces scénarios. Écoutez Matéo lire le scénario suivant afin d'évaluer ses compétences en lecture en tenant compte de sa reconnaissance et de sa prononciation des mots, ainsi que du décodage de la structure du texte.

Mégane est la capitaine de l'équipe de soccer de l'heure du dîner, Les Arrêteurs. C'est sa responsabilité de s'assurer que tous les membres de son équipe ont une chance de jouer. Mégane est une très bonne joueuse et elle connaît toutes les règles du jeu. Elle aime jouer, alors elle reste sur le terrain et, lorsqu'il y a changement de joueurs, elle continue de jouer. Ce qui veut dire qu'il n'y a de joueurs sur le terrain que pendant une période de jeu de 5 minutes, tandis que Mégane joue pendant toute la partie. Est-ce que c'est juste? Si non, pourquoi?

Le texte tel que réalisé par Matéo [linked audio « Mégane est la capitaine. De l'équipe de soccer à l'heure du dîner Les Arrêts. C'est sa res…pon…sa…blité de s'assurer, que les membres de son équipe ont une chance de jouer Mégane est une très bonne joueuse. Et elle, connaît toutes les règles du jeu elle aime jouer. Alors elle reste sur le terrain et lorsqu'il y a changement de joueurs, elle continue de jouer. Cela veut dire qu'il y a… des joueurs sur…. le terrain que pendant une période, de jeu de 5 minutes? Tandis que Mégane joue pendant toute la partie est-ce que c'est juste? Si, non pourquoi? »]

Question

Comment décririez-vous les compétences de Matéo en lecture à haute voix?

a. Matéo est un lecteur compétent démontrant une grande précision dans la lecture des mots. De plus, il respecte la ponctuation et a une bonne prononciation.
b. Matéo a de la difficulté à décoder plusieurs mots et sa prononciation n'est pas claire.
c. Matéo n'a pas trop de difficulté à lire les mots, mais il ne tient pas compte de la ponctuation. Cela cause un problème dans le décodage de la structure du texte, mais il est possible de comprendre le sens du texte lu.
d. Matéo lit mot à mot. Il ne fait pas attention à la ponctuation et sa prononciation n'est pas claire. Cela nuit à la compréhension du texte lu.

Expression orale

Résultats de rendement : Donner des directives pour s'assurer que les élèves suivent les règles de la classe en ce qui concerne le travail scolaire et le comportement.

Stimulus

Vous êtes en train d'enseigner une leçon sur l'impact de la Guerre froide au Canada à vos élèves de dernière année du secondaire. En plus de traiter des concepts qui seront évalués lors de l'examen provincial, vous avez comme objectif lors de cette leçon de favoriser chez vos élèves une compréhension globale des événements du XX^{e} siècle et du rôle que le Canada a joué dans les conflits internationaux de cette période.

Un de vos élèves, Éric, ne prend pas de notes pendant votre leçon; ses bras sont croisés et il a l'air contrarié.

Vous lui posez la question suivante : « Éric, peux-tu me dire pourquoi tu ne prends pas de notes pendant la leçon? Est-ce qu'il y a un problème? »

Il vous répond : « J'ai pas envie. Pourquoi est-ce qu'on doit apprendre ça, de toute façon? »

Directives

Vous devez répondre au commentaire de votre élève Eric afin de le convaincre de se remettre au travail et de lui faire percevoir l'importance de la leçon au-delà des exigences de l'examen provincial.

Temps accordé à la préparation : 3 minutes
Durée de l'intervention : 2 minutes

Critères d'évaluation : (9 points)

1. Le discours est cohérent, répond aux tâches exigées et est approprié à l'audience. (3 points)
2. Le discours est intelligible avec une prononciation claire. (3 points)
3. Le discours est fluide (l'accent tonique, l'articulation, le débit et le ton du discours sont appropriés à la tâche). (3 points)

Interface du test

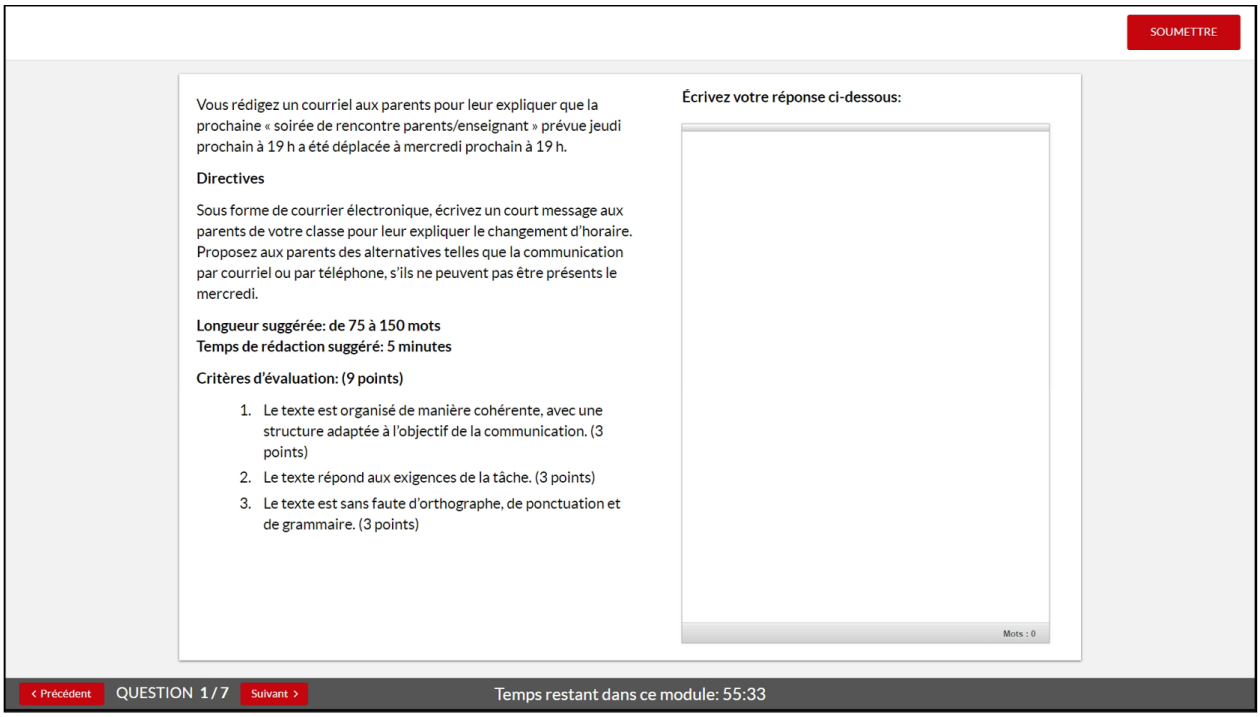
Les participantes et participants au test ont utilisé la plateforme XpressLab, qui est un logiciel de test des compétences linguistiques sur serveur qui peut être adapté sur mesure[9]. La surveillance des épreuves a été effectuée en direct à l'aide de Proctortrack de Verificient, service de surveillance à distance. L'évaluation comprend quatre modules d'une heure (soit une heure pour la compréhension écrite, une heure pour l'expression écrite, une heure pour l'expression orale et une heure pour la compréhension orale), avec au maximum 15 minutes de pause entre les modules. Les participantes et participants pouvaient lancer les modules dans l'ordre de leur choix.

[9] L'entreprise d'informatique Wired Solutions, qui a fourni le logiciel XpressLab pour la phase I, a été embauchée pour la mise à l'essai de la phase II.

Les formats de question suivants ont été utilisés dans l'évaluation :

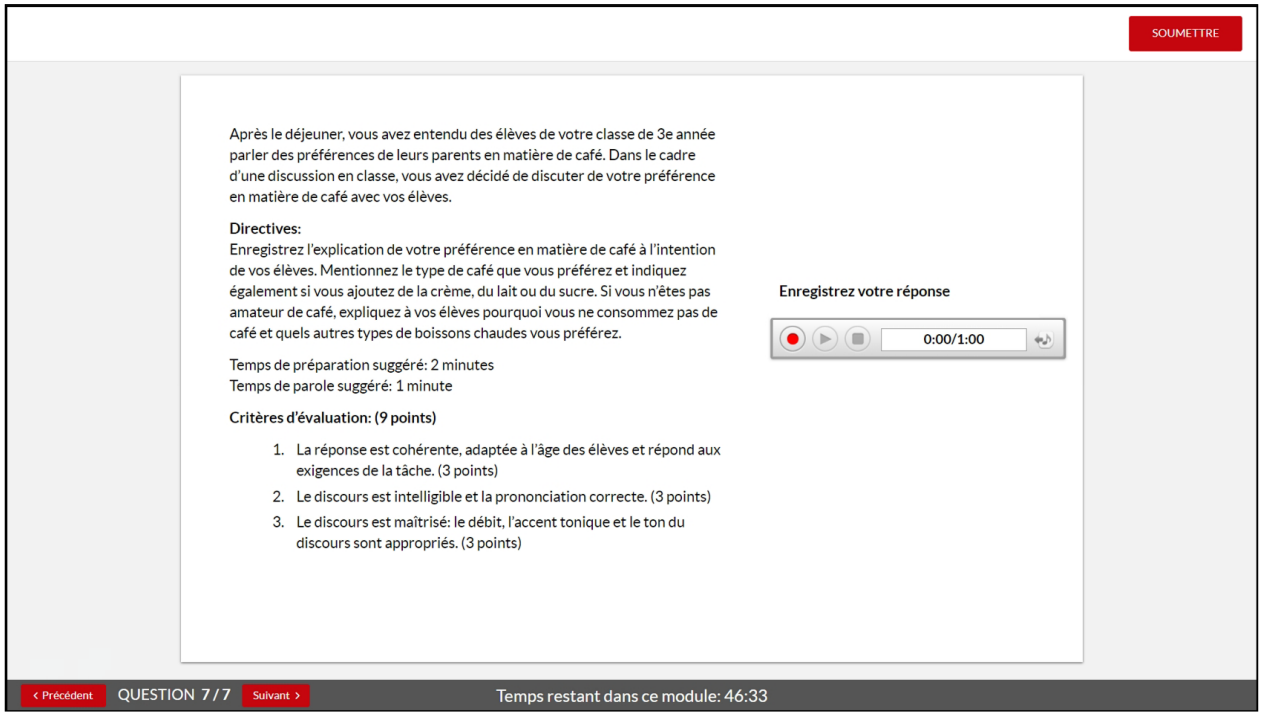
Expression écrite : Les tâches contenaient un stimulus, des directives et des critères d'évaluation. Les participantes et participants fournissaient une réponse écrite.

Figure 1. Interface pour la réponse écrite



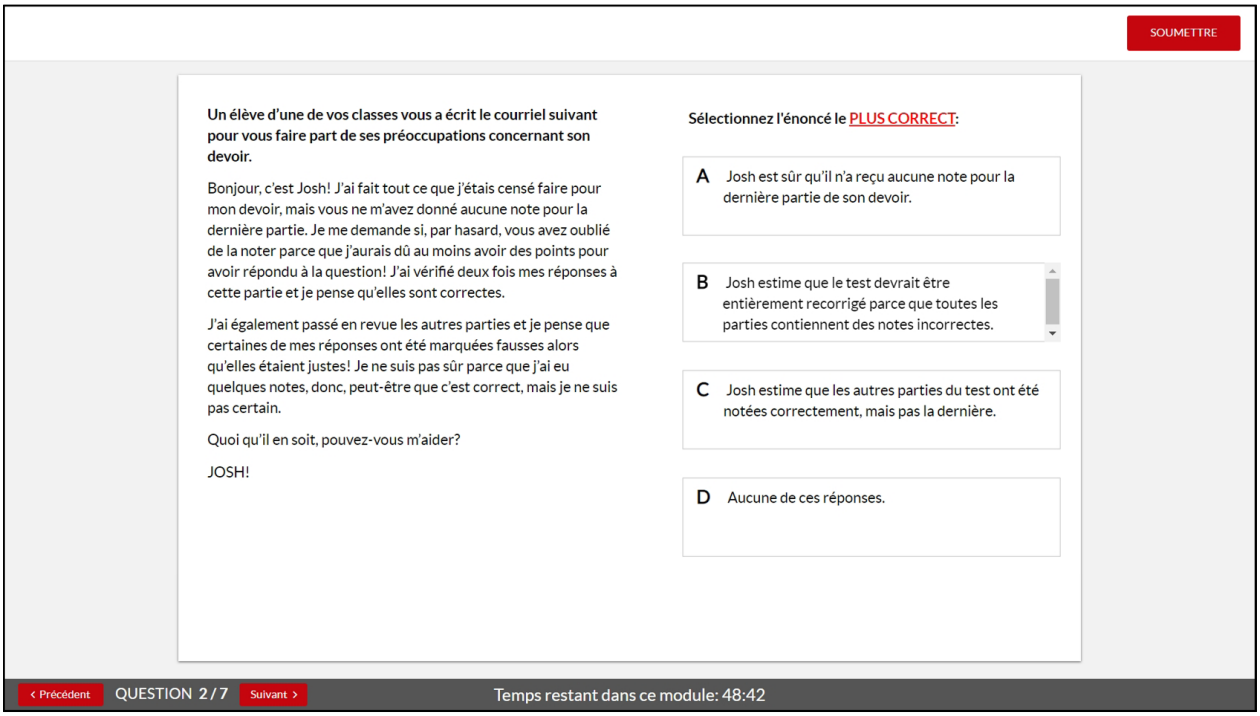
Expression orale : Les tâches contenaient un stimulus, des directives et des critères d'évaluation. Les participantes et participants enregistraient une réponse parlée.

