L'élaboration d'un test provincial pour le classement des étudiants en anglais langue seconde, au collégial

Michel LAURIER
Lydia FROIO
Charles PEARO
Michel FOURNIER

Note 1 du CDC: Le test est disponible par **une entente avec le collège** auprès de Lydia Froio (Collège de Maisonneuve) : Téléphone : (514) 254-7131 ext. 4553

Note 2 : Au sujet de cette version PDF, des pages blanches n'étaient pas présentes dans la copie papier.

Ministère de l'Éducation du Québec
Direction générale de l'enseignement collégial

PAREA

L'élaboration d'un test provincial pour le classement des étudiants en anglais langue seconde, au collégial

Michel LAURIER, *Université de Montréal*
Lydia FROIO, *Collège de Maisonneuve*
Charles PEARO, *Cegep André Laurendeau*
Michel FOURNIER, *Université de Montréal*

Ministère de l'Éducation du Québec
Direction générale de l'enseignement collégial
PAREA

TABLE DES MATIÈRES

Introduction

Avec les modifications aux programmes apportées en 1993, les cours de langue seconde sont devenus obligatoires en vue de l'obtention d'un Diplôme d'études collégiales (DEC). Tous les cégépiens, qu'ils fréquentent un établissement anglophone ou un établissement francophone, doivent maintenant suivre deux cours de langue seconde. Par ailleurs, étudiants et professeurs semblent d'accord pour dire qu'il est souhaitable que les étudiants d'un groupe démontrent, au début du cours, un niveau de maîtrise comparable pour que les apprentissages se déroulent d'une façon adéquate. C'est pourquoi, avant même qu'on ne décide de rendre obligatoires les cours d'anglais langue seconde (ALS), la majorité des collèges au Québec avait mis en place des mécanismes pour administrer un test de classement.

Au moment d'élaborer le test, l'opération de classement consistait à assigner aux élèves un des quatre niveaux de cours normalement offerts ou à les placer dans un cours d'appoint. L'outil de classement utilisé dans la majorité des collèges (le TCALS) datait des années 70. De plus, plusieurs collèges utilisaient d'autres moyens pour effectuer le classement. En imposant des cours de langue seconde aux étudiants, les modifications de 1993 comportaient des propositions quant à l'uniformisation des contenus d'un établissement à l'autre. Cette nouvelle exigence rendait nécessaire la mise au point d'un test qui tienne compte des réalités nouvelles dans lesquelles s'insère l'apprentissage de l'anglais dans les établissements francophones. Ayant toujours connu un milieu dynamique de développement pédagogique où les échanges sont nombreux, les professeurs d'anglais langue seconde du réseau collégial se sont rendu compte de la difficulté, sans référence commune, de communiquer entre eux à propos des nouveaux objectifs des cours. La construction d'un nouveau test de classement est apparue comme une condition à la poursuite du dialogue pédagogique autour des compétences à développer.

Grâce à l'appui financier de PAREA, il a été possible de mettre en marche un projet d'élaboration d'un instrument de classement qui satisfasse les exigences du milieu. Cette recherche-développement a commencé à l'automne 1995 et nous avons pu offrir l'instrument aux établissements au printemps 1997. Le test est présentement à l'essai dans plusieurs collèges et certaines discussions ont eu lieu notamment en ce qui a trait à la modification des scores de césure et aux modalités d'administration. Comme la gestion financière du projet relevait du Cegep André-Laurendeau, l'équipe de recherche était dirigée par Charles Pearo, professeur d'anglais langue seconde dans cet établissement. L'équipe incluait en outre Lydia Froio, professeure d'anglais langue seconde au Collège de Maisonneuve et Michel Laurier, professeur à l'Université de Montréal et spécialiste de l'évaluation en langue seconde. Ce dernier s'est adjoint les services de Michel Fournier, étudiant au doctorat en mesure et évaluation pour la réalisation des analyses.

1 - La fonction du test de classement

Apprendre une langue est un processus complexe, ce qui signifie que la compétence à développer comporte plusieurs facettes, que les habiletés évoluent dans le temps et que les apprentissages se déroulent différemment selon les individus. On ne peut donc pas s'attendre à ce que les groupes soient parfaitement homogènes. La plupart des programmes d'étude en langue proposent une séquence de cours qui se distinguent par le niveau d'entrée qu'on attend des étudiants. Dans ces conditions, la fonction d'un classement est de situer les apprenants sur une échelle unique composée de différents niveaux. Le degré d'homogénéité du groupe à la suite d'un classement dépend bien sûr du nombre de niveaux de cours effectivement offerts. On peut penser que plus l'on distingue de niveaux de cours, plus les groupes seront homogènes et plus le contenu du cours conviendra aux élèves.

Le professeur de langue s'attend à un groupe plutôt homogène afin que les activités qu'il prépare puissent être utiles aux élèves ; ceux-ci s'attendent à être capable de participer à ces activités, qu'elles soient écrites ou orales, et d'en tirer profit. Le test de classement en anglais sert d'abord et avant tout à évaluer le niveau d'anglais des étudiants afin de former des groupes les plus homogènes possible. Ce genre de test mesure un ensemble d'habiletés telles que la compréhension auditive, l'étendue du vocabulaire, la connaissance de la grammaire. On suppose que les aspects de la langue qui sont mesurés dans le test de classement permettent de faire des inférences justes quant au niveau général de maîtrise. Il faut toutefois accepter que, compte tenu de la multiplicité de dimensions que comporte l'apprentissage d'une langue, un classement parfait reste un idéal inaccessible[1].

Un test de classement n'est habituellement pas un outil diagnostique qui vise à identifier des lacunes dans l'apprentissage, ni une épreuve qu'on peut administrer à la fin d'un cours pour vérifier l'atteinte des objectifs de ce cours, ni un examen de compétence qui pourrait, par exemple, attester qu'un apprenant peut exercer certaines tâches professionnelles dans la langue seconde[2]. Puisqu'il s'agit essentiellement de départager les étudiants en fonction du niveau de maîtrise générale, on privilégie un test à interprétation normative. Contrairement au test à l'interprétation critériée qui vise à déterminer si l'apprenant peut réaliser la performance attendue, le test à interprétation normative sert à comparer les élèves entre eux[3]. Cette fonction est déterminante dans la conception de l'épreuve. En effet, il est essentiel que les scores d'un test de classement montrent suffisamment de variance pour que les comparaisons entre apprenants puissent s'établir. Le test doit donc comprendre des tâches qui discriminent bien et ce sur une gamme relativement large de niveaux. Des tests ciblés autour d'un niveau correspondant aux objectifs d'un cours ou aux exigences spécifiques d'une certification ne sont donc pas appropriés, car les scores ne

présentent pas la variance souhaitée. Il importe de garder à l'esprit la spécificité du test de classement en regard des décisions qu'il doit guider. Aussi importante soit-elle, la décision de classer un étudiant à un niveau plutôt qu'un autre ne devrait pas être irrémédiable, ni avoir des effets majeurs sur l'avenir des individus. Dans cette perspective, on ne saurait substituer un test de classement à des instruments parfois plus lourds qui conduisent à des décisions plus déterminantes telle la certification.

Sur le plan du contenu, le test de classement peut être construit en se référant aux objectifs de l'ensemble du programme pour lequel il est conçu ou en se référant à des habiletés transversales qu'on retrouve dans un programme ou un autre et qui s'avèrent déterminantes dans le jugement qu'on peut porter sur le niveau d'un apprenant par rapport à celui des autres apprenants. Étant donné les contenus des cours d'anglais langue seconde offerts dans les collèges, il apparaissait important que le test puisse tenir compte de la compétence de l'apprenant dans les quatre savoirs (lire, comprendre, parler et écrire). Toutefois, à cause des contraintes pratiques qui sont imposées lors de l'administration d'un test de classement, il est difficile d'obtenir des mesures directes des habiletés de production. Ainsi, s'il est possible de prétendre à une validité de contenu en ce qui a trait à la lecture et à l'écoute, on doit se satisfaire d'une validité prédictive en ce qui a trait à l'expression écrite ou orale.

2 -Un sondage auprès des intervenants

En lançant son appel d'offre, le Ministère de l'Éducation du Québec reconnaissait, le besoin de développer un nouveau test de classement en ALS au niveau collégial. Afin de préciser les besoins du milieu, nous avons commencé cette recherche par un sondage. Nous résumons ici cette consultation ; pour plus de détails, nous référons le lecteur au rapport de recherche réalisé par L. Froio[4].

Les contraintes reliées à l'administration du test affectent les administrateurs, le professeurs et les élèves. Afin de tracer un portrait complet de la situation dans le réseau, trois questionnaires distincts ont été élaborés et envoyés dans tous les collèges (publics ou privés). Chaque questionnaire s'adressait à des groupes d'intervenants qui sont affectés d'une façon ou d'une autre par la procédure de classement :

- la Direction des études ;
- les coordonnateurs des départements de langues ;
- les professeurs d'anglais.

Les questionnaires comportaient deux grandes sections : le test de classement présentement utilisé et le nouveau test de classement. Les questions portaient sur la pertinence d'un test de classement, les perceptions des intervenants vis-à-vis de leur propre test de classement, les changements de niveaux et d'autres problèmes tel que le rétro-classement (les étudiants qui feignent le comportement d'un étudiant moins avancé que leur niveau réel pour pouvoir suivre un cours plus «facile»). On touchait aussi l'administration du test.

Nous avons également demandé aux trois instances leurs opinions sur la façon dont elles envisageaient le nouveau test de classement et sur les caractéristiques qu'un tel test devrait présenter. Les questions portaient sur les thèmes suivants :

- l'uniformisation des contenus en regard du test et les scores de césure (seuils de classement d'un niveau à l'autre) ;
- le contenu du test en fonction des compétences attendues ;
- le type de test (administré à l'aide de l'ordinateur ou conventionnel) ;
- la responsabilité des divers intervenants en ce qui a trait à l'administration et à la correction du test.

2.1 - La réponse des intervenants

La participation à cette étape de la recherche a de loin dépassé nos attentes. Le tableau 1 indique les taux de réponses pour chacun des groupes consultés.

Établissements	Direction des études	Coordination départementale	Enseignants d'anglais
publics	39	39	213 réponses
privés	9	8	26 réponses
Taux de réponse	85,7%	80,3%	239 questionnaires reçus*

* Il est difficile d'estimer le taux de réponse, car on ne connaît pas le nombre exact de professeurs d'anglais langue seconde dans le réseau.

Tableau 1
Nombre de collèges et de professeurs qui ont répondu aux questionnaires

Nous croyons que le taux élevé de participation est la manifestation d'un désir chez les intervenants d'améliorer non seulement **l'outil** mais aussi la **procédure** de classement. De fait, ce désir se trouve confirmé, en partie, par les réponses à la question qui visait à déterminer si les directeurs des études et les enseignants voulaient un nouveau test de classement (tableau 2).

	Direction des études	Enseignants d'anglais
OUI	78,8%	77,1%
NON	12,5%	2,1%
Incertain	16,7%	20,8%

Tableau 2
Souhaitez-vous voir un nouveau test de classement ?

2.2 - Les outils en usage

Nous avons constaté que plusieurs outils servent à mesurer la compétence des étudiants pour les fins du classement. L'histogramme de la figure 1 montre la répartition des différents tests de classement utilisés dans les 45 collèges qui ont répondu. Il est à noter que même quand plusieurs collèges utilisent le même outil de classement, les scores de césure peuvent varier.

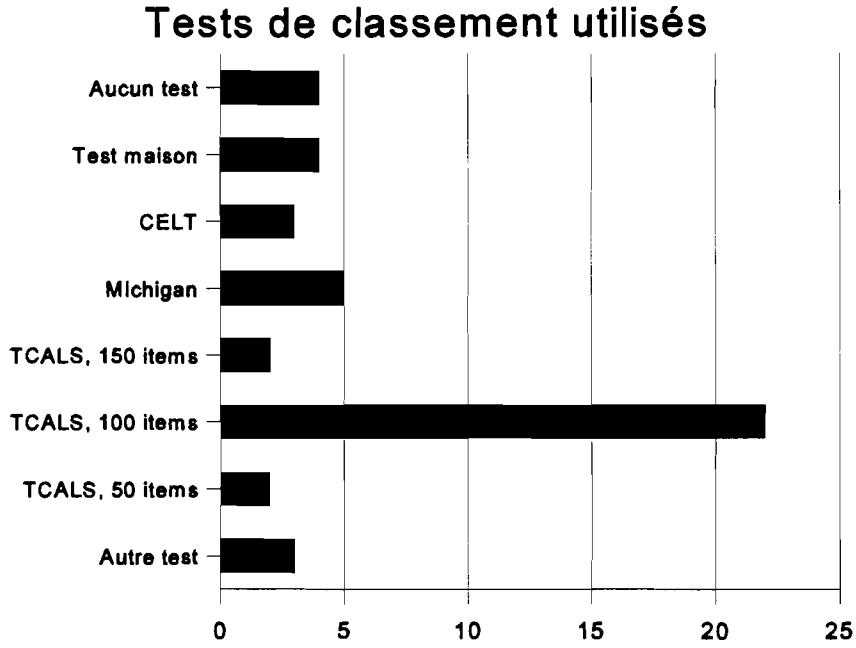

Figure 1
Outils de classement en usage dans les collèges

Une telle variété d'instruments ne permet pas un classement dont les résultats puissent être comparables d'un cégep à l'autre à travers la province. De plus, comme nous l'avons mentionné, certains collèges qui utilisent le même outil n'ont pas nécessairement les mêmes scores de césure. Un élève classé au niveau 101 dans un collège pourrait se retrouver à un niveau différent dans un collège voisin. On imagine les complications que pareille situation engendre pour les étudiants qui envisagent changer de collège. Nous avons interrogé les directeurs des études et les enseignants sur la pertinence d'un instrument commun pour le classement en anglais langue seconde. Les résultats que nous reproduisons dans le tableau 3 montrent qu'une vaste majorité des répondants favorise l'usage d'un test unique.

	Direction des études	Enseignants d'anglais
OUI	85,1%	81%
NON	6,4%	8,9%
Incertain	8,5%	10,1%

Tableau 3
Souhaitez-vous voir le même test de classement utilisé dans tous les cégeps ?

Non seulement les répondants favorisent-ils l'usage d'un test unique, mais ils souhaitent aussi l'harmonisation des scores de césure c'est-à-dire des seuils uniformes pour décider de classer un étudiant à un niveau plutôt qu'à un autre (tableau 4)

	Direction des études	Enseignants d'anglais
OUI	78,7%	65,5%
NON	10,6%	20%
Incertain	10,6%	14,5%

Tableau 4
Souhaitez-vous voir les même scores de césure utilisés dans tous les cégeps ?

Les professeurs qui sont en faveur d'un test commun avec les mêmes scores de césure donnent comme raison principale la nécessité d'assurer que les différents niveaux de cours correspondent aux objectifs et aux exigences ministériels dans tous les collèges. Selon eux, sans cette uniformisation, il sera impossible d'atteindre les mêmes objectifs. L'uniformisation permet non seulement le passage d'un étudiant d'un niveau à un autre, mais apparaît aussi comme une condition incontournable à des discussions pédagogiques intercollégiales fructueuses sur l'évaluation des apprentissages, sur le matériel didactique et sur le contenu des cours. Il faut cependant noter, en comparant les tableaux 3 et 4, que plusieurs enseignants qui se disent d'accord avec un instrument commun hésitent à endosser le principe de scores de césure uniformes. De fait, les commentaires recueillis indiquent que certains enseignants craignent que des niveaux communs ne conduisent à une sur-représentation des niveaux avancés dans la région de Montréal. Certains pensent même ne pas pouvoir offrir les cours les plus avancés dans leur collège. En effet, il faut s'attendre, avec l'administration d'un test unique utilisant des scores de césure

communs, à ce que la distribution du nombre de cours donnés à un certain niveau varie d'une région à une autre. Il faut cependant préciser que cela n'affecte pas pour autant le nombre total de cours offerts.

2.3 - Le degré de satisfaction

48% des professeurs se disent satisfaits du test qu'utilise l'établissement. Par contre, 15% n'ont pas pu répondre à cette question. Ce pourcentage élevé d'indécis s'explique principalement par le fait que beaucoup de professeurs ont peu d'expérience d'enseignement dans le réseau collégial et se disent alors incapable de juger des résultats de leur outil de classement. En effet, 50% des professeurs qui ont répondu au questionnaire ont trois ans ou moins d'expérience au collégial. Beaucoup ne sont pas familiers avec le test qui est utilisé. 60% de tous les professeurs n'ont jamais passé le test de classement eux-mêmes.

La plupart des intervenants semblent s'entendre pour dire que la présence d'étudiants mal classés au sein d'un groupe affecte la qualité des apprentissages. Par ailleurs, on ne peut pas miser sur la possibilité d'effectuer des changements de groupe une fois le cours commencé. En effet, il peut s'écouler trois semaines de cours avant que l'on s'aperçoive d'un problème de classement et il est alors trop tard pour changer de niveau. Même si les collèges acceptent de changer les élèves de niveau, les changements doivent habituellement être faits dans un délai d'une semaine (48%) ou deux semaines (33%). De plus, des conflits d'horaire empêchent souvent les changements. C'est pourquoi, parmi les élèves mal classés, seulement 55% changent de niveau.

On demandait aux enseignants d'ordonner, selon leur importance, les raisons qui font qu'un élève se trouve mal classé. Deux de ces raisons sont souvent évoquées :

- Le rétro-classement : 78% des enseignants affirment que des élèves font exprès de ne pas répondre correctement aux questions du test dans le but de se voir classés à un niveau inférieur à leur niveau réel. 69% des départements qui ont répondu n'ont pas de politique à l'égard de cette forme de «tricherie».
- Le contenu du test : 76% des enseignants mettent en cause la validité du test en soulignant que le test ne mesure pas les quatre savoirs.

Selon les réponses des enseignants concernant la composition de l'épreuve, il ressort que le nouveau test de classement devrait viser les quatre savoirs langagiers c'est-à-dire la lecture, l'écrit, l'écoute et l'expression orale. Plusieurs commentaires signalent l'importance que le test...

- soit adapté à la clientèle cible en tenant compte de l'âge des cégépiens, de leurs intérêts et de leur réalité culturelle ;
- contienne une partie d'auto-évaluation ;
- puisse détecter les cas frontières aussi bien que ceux et celles qui tentent délibérément de se classer à un niveau moins avancé que leur niveau réel.

2.4 - L'administration de l'épreuve

Outre les questions touchant la composition du test, le questionnaire abordait des aspects reliés à l'administration d'un test de classement. Les résultats du questionnaire destiné aux administrateurs révèlent que, dans 69% des collèges, c'est le registrariat qui a la responsabilité d'administrer le test. Dans l'éventualité d'un nouveau test, 63% des administrateurs souhaiteraient continuer à jouer ce rôle. Ils aimeraient tout de même trouver des moyens de réduire les coûts d'administration et de correction du test. La majorité des étudiants passent le test de classement en grand groupe, dans un auditorium, avant leur premier semestre d'étude. Cette opération peut poser des problèmes de logistique de sorte que les administrateurs cherchent à faciliter le processus.

Dès le début de la recherche, l'équipe croyait à la faisabilité et même à la nécessité de développer un test adaptatif[5]. Cette forme de test qui met à contribution le micro-ordinateur s'avère particulièrement intéressante dans le cas d'un test de classement, car elle permet de choisir les items les plus pertinents c'est-à-dire ni trop faciles ni trop difficiles. De cette façon, on réduit la longueur du test sans affecter la fidélité de la mesure. Par ailleurs, l'informatisation d'un test permet une correction basée sur la configuration des réponses plutôt que sur le simple cumul des réponses correctes. Il devient alors possible de détecter certaines combinaisons aberrantes qui peuvent signaler un comportement de rétro-classement. Le cas de ces étudiants qui, croyant ainsi réussir un cours sans grand effort pose des problèmes, parce qu'une fois en classe, ces étudiants s'avèrent peu motivés et peuvent même intimider leurs pairs par leur performance. Il faut cependant souligner que la solution définitive à ce problème ne se trouve pas dans le mode d'évaluation. Quelle que soit la forme de test retenue, le problème du rétro-classement devra faire l'objet de discussion dans les départements de langue.

Quand nous avons demandé aux intervenants s'ils souhaitaient voir un test administré à l'aide de l'ordinateur, ils se sont montrés, dans l'ensemble, réceptifs à l'idée (tableau 5)

	Direction des études	Enseignants d'anglais
OUI	44,7%	57,6%
NON	27,7%	16,5%
Incertain	27,7%	26%

Tableau 5
Souhaitez-vous voir un test de classement administré à l'aide de l'ordinateur ?

Toutefois, malgré la préférence qu'ont exprimées les répondants envers une version informatisée du test, les contraintes de temps imposées au projet (deux ans sans possibilité d'extension), nous ont amenés vers un test papier-crayon plutôt qu'un test administré à l'aide de l'ordinateur. Certaines techniques issues des recherches sur le testing adaptatif nous ont cependant inspirés dans la conception du nouveau test. On pourrait aussi envisager dans les prochaines années, l'informatisation de la correction avec l'application de procédures de détection des cas de rétro-classement.

3 - La rédaction des tâches de mesure

3.1 - Structure générale du test

Des contraintes pratiques incontournables dictent souvent les modalités d'application d'un outil de classement. Ainsi, il nous apparaissait clair que tout devait être mis en oeuvre afin de s'assurer que la passation du test puisse se faire rapidement (pas plus d'une heure et demie) sans imposer de conditions particulières (outre l'utilisation d'un magnétophone). De plus, la correction devait se faire rapidement, économiquement et sans erreur.

Étant donné ces contraintes, il était peu réaliste de suggérer l'utilisation d'épreuves de production que ce soit à l'oral ou à l'écrit. La décision de ne pas évaluer directement la production n'empêche cependant pas les établissements qui souhaiteraient le faire, de compléter l'épreuve de classement par une composition écrite ou une courte entrevue. Une telle initiative ne peut que fournir une information plus valide et plus fiable en vue de la décision de classement. Par ailleurs, comme les résultats doivent habituellement être disponibles dans un délai très court, nous avons opté pour des questions à choix multiple. Bien que ce type d'items ait souvent fait l'objet de critiques virulentes dans les milieux éducatifs, il permet la construction de tests qui départagent bien les étudiants en fonction de leur niveau général.

Le fait d'exclure la production et de limiter les tâches de mesure à des questions à choix multiple soulève certains problèmes sur le plan de la validité du test. Le long débat autour de l'hypothèse du trait unitaire, hypothèse selon laquelle il existerait une compétence générale[6] qui se situe en amont des manifestations plus spécifiques, a fait ressortir la complexité de la compétence langagière et nul ne saurait prétendre aujourd'hui pouvoir rendre compte d'une compétence générale dans la langue, par des mesures indirectes. Cependant s'il ne fait pas de doute que la compétence langagière est multidimensionnelle, nous avons vu que l'opération de classement conduit quant à elle à une décision unidimensionnelle qui consiste habituellement à situer l'individu sur un continuum qu'on pourrait associer à une compétence générale. Dans cette perspective, nous avons cherché à construire un instrument qui couvre plusieurs aspects de la compétence langagière en lien les uns avec les autres parce qu'ils impliquent des habiletés qui interviennent quelle que soit la situation de communication. Par exemple, il ne fait pas de doute que la connaissance de vocabulaire s'avère une habileté déterminante que ce soit à l'oral ou à l'écrit, que ce soit en production ou en compréhension.

La spécification des tâches s'est effectuée à partir des objectifs du programme d'anglais langue seconde. Ces objectifs se retrouvent dans un document auquel ont contribué les professeurs

membres de l'équipe de recherche[7]. Ce document, qui fixe les standards des programmes en ALS, établit clairement les aspects qui doivent être couverts par les professeurs pour chacun des cours pour lesquels un classement est requis. Nous nous sommes assurés d'inclure dans les versions initiales du test des items qui visent les principaux éléments des profils d'anglais langue seconde tel que définis dans ce document, à condition que leur évaluation soit compatible avec le type de tâches de mesure que nous avions retenu.

Cette forme d'échantillonnage nous assurait au départ d'une certaine validité de contenu bien que par la suite nous n'ayons jamais cherché à préserver cette représentativité. De fait, l'analyse que nous avons effectuée par la suite nous a amené à éliminer plusieurs aspects du contenu qui s'avéraient peu pertinents en regard de la décision de classement. L'analyse nous a de plus démontré que certains aspects appartenant à un niveau donné dans le programme pouvaient, dans le contexte d'une tâche de mesure réelle, être associés à un degré de difficulté très différent de celui que le niveau du programme aurait laissé présager. C'est d'abord la qualité métrologique de chaque item plutôt que son lien étroit avec le programme qui nous a amenés à le conserver ou à l'éliminer. Puisque que c'est le score qui est interprété plutôt que le résultat de chaque item, une telle situation ne pose pas de problème. On pourrait même arguer qu'une telle approche assure une viabilité de l'instrument au-delà des inévitables modifications de programmes et peut même fournir des données précieuses en vue de l'établissement de descriptions de performance sur une base empirique. À un moment où des modifications de programme sont à prévoir et où le courant des standards de performance se répand, une telle robustesse de l'instrument doit être considérée comme une caractéristique souhaitable.

Dans la perspective que nous venons de décrire quatre composantes de la compétence langagières nous semblaient devoir être évaluée dans ce test de classement :

- la compréhension auditive ;
- le vocabulaire ;
- les habiletés grammaticales ;
- la lecture.

L'évaluation de la compréhension auditive se fait à l'aide de trois types de tâche différents. Les contraintes inhérentes à la nature de l'épreuve nous empêchent de reproduire dans le présent rapport les items effectivement retenus. Toutefois, afin d'illustrer le type de tâche, nous reproduisons l'exemple qui est fourni dans le livret afin de s'assurer de la compréhension de la consigne. Il s'agit donc d'items particulièrement faciles qui montrent bien ce qui est attendu de l'élève. Dans un premier sous-test, l'étudiant doit écouter une phrase et répondre à une question de compréhension (exemple 1). Dans un deuxième sous-test, l'étudiant écoute un court dialogue et répond à une question (exemple 2). Enfin, un troisième sous-test requiert l'audition d'un monologue plus long, un «mini-exposé», sur lequel portent plusieurs questions (aucun exemple). Il faut souligner l'effort que nous avons mis à produire des enregistrements de qualité professionnelle. Nous avons choisi des personnes expérimentées pour les voix et nous avons réalisé les bandes maîtresse en studio avec l'aide d'un technicien de son.

On demande à l'élève de lire les énoncés, d'écouter la phrase (une seule fois) puis de choisir l'énoncé qui correspond le plus à la phrase entendue.

«John wants to go to the movies with Mary.»

A) He doesn't like to go to the movies
B) He want to go to the movies alone.
C) He prefers to go to the movies with Mary. *
D) He doesn't want to go to the movies.

Exemple 1
Item de compréhension auditive : phrase

On demande à l'élève de lire les énoncés, d'écouter le dialogue (une seule fois et de choisir l'énoncé qui répond à la question.

Dialogue où un des personnages demande à l'autre où se trouve son livre

Where is the man's book?

A) In the bedroom
B) In the dibing room
C) In the kitchen *
D) In the living room

Exemple 2
Item de compréhension auditive : dialogue

Comme les habiletés lexicales s'avèrent souvent d'excellent prédicteurs de la compétence générale, nous avions prévu que la section de compréhension écrite commencerait avec deux sous-tests de vocabulaire. Dans le premier type de tâche, l'étudiant devait identifier un synonyme (ou un mot sémantiquement apparenté) au mot souligné qui apparaissait dans la phrase. Nous n'avons cependant pas retenu ce type de tâches dans la version finale. Le sous-test que nous avons conservé est constitué de phrases lacunaires où l'étudiant doit identifier le mot qui convient le mieux dans l'espace (exemple 3).

On demande à l'élève de choisir le mot qui complète correctement la phrase présentée.

My sister plays the _____ .

A) concert
B) piano
C) store
D) table

Exemple 3
Item de vocabulaire 2 : phrase lacunaire

Les habiletés grammaticales se trouvent évaluées dans deux sous-tests. Dans un cas, il s'agit d'une tâche semblable à la précédente qui évaluait les connaissances lexicales, mais le remplacement suppose l'application de règles morpho-syntaxiques (exemple 4). Dans l'autre cas, l'étudiant doit détecter les erreurs grammaticales en identifiant la phrase qui est écrite correctement (exemple 5). Rappelons que les exemples présentés ici sont beaucoup plus simples que la majorité des items qui se trouvent effectivement dans l'épreuve puisqu'il s'agit des exemples fournis dans les directives à l'étudiant.

On demande à l'élève de choisir le mot qui complète correctement la phrase présentée.

My father _____ to work at 9 a.m.

A) go
B) goes *
C) going
D) gone

Exemple 4
Item de grammaire 1 : phrase lacunaire

On demande à l'étudiant d'indiquer l'énoncé qui ne comporte pas d'erreur grammaticale.

A) I never is playing football.
B) I never play football. *
C) I never player football.
D) In never plays football.

Exemple 5
Item de grammaire 2 : correction d'erreurs

La capacité à lire un texte en anglais est évaluée par les deux derniers sous-tests. Étant donné qu'il s'agit d'une tâche familière et qu'un exemple aurait alourdi inutilement l'administration, le livret de l'étudiant ne fournit pas d'exemple. Un des sous-tests implique une compréhension plutôt locale puisqu'il s'agit de répondre à une seule question portant sur un segment assez court (20-30 mots). L'autre fait davantage intervenir les phénomènes discursifs, car il est composé de passages plus longs (150-250 mots) qui sont suivis de plusieurs questions.