Figure 2. Interface pour la réponse parlée



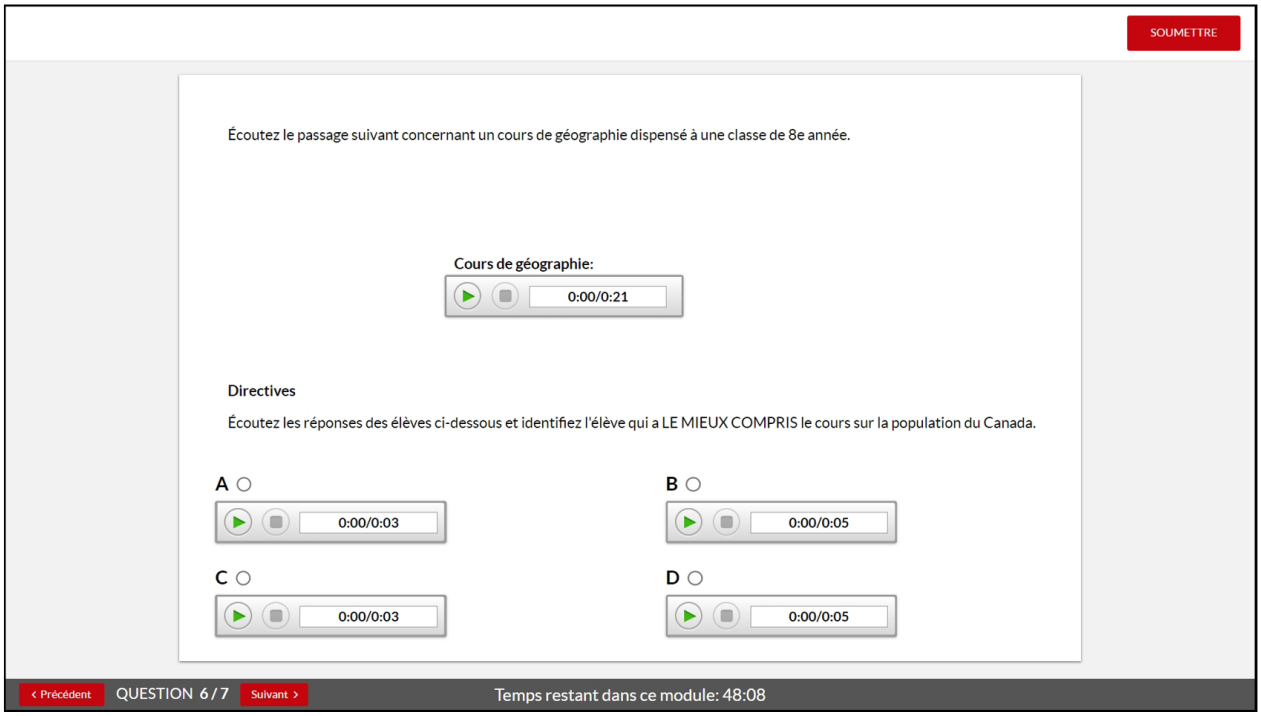
Compréhension écrite ou orale — options textuelles de réponse à choix multiples : Les tâches contenaient un stimulus textuel ou audio et des directives pour sélectionner une réponse parmi les options textuelles proposées.

Figure 3. Interface pour les réponses textuelles à choix multiples



Compréhension orale – options audio de réponse à choix multiples : Les tâches contenaient un stimulus audio ou textuel et des directives pour sélectionner une réponse parmi les options audio proposées.

Figure 4. Interface pour les réponses audio à choix multiples



Déploiement de la mise à l'essai du test et participantes et participants au test

Les RAPEC ont recruté des participantes et participants pour le test entre novembre 2020 et mars 2021 et une nouvelle fois entre août et décembre 2021.

La mise à l'essai du test s'est faite à distance dans le cadre de deux sessions. La première session (de janvier à mars 2021) a concerné 70 participantes et participants en français (V2 et V3) et 143 participantes et participants en anglais (V2 et V3). Si des analyses psychométriques ont été effectuées après cette première session, la faiblesse du nombre de participantes et participants a fait que la pertinence statistique des résultats était limitée, en particulier pour les participantes et participants de niveau inférieur, parce que la plupart des participantes et participants ont obtenu des résultats de bon niveau au test. Une deuxième session de mise à l'essai du test a donc été effectuée (entre septembre et décembre 2021). Ceci a permis d'améliorer le nombre total de participantes et participants, qui s'est élevé, sur l'ensemble des deux sessions, à 349 en français (V2 et V3) et à 589 en anglais (V2 et V3).

Les participantes et participants au test lors de la mise à l'essai étaient principalement des enseignantes certifiées ayant effectué leur formation à l'enseignement au Canada dans la langue du test. (Les participantes et participants au test en anglais, par exemple, étaient principalement des enseignantes ayant fait leur formation à l'enseignement au Canada et en anglais.) Les participantes et participants au test maîtrisaient mieux la langue du test que l'autre langue. (Autrement dit, les participantes et participants au test en français maîtrisaient mieux le français que l'anglais.)

Correction des items en compréhension orale et en compréhension écrite

Les modules de compréhension orale et de compréhension écrite étaient des tests à choix multiples corrigés automatiquement (0 pour les réponses incorrectes et 1 pour les réponses correctes) par le logiciel de test XpressLab. Le module de compréhension orale de la version 2 en anglais contenait un seul item avec des trous à remplir, qui était également corrigé automatiquement par le logiciel (0 pour les réponses incorrectes et 1 pour les réponses correctes).

Correction des items en expression orale et en expression écrite

Deux sessions de correction à distance pour les items en expression orale et en expression écrite ont eu lieu : (1) de mars à avril 2021 (correction des épreuves de janvier à mars 2021) et (2) en décembre 2021 (correction des épreuves de septembre à décembre 2021). La correction des réponses s'est faite en ligne à l'aide de l'interface fournie par la plateforme XpressLab.

Correctrices et correcteurs

Les correctrices et correcteurs et les chefs de table étaient des enseignantes et enseignants expérimentés, ayant de préférence au moins cinq années d'expérience dans l'enseignement. Toutes sortes de matières scolaires étaient représentées (éducation artistique, musique, langue et littérature, sciences, mathématiques, sciences humaines, etc.) et les différents niveaux scolaires étaient également représentés (école primaire, école intermédiaire, école secondaire). Bon nombre des correctrices et correcteurs étaient des personnes expérimentées. Les correctrices et correcteurs comprenaient des personnes nommées par les provinces et les territoires. Au total, 91 personnes ont participé à l'une des sessions de correction ou aux deux. Ces personnes venaient de l'Alberta, du Manitoba, de Terre-Neuve-et-Labrador, des Territoires du Nord-Ouest, de l'Ontario, de l'Île-du-Prince-Édouard et du Québec.

Formation et correction

Les correctrices et correcteurs ont participé à une session de formation de deux journées en anglais et en français, selon le cas, avec les éléments suivants : explication du processus de correction; travail indépendant de correction d'un échantillon commun de réponses à des items en expression orale et en expression écrite tirées d'épreuves authentiques; et discussions autour d'une table (groupes de 4 à 10 correctrices ou correcteurs avec un chef de table), entre chefs de table et en grand groupe sur les problèmes, les désaccords dans la correction et les difficultés rencontrées. Les documents de formation utilisés étaient les suivants : vidéos sur l'évaluation, sur le processus de correction et sur le logiciel de correction; échantillons d'items du test et de réponses; et manuel technique pour la correction et les arbitrages.

Nous avons utilisé un processus à plusieurs étapes pour renforcer la cohérence des corrections. La première étape pour les correctrices et correcteurs consistait à examiner trois réponses de qualité faible, moyenne et élevée à un item en expression orale et à un item en expression écrite. Ceci a servi de point de départ à la conversation et à la mise en évidence de ce qui distingue les différents niveaux de qualité des réponses. La deuxième étape était de fournir aux correctrices et correcteurs un ensemble commun de 10 items en expression orale et de 10 items en expression écrite à corriger. Les correctrices et correcteurs et les chefs de table ont reçu les données des corrections afin de pouvoir discuter des sources de désaccord et décider de la façon de procéder et des critères de correction à utiliser. Après cette mise au point, les correctrices et correcteurs ont reçu 30 à 50 réponses authentiques à des items en expression orale et en expression écrite à corriger. Une fois qu'ils ont corrigé ces items, la session de correction a marqué une pause. L'équipe du groupe *Directions* a rencontré les chefs de table (qui avaient eux-mêmes corrigé un choix de réponses, afin de comprendre le processus et les difficultés auxquelles leurs correctrices et correcteurs pouvaient avoir fait face) pour discuter des questions soulevées. Les analyses et discussions initiales ont fourni des informations que nous avons utilisées pour donner aux correctrices et correcteurs des consignes supplémentaires par écrit sur l'application des critères de correction. À ce stade, les correctrices et correcteurs pouvaient travailler à leur propre rythme pour faire leur quota avant la fin de la session. Tout au long du processus de formation et de correction, nous avons encouragé les correctrices et correcteurs à discuter fréquemment avec les autres membres de leur table, à l'aide de l'outil de communication en ligne Slack, sur les questions que le processus pouvait soulever.

Quand une évaluation a des points de coupure bien définis, les correctrices et correcteurs ont des guides détaillés pour la correction. Comme la présente évaluation ne comprenait aucun point de coupure, cependant, il n'y avait pas de guide de correction détaillé pour chaque item, avec des exemples de réponses correspondant aux normes établies. Dans le travail d'élaboration de l'évaluation, les correctrices et correcteurs avaient en fait pour tâche de mettre en pratique les « normes » de correction et de les perfectionner à mesure qu'ils progressent dans leur travail de correction. Il s'agissait donc d'un processus d'interprétation plus complexe que la correction d'une évaluation avec des normes bien établies et ce processus exigeait un dialogue en continu tout au long du processus, entre les correctrices et correcteurs, les chefs de table et le groupe *Directions*. Il s'agit là d'un processus normal lors de la mise à l'essai d'un instrument d'évaluation, qui éclaire, par la suite, l'élaboration d'un guide de correction et l'établissement de normes.

Nous avons demandé aux correctrices et correcteurs d'attribuer à chaque réponse en expression orale ou en expression écrite un score entre 0 et 9, en tenant compte des trois critères de correction accompagnant chaque item. Les trois critères variaient d'un item à l'autre, mais se présentaient sous le format général suivant :

Expression écrite : critères d'évaluation

- 1. La réponse répond-elle aux exigences de la tâche? (3 points)
- 2. La communication est-elle appropriée aux lecteurs? (3 points)
- 3. Conventions, grammaire, ponctuation, orthographe, etc. (3 points)

Expression orale : critères d'évaluation

- 1. Le discours est cohérent, répond aux tâches exigées et est approprié à l'auditoire. (3 points)
- 2. Le discours est intelligible avec une prononciation correcte. (3 points)
- 3. Le débit, l'accent tonique et le ton du discours sont appropriés. (3 points)

Nous demandions aux correctrices et correcteurs de fournir une justification écrite pour chaque score, afin de pouvoir analyser la façon dont les scores avaient été attribués.