Sous-test	nombre d'items
Compréhension auditive	
Phrases	14
Dialogue	14
Mini-exposés	8
Anglais écrit	
Vocabulaire 1	17
Vocabulaire 2	8
Grammaire	25 ou 26
Analyse d'erreurs	11
Lecture 1	5
Lecture 2	8
Total	110 ou 111

Tableau 6
Nombre d'items par sous-test

Nous avons lancé un appel aux professeurs d'anglais des différents cégeps de la province afin qu'ils rédigent des items en relation avec les profils de niveau. Plusieurs professeurs nous ont envoyé des items, particulièrement en ce qui a trait aux habiletés grammaticales. L'équipe a également rédigé un grand nombre d'items. Les items susceptibles d'être retenus ont été minutieusement vérifiés afin de déterminer s'ils se conformaient aux règles habituelles de rédaction d'items à choix multiple. Les items ont été par la suite soumis à quelques collègues pour qu'ils en fassent une lecture attentive. Après quelques corrections, les items ont été répartis afin de créer les versions expérimentales. Nous avons retenu 244 items pour composer trois versions expérimentales, deux versions de 111 items et une autre de 110. Ainsi, 44 items se retrouvaient dans les trois versions. Un tel chevauchement nous a permis de vérifier si ce type d'ancrage permettait effectivement de fixer l'échelle en vue d'une calibration commune. Le recours à des items d'ancrage permet d'expérimenter un grand nombre d'items sans devoir soumettre aux étudiants des épreuves d'une longueur excessive. Le tableau 6 permet de voir la distribution des items en fonction des différents sous-tests.

3.2 - Pré-expérimentation

La pré-expérimentation s'est effectuée à l'automne 1996 auprès d'un échantillon relativement réduit de sujets, soit 258 étudiants provenant de différents collèges où les professeurs avaient accepté de participer à l'opération avec leur groupe. Sans nécessairement viser une distribution tout à fait comparable pour les trois versions, nous nous sommes assurés de toujours trouver une grande variété de niveaux et de réunir environ 75 étudiants par version. La pré-expérimentation visait deux buts principaux :

- détecter les items déficients pour les éliminer ou les corriger ;
- examiner la structure interne du test (le construit).

Comme les étudiants répondaient sur des feuilles de réponses lisibles par lecteur optique, nous avons pu procéder à une saisie automatisée des données. Ces données ont par la suite été analysées au *Laboratoire d'analyse des données de la Faculté des sciences de l'éducation de l'Université de Montréal.* Nous avons d'abord procédé à une analyse classique des items à l'aide de la procédure ITEMAN du logiciel *MicroCAT®*. ITEMAN permet de réaliser une analyse d'items classique en fournissant des statistiques sur l'ensemble du test et sur chacun des items. Les données de chaque version ont fait l'objet d'une analyse particulière.

	Version A	Version B	Version C
Nombre d'items	111	111	110
Nombre de sujets	105	87	66
Moyenne	44,2	65,3	61,4
Médiane	40	66	64
Écart-type	17,8	17,3	25,6
Alpha	0,94	0,94	0,97

Tableau 7
Résultats globaux de la pré-expérimentation

Le tableau 7 rend compte des résultats obtenus aux trois versions pré-expérimentales soumises respectivement à 105, 87 et 66 étudiants. En comparant les moyennes et les médianes, on constate que la première version semblait plus difficile que les deux autres, mais il est possible que la différence tienne à l'échantillon lui-même plutôt qu'aux items. De fait, ces premiers résultats ont confirmé les inquiétudes que nous avaient communiquées plusieurs professeurs quant à la difficulté de l'épreuve. Par ailleurs, toutes les versions présentaient une variance assez grande

(indiquée par l'écart-type) ce qui est essentiel dans un test de classement puisqu'il s'agit de maximiser les écarts entre les étudiants. Cette variance, associée au nombre relativement grand d'items, se traduit par un degré de fidélité (alpha) très élevé. Le coefficient alpha permet quantifier la marge d'erreur du test. Puisqu'il est d'usage de le rapporter, nous l'avons inclus dans le tableau, bien qu'il faille être prudent dans l'interprétation d'un indice de fidélité calculé sur l'ensemble d'une épreuve qui se compose de sous-tests dont les contenus sont différents. Les coefficients obtenus nous ont tout de même rassurés sur la précision de la mesure que nous pouvions attendre de l'instrument.

Avant d'examiner les résultats pour chaque item, nous avons procédé à une calibration selon un modèle de théorie de réponses aux items (TRI) à un paramètre (modèle de Rasch) en utilisant le programme *BILOG®*. Nous reviendrons plus loin sur la technique de calibration, car elle a joué un rôle déterminant dans la composition finale du test. Lors de la pré-expérimentation, la calibration nous a servi simplement à vérifier l'efficacité des 44 items d'ancrage. Le procédé s'avérant efficace, nous avons pu tenir compte des différences entre les échantillons des trois versions dans l'interprétation des degrés de difficulté de chacun des items. Nous avons procédé à une analyse systématique des probabilités de bonne réponse (p) de chaque item. Nous avons éliminé les items qui s'avéraient beaucoup trop difficiles ($p < 0,3$) ou beaucoup trop faciles ($p > 0,95$). Dans certains cas, une modification a permis de rectifier le niveau de difficulté. Nous avons également tenu compte des indices de discrimination que représentent les corrélations bisérielles (r_{bis}). Nous recherchions un indice fortement positif pour la bonne réponse ($r_{bis} > 0,4$, en prenant en considération p) et, inversement, des indices négatifs du côté des leurres (les choix de mauvaises réponses). Nous avons modifié les leurres qui se révélaient trop attirants ou, au contraire, qui n'avaient aucun effet.

L'ensemble de l'opération a conduit à des révisions importantes dans la banque. À cet égard, comme nous le constaterons par la suite, la pré-expérimentation a permis de réduire substantiellement la «mortalité» des items à la suite de l'expérimentation. Les items retenus après révision ont servi à former trois nouvelles versions du test, comportant chacune 110 items. Parmi ces items, 42 étaient communs aux trois versions et allaient servir d'ancrage pour les analyses subséquentes. Nous avons donc conservé 246 items différents.

Le but de la pré-expérimentation était aussi d'obtenir des informations préliminaires sur la structure interne du test, c'est-à-dire sur les liens entre les sous-tests, de façon à pouvoir inférer les traits qui sont effectivement mesurés. Il s'agissait donc d'une première opération en vue de vérifier la validité conceptuelle (validité de construit) de l'instrument. Au moyen de la procédure CORRELATION de *SPSS®*, nous avons créé les matrices de corrélation entre les sous-tests de chaque version. Ces matrices sont reproduites dans le tableau 8. Evidemment, étant donné le nombre limité de sujets en pré-expérimentation, nous avons dû limiter l'analyse à un examen visuel des matrices. À ce moment, nous avons pu faire les constatations suivantes :

Version A
(105 sujets)

	Phrases	Dial	Mini	Voc1	Voc2	Gramm	Erreurs	Lec1	Lec2
Phrases	-								
Dialogues	0,44	-							
Mini-exposés	0,47	0,29	-						
Vocabulaire 1	0,59	0,63	0,35	-					
Vocabulaire 2	0,33	0,43	0,30	0,45	-				
Grammaire	0,61	0,63	0,41	0,75	0,57	-			
Analyse d'erreurs	0,43	0,47	0,28	0,59	0,31	0,74	-		
Lecture 1	0,35	0,51	0,36	0,60	0,29	0,64	0,54	-	
Lecture 2	0,45	0,61	0,37	0,61	0,50	0,67	0,51	0,57	-
Alpha	0,60	0,66	0,14	0,71	0,43	0,84	0,65	0,56	0,73

Version B
(87 sujets)

	Phrases	Dial	Mini	Voc1	Voc2	Gramm	Erreurs	Lec1	Lec2
Phrases	-								
Dialogues	0,55	-							
Mini-exposés	0,49	0,45	-						
Vocabulaire 1	0,59	0,66	0,52	-					
Vocabulaire 2	0,48	0,44	0,49	0,56	-				
Grammaire	0,60	0,69	0,61	0,73	0,49	-			
Analyse d'erreurs	0,53	0,54	0,43	0,62	0,50	0,69	-		
Lecture 1	0,52	0,63	0,46	0,50	0,43	0,60	0,48	-	
Lecture 2	0,53	0,55	0,52	0,55	0,55	0,56	0,53	0,57	-
Alpha	0,60	0,71	0,50	0,71	0,49	0,82	0,46	0,57	0,64

Version C
(66 sujets)

	Phrases	Dial	Mini	Voc1	Voc2	Gramm	Erreurs	Lec1	Lec2
Phrases	-								
Dialogues	0,80	-							
Mini-exposés	0,57	0,58	-						
Vocabulaire 1	0,90	0,83	0,67	-					
Vocabulaire 2	0,57	0,44	0,42	0,56	-				
Grammaire	0,87	0,83	0,65	0,91	0,52	-			
Analyse d'erreurs	0,77	0,72	0,56	0,82	0,50	0,80	-		
Lecture 1	0,71	0,55	0,44	0,63	0,31	0,65	0,60	-	
Lecture 2	0,78	0,73	0,53	0,79	0,32	0,76	0,75	0,74	-
Alpha	0,85	0,85	0,61	0,89	0,33	0,90	0,82	0,53	0,75

Tableau 8
Corrélations entre les sous-tests pour chacune des versions pré-expérimentales

- Dans l'ensemble de la grille les corrélations sont relativement élevées. On peut donc penser que le test mesure, avec une certaine précision, des aspects interreliées qui convergent vers une compétence générale complexe. C'est précisément ce que nous recherchons dans un test de classement.
- Bien qu'elles ne ressortent pas autant que nous l'aurions souhaité, les corrélations entre les sous-tests associés à une même composante (compréhension auditive, vocabulaire, habiletés grammaticales et lecture) sont relativement élevées (sauf pour ce qui est des mini-exposés de la première version). On peut y voir une indication du fait qu'on mesure effectivement un trait commun à ces sous-tests.
- Les sous-tests «Vocabulaire 1» et «Grammaire» sont fortement associés à la plupart des autres sous-tests. Cela est particulièrement évident dans la troisième version. On peut expliquer ce phénomène par le caractère transversal des habiletés grammaticales et lexicales. On note également des corrélations assez élevées entre les sous-tests «Phrases» et «Lecture 2» avec les autres sous-tests.

Une telle analyse n'est qu'indicative, étant donné la multiplicité des interactions et l'effet variable de l'erreur de mesure d'un sous-test à l'autre. Les résultats ont toutefois permis, au terme de la pré-expérimentation, de confirmer, jusqu'à un certain point, la cohérence et la pertinence du contenu du test pour des fins de classement.

4 - L'expérimentation des items

4.1 - L'échantillonnage

Un test destiné à l'ensemble des établissements d'enseignement collégial du Québec doit évidemment être expérimenté auprès d'un échantillon qui partage les caractéristiques dominantes de la population. Au moment de l'expérimentation des items, nous avons apporté une attention particulière aux variations régionales et nous avons visé une représentation des niveaux conforme à la réalité. La composition de l'échantillon a donc été dictée par une double stratification : d'une part, selon la région et d'autre part, selon le niveau.

Cégep	sujets	pourcentage
Abitibi	189	9,3%
Alma	42	2,1%
Granby	80	4,0%
Lasalle	139	6,9%
Lévis-Lauzon	179	8,8%
Limoilou	47	2,3%
Lionel-Groulx	278	13,7%
Matane	40	2,0%
Mérici	47	2,3%
Rimouski	23	1,1%
Rivière du Loup	67	3,3%
Rosemont	635	31,4%
St-Laurent	171	8,5%
Victoriaville	86	4,3%
Total	2023	100,0%

Tableau 9
Répartition des sujets par établissement

Afin d'expérimenter les items du test de classement, nous avons communiqué avec les coordonnateurs de tous les établissements d'enseignement collégial qui avaient répondu positivement à notre sondage initial et qui utilisaient le TCALS. L'expérimentation s'est déroulée

à l'automne 1997. Grâce à la collaboration admirable des coordonnateurs et malgré les débrayages et fermetures qui ont affligé beaucoup de cégeps à ce moment, nous avons pu constituer un échantillon de 2023 sujets. Le tableau 9 présente la répartition des sujets à qui une des trois versions a été administrée.

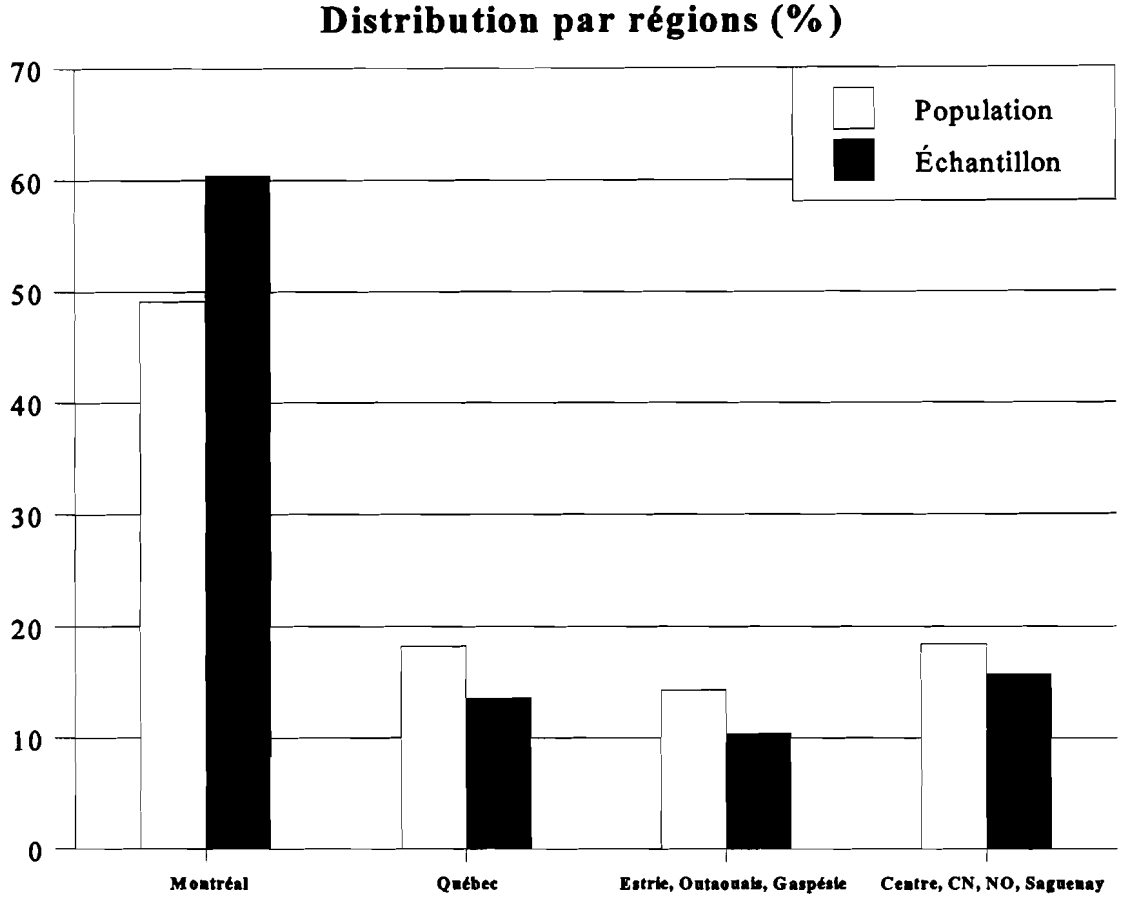

Figure2
Distribution de l'échantillon et de la population des Cégeps par région, en pourcentage

L'importance de la région tient aux grandes disparités qu'on observe en ce qui a trait aux occasions d'utiliser la langue seconde. On sait que dans des régions où le contact avec la langue anglaise est plus fréquent et où sa maîtrise apparaît plus nécessaire, le niveau général des étudiants est plus élevé. De plus, il est possible, qu'à compétence égale, selon leur degré d'exposition à l'anglais et leur contexte socio-géographique, les sujets répondent différemment à certains items. Afin de contrôler cette variable, nous avons regroupé les sujets en quatre grandes catégories :

- Montréal : milieu urbain, pluriethnique où le contact avec la langue anglais est fréquent ;
- Québec : milieu urbain, relativement homogène où le contact avec l'anglais est plutôt limité ;
- Estrie, Outaouais et Gaspésie : régions où l'exposition à l'anglais est en général importante ;
- Centre, Côte Nord, Nord-Ouest et Saguenay : régions où l'exposition à l'anglais est en général réduite.

L'histogramme de la figure 2 rend compte de la représentativité de l'échantillon en ce qui a trait à la région. On observe que malgré une légère surreprésentation de la région montréalaise, la distribution des sujets se conforme assez bien à celle de la population.

Nous avons mentionné que la mise au point d'un outil destiné à classer tous les étudiants qui ont à suivre des cours d'ALS suppose que l'échantillon se caractérise par une répartition des niveaux similaire à celle qu'on trouve habituellement dans les établissements. Il faut souligner que cette variable est plus difficile à contrôler étant donné l'absence d'outil de classement commun. C'est pour cette raison que nous avons limité l'échantillon à des sujets inscrits dans des collèges qui effectuaient leur classement à l'aide de l'instrument le plus répandu (le TCALS). Par ailleurs, comme nous le verrons, le type d'analyse que nous avons utilisé fournissait des indices assez robustes du degré de difficulté des items à la condition que nous ayons un nombre suffisant de sujets à chacun des niveaux. Qui plus est, avec ce type d'analyse nous avons pu tirer profit des items d'ancrage et ne pas fausser les résultats à cause d'une distribution inégale du nombre de sujets pour chaque version. De fait, nous avons recueilli beaucoup plus de réponses pour la version A que pour les versions B ou C. Le tableau 10 montre que, toutes versions confondues, chaque niveau était effectivement bien représenté dans notre échantillon.

Niveau	sujets	pourcentage
MN	253	12,5%
100	658	32,6%
101	421	20,8%
102	423	21,4%
103	268	13,6%
Total	2023	100,0%

Tableau 10
Répartition des sujets par niveau

L'histogramme de la figure 3 montre que la distribution des niveaux de l'échantillon s'approche de celle qu'on trouve dans la population, du moins pour l'année durant laquelle

l'expérimentation a eu lieu. Le fait que le niveau 100 soit un peu surreprésenté n'altère en rien la qualité de l'échantillon.

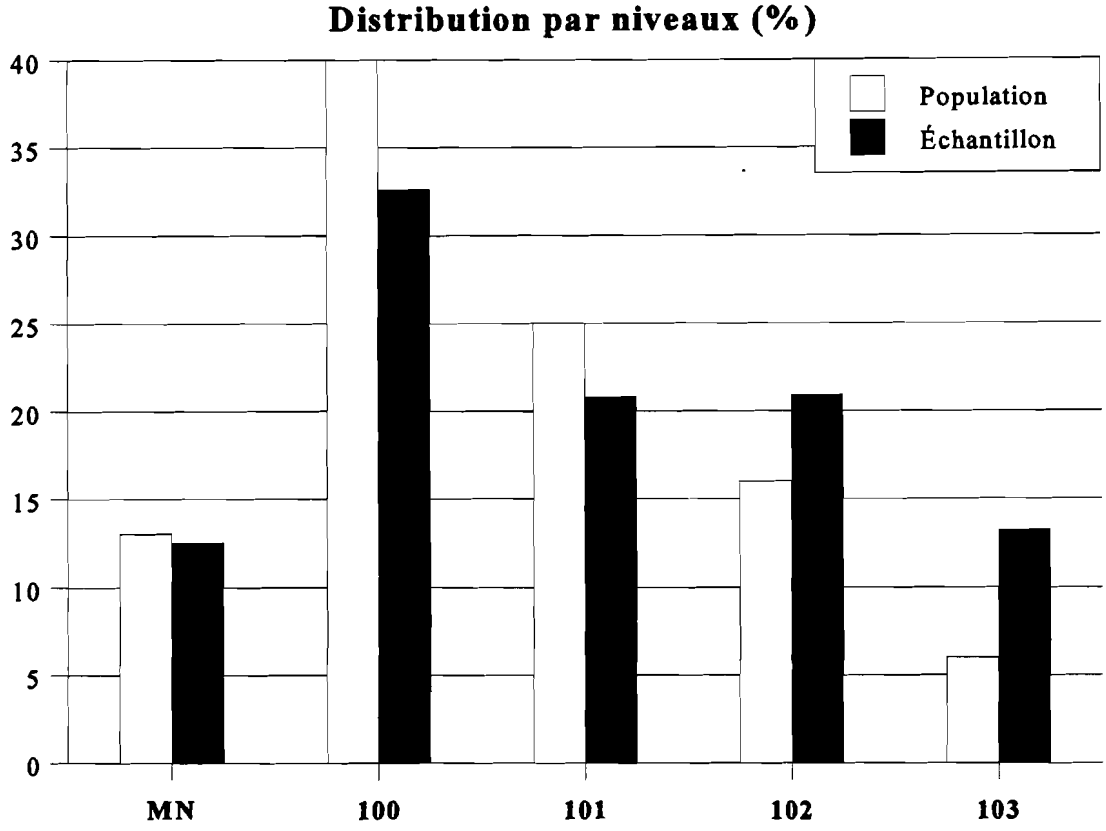

Figure 3
Distribution de l'échantillon et de la population par niveau, en pourcentage

Les conditions d'administration ont été soigneusement contrôlées de façon à ne pas fausser les résultats. Nous nous sommes assurés de récupérer tout le matériel (cassette, questionnaires feuilles de réponses) aussitôt l'administration terminée. Les enseignants ont reçu des directives claires et détaillées concernant la passation du test et ils étaient invités à noter toutes circonstances particulière sur une feuille prévue à cet effet. La collaboration a été exceptionnelle et peu de cas spéciaux nous ont été signalés.

Avant de commencer le test, les étudiants avaient à remplir un court questionnaire d'auto-évaluation dans lequel ils devaient indiquer jusqu'à quel point ils s'estimaient capables d'utiliser l'anglais dans diverses situations de communication. Ces informations devaient servir à décrire l'échantillon et à comparer les résultats du test avec des variables reliés à l'usage de la langue.

Afin de valider le test par rapport à certaines mesures qui peuvent servir de critères, nous avons recueilli des données supplémentaires relativement au niveau des étudiants. D'abord, nous avons pu récupérer les résultats au TCALS de 1336 sujets. Ces résultats se distribuent à peu près normalement ainsi que le montre la figure 4. La moyenne se situe à 61 (sur 100) alors que l'écart type est de 18,37. Ensuite, dans un collège de la région montréalaise, nous avons mené des épreuves d'évaluation de la compétence à l'oral auprès de 28 étudiants. Le test se faisait à l'aide du matériel *SPEAK®*, un test semi-direct de douze questions, mis au point par *Educational Testing Services*. Enfin, nous avons aussi analysé les productions écrites de 58 étudiants à qui nous avions demandé de produire un texte d'environ 100 mots, sur des vacances passées ou futures. Ces textes ont par la suite été corrigés par deux juges. La responsabilité de l'évaluation des productions écrites et orales a été confiée à Laurence Myles, un étudiant de doctorat, spécialiste du testing en ALS.

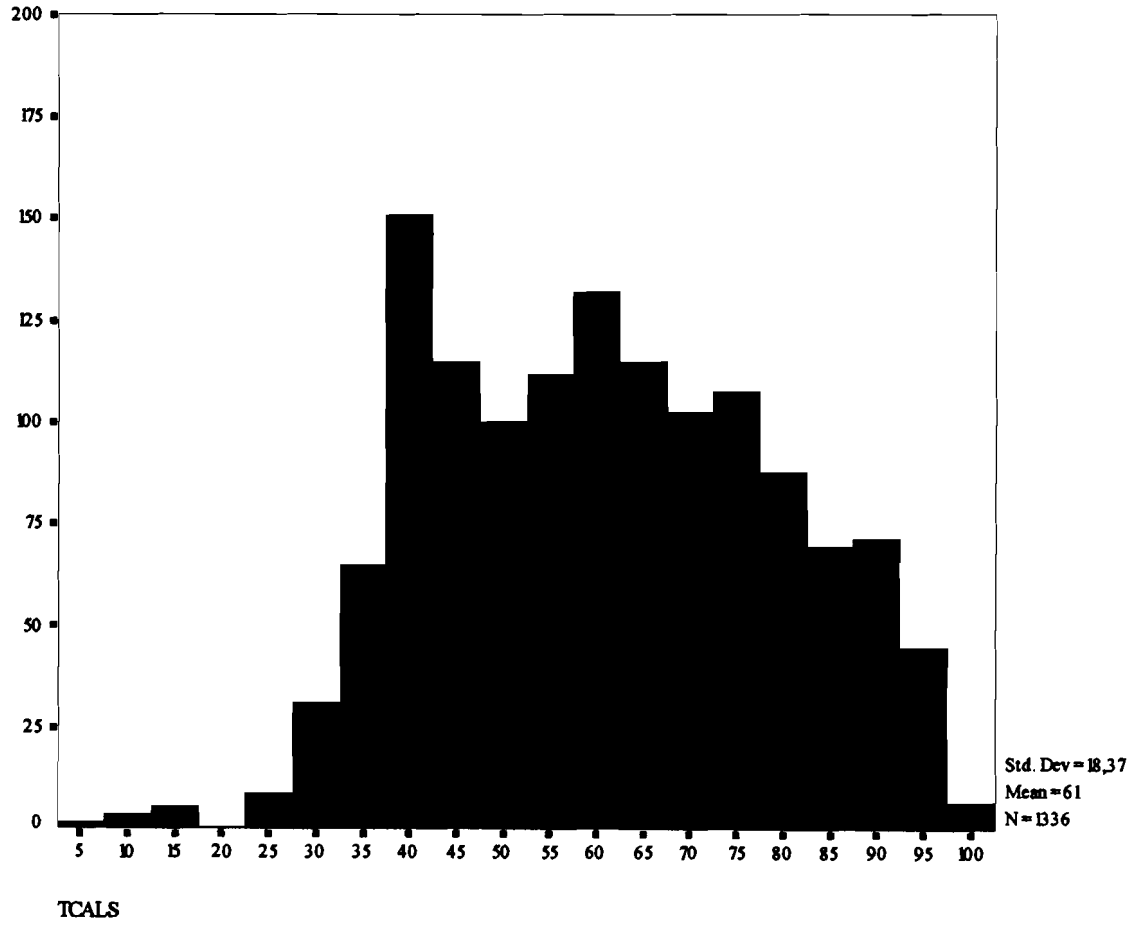

Figure 4
Distribution des résultats au TCALS

4.2 - Analyse classique des données

Avant de procéder à une analyse à l'aide de moyens plus sophistiqués, nous avons obtenu, au moyen de la procédure ITEMAN de *MicroCAT®*, les indices classiques auxquels nous avions eu recours pour le traitement des données de la pré-expérimentation.

		Version A	Version B	Version C	Total
Mise à niveau					
	Sujets	191	21	41	253
	Moyenne	33,3	30,0	35,5	33,4
	Médiane	33	28	37	33
	Écart-type	12,5	8,6	5,6	11,4
Niveau 100					
	Sujets	262	182	214	658
	Moyenne	45,5	47,9	47,5	46,8
	Médiane	46	48	45	46
	Écart-type	23,1	22,0	21,5	22,8
Niveau 101					
	Sujets	134	168	119	421
	Moyenne	62,6	64,6	62,2	63,3
	Médiane	63	63	63	63
	Écart-type	14,2	12,6	15,3	13,9
Niveau 102					
	Sujets	185	128	110	423
	Moyenne	78,3	84,7	79,6	80,5
	Médiane	80	86	81	82
	Écart-type	23,1	22,0	21,5	22,8
Niveau 103					
	Sujets	109	58	49	216
	Moyenne	75,6	101,2	87,4	85,1
	Médiane	91	102	92	95
	Écart-type	28,0	4,6	17,4	24,2
Tous les niveaux					
	Sujets	881	557	533	1971
	Moyenne	56,1	66,3	60,1	60,1
	Médiane	52	63	57	57
	Écart-type	23,1	22,0	21,5	22,8
	Alpha	0,96	0,96	0,96	

Tableau 11
Résultats par niveau et par version.

Le tableau 11 présente les résultats par niveau pour nos trois versions. En examinant la partie inférieure du tableau qui rapporte les résultats généraux, toutes versions confondues, on pourrait s'inquiéter du fait que les versions ne sont pas de difficulté égale : une moyenne de 56,1 pour la version A, 66,3 pour la version B et 60,1 pour la version C. La moyenne de la version B est donc de 10 points plus élevée que celle de la version A. Cela ne pose pas de problème particulier puisque l'opération visait à décrire chacun des items afin de reconstituer des versions qui soient effectivement comparables. Par ailleurs, avec un écart-type moyen de 22,8, on peut être assuré d'une variance assez grande pour pouvoir interpréter les résultats de façon normative. Cette variance se traduit par un coefficient général de fidélité de 0,96 qui, dans la mesure où il est interprétable, se révèle tout à fait satisfaisant

En comparant les résultats par niveau, on observe des différences importantes qui s'expliquent notamment par le fait que les niveaux sont loin d'être équivalents d'un établissement à un autre. Ainsi parmi les 109 sujets du niveau 4 qui ont reçu la version A, on retrouve 30 sujets du Cégep Saint-Laurent dont la moyenne est de 90,0, alors que les 27 sujets du Cégep d'Alma n'obtiennent qu'une moyenne de 35,5. Aussi les écarts-types sont-ils parfois assez élevés. Malgré cela, la médiane montre bien la progression dans les scores d'un niveau à l'autre.

On remarque que le tableau tient compte des résultats de seulement 1971 sujets au lieu des 2023 que nous avions à l'origine. Cela tient au fait qu'une première analyse des données a fait apparaître qu'un sous-groupe de 52 personnes, provenant de deux classes du niveau 103, avait répondu au test de manière aberrante. Leurs réponses à une bonne quantité d'items étaient à la fois identiques et erronées, sans qu'il ne s'agisse d'une erreur dans la clé de correction. En conséquence, afin de ne pas contaminer les données, ces sujets ont été écartés des analyses ultérieures, ne laissant plus qu'un groupe de 1971 sujets valides.

L'analyse à l'aide de ITEMAN nous permet d'obtenir, pour chacun des items, un indice de difficulté (p)qui correspond à probabilité d'une réponse correcte dans l'échantillon c'est-à-dire à la proportion (exprimée en décimale) du nombre de sujets qui réussissent un item. Il est peu utile dans les limites du présent rapport de donner la liste des indices de difficulté pour chaque item. La figure 5 permet cependant de visualiser la distribution des indices de facilité pour les 246 items des tests. On voit que la grande majorité des items ont un indice de difficulté qui varie entre 0,4 et 0,75. La plupart des items très difficiles ($p < 0,3$) étaient des items qui présentaient des problèmes : il pouvait, par exemple, s'agir d'items que seuls des sujets parfaitement bilingues pouvaient réussir ou encore des items dont la formulation était ambiguë.