Chaque réponse aux items en expression orale et en expression écrite a été corrigée par deux correctrices ou correcteurs différents de façon indépendante. Si l'intervalle entre les deux scores attribués était de deux points ou plus, le processus considérait qu'il y avait un désaccord et l'item, les réponses, les scores et les commentaires de la correctrice ou du correcteur étaient examinés par un arbitre (le chef de table), qui attribuait le score définitif. Si l'intervalle entre les deux scores n'était que d'un point, le score définitif était automatiquement la moyenne des deux scores.

Établissement des normes

Les données du test ont fait l'objet d'une analyse psychométrique avant le travail d'établissement des normes.

Les normes (points de coupure indiquant le seuil de réussite pour chaque module) ont été établies par deux comités distincts d'établissement des normes (un pour la version anglaise et un pour la version française du test). Un processus Angoff modifié a été utilisé pour définir les points de coupure pour les quatre modules, tant pour le test en anglais que pour le test en français.

Participantes et participants

Nous avons invité les registraires des différentes provinces et des différents territoires à nommer une personne responsable de l'établissement des normes en anglais et en français. Il fallait que ces personnes aient au moins cinq années d'expérience dans l'enseignement dans la salle de classe au Canada et que les tests normalisés lui soient familiers. Pour l'établissement des normes en anglais, cinq individus (nommés par l'Alberta, la Nouvelle-Écosse, les Territoires du Nord-Ouest, l'Ontario et la Saskatchewan) ont participé. Pour l'établissement des normes en français, nous avons également disposé de cinq participantes et participants (nommés par l'Alberta, le Nouveau-Brunswick, l'Ontario, le Québec et les Territoires du Nord-Ouest); l'une de ces personnes a également participé à l'établissement des normes en anglais.

Méthodes

Les responsables de l'établissement des normes ont été informés du fait que l'objectif du travail d'établissement des normes était de fixer des normes minimales, dans l'évaluation des compétences linguistiques des RAPEC destinée aux EEFE n'ayant pas suivi de programme de formation à l'enseignement en anglais ou en français, pour que ces personnes soient autorisées à intégrer la profession.

Travail préliminaire : Avant de se réunir en groupe, les responsables de l'établissement des normes se sont vu remettre des documents sur la méthode Angoff modifiée et ont répondu à tous les items des versions V2 et V3 en anglais et en français en tant que participantes et participants au test. Nous leur avons également demandé de fournir, immédiatement après avoir fait chaque item, une estimation initiale du niveau de rendement d'une personne candidate ayant les compétences minimales nécessaires pour l'item. Pour les items en compréhension écrite et en compréhension orale, la question était : « Quel pourcentage de candidats peu compétents répondrait correctement à cette question? ». Pour les items en expression écrite et

en expression orale, la question était : « Quelle note (sur 9) vous attendez-vous à ce qu'un candidat ayant une compétence minimale obtienne sur cet item? »

Réunions d'établissement des normes : Les réunions ont eu lieu en mode virtuel sur deux jours pour chaque langue. Pour commencer chaque réunion, les personnes participantes ont passé en revue le processus d'établissement des normes et se sont vu rappeler qu'il fallait qu'ils songent aux candidates et candidats ayant une compétence minimale quand ils fixaient le point de coupure et non aux candidates et candidats de niveau moyen ou de niveau compétent.

Les participantes et participants ont d'abord discuté des modules de compréhension orale, puis des modules d'expression orale, d'expression écrite et de compréhension écrite. Pour chaque module, la phase initiale de discussion utilisait les estimations initiales des points de coupure afin de favoriser la discussion sur les items, leurs caractéristiques et le type de travail ou d'erreur que les candidates et candidats ayant une compétence minimale étaient susceptibles de produire pour ces items. Après avoir vu les estimations des scores pour chaque item fournies par les autres responsables de l'établissement des normes et discuté de la justification de ces estimations, les gens ont eu l'occasion de réviser leurs points de coupure.

Lors de la deuxième phase de discussion sur chaque module, les nouveaux choix pour les points de coupure ont été présentés, avec des données statistiques sur le rendement des participantes et participants dans chaque item (score moyen et écart-type, rendement d'une locutrice ou d'un locuteur natif et d'une locutrice ou d'un locuteur non natif pour chaque item individuellement et pour le module dans son ensemble). Les responsables de l'établissement des normes se sont également vu expliquer qu'aucun point de coupure ne saurait être une frontière parfaite entre les personnes ayant une compétence minimale et les personnes ayant une compétence insuffisante et qu'il fallait donc qu'ils se demandent s'il était préférable de pécher par excès d'indulgence (en accordant aux participantes et participants la note minimum pour réussir alors qu'ils n'ont pas forcément le niveau minimum de compétence) ou par excès de sévérité (en faisant échouer des participantes et participants qui ne se situent pas forcément en dessous du niveau minimum de compétence). Lors de la discussion sur les points de coupure, les responsables de l'établissement des normes ont été informés du nombre de participantes et participants à la mise à l'essai du test qui réussiraient ou qui échoueraient pour chaque point de coupure évoqué.

Pour les modules de compréhension orale et de compréhension écrite, la discussion s'est achevée à ce stade et les points de coupure ont été établis. Pour les modules d'expression orale et d'expression écrite, les responsables de l'établissement des normes ont reçu un choix de réponses se situant en deçà du point de coupure, au point de coupure ou au-dessus du point de coupure établi à l'heure actuelle. Ils ont pu, à partir de ces réponses, effectuer un travail de réflexion supplémentaire et modifier les points de coupure pour les modules d'expression orale et d'expression écrite.

Réflexion après les réunions : Une fois que les points de coupure ont été établis à l'issue des deux journées de réunions, les responsables de l'établissement des normes se sont vu accorder une quinzaine de jours pour discuter de leur expérience et des points de coupure établis avec leur registraire et pour réfléchir à la question de savoir s'ils étaient satisfaits des points de coupure établis. Les responsables de l'établissement des normes en anglais se sont à nouveau réunis et ont procédé à un léger ajustement du point de coupure pour le module de compréhension orale V2, mais ils n'ont pas modifié les autres points de coupure. Les responsables de l'établissement des normes en français n'ont apporté aucun changement aux points de coupure à l'issue de cette période de réflexion.

Définition de « compétence » : Les responsables de l'établissement des normes ont décrit la personne candidate ayant le niveau minimal de compétence en fonction de la capacité qu'elle a de respecter les obligations professionnelles d'un membre de la profession enseignante. Il s'agit d'être capable de communiquer de façon claire et efficace avec différents auditoires, dont les parents d'élèves, les élèves et les membres du personnel administratif, mais aussi de donner l'exemple sur le plan linguistique à ses élèves. Ces descriptions d'une personne ayant le niveau minimal de compétence correspondent aux éléments du cadre conceptuel comme « modeler *[sic]* un bon usage de la langue », « faire des exposés pour de petits et

de grands groupes de parents et de professionnels du milieu scolaire » et « rédiger, en utilisant un langage à caractère non technique, des courriers électroniques ou des lettres aux parents ». Les points de coupure ont été fixés à un niveau correspondant au niveau 2 des « niveaux de maîtrise de la langue » décrits dans le cadre des compétences linguistiques et des niveaux de compétence linguistique.

Points de coupure : Lors de la définition des points de coupure, les responsables de l'établissement des normes savaient qu'aucun point de coupure ne saurait représenter une parfaite frontière entre les personnes faisant une utilisation compétente de la langue et les personnes non compétentes. Les responsables de l'établissement des normes ont convenu, en français comme en anglais, qu'il était préférable de risquer d'inclure les candidates et candidats ayant une maîtrise marginale de la langue que de risquer d'exclure des candidates et candidats compétents. La justification était ici que (a) la maîtrise de la langue de l'individu s'améliore souvent avec le temps et avec l'expérience sur le terrain et que (b) la réussite au test n'est pas une garantie d'obtention d'un emploi. Les points de coupure représentent ainsi le consensus des responsables de l'établissement des normes sur le score le plus bas qui soit acceptable de la part d'une enseignante ou d'un enseignant aux compétences linguistiques minimales.

Les points de coupure issus des délibérations sur l'établissement de normes sont présentés dans la partie intitulée « Question 7 – Est-ce que les normes (points de coupure) sont établies de façon appropriée? » (voir le tableau 9. Points de coupure).

Étude de la validité de l'évaluation des compétences linguistiques des RAPEC

Les tests sont créés pour remplir une fonction. L'étude de validité détermine dans quelle mesure le test correspond à la fonction prévue. Elle ne permet pas de répondre par oui ou non à la question de la validité du test; au lieu de cela, elle présente une argumentation. Cette argumentation guide l'utilisatrice ou utilisateur du test et les autres parties prenantes dans leur réflexion et dans la prise de décisions sur ce que le test mesure, sur la fiabilité de ce qu'il mesure et sur les décisions qui peuvent être justifiées à partir des scores obtenus au test.

Il existe plusieurs cadres pouvant servir d'ancrage à une étude de validité. Les travaux de recherche ayant le plus influencé la présente étude de validité sont ceux de Kane[10]. Dans le cadre défini par cet auteur, les arguments relatifs à la validité s'appuient sur diverses données objectives et analyses. Il est utilisé couramment en raison de sa nature pragmatique et du fait qu'il s'applique directement aux examens utilisés pour l'attribution du permis d'exercer une profession.

L'objectif de l'évaluation des compétences linguistiques des RAPEC est de déterminer si les EEFE qui n'ont pas fait leur formation à l'enseignement en anglais ou en français ont les compétences linguistiques exigées pour qu'ils puissent faire un bon travail dans la salle de classe au primaire-secondaire au Canada. Il est, par conséquent, important de savoir que les tests mesurent des compétences linguistiques appropriées, que les mesures sont fiables et impartiales et que les normes sont fixées de façon appropriée. La présente étude de validité a, dans cette optique, examiné les questions suivantes :

1. Est-ce que les tests sont fondés sur un cadre de compétence linguistique approprié?
2. Est-ce que les items des tests correspondent à l'utilisation que le personnel enseignant fait de la langue dans les écoles du Canada?
3. Qui étaient les participantes et participants à la mise à l'essai du test?
4. Quelles sont les propriétés psychométriques du test?
   A. Est-ce que chaque module pris individuellement présente un niveau acceptable de fiabilité?
   B. Est-ce que les items en expression orale et en expression écrite présentent une bonne fiabilité entre membres du groupe de correction?

[10] M. T. Kane, « Validating Interpretive Arguments for Licensure and Certification Examinations », *Evaluation and the Health Professions*, vol. 17, nº 2, 1994, p. 133-159.

C. Quelles sont les propriétés des items?
D. À quels niveaux d'aptitude les tests fournissent-ils de bonnes informations?
E. Est-ce que les tests mesurent ce qu'ils prétendent mesurer?

5. Est-ce que certains groupes particuliers de participantes ou participants sont avantagés ou désavantagés par les tests?
6. Est-ce que les résultats des tests sont en corrélation avec d'autres indicateurs de la maîtrise de la langue?
7. Est-ce que les normes (points de coupure) sont fixées de façon appropriée?

Le but affiché de l'étude est de permettre aux registraires des certificats d'aptitude à l'enseignement de prendre une décision éclairée sur la question de savoir si ce test devrait être déployé afin de déterminer si les EEFE possèdent les compétences linguistiques exigées pour connaître la réussite dans la salle de classe au Canada. La réussite à ce test ne serait pas une garantie et ne permettrait même pas de prédire que la personne ferait un bon travail dans la salle de classe. Elle indiquerait simplement que les compétences linguistiques de la personne ne l'*empêchent* pas d'être une bonne enseignante ou un bon enseignant dans la salle de classe primaire ou secondaire au Canada.

Question 1 – Est-ce que les tests sont fondés sur un cadre de compétence linguistique approprié?

À la phase I du projet, les responsables de l'élaboration du test ont effectué une analyse des travaux de recherche en vue d'éclairer la définition du cadre des compétences linguistiques et des niveaux de compétence linguistique (voir la section « Phase I : Mise au point de l'évaluation des compétences linguistiques des RAPEC » dans le présent rapport). L'analyse des travaux de recherche et le cadre lui-même ont été publiés par le CMEC. Le cadre fournit des recommandations claires sur l'incorporation de tout un éventail d'éléments propres à la profession dans les items du test. Ceci a permis de garantir que les tests fournissent une évaluation authentique de l'utilisation de la langue dans le contexte de l'enseignement dans la salle de classe au Canada. L'examen des processus d'élaboration du cadre[11], des items et des tests montre que ces derniers sont fondés sur un cadre théorique qui est éclairé par les travaux de recherche, qui est pertinent par rapport à la pratique et qui est à la disposition de toutes les parties prenantes.