Dans un test à interprétation normative, la discrimination d'un item est une caractéristique appréciable. Comme l'avions fait pour les données de la pré-expérimentation, nous avons pris en considération un indice de discrimination, la corrélation bisérielle (r_{bis}). Ce coefficient représente le lien entre la réponse à un item (correcte ou incorrecte) et le score total pour chaque sujet ; ce lien permet de rendre compte de la façon dont un item départage les sujets plus avancés des sujets moins avancés. Lorsque la valeur du coefficient s'approche de 1, l'item est en général réussi par les meilleurs sujets et échoué par les plus faibles. Bien que nous l'ayons calculé d'abord pour

l'ensemble du test, puis par composante et par sous-test, nous nous sommes surtout référés aux deux derniers calculs puisque les sous-tests ne portent pas tous sur les mêmes contenus. La figure 6 permet de voir la distribution des indices de discrimination calculés pour chaque sous-test. On constate que la très grande majorité des items présente des indices de discrimination qui varient ntre 0,45 et 0,8 ce qui nous assure d'une bonne discrimination. Les quelques items dont l'indice se rapproche de 0 étaient souvent des items très difficiles ou très faciles pour lesquels il est rare de trouver des indices très élevés. Une dizaine d'items dont la discrimination a été jugée insatisfaisante ont été rejetés à cette étape.

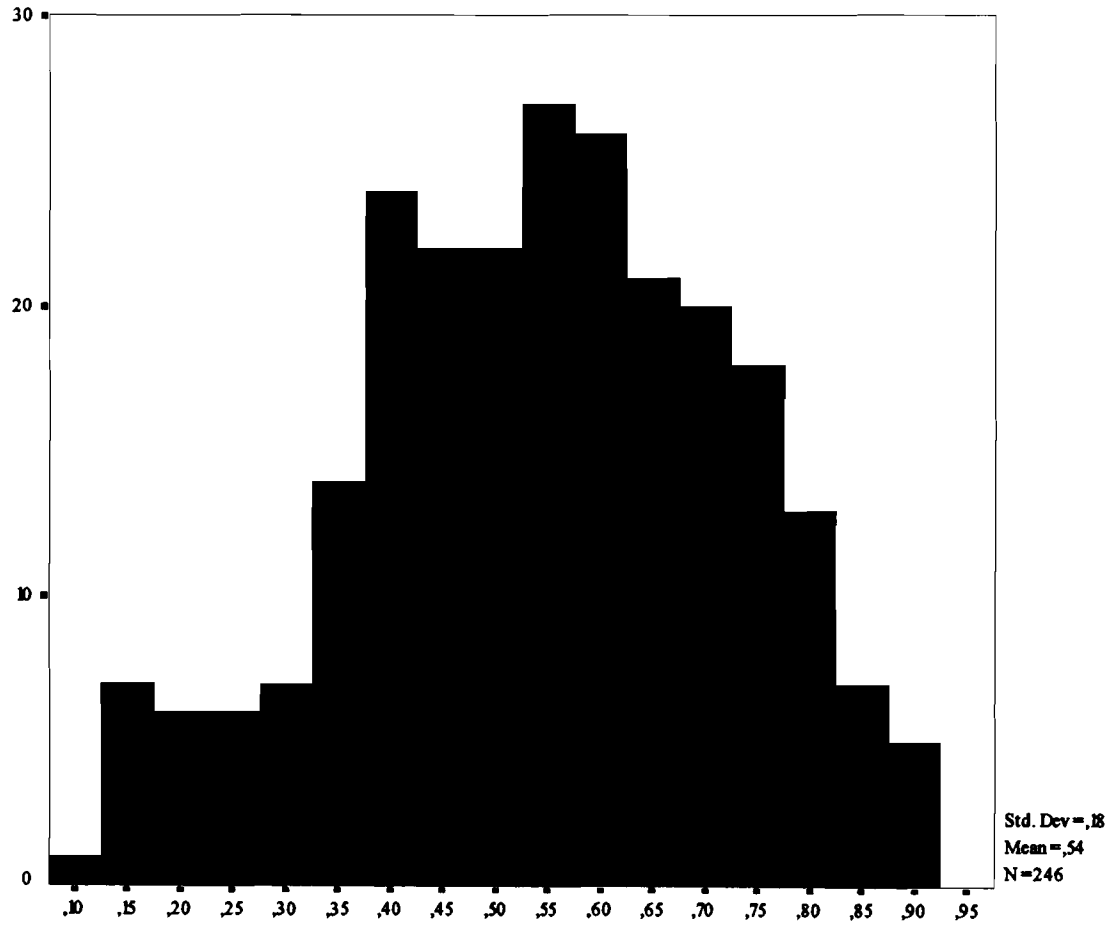

Figure 5
Distribution des indices de difficulté des items

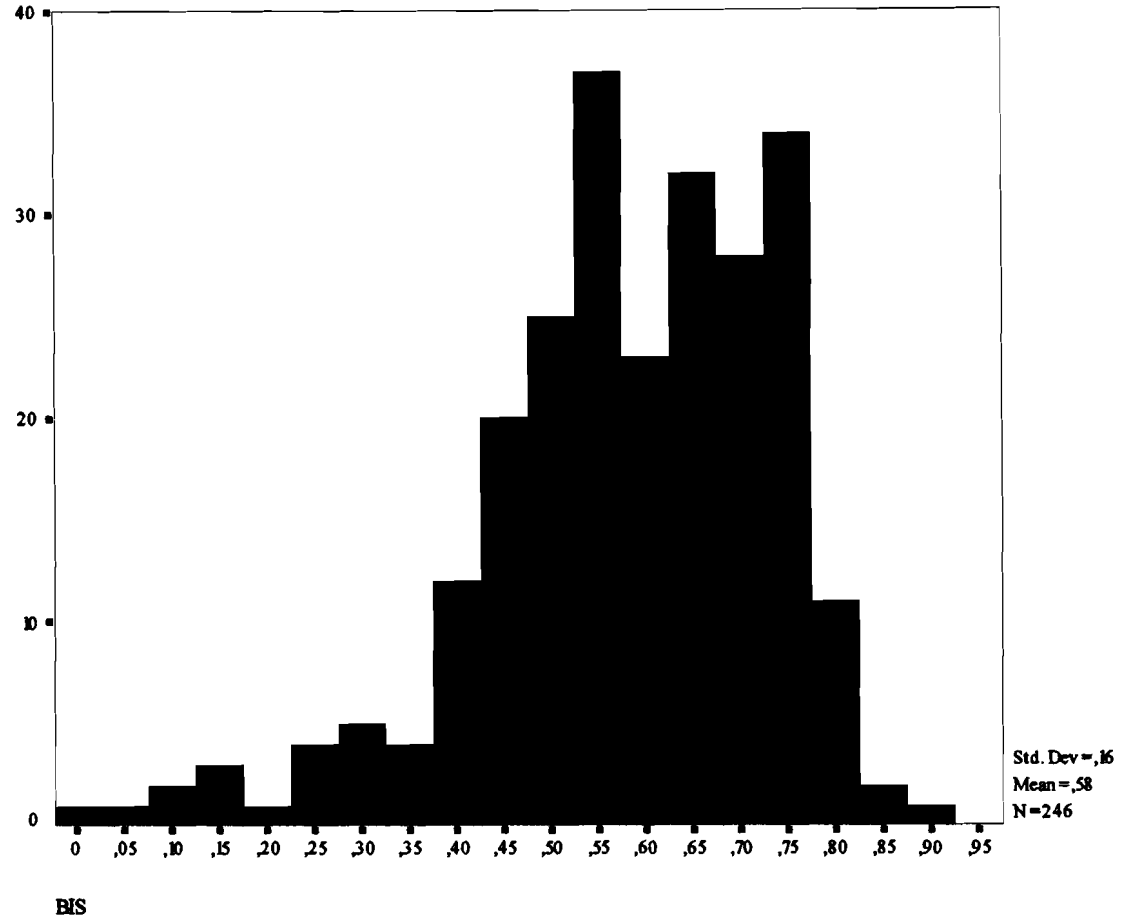

Figure 6
Distribution des coefficients bisériels des items.

4.3 - La structure interne du test

Ainsi que nous l'avions fait avec les données de la version pré-expérimentale, nous avons cherché à découvrir la structure interne du test en analysant les corrélations entre les sous-tests. Les matrices reproduites au tableau 12 montrent que les coefficients de corrélation sont encore plus élevés dans la version expérimentale : environ les deux tiers des coefficients sont supérieurs à 0,6 (malgré des coefficients alpha qui se situent sous ce seuil). Il faut y voir le fait qu'il y a souvent concordance, pour un étudiant donné, entre la proportion de réponses correctes d'un sous-test à l'autre. Cela pourrait poser problème dans l'optique d'un test diagnostique où l'on s'attendrait à ce que chaque sous-test apporte une information spécifique, mais cela est plutôt heureux pour un test de classement dont on retient généralement le score global.

Version A
(881 sujets)

	Phrases	Dial	Mini	Voc1	Voc2	Gramm	Erreurs	Lec1	Lec2
Phrases	-								
Dialogues	0,75	-							
Mini-exposés	0,63	0,68	-						
Vocabulaire 1	0,77	0,78	0,64	-					
Vocabulaire 2	0,56	0,61	0,53	0,65	-				
Grammaire	0,77	0,76	0,67	0,82	0,65	-.			
Analyse d'erreurs	0,70	0,69	0,61	0,76	0,61	0,84	-		
Lecture 1	0,54	0,52	0,46	0,57	0,45	0,59	0,55	-	
Lecture 2	0,55	0,58	0,54	0,57	0,50	0,62	0,58	0,51	-
Alpha	0,76	0,77	0,63	0,83	0,50	0,90	0,82	0,41	0,71

Version B
(557 sujets)

	Phrases	Dial	Mini	Voc1	Voc2	Gramm	Erreurs	Lec1	Lec2
Phrases	-								
Dialogues	0,81	-							
Mini-exposés	0,71	0,70	-						
Vocabulaire 1	0,80	0,81	0,69	-					
Vocabulaire 2	0,57	0,53	0,52	0,59	-				
Grammaire	0,79	0,79	0,70	0,80	0,59	-			
Analyse d'erreurs	0,68	0,67	0,63	0,70	0,51	076	-		
Lecture 1	0,60	0,58	0,54	0,60	0,50	0,64	0,58	-	
Lecture 2	0,57	0,56	0,55	0,61	0,44	0,60	0,59	0,58	-
Alpha	0,81	0,80	0,65	0,82	0,47	0,84	0,76	0,53	0,74

Version C
(533 sujets)

	Phrases	Dial	Mini	Voc1	Voc2	Gramm	Erreurs	Lec1	Lec2
Phrases	-								
Dialogues	0,74	-							
Mini-exposés	0,63	0,63	-						
Vocabulaire 1	0,73	0,78	0,64	-					
Vocabulaire 2	0,57	0,54	0,48	0,58	-				
Grammaire	0,73	0,74	0,64	0,81	0,59	-			
Analyse d'erreurs	0,60	0,62	0,55	0,70	0,50	0,76	-		
Lecture 1	0,59	0,57	0,52	0,59	0,47	0,62	0,52	-	
Lecture 2	0,60	0,62	0,51	0,67	0,51	0,68	0,63	0,59	-
Alpha	0,78	0,79	0,58	0,81	0,34	0,84	0,74	0,48	0,74

Tableau 12
Corrélations entre les sous-tests pour chacune des versions expérimentales

Il faut également souligner, dans le tableau 12, les corrélations particulièrement élevées qu'on trouve entre le sous-test «Vocabulaire 1» et les autres sous-tests, de même qu'entre «Grammaire 1» et les autres sous-tests. Cela tient d'une part au niveau de fidélité élevé ($\alpha > 0,8$) de ces sous-tests et sans doute au rôle que joue le vocabulaire et la grammaire dans la réalisation des tâches qu'on trouve dans les autres sous-tests.

Afin de mieux saisir les liens entre les éléments du test nous avons regroupé les sous-tests qui mesurent des composantes communes pour constituer quatre scores composites : compréhension auditive, vocabulaire, habiletés grammaticales et lecture. Le tableau 13 présente la matrice de corrélation ainsi obtenue. La plupart des coefficients dépassent 0,7 ce qui correspond à 50% de variance commune. De fait, on constate que les corrélations les plus basses impliquent le sous-test de lecture, ce qui indique que ce sous-test comporte des aspects particuliers (une variance spécifique) que ne mesurent pas les autres sous-tests. Il reste que, dans l'ensemble, ces résultats tendent à démontrer que le test mesure des aspects suffisamment apparentés pour permettre des jugements sur une hypothétique compétence générale en anglais langue seconde.

Version A
(881 sujets)

	Compréhension	Vocabulaire	Grammaire	Lecture
Compréhension	-			
Vocabulaire	0,84	-		
Grammaire	0,83	0,85	-	
Lecture	0,68	0,67	0,71	-

Version B
(557 sujets)

	Compréhension	Vocabulaire	Grammaire	Lecture
Compréhension	-			
Vocabulaire	0,85	-		
Grammaire	0,85	0,82	-	
Lecture	0,69	0,70	0,72	-

Version C
(533 sujets)

	Compréhension	Vocabulaire	Grammaire	Lecture
Compréhension	-			
Vocabulaire	0,82	-		
Grammaire	0,76	0,81	-	
Lecture	0,71	0,73	0,71	-

Tableau 13
Corrélations entre les composantes pour chacune des versions

Ces considérations sont reliées à une problématique qui a fait l'objet de nombreux débats chez les spécialistes de la psychométrie. Il s'agit de la question de la dimensionalité d'un test. Cette question est importante car l'interprétation d'un score unique, comme c'est le cas pour le résultat d'un test de classement, suppose que l'ensemble du test mesure un attribut unique ou tout ou moins des aspects étroitement associés. Qui plus est, beaucoup de techniques d'analyse d'instruments de mesure s'appuient sur une certaine forme d'unidimensionalité de l'épreuve. C'est pourquoi, diverses approches ont vu le jour afin de déterminer jusqu'à quel point une épreuve est unidimensionnelle[8]. Une des approches consiste à procéder à procéder à une analyse factorielle des données obtenues. Compte tenu de la particularité des données que représentent des réponses à un test composé d'items à correction dichotomique (correct/incorrect) de divers niveaux de difficulté, une analyse factorielle non linéaire est recommandée. Nous avons donc procédé à une analyse factorielle non linéaire à l'aide du logiciel *TESFACT®*.