Question 2 – Est-ce que les items des tests correspondent à l'utilisation que le personnel enseignant fait de la langue dans les écoles du Canada?

L'enseignement est une activité complexe, qui présente des exigences complexes pour ce qui est de l'utilisation de la langue. À la phase I du projet, les items du test ont été créés par des personnes dotées de connaissances directes et approfondies sur l'enseignement au Canada, pour qu'il y ait le plus de chances que les items correspondent à la façon dont le personnel enseignant au Canada utilise la langue pour faire un bon travail (voir « Mise au point des items du test et construction du test » dans ce rapport). Dix versions de l'évaluation (cinq en anglais et cinq en français) ont été créées à partir de la banque d'items et examinées par des spécialistes externes reconnus sur la scène internationale dans le domaine de l'évaluation dans l'enseignement, du perfectionnement professionnel du personnel enseignant, de la langue dans l'éducation et de l'élaboration de programmes d'études. Ces spécialistes externes ont estimé que, dans l'ensemble, les versions du test présentaient des caractéristiques positives pour ce qui est de leur authenticité, de la validité apparente et de la validité du contenu. Ils ont fait des suggestions en vue d'améliorer certains items[12]. L'équipe a révisé ses items en incorporant les commentaires et suggestions à la fois de ces spécialistes

[11] *Parlons d'excellence : Compétences linguistiques pour un enseignement efficace*, Conseil des ministres de l'Éducation (Canada), 2013; sur Internet : https://www.cmec.ca/Publications/Lists/Publications/Attachments/320/Parlons_dexcellence.pdf.

[12] *Authenticité :* les items représentent-ils de façon réaliste des tâches que l'enseignante ou enseignant débutant est susceptible de rencontrer dans le cadre de l'exécution de ses responsabilités? *Validité apparente :* les items semblent-ils mesurer les résultats de rendement ou les compétences auxquels ils sont associés? *Validité du contenu :* est-ce que, dans leur ensemble, les items comprennent bien les compétences qu'il est raisonnable d'évaluer, étant donné les contraintes de l'évaluation?

externes (améliorations de l'authenticité, de la validité apparente et de la validité du contenu), des membres du Sous-comité des RAPEC chargé des compétences linguistiques (pour veiller à ce que le test n'exige aucune connaissance particulière relative à la pédagogie ou aux matières enseignées et à ce qu'il n'y ait pas de test distinct pour le personnel enseignant du primaire et pour le personnel enseignant du secondaire) et de la phase de test alpha (améliorations de la structure, du format et du contenu des items). Au cours de ce processus, les items ont été révisés conformément aux suggestions ou remplacés par des items considérés comme respectant les normes d'évaluation. Les révisions de la phase II (augmentation du nombre d'items dans chaque module, simplification des critères de correction pour les items en expression écrite et en expression orale, révisions grammaticales et stylistiques) n'ont pas eu d'incidence sur le contenu des questions du test.

Lors de la mise à l'essai de la phase II, deux méthodes de rassemblement de données ont été utilisées pour déterminer avec plus de précision si les items étaient représentatifs de l'éventail des usages linguistiques dans le contexte de l'enseignement au Canada. La première était un sondage adressé par le CMEC aux participantes et participants au test (n = 275). L'une des questions du sondage leur demandait d'évaluer le contenu du test sur une échelle à cinq points[13]. La note moyenne a été de 4,04 et 82 p. 100 des participantes et participants au test ont indiqué que les items étaient soit « excellents » soit « bons » (soit aux deux rangs les plus élevés dans l'échelle à cinq points). Le sondage les invitait également à faire des commentaires généraux sur le test. Sur les 275 personnes ayant répondu au sondage, 28 ont choisi de répondre à l'invitation et ont fait des commentaires sur les liens entre ces items et la réalité de l'enseignement au Canada; 26 de ces 28 personnes ont indiqué que les items correspondaient bien à l'usage linguistique dans la profession enseignante. L'une des personnes (en anglais) a indiqué que, selon elle, les items ne correspondaient pas au « contexte de la situation réelle dans la salle de classe dans toutes ses nuances » et une autre a indiqué qu'elle craignait que « le contenu ne corresponde pas parfaitement à la réalité sur le terrain ».

Les autres données ont été recueillies dans le cadre de deux réunions (l'une en anglais, l'autre en français) des chefs de table. Ceux-ci se sont vu demander de recueillir les commentaires des correctrices et correcteurs de leur table et de les présenter. Tant dans la réunion en anglais que dans la réunion en français, tous les chefs de table ont convenu que les items correspondaient bien aux tâches typiques du personnel enseignant et représentaient un échantillon adéquat de l'éventail des compétences linguistiques.

Aucun test ne peut incorporer toutes les manières dont la langue est utilisée dans une profession donnée, mais il semble que les items du test soient réalistes pour ce qui est de la représentation de l'éventail des manières dont les enseignantes et enseignants utilisent la langue dans leur travail.

Question 3 – Qui étaient les participantes et participants à la mise à l'essai du test?

La population ciblée par l'évaluation des compétences linguistiques des RAPEC est celle des enseignantes et enseignants qui ont fait leur formation à l'enseignement à l'étranger dans une langue autre que l'anglais ou le français. Pour pouvoir généraliser du mieux possible les résultats de la mise à l'essai à l'intention de cette population ciblée, il est utile de faire en sorte que les caractéristiques des participantes et participants à la mise à l'essai correspondent à celles de la population ciblée.

Les tests de la mise à l'essai se sont déroulés lors de deux sessions, pour un total de 349 tests en français (versions V2 et V3 combinées) et de 589 tests en anglais (versions V2 et V3 combinées). Dans la population des participantes et participants au test, 73 p. 100 des individus se sont identifiés comme étant de sexe féminin, pourcentage proche de la proportion de 75 p. 100 d'enseignantes dans la population enseignante au Canada[14]. Trois quarts (75 p. 100) des participantes et participants au test étaient des enseignantes et enseignants certifiés, les 25 p. 100 restants étant des individus inscrits à un programme de formation

[13] L'énoncé de la question était le suivant : « Comment évalueriez-vous le CONTENU de l'outil en ligne d'évaluation linguistique pour la profession enseignante? »

[14] « La rentrée scolaire… en chiffres », Statistique Canada, 2018; sur Internet : https://www.statcan.gc.ca/fr/quo/smr08/2018/smr08_220_2018.

à l'enseignement. Même si 34 p. 100 (*n* = 314) des tests ont été faits par des personnes ayant suivi leur formation à l'enseignement à l'étranger (sur l'ensemble des deux langues et sur l'ensemble des versions du test), 5 p. 100 seulement (*n* = 50) des tests (sur l'ensemble des deux langues et sur l'ensemble des versions du test) ont été faits par des personnes ayant suivi leur formation à l'enseignement dans une langue autre que l'anglais ou le français. Il s'agit là d'une différence importante entre la mise à l'essai et la population ciblée par le test proprement dit. L'autre différence importante est que 4 p. 100 seulement des participantes et participants au test ont dit que la langue dans laquelle ils avaient le meilleur niveau était une langue autre que l'anglais ou le français

Les différences entre la population ciblée par le test lui-même et l'échantillon de personnes ayant participé à la mise à l'essai laissent à penser qu'il est fort possible que les scores moyens obtenus par les personnes de l'échantillon dans les items soient plus élevés que ceux qu'obtiendront les membres de la population ciblée. Ceci est susceptible d'avoir deux effets : le premier est qu'il est possible qu'il y ait un effet de plafonnement pour certains items dans lesquels le score moyen est très élevé, ce qui réduit la variance et la capacité qu'a l'item de discriminer les participantes et participants au test ayant des aptitudes linguistiques différentes. Le travail régulier de rassemblement et d'analyse des données lors de la mise en œuvre du test est un aspect important de cette mise en œuvre. Ce processus permettra de disposer régulièrement de données utiles pour surveiller les résultats produits par les items et le caractère approprié des points de coupure.

Le deuxième effet est que les personnes responsables de l'établissement des normes risquent d'être influencées par la moyenne élevée des scores pour les items et de fixer des points de coupure plus élevés que ce qui est souhaitable. Nous avons adressé à ces personnes des directives au départ et des rappels les invitant à fixer les points de coupure en s'appuyant sur l'idée qu'elles se faisaient d'une candidate ou d'un candidat ayant les compétences minimales nécessaires. Les personnes responsables de l'établissement des normes ont été informées des scores moyens pour les items, mais elles ont également été informées des caractéristiques des participantes et participants à la mise à l'essai et des résultats obtenus par différents groupes de participantes et participants pour chaque item et pour chaque module (locutrices et locuteurs natifs ou non, personnes ayant suivi leur formation à l'enseignement au Canada ou à l'étranger, etc.).

Question 4 – Quelles sont les propriétés psychométriques du test?

Nous avons effectué deux analyses psychométriques indépendantes. Le groupe *Directions* a effectué une première série d'analyses et utilisé le cadre de la théorie classique des tests (TCT). Le cadre TCT est un modèle psychométrique plus ancien mais bien établi et souvent utilisé dans la mise au point de tests. Il s'agit d'un cadre qui est bien compris et utilisé quand les échantillons sont de petite taille. Une psychométricienne externe a effectué la deuxième série d'analyses et utilisé le cadre de la théorie des réponses aux items (TRI)[15]. Le cadre TRI est un cadre puissant capable de fournir des informations psychométriques détaillées sur les items du test, mais il exige un échantillon de grande taille. Il existe plusieurs modèles TRI différents, chacun fonctionnant selon sa propre série de suppositions et ses propres exigences pour les données. Nous présentons ici les résultats des analyses à la fois selon le cadre TCT et selon le cadre TRI. Dans la plupart des cas, les deux analyses ont débouché sur les mêmes conclusions, mais, quand il existe des différences, nous le notons.

A. Est-ce que chaque module pris individuellement présente un niveau acceptable de fiabilité?

Les analystes ont utilisé, pour examiner la fiabilité de chaque module du test, un indicateur appelé « coefficient alpha de Cronbach ». Ce coefficient est un indicateur de cohérence interne couramment utilisé, qui indique dans quelle mesure les items du test mesurent le même concept. L'analyse de la psychométricienne externe s'est appuyée sur un modèle TRI pour mesurer la fiabilité du test. Le tableau 2

[15] La psychométricienne externe a effectué ses analyses TRI, mais les conclusions et recommandations présentées dans ce rapport proviennent du groupe *Directions*. Aucun élément du contenu du présent rapport n'émane de la psychométricienne externe, qui est dégagée de toute responsabilité à cet égard. Le rapport psychométrique externe est disponible sur demande auprès du CMEC.

présente les deux séries de résultats. Les valeurs du coefficient alpha de Cronbach inférieures à 0,70 sont généralement considérées comme problématiques, de même que les valeurs inférieures à 0,75 dans l'analyse TRI.

Tableau 2. Fiabilité du test

Version du test	Module	Anglais		Français	
		Coefficient alpha de Cronbach	Fiabilité TRI	Coefficient alpha de Cronbach	Fiabilité TRI
V2	Compréhension orale	0,25	0,42	0,42	0,42
	Compréhension écrite	0,65	0,67	0,78	0,60
	Expression orale	0,87	0,94	0,88	0,87
	Expression écrite	0,85	0,90	0,77	0,79
V3	Compréhension orale	0,46	0,54	0,42	0,47
	Compréhension écrite	0,61	0,71	0,70	0,65
	Expression orale	0,87	0,88	0,85	0,85
	Expression écrite	0,82	0,82	0,73	0,75

Comme le montre le tableau 2, les modules de compréhension orale ont un faible niveau de fiabilité dans toutes les versions du test. Les chiffres de la fiabilité pour les modules de compréhension écrite sont meilleurs, mais restent problématiques. Les modules d'expression orale et d'expression écrite ont des valeurs acceptables pour la fiabilité dans toutes les versions du test.

À la suite de ces analyses, certains items particuliers du test ont été réexaminés en raison de leurs propriétés psychométriques[16]. En outre, certains items ont été réexaminés pour s'assurer qu'ils étaient dépourvus de toute erreur et appropriés sur le plan culturel. Cette recommandation repose sur les conclusions d'un sondage auprès des participantes et participants à la fin du test et des réunions avec les chefs de table.

Les modules de compréhension orale et de compréhension écrite avaient un niveau problématique de fiabilité et leurs items n'étaient pas cohérents sur une échelle à une seule dimension sous leur forme actuelle. Ils étaient également plutôt faciles pour les participantes et participants à la mise à l'essai. Les items des modules d'expression orale et d'expression écrite ont servi d'échelles présentant une bonne fiabilité.

B. Est-ce que les items en expression orale et en expression écrite présentent une bonne fiabilité entre membres du groupe de correction?

Avant de décrire la fiabilité entre membres du groupe de correction, il est important de rappeler à la lectrice ou au lecteur la façon dont les items ont été corrigés. Les items en expression orale et en expression écrite ont été corrigés selon une échelle à 10 points, de 0 à 9. Chaque item a été corrigé selon trois critères; chacun de ces critères a été évalué selon une échelle à quatre points allant de 0 à 3. Toutes les réponses ont été corrigées par deux personnes différentes. Si l'intervalle entre les deux scores n'était que d'un point, le score définitif était automatiquement la moyenne des deux scores. (Par exemple, si les scores étaient 6 et 7, le score définitif était de 6,5.) Si l'intervalle entre les deux scores attribués était de deux points ou plus, le processus considérait qu'il y avait un désaccord et le désaccord était examiné par un arbitre (le chef de table). Après avoir examiné les deux scores et les arguments avancés, l'arbitre attribuait le score définitif. Tous les scores des items en expression orale et en expression écrite sont donc le fruit des jugements combinés de deux ou trois correctrices ou correcteurs.

[16] Les analyses psychométriques détaillées se trouvent dans le texte intégral du rapport psychométrique, disponible sur demande.

Trois indicateurs ont été utilisés pour examiner la fiabilité entre membres du groupe de correction (tableau 3) :

1. **Pourcentage du degré d'accord** – Nous avons considéré que les deux correctrices ou correcteurs étaient d'accord si l'intervalle entre les deux scores totaux pour la réponse était au maximum d'un point. Sur l'ensemble des modules, le pourcentage du degré d'accord s'est avéré être encourageant, mais ce chiffre inclut les absences de réponse et les réponses vides, pour lesquelles le score était de zéro. Il est très facile pour les correctrices et correcteurs d'être d'accord sur l'attribution d'un score de zéro pour l'absence de réponse, ce qui exagère le niveau d'accord entre les correctrices et correcteurs.
2. **Coefficient kappa de Cohen** – Ce paramètre mesure le degré d'accord exact entre les correctrices et correcteurs, en prenant en compte la possibilité que l'attribution de la même note soit purement le fruit du hasard. Les valeurs du coefficient kappa de Cohen inférieures à 0,20 sont considérées comme problématiques.
3. **Coefficient de corrélation au sein d'une même classe** – Ce paramètre examine la corrélation entre les scores des deux correctrices ou correcteurs pour chaque item. Les valeurs inférieures à 0,80 sont considérées comme problématiques.

Tableau 3. Fiabilité entre membres du groupe de correction : éventail des valeurs

Langue	Version du test	Module	Pourcentage du degré d'accord	Coefficient kappa de Cohen	Coefficient de corrélation au sein d'une même classe
Anglais	V2	Compréhension orale	71 % – 81 %	0,20 – 0,40	0,72 – 0,90
		Compréhension écrite	62 % – 77 %	0,22 – 0,30	0,76 – 0,88
	V3	Expression orale	70 % – 85 %	0,22 – 0,43	0,80 – 0,92
		Expression écrite	68 % – 76 %	0,19 – 0,33	0,80 – 0,88
Français	V2	Compréhension orale	61 % – 79 %	0,18 – 0,37	0,82 – 0,91
		Compréhension écrite	62 % – 72 %	0,22 – 0,35	0,80 – 0,87
	V3	Expression orale	69 % – 80 %	0,24 – 0,42	0,83 – 0,92
		Expression écrite	61 % – 76 %	0,17 – 0,34	0,71 – 0,87

Les détails sur la fiabilité entre membres du groupe de correction, en particulier en ce qui a trait aux items pris individuellement, se trouvent dans le texte intégral du rapport psychométrique. De façon générale, les items des modules d'expression orale et d'expression écrite présentaient une bonne fiabilité entre membres du groupe de correction.

C. Quelles sont les propriétés des items?

Le présent résumé a pour but de permettre à la lectrice ou au lecteur de comprendre les forces et les faiblesses psychométriques de chaque module et de se faire une idée globale des résultats produits par chaque module. La description complète de toutes les propriétés des items se trouve dans le texte intégral du rapport psychométrique du groupe *Directions* et du rapport psychométrique externe. Ces rapports contiennent des descriptions détaillées des propriétés pour tous les items de tous les tests.

Compréhension orale

Les modules de compréhension orale comprenaient 13 items et se sont avérés être faciles pour les participantes et participants à la mise à l'essai. Le score moyen pour les modules de compréhension orale en anglais a été de 77 p. 100 pour V2 et de 83 p. 100 pour V3. Le score moyen pour les modules de compréhension orale en français a été de 71 p. 100 pour V2 et de 77 p. 100 pour V3. Un sous-ensemble d'items a été retenu pour un examen dans les tests en anglais et en français. Pour être retenu, il fallait que l'item ait un score moyen supérieur à 95 p. 100 ou inférieur à 50 p. 100 ou produise de faibles résultats pour

ce qui est de discriminer les individus (c'est-à-dire que les participantes et participants au test ont tous obtenu des scores comparables pour l'item, indépendamment de leur niveau d'aptitude). Certains items se sont avérés avoir un effet discriminatoire négatif, c'est-à-dire que les participantes et participants ayant des aptitudes de niveau élevé ont obtenu un moins bon score pour l'item que les participantes et participants ayant des aptitudes de niveau faible. Il est généralement considéré comme souhaitable d'avoir une corrélation item-total corrigée supérieure à 0,20; les concepteurs du test ont donc examiné les items pour lesquels la corrélation item-total corrigée était inférieure à 0,20 et en particulier ceux pour lesquels la corrélation item-total corrigée était négative[17]. Lorsque la corrélation item-total corrigée et le coefficient de discrimination ont des valeurs plus élevées, cela indique que l'item a un meilleur effet discriminatoire.

Il n'y a rien d'intrinsèquement problématique dans les scores moyens supérieurs à 95 p. 100 ou inférieurs à 50 p. 100. Ces items ont été retenus pour un examen parce que les items très faciles tendent à avoir un faible effet discriminatoire et le fait d'avoir trop d'items faciles dans un test a tendance à réduire la quantité d'informations fournies par le test. Nous avons donc examiné les items faciles pour déterminer s'il était possible de les améliorer ou de les éliminer du test. De même, il n'y a rien d'intrinsèquement problématique dans les scores moyens inférieurs à 50 p. 100, mais, comme les participantes et participants au test ont trouvé que la plupart des items du test étaient faciles, nous avons examiné les items plus difficiles afin de vérifier que cette difficulté n'était pas due à un manque de clarté dans l'énoncé ou à des leurres trop plausibles.

Compréhension écrite

Les modules de compréhension écrite se sont également avérés faciles pour les participantes et participants au test. Les modules anglais comptaient 22 items et le score moyen a été de 87 p. 100 pour V2 et de 89 p. 100 pour V3. Les modules français comptaient 19 items et le score moyen a été de 82 p. 100 pour V2 et de 78 p. 100 pour V3[18]. Un sous-ensemble d'items a été retenu pour un examen dans les tests en anglais et en français. Pour être retenu, il fallait, comme pour les modules de compréhension orale, que l'item ait un score moyen supérieur à 95 p. 100 ou inférieur à 50 p. 100 ou une corrélation item-total corrigée négative. Les résultats pour la compréhension orale et la compréhension écrite proviennent des analyses TCT. Les analyses TRI effectuées par la psychométricienne externe ont produit des résultats qui concordaient, de façon générale, avec ceux des analyses TCT.

Expression orale

Le module d'expression orale V2 en anglais comptait 12 items et les scores moyens sont allés de 7,64 à 7,97 sur l'échelle de 0 à 9; le module d'expression orale V3 en anglais comptait 13 items et les scores moyens sont allés de 7,18 à 8,27; les scores 0 dus à l'absence de réponse ont été exclus. La corrélation item-total corrigée se situait entre 0,49 et 0,64 pour V2 et entre 0,49 et 0,59 pour V3. Ces chiffres indiquent que les modules d'expression orale avaient un bon effet discriminatoire.

Les analyses TRI effectuées par la psychométricienne externe ont montré que la plupart des participantes et participants ont obtenu de bons résultats dans les items en expression orale, de sorte qu'il y avait une forte probabilité qu'ils obtiennent un 9 sur 9 pour ces items. Ceci est probablement dû au fait qu'il y avait peu de participantes et participants avec un score faible, de sorte qu'il y avait des données insuffisantes aux niveaux

[17] *Corrélation item-total corrigée* : valeur indiquant l'effet discriminatoire de chaque item, calculé sous la forme d'un coefficient de corrélation de Pearson entre le score pour l'item et le score à l'échelle avec la suppression d'un item donné. Le coefficient de corrélation Pearson est un indice du degré de relation linéaire entre deux variables. Il est souvent connu sous le nom de coefficient de corrélation produit-moment de Pearson (ou coefficient *r* de Pearson) et est l'un des coefficients de corrélation le plus souvent utilisés pour les échantillons. Il est rapporté à l'échelle, de sorte que la valeur +1 indique une relation parfaitement positive (l'obtention de scores élevés pour la variable *x* est liée à l'obtention de scores élevés pour la variable *y*), la valeur –1 indique une relation parfaitement négative (l'obtention de scores élevés pour la variable *x* est liée à l'obtention de scores faibles pour la variable *y*, ou vice-versa) et la valeur 0 indique l'absence de relation.

[18] Les différents modules dans une seule et même modalité ne comptaient pas tous le même nombre de questions. Ceci est dû au fait que, lors de l'élaboration des items, une estimation du temps nécessaire pour la participante ou le participant était attribuée à chaque item et que, lors de la construction du test, chaque module a été construit de façon à ce que tous les modules aient approximativement la même longueur.

d'aptitude faibles pour pouvoir produire des informations statistiques exactes sur l'effet discriminatoire du module pour les personnes à faible aptitude.

Les modules en français comptaient dans les deux cas 12 items. Les scores moyens sont allés de 5,98 à 6,88 pour le module d'expression orale V2 en français et de 5,77 à 6,83 pour le module d'expression orale V3 en français. La corrélation item-total corrigée se situait entre 0,45 et 0,56 pour V2 et entre 0,41 et 0,48 pour V3.

Les analyses TRI effectuées par la psychométricienne externe ont montré une tendance analogue en français à la tendance en anglais. Lorsque la personne avait un niveau d'aptitude élevé, elle obtenait un score élevé, mais l'effet discriminatoire était faible pour les personnes à faible aptitude. Ici encore, ceci est probablement dû au faible nombre de participantes et participants ayant obtenu un score faible. À titre d'exemple, le module d'expression orale V3 en français ne comptait que trois items avec des réponses atteignant un score total de 1 (et une personne seulement ayant obtenu ce score pour chacun de ces trois items) et que sept items avec des réponses atteignant un score total de 2 (et trois personnes au maximum ayant reçu ce score). Avec si peu de données au bas de l'échelle des scores, il est impossible de produire des statistiques solides pour ce niveau.

Pour résumer, les analyses TCT indiquent que les items d'expression orale étaient faciles pour les participantes et participants à la mise à l'essai, présentaient un bon niveau de fiabilité entre membres du groupe de correction et avaient un bon effet discriminatoire. Les analyses TRI montrent que les items étaient faciles pour les participantes et participants au test, mais avaient du mal à produire un effet discriminatoire pour les personnes à faible aptitude.

Expression écrite

Les modules d'expression écrite V2 et V3 en anglais comptaient chacun sept items et les scores moyens sont allés de 6,72 à 7,52 (V2) et de 6,89 à 7,56 (V3); les scores 0 dus à l'absence de réponse ont été exclus. Si cette cohorte de participantes et participants au test a obtenu de bons résultats dans les modules d'expression écrite, les scores moyens étaient inférieurs à ceux pour les modules d'expression orale. La corrélation item-total corrigée se situait entre 0,56 et 0,65 pour V2 et entre 0,53 et 0,59 pour V3; ces chiffres indiquent que les modules d'expression écrite avaient un bon effet discriminatoire.