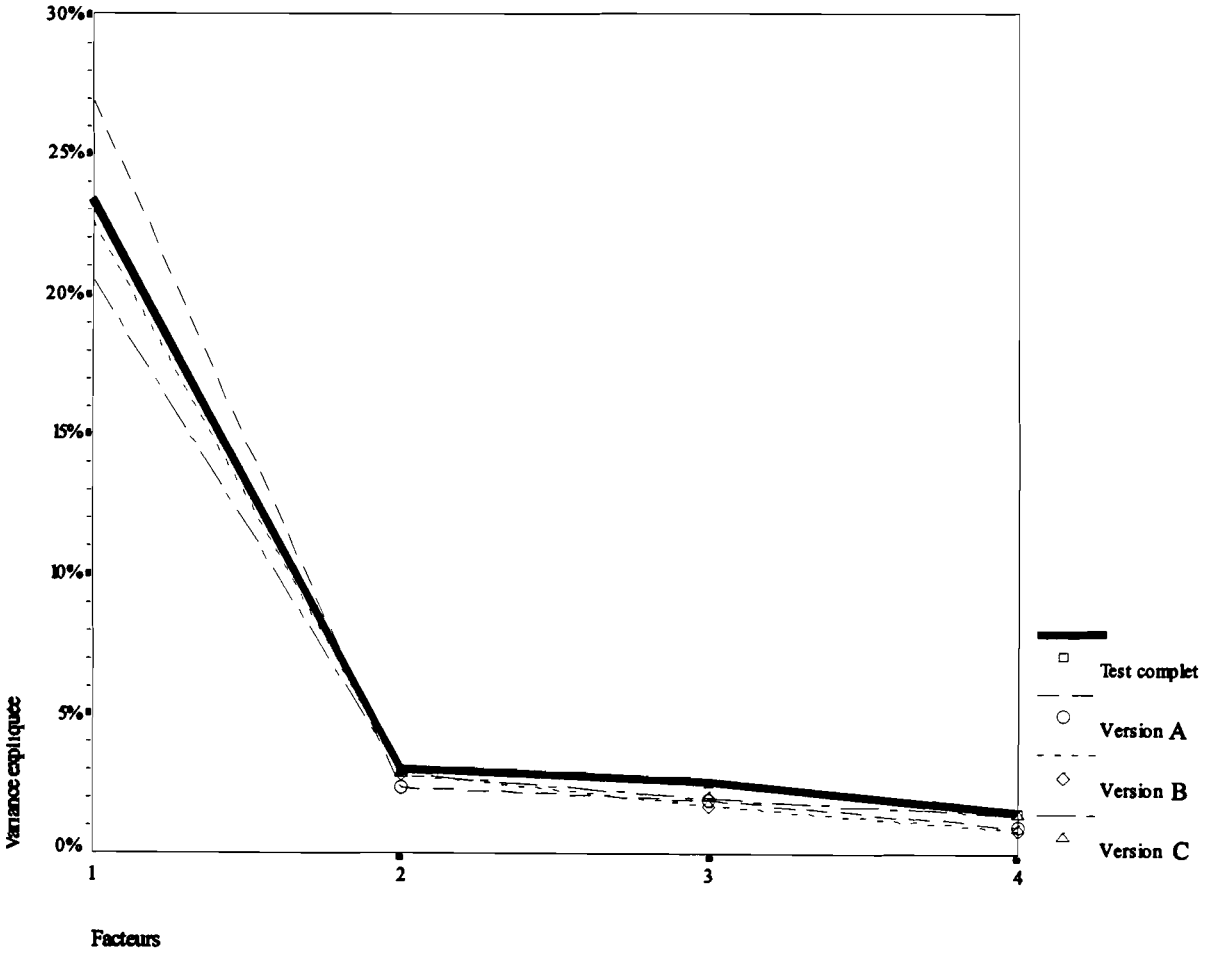

Figure 7
Contribution des quatre facteurs

Le concept d'unidimensionalité est relié au postulat d'indépendance locale. Le postulat d'indépendance locale requiert que les réponses, pour un niveau d'habileté fixé, soient statistiquement indépendantes. Autrement dit, parmi les sujets de même habileté, la probabilité de succès à un item ne doit pas être conditionnelle à la réussite d'un autre item. En apparence, la présence d'items «groupés» (les mini-exposés et la lecture de documents) contredirait ce postulat. On pourrait également penser que l'organisation du test autour de quatre composantes peut contribuer à l'émergence de quatre dimensions correspondant à chacune de ces composantes. C'est dans cette perspective que nous avons procédé à une analyse factorielle qui recherchait quatre facteurs. Au terme de cette analyse, nous avons dû conclure que les items du test n'avaient pas tendance à se regrouper autour des quatre dimensions attendues. La figure 7 montre le gain qu'on réalise en ajoutant des facteurs supplémentaires. Il ressort que le premier facteur explique environ 23% de la variance. C'est relativement peu, mais il faut admettre que l'ajout de facteurs supplémentaires ne contribue guère à hausser cette proportion. En conséquence, nous pouvons croire que malgré une certaine complexité structurelle, le test est statistiquement unidimensionnel...

4.4 - La théorie des réponses aux items

La théorie des réponses aux items (TRI) offre une alternative avantageuse à l'analyse classique des items. Il s'agit d'une modélisation probabiliste dont les procédures relativement complexes sont issues des recherches dans le domaine de la psychométrie. Pour une introduction à la théorie de réponse aux items avec des exemples d'application pour l'évaluation en langue seconde, nous suggérons les ouvrages de R. Baker[9] ou de G. Henning[10]. Dans le cadre de la TRI, on essaie d'estimer la probabilité de succès de chaque item en fonction du niveau d'habileté du répondant. Lorsque les conditions d'application sont satisfaites, ce type de modélisation jouit d'une double propriété d'invariance : d'abord, les paramètres des items sont indépendants de l'échantillon sur lequel ils sont calibrés, ensuite, les estimations d'habileté des sujets sont indépendantes des items qui ont été présentés. Cette double invariance permet notamment de déterminer, une fois les items calibrés, le niveau de sujets qui n'ont pas nécessairement répondu aux mêmes items. L'intérêt de l'approche réside aussi dans le fait qu'on peut tenir compte de certaines irrégularités dans l'échantillon et qu'on peut porter le niveau d'habileté des sujets et la difficulté des items sur une même échelle. Il faut toutefois souligner que la calibration postule l'unidimensionalité et requiert un grand nombre de sujets particulièrement avec les modèles plus complexes. La précision de la calibration peut varier selon les algorithmes de solution ; pour notre part, nous avons utilisé la procédure d'estimation par maximum de vraisemblance marginale implantée dans le progiciel *PC-BILOG3*®.

Nous avons vu que le test, dans son ensemble, était raisonnablement unidimensionnel ce qui justifierait l'application de la TRI. Par ailleurs, avec un minimum 500 réponses valides pour chaque item, on pouvait envisager l'application d'un modèle à trois paramètres bien que l'application d'un modèle plus simple à un seul paramètre fût plus sûre. La figure 8 illustre les courbes caractéristiques de cinq items différents selon un modèle à un paramètre (modèle de Rasch). On trouve en abscisse l'échelle de difficulté/habileté, le niveau le plus faible étant à gauche et le

niveau le plus fort à droite. L'échelle utilisée avec le modèle de Rasch est habituellement graduée en *logits* de sorte que la plupart des valeurs se situent entre -3 et 3. L'axe des ordonnées correspond à la probabilité d'obtenir une bonne réponse. Ainsi, à mesure que l'habileté augmente, la probabilité d'obtenir une bonne réponse augmente. La courbe la plus à gauche représente donc la modélisation d'un item très facile (b = -2) alors que la courbe la plus a droite représente un item très difficile (b = 2).

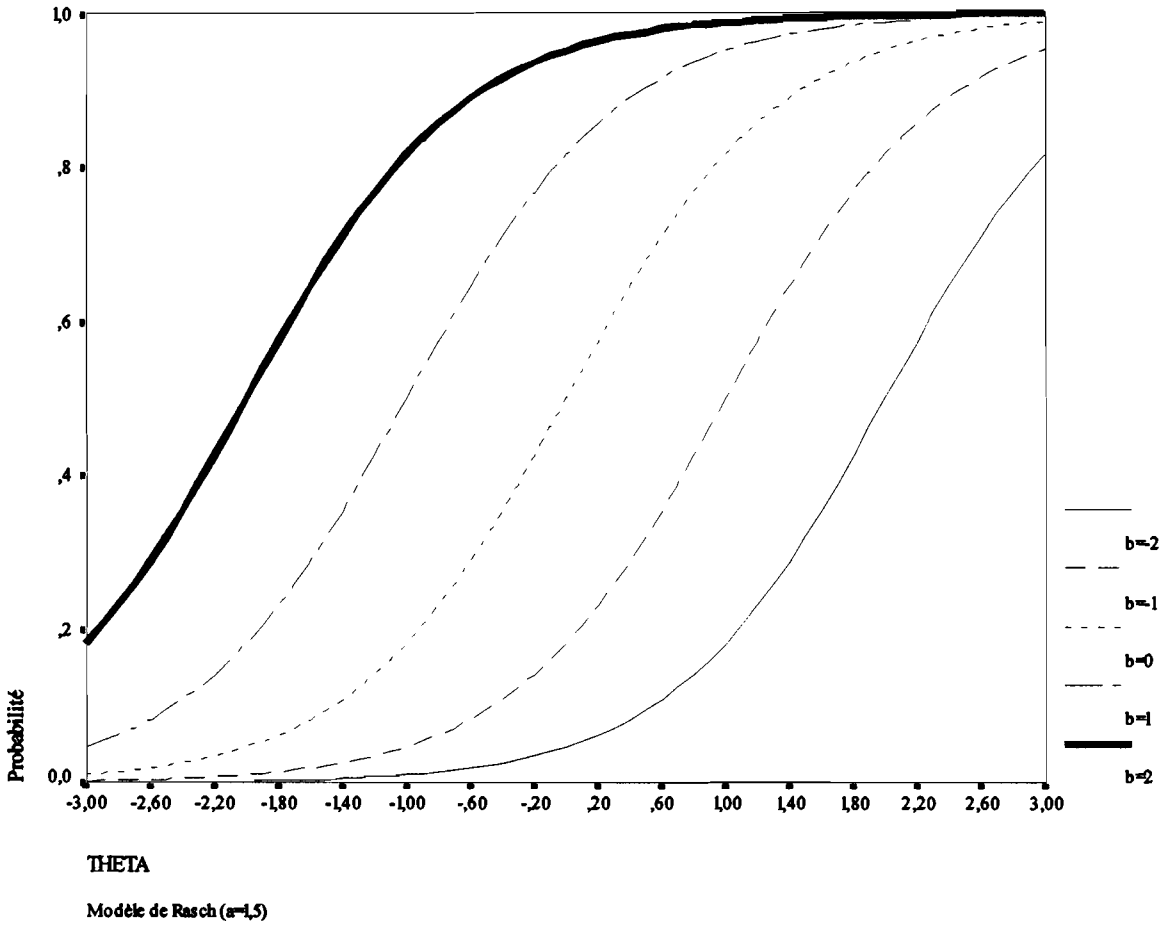

Figure 8
Courbes caractéristiques de cinq items

Le modèle à trois paramètres convient généralement mieux pour les items à choix multiple parce qu'il tient compte non seulement de la difficulté des items (le paramètre b) mais aussi de leur discrimination (le paramètre a) et de l'effet du hasard (le paramètre c). Par contre, le modèle de Rasch est plus simple, requiert moins de sujets et convient mieux quand on prévoit estimer le niveau à partir du nombre de réponses correctes. De fait, l'estimation qu'on obtient en considérant la configuration des réponses compte tenu des paramètres des items est plus précise que le simple cumul des réponses exactes. Ce cumul, le score brut, est cependant proche de l'estimation de la

TRI quand on utilise le modèle de Rasch. C'est ce qui se dégage de l'observation de la figure 9 qui compare l'estimation du niveau des sujets expérimentaux réalisée au moyen de la TRI et le score brut à chacune des trois versions. Les points qui s'écartent sensiblement du tracé de la courbe correspondent aux sujets qui ont beaucoup de réponses omises ou d'items non atteints.

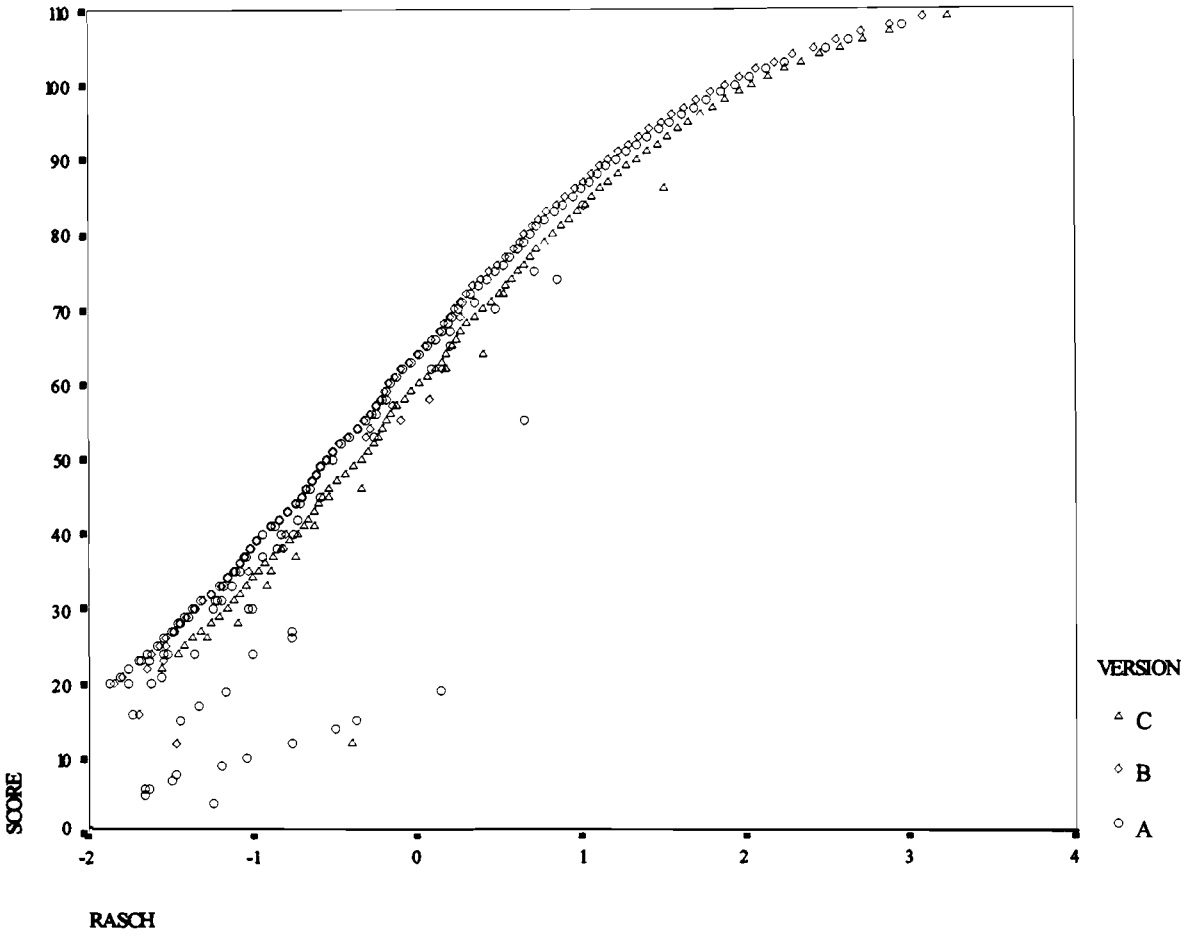

Figure 9
Estimation du niveau selon TRI par rapport au score

En ce qui a trait au choix du modèle, nous avons opté pour une stratégie mixte. Nous avons procédé à deux calibrations, une avec un modèle à un paramètre (modèle de Rasch) et l'autre avec un modèle à trois paramètres. Le modèle de Rasch, en vertu de la relation monotone qu'il entretient avec le score total d'un test, nous a fourni les scores de césure. L'histogramme de la figure 10 rend compte de la distribution des sujets en fonction leur habileté telle qu'établie par la calibration avec le modèle de Rasch (sur une échelle de *logits*). Comme nous le verrons, nous

avons déterminé les scores de césure en faisant correspondre les scores avec les points de transition d'un niveau à l'autre sur l'échelle de Rasch.

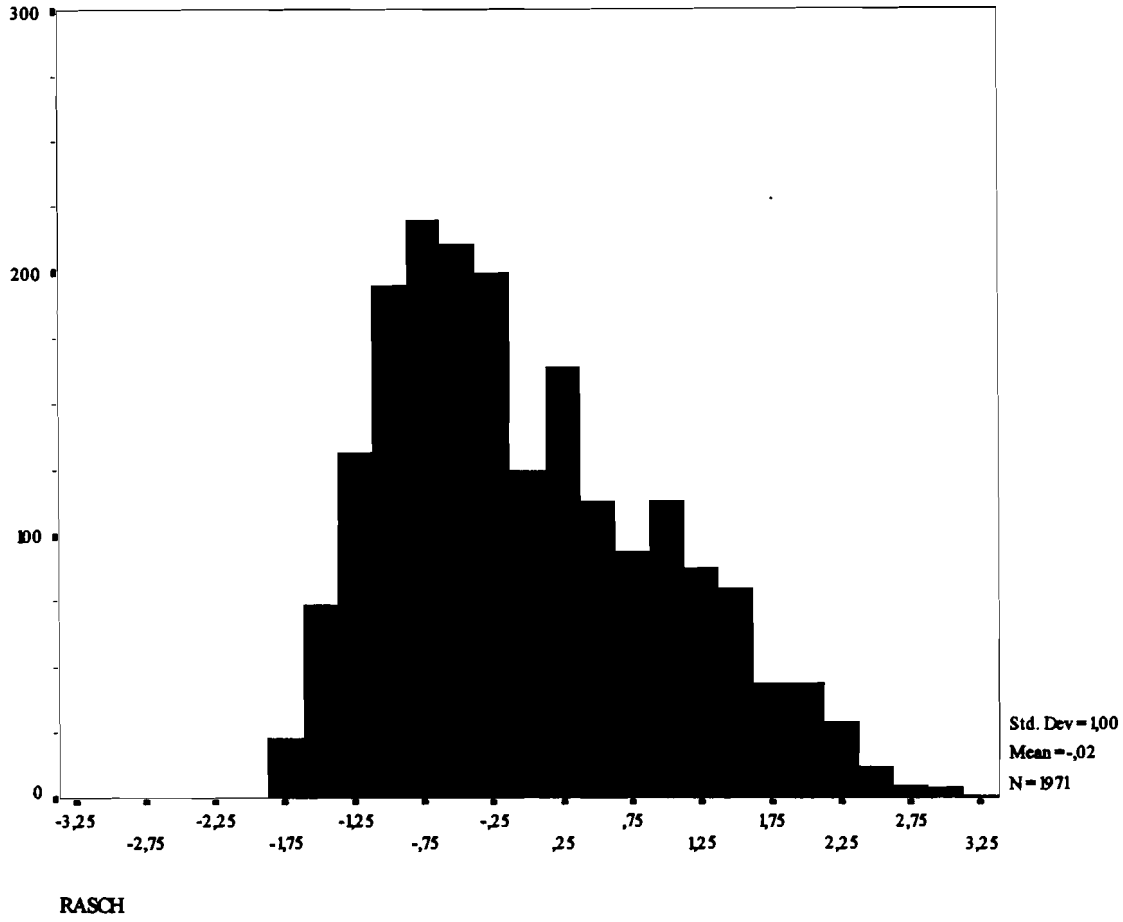

Figure 10
Distribution de l'habileté selon le modèle de Rasch

En ce qui a trait au choix du modèle, nous avons opté pour une stratégie mixte. Nous avons procédé à deux calibrations, une avec un modèle à un paramètre (modèle de Rasch) et l'autre avec un modèle à trois paramètres. Le modèle de Rasch, en vertu de la relation monotone qu'il entretient avec le score total d'un test, nous a fourni les scores de césure. L'histogramme de la figure 10 rend compte de la distribution des sujets en fonction leur habileté telle qu'établie par la calibration avec le modèle de Rasch (sur une échelle de *logits*). Comme nous le verrons, nous avons déterminé les scores de césure en faisant correspondre les scores avec les points de transition d'un niveau à l'autre sur l'échelle de Rasch.

Une autre façon de s'assurer de la justesse d'un modèle est de vérifier ses prédictions. L'une de ces prédictions est que les paramètres obtenus sont indépendants de l'échantillon utilisé lors de

la calibration. Nous avons donc repris la calibration de nos items sur deux sous-échantillons distincts, d'abord avec les sujets provenant de la région de Montréal, puis avec tous les autres. La figure 11 montre que les paramètres de difficulté du modèle de Rasch (ceux qui ont servi à établir les scores de césure) sont stables d'une calibration à l'autre puisqu'ils s'alignent le long de la droite oblique. On peut donc penser que le modèle est adéquat pour l'utilisation que nous en ferons par la suite. Par ailleurs, le fait que les items s'écartent peu de la droite est aussi une indication que les items ne favorisent pas les étudiants de la région montréalaise au détriment des étudiants d'ailleurs (ou vice-versa).

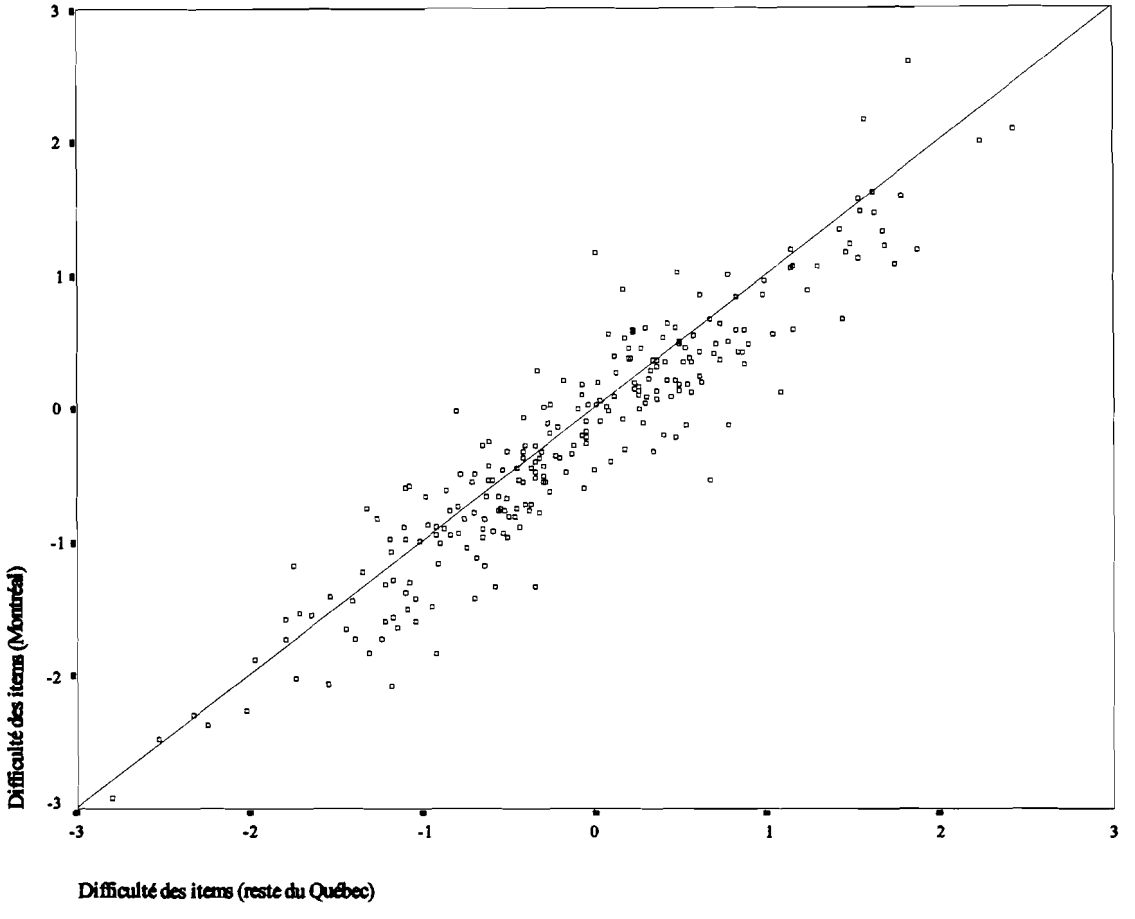

Figure 11
Comparaison entre la calibration pour Montréal et la calibration pour le reste du Québec

Le modèle à trois paramètres a été utilisé dans la sélection des items, car il permet de construire, à partir des valeurs des trois paramètres, des courbes d'information plus précises. Il faut souligner que la notion d'information joue, dans la TRI, un rôle similaire à celui joué par la fidélité dans la théorie classique. La quantité d'information recueillie nous renseigne sur la

précision des estimations de l'habileté d'un sujet compte tenu des items qui lui ont été soumis. L'avantage par rapport aux indices de la théorie classique et qu'on peut voir comment la précision de l'estimation varie en fonction du niveau d'habileté des sujets. Un item trop facile est peu informatif pour des sujets très habiles : presque tous vont le réussir. De même, un item de grande difficulté nous apportera peu d'information pour des sujets plus faibles : les réponses y seront incorrectes dans la plupart des cas. L'information varie selon la difficulté et la discrimination d'un item ou, dit autrement, elle est inversement proportionnelle à l'erreur commise sur l'estimation de l'habileté des sujets. L'information ainsi calculée a la particularité d'être une fonction additive. On peut donc cumuler l'information apportée par chacun des items et tracer des courbes d'information pour un test complet. Plus l'information recueillie à un niveau donné est grande, plus précise sera l'estimation pour les sujets qui se retrouvent à ce niveau.

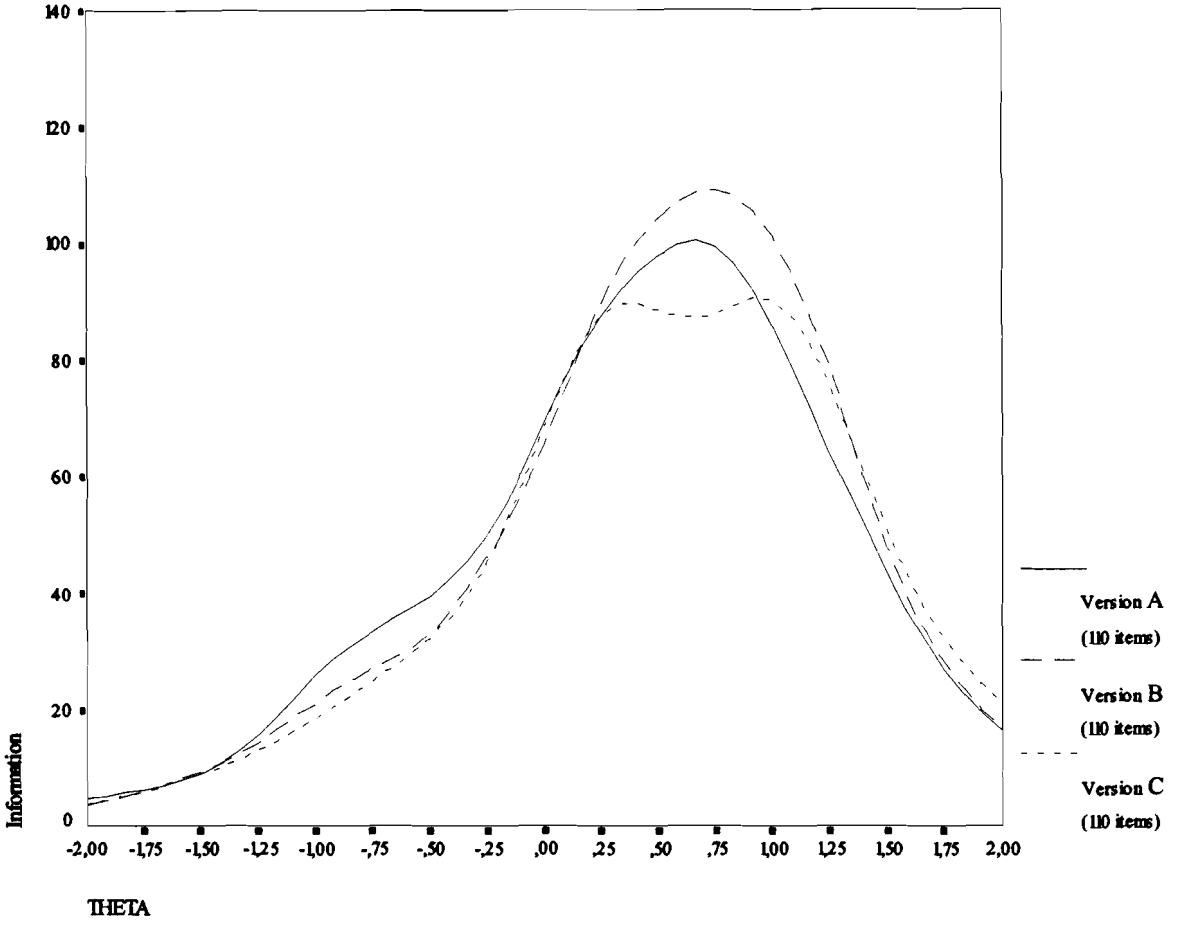

Figure 12
Courbes d'information des trois versions expérimentales

La figure 12 présente les courbes d'information de nos trois versions. Il faut d'abord noter que le modèle à trois paramètres utilise une échelle différente de celle du modèle de Rasch. Au

lieu de l'échelle en *logits*, on trouve plutôt une échelle dont les points correspondent à ceux d'une courbe normale. L'étalonnement des sujets est différent mais, avec les deux échelles, le point 0 correspond à un niveau moyen. On constate que les versions A, B et C sont toutes les trois particulièrement efficaces dans l'intervalle d'habileté 0,25 ~ 1,25, ce qui correspond plus ou moins au niveaux «Intermédiaire fort» et «Avancé». L'information décroît rapidement vers le niveau «Très avancé» à l'extrémité droite de la courbe, ce qui, du point de vue du classement ne pose pas de problème majeur puisque les sujets qui s'y trouvent seront classés dans le cours le plus avancé. Inversement, la faible quantité d'information recueillie entre -2,0 et -1,25 ne doit pas inquiéter outre mesure puisque que les sujets ne peuvent être placés plus bas que le premier niveau. La faible quantité d'information recueillie entre -1,25 et -0,25 («Faux débutant» et «Intermédiaire faible») est plus préoccupante et on peut penser que tous les items qui ciblent cet intervalle seront nécessaires dans la construction des versions finales. Le peu d'information qu'on obtient à ces niveaux avec les versions expérimentales reflète la difficulté générale du test, difficulté que n'ont pas manqué de nous signaler les enseignants et les étudiants. Par ailleurs, on peut voir que, contrairement à ce que pouvaient laisser croire les moyennes, les trois versions sont à peu près de difficulté égale puisque leurs tracés sont similaires. La version B s'avère un plus précise alors que la version C, avec son sommet tronqué, apporte moins d'information autour du point 0,5.