Les analyses TRI effectuées par la psychométricienne externe ont montré que la plupart des participantes et participants ont obtenu de bons résultats dans les items en expression écrite en anglais, de sorte qu'il y avait une forte probabilité qu'ils obtiennent un 9 sur 9 pour ces items. Pour les personnes d'aptitude moyenne, il y avait une plus grande probabilité d'avoir un 8 plutôt qu'un 9 pour l'item. Cette situation est différente de la situation pour les items en expression orale et laisse à penser que les items en expression écrite ont un meilleur effet discriminatoire à un niveau d'aptitude moyen. Comme pour les modules d'expression orale, les analyses indiquent que, lorsque la personne avait un niveau d'aptitude élevé, elle obtenait un score élevé, mais que l'effet discriminatoire des items est faible quand le niveau d'aptitude des personnes est plus bas.

Comme les modules en anglais, les modules d'expression écrite en français comptaient chacun sept items; les scores moyens sont allés de 5,98 à 6,75 pour V2 et de 5,77 à 6,83 pour V3; les scores 0 dus à l'absence de réponse ont été exclus. Si cette cohorte de participantes et participants au test a obtenu de bons résultats dans les modules d'expression écrite en français, les scores moyens étaient inférieurs à ceux pour les modules d'expression orale en français. La corrélation item-total corrigée se situait entre 0,45 et 0,56 pour V2 et entre 0,41 et 0,48 pour V3; ces chiffres indiquent que les modules d'expression écrite avaient un bon effet discriminatoire.

Les analyses TRI effectuées par la psychométricienne externe ont montré que la plupart des participantes et participants ont obtenu de bons résultats pour les items en expression écrite, mais, par comparaison aux tests en anglais, la probabilité était plus élevée que les participantes et participants aient un score de 8 sur 9 pour les items en français aux niveaux d'aptitude élevés. Ceci semble indiquer que les items en expression écrite étaient plus difficiles pour les participantes et participants au test en français que pour les

participantes et participants au test en anglais et que, par rapport aux items en expression orale, lorsque la personne avait un niveau d'aptitude élevé, elle obtenait un score élevé, mais l'effet discriminatoire était faible pour les personnes à faible aptitude. Pour tous les items, les analyses TRI se trouvent dans le rapport de la psychométricienne externe.

D. À quels niveaux d'aptitude les tests fournissent-ils de bonnes informations?

Les analyses TCT effectuées par le groupe *Directions* ne permettent pas de bien répondre à cette question, mais les analyses TRI effectuées par la psychométricienne externe apportent certains éclairages. (Vous trouverez dans le rapport psychométrique externe des figures complètes avec les fonctions d'information du test.) D'après les conclusions générales de ce travail, les modules de compréhension orale et de compréhension écrite sont le plus précis pour les niveaux d'aptitude faibles, mais fournissent des informations de piètre qualité pour les niveaux d'aptitude moyens ou élevés. Cela est vrai sur l'ensemble des quatre tests. La fonction d'information (du test) pour le module de compréhension orale V3 en anglais est tout particulièrement limitée, ce qui signifie que le module ne fournit de bonnes informations sur les participantes et participants au test que dans un intervalle relativement limité d'apprenantes et apprenants ayant un faible niveau d'aptitude. Ces conclusions ne sont pas forcément problématiques si le point de coupure est établi à un niveau où le module fournit de bonnes informations et l'erreur-type dans les mesures est faible. Si le point de coupure est établi en dehors de l'intervalle de niveaux d'aptitude dans lequel le module fournit de bonnes informations, alors le module ne fournira pas de bons résultats au point de coupure. Le point de coupure fixé par les responsables de l'établissement des normes (voir « Question 7 – Est-ce que les normes (points de coupure) sont établies de façon appropriée? ») se situe à un niveau inférieur au niveau auquel le module fournit de bonnes informations. Cela n'est certes pas idéal du point de vue psychométrique, mais il est très peu probable que les participantes et participants au test soient désavantagés parce que le point de coupure est fixé à un niveau très bas.

Pour les modules d'expression orale, les fonctions d'information indiquent que le test produit les meilleurs résultats quand le niveau d'aptitude se situe autour de la moyenne, mais qu'il ne produit pas de bons résultats pour les participantes et participants au test qui se situent aux extrêmes (c'est-à-dire à deux écarts-types au-dessus ou en dessous de la moyenne). Il s'agit là d'un intervalle approprié de niveaux d'aptitude dans lequel le module d'expression orale produit de bons résultats, mais il faut noter que les erreurs-types dans les mesures sont plus élevées qu'il serait souhaitable dans l'idéal. La situation est la même pour les modules d'expression écrite, même si la fonction d'information pour le module d'expression écrite V2 en français est impossible à interpréter, ce qui fait qu'il est impossible de tirer des conclusions sur les résultats produits par ce module aux différents niveaux d'aptitude.

E. Est-ce que les tests mesurent ce qu'ils prétendent mesurer?

Deux types différents d'analyse factorielle ont été utilisés pour déterminer la structure des tests. Pour les tests linguistiques, l'attente est que les scores soient en corrélation d'une modalité à l'autre. C'est le cas parce que la compréhension orale, la compréhension écrite, l'expression orale et l'expression écrite sont certes des domaines de compétence distincts, mais que les gens qui sont doués dans l'un de ces domaines ont tendance à être également doués dans tous les autres. À titre d'exemple, il serait difficile d'être d'un excellent niveau en expression écrite dans une langue donnée sans être également excellent en compréhension écrite.

Le premier type d'analyse factorielle – l'analyse factorielle exploratoire (AFE) – est utile quand la structure du test est inconnue. La conclusion de l'AFE est que les tests sont unidimensionnels, avec des corrélations de niveau modéré ou élevé entre les modules[19]. Le tableau 4 montre la saturation factorielle pour chaque module pour un seul facteur. Cette saturation factorielle indique dans quelle mesure les scores obtenus dans le module correspondent au niveau général d'aptitude linguistique dans le contexte de l'enseignement. Dans l'idéal, il faut que les saturations factorielles soient supérieures à 0,30.

[19] Les échelles unidimensionnelles mesurent un seul concept, une seule caractéristique ou un seul attribut.

Tableau 4. Saturation factorielle des quatre modules dans chaque test

Langue	Version du test	Compréhension orale	Compréhension écrite	Expression orale	Expression écrite
Anglais	V2	0,54	0,75	0,82	0,82
	V3	0,61	0,70	0,66	0,75
Français	V2	0,50	0,56	0,70	0,84
	V3	0,49	0,53	0,63	0,75

Le tableau 5 montre les coefficients de corrélation de Pearson entre différents modules pour les tests V2 et V3 en anglais; le tableau 6 les présente pour les tests V2 et V3 en français. Si le coefficient de corrélation est faible (inférieur à 0,20), cela indique que les deux domaines de compétence sont indépendants l'un de l'autre; en revanche, si le coefficient est très élevé (supérieur à 0,80), cela indique que les deux modules mesurent les mêmes aptitudes. L'étroitesse de la corrélation entre modules indique que les modules mesurent des aptitudes apparentées, sans être identiques.

Tableau 5. Coefficients de corrélation de Pearson entre les modules pour le test en anglais

		V2			
		Compréhension orale	Compréhension écrite	Expression orale	Expression écrite
V3	Compréhension orale		0,49	0,36	0,40
	Compréhension écrite	0,61		0,60	0,56
	Expression orale	0,28	0,38		0,73
	Expression écrite	0,38	0,45	0,64	

Toutes les corrélations sont significatives à $p < 0,01$.

Tableau 6. Coefficients de corrélation de Pearson entre les modules pour le test en français

		V2			
		Compréhension orale	Compréhension écrite	Expression orale	Expression écrite
V3	Compréhension orale		0,35	0,38	0,37
	Compréhension écrite	0,39		0,30	0,49
	Expression orale	0,23	0,33		0,59
	Expression écrite	0,41	0,43	0,53	

Toutes les corrélations sont significatives à $p < 0,01$.

La psychométricienne externe a effectué une analyse factorielle confirmatoire (AFC) pour trouver, parmi plusieurs modèles prédéterminés, celui qui correspond le mieux aux données. Son constat est que, pour V2 en anglais, V2 en français et V3 en français, le meilleur modèle a un facteur linguistique général avec quatre composantes (compréhension orale, compréhension écrite, expression orale, expression écrite). Pour le test V3 en anglais, le meilleur modèle a un facteur sous-jacent pour l'ensemble du test (c'est-à-dire que tous les modules mesurent la même aptitude linguistique générale). Ces résultats semblent indiquer que les quatre modules mesurent des aptitudes distinctes, mais apparentées. Ceci correspond au cadre conceptuel utilisé pour concevoir l'évaluation des compétences linguistiques des RAPEC. Ce cadre décrit une aptitude globale (maîtrise de la langue dans le contexte de l'enseignement) ayant quatre dimensions

ou « modalités » (compréhension orale, compréhension écrite, expression orale, expression écrite). Pour consulter une description complète du modèle et des indices d'ajustement, veuillez consulter le rapport de la psychométricienne externe.

Question 5 – Est-ce que certains groupes particuliers de participantes ou participants sont avantagés ou désavantagés par les tests?

Nous avons effectué deux analyses pour déterminer si certains groupes particuliers de participantes ou participants étaient avantagés ou désavantagés par les tests : une analyse de l'impact du test et une analyse du fonctionnement différentiel des items (FDI).

Impact du test

L'analyse de l'impact du test fait des comparaisons entre les résultats de différents groupes au test (c'est-à-dire du score moyen des membres de ces groupes). Pour cette analyse, deux facteurs démographiques différents ont été pris en compte : le sexe de la personne et la langue qu'elle maîtrisait le mieux (tableau 7). Des tests T ont été utilisés pour mettre en évidence les différences statistiquement significatives dans le score moyen par item effectué.

Tableau 7. Différences dans les résultats selon le sexe et selon la langue

Langue du test	Version du test	Différences selon le sexe	Différences selon la langue
Anglais	V2	Aucune différence, dans aucun des modules	Les personnes ayant indiqué que l'anglais était la langue qu'elles maîtrisaient le mieux ont obtenu un meilleur score que les personnes n'ayant pas indiqué que l'anglais était la langue qu'elles maîtrisaient le mieux. Il s'agit d'une conclusion attendue, qui ne prouve pas qu'il y a un biais dans le test.
	V3	Les femmes ont obtenu de meilleurs scores que les hommes dans les items en expression orale (*d* de Cohen = 0,31).	
Français	V2	Les femmes ont obtenu de meilleurs scores que les hommes dans les items en compréhension écrite (*d* de Cohen = 0,35) et dans les items en expression écrite (*d* de Cohen = 0,42).	Les personnes ayant indiqué que le français était la langue qu'elles maîtrisaient le mieux ont obtenu un meilleur score que les autres participantes et participants au test dans les items en expression orale et en expression écrite, mais pas dans les items en compréhension orale et en compréhension écrite.
	V3	Les femmes ont obtenu de meilleurs scores que les hommes dans les items en compréhension écrite (*d* de Cohen = 0,42) et dans les items en expression écrite (*d* de Cohen = 0,49)	

Fonctionnement différentiel des items (FDI)

L'analyse de l'impact du test permet parfois de dégager certains indices sur les biais susceptibles d'être présents dans un test, mais comme elle ne prend pas en compte le niveau d'aptitude de la participante ou du participant au test, elle produit aussi parfois des résultats trompeurs. Pour les tests linguistiques, par exemple, les locutrices et locuteurs natifs ont habituellement de meilleurs résultats que les locutrices et locuteurs non natifs. Cela ne prouve pas nécessairement l'existence d'un biais, mais cela peut être dû au fait que les locutrices et locuteurs natifs ont un meilleur niveau d'aptitude linguistique que les locutrices et locuteurs non natifs. L'analyse du FDI est une approche plus subtile de l'examen des biais, qui incorpore le niveau d'aptitude du participant ou de la participante au test dans l'analyse.

Trois approches statistiques ont été utilisées pour trouver le FDI dans les items du test. Le groupe *Directions* a utilisé les statistiques de Mantel-Haenszel et la régression logistique, tandis que la psychométricienne externe a utilisé les méthodes TRI. Les statistiques de Mantel-Haenszel fonctionnent bien pour les échantillons de petite taille et sont utiles pour trouver les cas de FDI uniforme, c'est-à-dire les cas où l'item présente un biais défavorable à un groupe donné à tous les niveaux d'aptitude. La régression logistique est également utilisée pour les échantillons de petite taille, mais il s'agit d'une technique plus subtile, qui permet de trouver à la fois les cas de FDI uniforme et les cas de FDI non uniforme. Les cas de FDI non uniforme sont les cas où l'item a un comportement différent selon le niveau d'aptitude. Il est possible, par exemple, d'avoir un item qui ne présente pas de FDI pour les participantes et participants au test ayant un niveau d'aptitude faible, mais qui est favorable aux participants au test de sexe masculin ayant un niveau d'aptitude élevé. Les analyses TRI effectuées par la psychométricienne externe sont une approche très utile pour trouver les cas de FDI, mais elles exigent des échantillons de grande taille et le respect des présuppositions du modèle TRI.

Les différentes méthodes de calcul du FDI produisent différents résultats. (Par exemple, pour un item particulier, l'analyse de régression logistique a montré un FDI selon le sexe, tandis que, pour un autre item, le test de Mantel-Haenszel et l'analyse de régression logistique ont tous deux fourni des données indiquant un FDI selon le sexe.) Nous avons donc traité en priorité les items pour lesquels deux ou plusieurs méthodes différentes montraient un FDI. Veuillez noter que les analyses du FDI présentées ici n'examinent pas la question de savoir quel groupe est avantagé ou désavantagé par les items présentant un FDI. Il s'agit là d'une analyse supplémentaire qu'il faudrait effectuer pour tous les items présentant un FDI. Les items présentant un FDI n'étaient pas forcément à éliminer, mais nous avons effectué un examen de ces items pour vérifier que les différences dans les résultats entre les groupes n'étaient pas trop prononcées et que ce n'était pas le même groupe qui était toujours désavantagé.

Nous n'avons, dans l'ensemble, trouvé qu'un nombre relativement réduit d'items présentant un FDI et en particulier d'items présentant un FDI selon plus d'une méthode d'analyse. Nous avons examiné ces cas particuliers lors de l'examen des items, parce qu'il est important, du point de vue de l'équité et de l'impartialité, que les items présentant un FDI soient examinés de façon attentive, afin de déterminer le groupe qui est désavantagé et la mesure dans laquelle il est désavantagé, ainsi que les raisons de l'existence du FDI.

L'étude ne recueillait pas d'informations démographiques relatives à la race, à l'orientation sexuelle ou aux groupes en quête d'équité, de sorte qu'il n'a pas été possible de faire des analyses dans ces catégories. Il faudrait recueillir plus de données démographiques lors de la mise en œuvre des tests pour pouvoir faire des analyses du FDI plus approfondies, afin de vérifier que les tests sont impartiaux et ne désavantagent aucun groupe particulier de candidates et candidats.

Question 6 – Est-ce que les résultats des tests sont en corrélation avec d'autres indicateurs de la maîtrise de la langue?

Dans l'idéal, dans une étude de validation du test, les scores obtenus au test sont comparés à d'autres indicateurs relatifs à la dimension concernée. Dans cette mise à l'essai, le seul autre indicateur mesurant le niveau de maîtrise de la langue était une question dans laquelle les candidates et candidats évaluaient eux-mêmes leur maîtrise de la langue sur une échelle à six points. Nous avons examiné le lien entre les résultats des candidates et candidats au test et leur évaluation de leur maîtrise de la langue selon une régression linéaire (tableau 8). Pour les deux tests en anglais, il y a un lien significatif entre leur évaluation de leur maîtrise de la langue et leur score moyen dans tous les modules. Pour les deux tests en français, il y a un lien significatif entre l'évaluation que fait la participante ou le participant de sa maîtrise de la langue et son score moyen pour les items en expression orale et en expression écrite. Il n'y a aucun lien significatif entre l'évaluation que fait la participante ou le participant de sa maîtrise de la langue et son score moyen pour les items en compréhension orale dans les deux tests en français, et il n'y a un lien significatif entre l'évaluation que fait la participante ou le participant de sa maîtrise de la langue et son score moyen pour les items en compréhension écrite que dans le test V2 en français.

Tableau 8. Résultats de la régression linéaire avec pour variable indépendante l'aptitude linguistique selon la candidate ou le candidat lui-même et pour variable dépendante le score moyen dans les items

Version du test	Compréhension orale	Compréhension écrite	Expression orale	Expression écrite
V2 anglais	Constante = 0,570 Pente = 0,040*	Constante = 0,692 Pente = 0,035*	Constante = 3,547 Pente = 0,772*	Constante = 1,059 Pente = 1,102*
V3 anglais	Constante = 0,552 Pente = 0,052*	Constante = 0,682 Pente = 0,040*	Constante = 3,441 Pente = 0,782*	Constante = 1,987 Pente = 0,920*
V2 français	Constante = 0,604 Pente = 0,019	Constante = 0,777 Pente = 0,013*	Constante = 3,017 Pente = 0,814*	Constante = 3,772 Pente = 0,503*
V3 français	Constante = 0,660 Pente = 0,021	Constante = 0,688 Pente = 0,000	Constante = 4,547 Pente = 0,546*	Constante = 3,303 Pente = 0,580*

La constante est le score attendu quand la personne indique que son niveau d'aptitude est nul. La pente est l'augmentation du score attendu pour chaque augmentation d'un point de l'évaluation par la personne de son niveau d'aptitude. Pour le test V2 en anglais, par exemple, si la personne indique que son niveau d'aptitude est 5, le score moyen attendu pour les items est 0,570 + 5 x 0,040 = 0,77. Pour le test V3 en français en expression orale, si la personne indique que son niveau d'aptitude est 4, le score moyen attendu pour les items est 4,547 + 4 x 0,546 = 6,731.

* Lien statistiquement significatif entre le score moyen dans les items du module et l'évaluation par la personne elle-même de son niveau d'aptitude linguistique.

À l'avenir, il serait utile d'envisager de demander aux participantes et participants au test d'indiquer les résultats qu'ils ont obtenus à d'autres tests linguistiques auxquels ils ont pu participer (DELF, TOEFL, tests linguistiques dans le cadre des démarches d'immigration, etc.), afin de disposer d'un meilleur ensemble de données pour examiner les corrélations entre les résultats à l'évaluation des compétences linguistiques des RAPEC et les autres tests linguistiques.

Question 7 – Est-ce que les normes (points de coupure) sont établies de façon appropriée?

Le groupe *Directions* a utilisé une méthode Angoff modifiée pour établir les normes[20]. La méthode Angoff est une méthode solide et défendable sur le plan juridique, utilisée couramment pour établir des normes, en particulier pour des examens avec différents types d'items. Les méthodes Angoff modifiées d'établissement des normes (points de coupure) s'appuient sur le jugement d'un comité de spécialistes. La défendabilité des points de coupure dépend donc de l'expertise des membres du comité. Dans ce cas-ci, les membres du comité responsable de l'établissement des normes avaient au minimum cinq années d'expérience d'enseignement dans la salle de classe au Canada et étaient nommés comme spécialistes par le registraire de leur province. Les membres du comité avaient donc tous une expérience de l'enseignement en salle de classe, ainsi qu'une expérience de l'agrément du personnel enseignant. Les points de coupure initiaux étaient fondés sur le jugement des membres du comité concernant la proportion de participantes et participants ayant les compétences minimales qui répondraient correctement aux items en compréhension orale et en compréhension écrite et sur le score attendu pour une candidate ou un candidat ayant les compétences minimales dans les items en expression orale et en expression écrite.

Le processus d'établissement des normes supposait implicitement que les participantes et participants feraient tous les items dans le module. L'analyse des résultats des participantes et participants, combinés aux résultats du sondage rempli par les participantes et participants à la fin du test, a révélé que bon nombre de participantes et participants avaient eu du mal à terminer les modules d'expression orale et d'expression écrite dans le délai accordé de 60 minutes. Par souci d'équité, le groupe *Directions* a réduit le nombre d'items dans les modules d'expression orale et d'expression écrite et ajusté les points de coupure en conséquence. Les nouveaux points de coupure pour l'expression orale et l'expression écrite ont été établis de manière à ce que le score moyen par item reste cohérent par rapport aux points de coupure initialement fixés. À la suite de

[20] Vous trouverez la description du processus de définition des points de coupure à la partie « Établissement des normes ».

l'examen des items[21], certains items ont également été enlevés des modules de compréhension écrite. Les points de coupure pour la compréhension écrite ont ensuite été révisés en fonction du nombre plus réduit d'items, afin qu'ils restent cohérents par rapport aux points de coupure initialement fixés pour la proportion de participantes et participants ayant les compétences minimales requises pour pouvoir répondre correctement aux items en compréhension écrite.

Les responsables de l'établissement des normes ont ensuite examiné et approuvé les points de coupure révisés et les justifications des révisions. Lors du déploiement de l'évaluation des compétences linguistiques des RAPEC, il faudra inclure dans le travail régulier de rassemblement et d'analyse des données un examen et une révision des points de coupure, ce qui est une démarche normale dans les efforts visant à garantir que le test reste de bonne qualité. Le tableau 9 ci-dessous indique les résultats des discussions et des délibérations pour l'établissement des normes.

Tableau 9. Points de coupure

Langue	Module	Test V2	Test V3
Anglais	Compréhension orale (13 items)	6,5 (max. = 13)* 20 sur 288 participantes et participants échouent	6,5 (max. = 13) 7 sur 291 participantes et participants échouent
	Compréhension écrite (18 items)	12,5 (max. = 18) 26 sur 283 participantes et participants échouent	12,5 (max. = 18) 24 sur 293 participantes et participants échouent
	Expression orale (8 items)	47,7 (max. = 72) 48 (max.) sur 271 participantes et participants échouent	47,7 (max. = 72) 51 (max.) sur 277 participantes et participants échouent
	Expression écrite (5 items)	24,7 (max. = 45) 49 (max.) sur 272 participantes et participants échouent	24,7 (max. = 45) 58 (max.) sur 283 participantes et participants échouent
Français	Compréhension orale (13 items)	7,5 (max. = 13) 26 sur 163 participantes et participants échouent	7,5 (max. = 13) 18 sur 176 participantes et participants échouent
	Compréhension écrite (18 items)	12,5 (max. = 18) 17 sur 161 participantes et participants échouent	12,5 (max. = 18) 34 sur 176 participantes et participants échouent
	Expression orale (8 items)	55,7 (max. = 72) 66 (max.) sur 153 participantes et participants échouent	55,7 (max. = 72) 87 (max.) sur 163 participantes et participants échouent
	Expression écrite (5 items)	29,7 (max. = 45) 70 (max.) sur 158 participantes et participants échouent	29,7 (max. = 45) 87 (max.) sur 165 participantes et participants échouent

* Dans chaque case du tableau pour les tests, la première ligne indique le point de coupure et la deuxième ligne donne le nombre de participantes et participants au test qui échoueraient dans ce module avec ce point de coupure.

Les points de coupure sont identiques pour V2 et V3, en anglais et en français. Comme les deux versions du test semblent être d'un niveau de difficulté équivalent, cela est approprié.

[21] Après l'achèvement des analyses psychométriques, l'équipe du groupe *Directions* a examiné les données de son propre rapport psychométrique et du rapport de la psychométricienne externe, dans le cadre d'un examen global des items, en vue de faire des suggestions sur les items à cibler en priorité pour un réexamen. Ces items ont ensuite été réexaminés par une équipe de spécialistes de l'enseignement de la langue, qui ont fait des recommandations de révision des items ou d'élimination des items quand il n'était pas possible de faire les révisions appropriées.

Avec les points de coupure actuels, la proportion de participantes et participants au test en français qui échoueraient au test serait plus élevée que la proportion de participantes et participants au test en anglais. Cela n'est pas nécessairement problématique (sachant que, par exemple, il peut y avoir de bonnes raisons, comme des cohortes différentes de participantes et participants au test en anglais et en français lors de la mise à l'essai), mais, dans une perspective d'équité, il serait important de surveiller les taux de réussite dans les deux langues, afin de veiller à ce que les participantes et participants au test en français ne soient pas désavantagés par rapport à leurs homologues participant au test en anglais (ou inversement).

Recommandations et considérations pour l'amélioration de la mobilité pancanadienne de la main-d'œuvre dans la profession enseignante

L'adoption de l'évaluation des compétences linguistiques des RAPEC apporterait une contribution utile à la coopération entre provinces et territoires en vue de garantir l'équité dans la mobilité de la main-d'œuvre au Canada. Le travail a commencé quand les registraires responsables des certificats d'aptitude à l'enseignement se sont réunis pour mettre en évidence les obstacles freinant la mobilité de la main-d'œuvre à l'échelle pancanadienne[22]. Le premier jalon important a été une entente visant à garantir que le certificat accordé à l'enseignante ou enseignant dans une province ou un territoire donné lui soit également accordé dans l'autre province ou l'autre territoire quand la personne y déménage.

Les registraires ont atteint un autre jalon lorsqu'ils ont décidé d'examiner et d'éliminer les obstacles empêchant les enseignantes et enseignants formés à l'étranger (EEFE) d'obtenir le certificat d'aptitude à l'enseignement. Ils ont cherché à déterminer si l'évaluation des compétences linguistiques des EEFE faisait obstacle à leur obtention du certificat.

Sous les auspices du CMEC et avec l'appui d'Emploi et Développement social Canada – EDSC (auparavant Ressources humaines et Développement des compétences Canada), les registraires ont pris l'engagement de mettre au point des évaluations des compétences linguistiques des EEFE qui sont dans l'incapacité de prouver qu'ils ont suivi une formation à l'enseignement acceptable en anglais ou en français.

Les registraires ont embauché le groupe *Directions* pour effectuer une analyse des travaux de recherche en vue de mettre en évidence les compétences particulières dont les enseignantes et enseignants du primaire et du secondaire dans les écoles d'anglais langue maternelle et de français langue maternelle ont besoin pour pouvoir prodiguer un bon enseignement. L'analyse des travaux de recherche et les Niveaux de compétence linguistique canadiens (NCLC) ont servi à éclairer la mise au point par le groupe *Directions* du cadre des compétences linguistiques et des niveaux de compétence linguistique, au nom des registraires. Ce cadre définit des compétences dans les quatre modalités suivantes : expression orale, compréhension orale, compréhension écrite et expression écrite. Chaque compétence dans le cadre précise des résultats de rendement dans trois domaines du travail de l'enseignante ou enseignant : enseignement et évaluation; gestion de la classe et du comportement des élèves; et communications avec les parents et avec d'autres spécialistes professionnels.

Les registraires ont également embauché le groupe *Directions* pour passer en revue les évaluations linguistiques utilisées à l'époque (en 2011) et effectuer un tour d'horizon des provinces et des territoires pour voir s'il existait une évaluation des compétences linguistiques adaptée aux EEFE cherchant à obtenir le certificat d'aptitude à l'enseignement au Canada. La conclusion a été qu'aucune évaluation existante ne répondait à toutes les exigences pour la profession enseignante (voir la partie « Phase I : Mise au point de l'évaluation des compétences linguistiques des RAPEC »).

Après avoir atteint un autre jalon dans le processus visant à garantir l'équité dans l'attribution du certificat d'aptitude à l'enseignement avec la production du cadre des compétences linguistiques et des niveaux de compétence linguistique, les registraires ont demandé au groupe *Directions* de mettre au point des évaluations des compétences dans chaque modalité linguistique et dans chaque domaine de l'exercice de la profession enseignante. Le groupe *Directions* a mis au point et mis à l'essai l'évaluation des compétences linguistiques des RAPEC et il a vérifié sa validité pour l'évaluation des compétences linguistiques pour l'enseignement des EEFE dont la formation à l'enseignement ne s'est faite ni en anglais ni en français.

[22] L'Accord de libre-échange canadien (ALEC) réaffirme les dispositions et obligations en matière de mobilité de la main-d'œuvre établies dans le cadre de l'Accord sur le commerce intérieur (ACI) de 1995. Les dispositions de l'ALEC sur la mobilité de la main-d'œuvre (chapitre 7) stipulent que les travailleurs réglementés dans une province ou un territoire doivent être reconnus comme qualifiés par un organisme de réglementation d'une autre province ou d'un autre territoire sans exigence supplémentaire significative en matière de formation, d'expérience, d'examens ou d'évaluations, à moins qu'une exception ait été affichée.

Le prochain jalon important pour l'équité dans l'attribution du certificat d'aptitude à l'enseignement est d'adopter et de se mettre à utiliser l'évaluation des compétences linguistiques des RAPEC. La recommandation qui suit et les considérations qui l'accompagnent se fondent sur les données psychométriques résumées dans le présent rapport, ainsi que sur les données des rapports psychométriques complets produits par le groupe *Directions* et par la psychométricienne externe. Cette recommandation et ces considérations ont pour but d'aider les registraires à décider de déployer ou non ce test et, si oui, à définir les conditions s'appliquant à ce déploiement.

Recommandation

Le groupe *Directions* recommande vivement l'adoption de l'évaluation des compétences linguistiques des RAPEC (en français et en anglais), car l'utilisation réfléchie des deux versions de cette évaluation constitue une façon équitable et défendable de sélectionner les EEFE qui respectent les normes fixées par les RAPEC pour les compétences linguistiques exigées dans l'enseignement au Canada.

Considérations relatives au déploiement de l'évaluation

Considération 1 : Utiliser le cadre de compétences linguistiques qui a servi à éclairer la mise au point des évaluations pour poursuivre le travail de perfectionnement de ces évaluations

L'évaluation des compétences linguistiques des RAPEC se fonde sur un cadre de compétences linguistiques utile et éclairé par les travaux de recherche, qui a constitué un guide pertinent lors de l'élaboration et de l'interprétation des tests. Les items de l'évaluation correspondent aux compétences linguistiques et aux contextes auxquels les enseignantes et enseignants du primaire et du secondaire au Canada font face dans leur milieu de travail. Il faut que le cadre continue de guider la poursuite du travail sur l'évaluation.

Considération 2 : Créer un guide pour la correction

Il est essentiel, pour le déploiement du test, de disposer d'un guide pour la correction des items en expression orale et en expression écrite. Il semble que les conseils prodigués initialement aux correctrices et correcteurs aient permis de produire un travail de correction cohérent et fiable, mais l'offre d'un guide pour la correction serait conforme aux pratiques en vigueur connues pour leur utilité dans l'utilisation de tests et elle déboucherait aussi probablement sur un renforcement de la fiabilité des items et de la fiabilité entre membres du groupe de correction. Le guide pour la correction apporterait également aux responsables de l'établissement des normes une plus grande clarté chaque fois qu'ils ont à réviser les normes. Les conseils prodigués aux correctrices et correcteurs pendant les sessions de correction de la mise à l'essai constituent un bon point de départ pour l'élaboration de ce guide, mais il faut que le document prenne une tournure formelle et qu'il soit examiné par des spécialistes des évaluations des compétences linguistiques, par les utilisatrices et utilisateurs prévus du test et par les correctrices et correcteurs (ou chefs de table) des sessions de correction des tests.

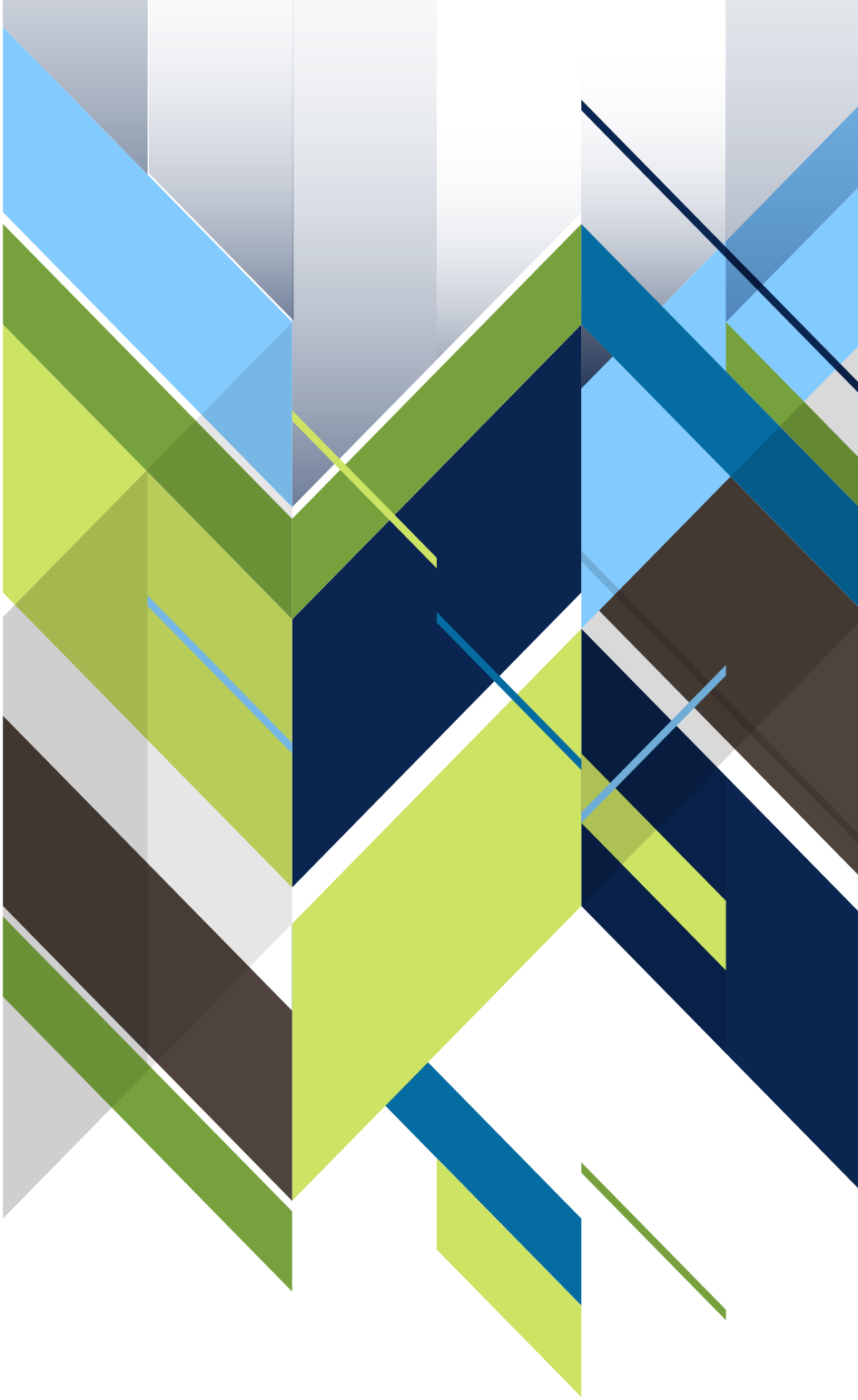
Considération 3 : Faire un travail régulier de rassemblement et d'analyse des données sur le test

Dans tous les programmes de test de qualité, il est normal d'avoir un travail régulier de rassemblement et d'analyse des données émanant du test. C'est tout particulièrement important pour l'évaluation des compétences linguistiques des RAPEC, parce que la population concernée par la mise à l'essai ne

correspondait pas à la population ciblée par le test. Il est donc possible que les propriétés psychométriques des items évoluent lors du déploiement. Il serait utile de demander aux participantes et aux participants au test, lors du déploiement, d'indiquer les scores qu'ils ont obtenus à d'autres tests linguistiques (DELF, IELTS, TOEFL, etc.). Ces scores fourniraient des informations utiles en vue de comparer l'évaluation des compétences linguistiques des RAPEC aux autres tests linguistiques et de comparer les points de coupure de cette évaluation aux autres normes relatives à la maîtrise de la langue. Le travail régulier de rassemblement et d'analyse des données porterait sur le rendement des items, le rendement des correctrices et correcteurs, les biais dans le test et la question de savoir si les points de coupure sont appropriés.

Considération 4 : Format du rapport pour les participantes et participants au test

Les participantes et participants au test d'évaluation des compétences linguistiques recevront un rapport indiquant s'ils ont atteint le score minimal exigé (la norme) pour chacune des modalités. Pour réussir à l'évaluation, il est obligatoire d'atteindre le score minimal dans l'ensemble des quatre modalités.



Évaluation des compétences linguistiques des RAPEC

Phase II : Résultats de la mise à l'essai

RAPPORT FINAL

Financé en partie par le gouvernement du Canada par le biais du Programme de reconnaissance des titres de compétences étrangers

Canada

www.cmec.ca