4.5 - La validation

La validation d'une mesure en éducation est un processus à long terme qui n'est jamais tout à fait complété. Avec l'accumulation des données relativement à une épreuve, on peut juger avec toujours plus de justesse de ce que mesure effectivement un instrument et de la portée de son utilisation.

	Test (Rasch) N=1971	Oral N=28	Ecrit N=58	TCALS N=1336
Test	-			
Oral	0,89	-		
Écrit	0,68	0,71	-	
TCALS	0,85	0,86	0,67	-
Fidélité	0,96	0,95	0,72	n.d.

Tableau 14
Corrélations entre les différentes mesures

Dans un premier temps, nous avons comparé les résultats de notre test de classement avec ceux qu'avaient obtenu les 1336 étudiants dont on avait récupéré les résultats au test TCALS. Pour les fins de cette comparaison, nous avons utilisé les estimations de niveaux basées sur la

configuration des réponses à partir de la calibration selon le modèle de Rasch. Comme nous l'avons vu, ces estimations se rapprochent de celles que fournit le cumul des bonnes réponses ; elles permettent cependant de contrôler le fait que les trois versions contiennent des items différents et que quelques étudiants n'ont pas fait tout le test. La corrélation entre notre test de classement et le TCALS fournit une indication sur la validité concurrente. Le coefficient de 0,86 qui apparaît dans la matrice du tableau 14 est une indication que les deux tests mesurent sensiblement la même compétence mais que chacun a ses caractéristiques propres. C'est d'ailleurs ces caractéristiques qui justifient le remplacement d'un instrument que d'aucuns jugent désuet et inadéquat. Par contre, le lien entre les deux instruments nous autorise à nous référer aux résultats du TCALS pour déterminer les scores de césure (*cf* chapitre suivant).

Intéressés à voir la relation entre une mesure indirecte et axée sur des habiletés réceptives, telle que nous la proposons, et une évaluation basée sur la production, nous avons cherché à comparer les résultats du test avec les résultats d'évaluation de la production orale et de la production écrite. Pour l'oral, nous avions construit un sous-échantillon de 28 sujets de tous les niveaux à qui nous avions administré le test *SPEAK®*. Les réponses aux 12 questions du test étaient enregistrées pour être évaluées par deux juges. La moyenne des scores du premier juge s'établissait à 40,43 (avec un écart-type de 7,75) et celle du deuxième à 39,22 (avec un écart-type de 7,3). La corrélation inter-juge peut être interprétée comme un indice de la fidélité ; le coefficient de 0,95 indique qu'il s'agit d'une mesure remarquablement fidèle compte tenu des problèmes de fidélité qui affectent souvent l'évaluation de l'oral. Comme le montre le tableau 14, la corrélation avec le test est très élevée (r = 0,89), ce qui indique que le test permet de faire certaines inférences sur le niveau à l'oral. Une telle corrélation signifie que, *grosso modo*, les résultats à l'oral recoupent dans la moitié des cas les résultats du test de classement. En ce qui a trait à l'écrit, nous avions construit un échantillon de 58 sujets de tous les niveaux à qui nous avons demandé de rédiger un texte narratif d'environ une centaine de mots. Les textes ont été évalués de façon holistique par deux juges. La moyenne des scores du premier juge se situait à 35,38 (avec un écart-type de 8,3) tandis que celle du second juge, nettement plus basse, se situait à 32,46 (avec un écart-type de 7,91). Le coefficient de corrélation inter-juge de 0,72 indique une fidélité beaucoup moins grande pour l'écrit que pour l'oral bien qu'elle reste d'un niveau acceptable pour des fins de validation. Cela explique pourquoi le coefficient de corrélation ne dépasse pas 0,68. Compte tenu de la marge d'erreur, on peut penser qu'il existe un lien entre les résultats à l'écrit et le résultat du test.

Le tableau 15 permet de mettre ces corrélations en perspective puisqu'on y trouve les coefficients obtenus pour chacun des sous-tests. Il faut cependant être prudent dans l'interprétation, car l'erreur de mesure varie d'une épreuve à l'autre et d'un sous-test à l'autre. La première ligne de tableau présente la corrélation de chaque sous-test avec l'ensemble du test. Les sous-tests «Phrases», «Dialogue», «Vocabulaire 1» et «Grammaire» montrent les corrélations les plus élevées de sorte qu'on peut conclure que les aspects qui y sont mesurés contribuent grandement au score général. Ce sont ces aspects qui sont également en lien le plus étroit avec l'épreuve orale et il est clair que les aspects lexicaux et grammaticaux jouent un rôle important à l'oral. Les corrélations avec l'épreuve écrite sont plus faibles qu'avec l'oral de sorte qu'on peut mettre en doute la

pertinence de compléter le test de classement par une composition. Enfin, la dernière ligne indique que le TCALS mesure des aspects que mesure notre test même si sa fidélité semble plus faible.

	Phrases	**Dialogues**	**Mini-exposés**	**Vocab. 1**	**Vocab. 2**	**Gram-maire**	**Analyse d'erreurs**	**Lecture 1**	**Lecture 2**
Test (n=1971)	0,85	0,87	0,77	0,89	0,70	0,90	0,81	0,69	0,73
Oral (n=28)	0,77	0,77	0,64	0,84	0,55	0,82	0,57	0,63	0,43
Écrit (n=58)	0,64	0,54	0,58	0,65	0,39	0,61	0,54	0,39	0,36
TCALS (n=1336)	0,77	0,79	0,63	0,77	0,56	0,77	0,69	0,57	0,59

Tableau 15
Corrélation entre les sous-tests et les autres mesures

Toujours dans l'optique de comparer le résultat du test avec d'autres informations relatives au niveau des étudiants, nous avons cherché à voir le lien entre le résultat du test et la réponse à onze questions auxquelles devaient répondre les étudiants avant de commencer le test. Ces questions visaient à obtenir des informations susceptibles de prédire le niveau de l'étudiant. Le libellé de chaque question se trouve dans le tableau 16. On y trouve aussi le coefficient de corrélation que nous avons calculé entre les réponses à chaque question et le score au test. Il apparaît clairement que les questions qui concernent les occasions d'utiliser l'anglais en dehors de la classe (de A5 à A10) ne permettent guère de prédire le niveau en anglais. Les corrélations avec les notes obtenues en Secondaire V (A3 et A4) sont plus élevées (le signe négatif n'est que le simple effet de la direction de l'échelle) ; la corrélation nous apparaît cependant trop faible pour qu'on puisse raisonnablement classer les étudiants à partir de cette seule information. Par contre, les questions qui amènent l'étudiant à s'auto-évaluer en anglais (A1, A2 et A11), c'est-à-dire à porter un jugement sur son propre niveau, s'avèrent beaucoup plus proches du score au test. Cela est particulièrement vrai de la dernière question où l'étudiant doit se classer lui-même dans l'un des cours offerts ($r = 0,73$). On peut dire que, dans une situation où les étudiants n'auraient pas de raison de se sous-évaluer ou de se sur-évaluer, l'auto-classement conduirait à assigner le groupe approprié à la majorité des étudiants.

	Corrélation avec habileté	
A1. Quand je parle en anglais :	0,64	(n=1891)
a. je ne peux pas m'exprimer.		
b. j'arrive à dire des bouts de phrases.		
c. je peux parler en phrases simples.		
d. j'arrive assez bien à m'exprimer.		
e. je parle couramment.		
A2. Quand j'entends de l'anglais :	0,65	(n=1888)
a. je ne comprends pas beaucoup.		
b. je comprends quelques bouts de phrase.		
c. je comprends le sens général.		
d. je comprends presque tout.		
e. je n'ai aucun problème à comprendre.		
A3. À quel rang cinquième vous classiez-vous le plus souvent en anglais, en Secondaire V ?	-0,57	(n=1780)
a. premier rang cinquième.		
b. second rang cinquième.		
c. troisième rang cinquième.		
d. quatrième rang cinquième.		
e. cinquième rang cinquième.		
A4. Quelle était votre note moyenne en anglais en secondaire V ?	-0,61	(n=1890)
a. 90% à 100%		
b. 80% à 90%		
c. 70% à 80%		
d. 60% à 70%		
e. au-dessous de 60%.		
A5. Parlez-vous anglais à la maison ?	0,30	(n=1885)
a. jamais.		
b. presque jamais.		
c. parfois.		
d. souvent.		
e. toujours.		
A6. Parlez-vous une autre langue que le français à la maison ?	0,23	(n=1885)
a. jamais.		
b. presque jamais.		
c. parfois.		
d. souvent.		
e. toujours.		

A7. Parlez-vous anglais au travail ? 0,27 (n=1872)
- a. jamais.
- b. presque jamais.
- c. parfois.
- d. souvent.
- e. toujours.

A8. Parlez-vous anglais avec des amis ? 0,26 (n=1872)
- a. jamais.
- b. presque jamais.
- c. parfois.
- d. souvent.
- e. toujours.

A9. Regardez-vous la télévision en anglais ? 0,50 (n=1870)
- a. jamais.
- b. presque jamais.
- c. parfois.
- d. souvent.
- e. toujours.

A.10 Lisez-vous en anglais ? 0,44 (n=1865)
- a. jamais.
- b. presque jamais.
- c. parfois.
- d. souvent.
- e. toujours.

A.11 À quel niveau vous classeriez-vous ? 0,73 (n=1779)
- a. mise à niveau (cours d'appoint)
- b. 104, transitoire (pour débutants)
- c. 101, 1er niveau collégial (intermédiaire)
- d. 102, 2e niveau collégial (avancé)
- e. 103, 3e niveau collégial (étudiants bilingues)

Tableau 16

Corrélations entre les informations relatives au niveau et le test

5 - La composition des versions finales

5.1 - Version commune et versions adaptées

Un des problèmes qui se présente souvent dans l'administration d'un test de classement est lié aux différences importantes de niveaux entre les candidats. En effet, étant donné que débutants et avancés sont soumis aux mêmes tâches, plusieurs de ces tâches n'apportent que peu d'information sur le niveau de l'apprenant tant la réponse est prévisible. Il y a par exemple peu d'information à tirer de l'administration d'un item facile à un candidat avancé tout comme il y en a peu à tirer de l'administration d'un item difficile à un débutant. De plus, dans ce dernier cas, l'étudiant risque d'éprouver une certaine frustration voire se décourager. Ainsi autant pour des raisons métrologiques que psychologiques, il est avantageux de chercher à cibler davantage le test. Cela est possible lorsqu'on dispose d'une information préalable sur le niveau de l'étudiant.

Dans notre cas, certaines des questions qu'on posait avant le test proprement dit permettent un classement préliminaire, notamment les questions qui font appel à l'auto-évaluation. De fait, une analyse discriminante que nous avons menée en complément à l'examen des corrélations a révélé qu'il suffit de se baser sur la variable la plus significative, c'est-à-dire de poser la dernière question (A11). En ajoutant à cette variable, la deuxième la plus significative, soit la question sur les résultats au secondaire (A4), on n'améliore que faiblement la prédiction en faisant passer la corrélation à 0,77. En retenant les huit variables significatives, on ne peut dépasser 0,8.

Dans cette perspective, nous proposons trois versions du test : une version commune s'adressant à tous les candidats quel que soit leur niveau et deux versions adaptées, une facile et une difficile, la première conçue pour les plus débutants et la seconde pour les plus avancés. Le tableau 17 décrit la répartition des types de tâche dans les trois versions. Les versions adaptées comportent moins d'items ce qui permet de réduire le temps d'administration : 65 items plutôt que 85. On suggère une heure et demie pour la version commune et une heure pour les versions adaptées. Ainsi, la lourdeur que suppose l'utilisation de deux versions est compensée par la rapidité de l'administration et la réaction plus favorable que suscite une version convenant mieux au niveau de l'élève. Il suffit de poser une seule question («À quel niveau vous classeriez-vous ?») pour départager les candidats et les orienter vers l'une au l'autre des versions adaptées. On suggère de soumettre la version facile aux étudiants qui se situent au cours de mise à niveau ou au cours 104 et la version plus difficile aux autres. Le fait de soumettre la version qui convient le moins (par exemple, quand le candidat évalue mal son niveau) ne pose pas de problème majeur si ce n'est, comme nous le verrons par la suite, que l'erreur de mesure augmente. Le cas le plus plausible soit l'administration de la version facile à un étudiant qui serait plutôt de niveau 101, ne

représente pas une situation où il serait hasardeux d'inférer le niveau de l'étudiant à partir de son résultat.

Versions adaptées		**Version commune**	
Sous-test	nombre d'items	Sous-test	nombre d'items
Compréhension auditive		Compréhension auditive	
Phrases	10	Phrases	12
Dialogues	11	Dialogues	13
Mini-exposés	4	Mini-exposés	8
	38,5%		38,8%
Anglais écrit		Anglais écrit	
Vocabulaire 1	10	Vocabulaire 1	13
Grammaire	14	Grammaire	17
Analyse d'erreurs	6	Analyse d'erreurs	7
Lecture 1	6	Lecture 1	7
Lecture 2	4	Lecture 2	8
	61,5%		61,2%
Total	65	Total	85

Tableau 17
Nombre d'items par sous-test pour les versions finales

L'utilisation de versions adaptées s'inspire des techniques de testing adaptatif où c'est l'ordinateur qui choisit les items à soumettre au candidat à partir d'une estimation du niveau qui est revue après chaque réponse. La sélection des items les plus appropriées de même que l'estimation du niveau à partir d'items qui peuvent varier d'un étudiant à l'autre sont rendues possibles grâce à l'application des principes de la TRI, en misant sur la propriété d'invariance des items. À partir des analyses précédentes, nous avons pu constituer de nouvelles versions du test comprenant moins d'items et mieux ciblées par rapport à l'habileté. Les résultats restent comparables d'une version à l'autre puisque nous avons établi des échelles de correspondance et que la proportion attribuée aux différentes tâches varie peu.

Les courbes d'information de ces nouvelles versions sont présentées dans la figure 13. On peut voir que la version commune présente une courbe plus aplatie puisqu'elle doit apporter de l'information à tous les niveaux. La version facile apporte autant d'information que la version commune pour les étudiants de niveau inférieur à la moyenne ($\theta < 0$) ; par contre quand le niveau s'écarte de la moyenne vers les niveaux plus avancés, l'information décroît rapidement. Cette version présente donc des tâches plus réalisables pour les étudiants débutants, mais il est préférable de ne pas l'utiliser avec des étudiants avancés. La version avancée s'avère remarquablement efficace auprès des étudiants dont le niveau est supérieur à la moyenne. Cela tient au fait

que nous disposions, lors de la composition des versions, d'items difficiles dont la discrimination était très élevée. Bien qu'elle comporte moins d'items que la version commune, on constate que la version difficile apporte plus d'information dans la zone d'habileté visée. Il faut toutefois souligner que pour les étudiants dont le niveau est très inférieur à la moyenne, l'information n'est pas suffisante pour permettre des inférences justes.

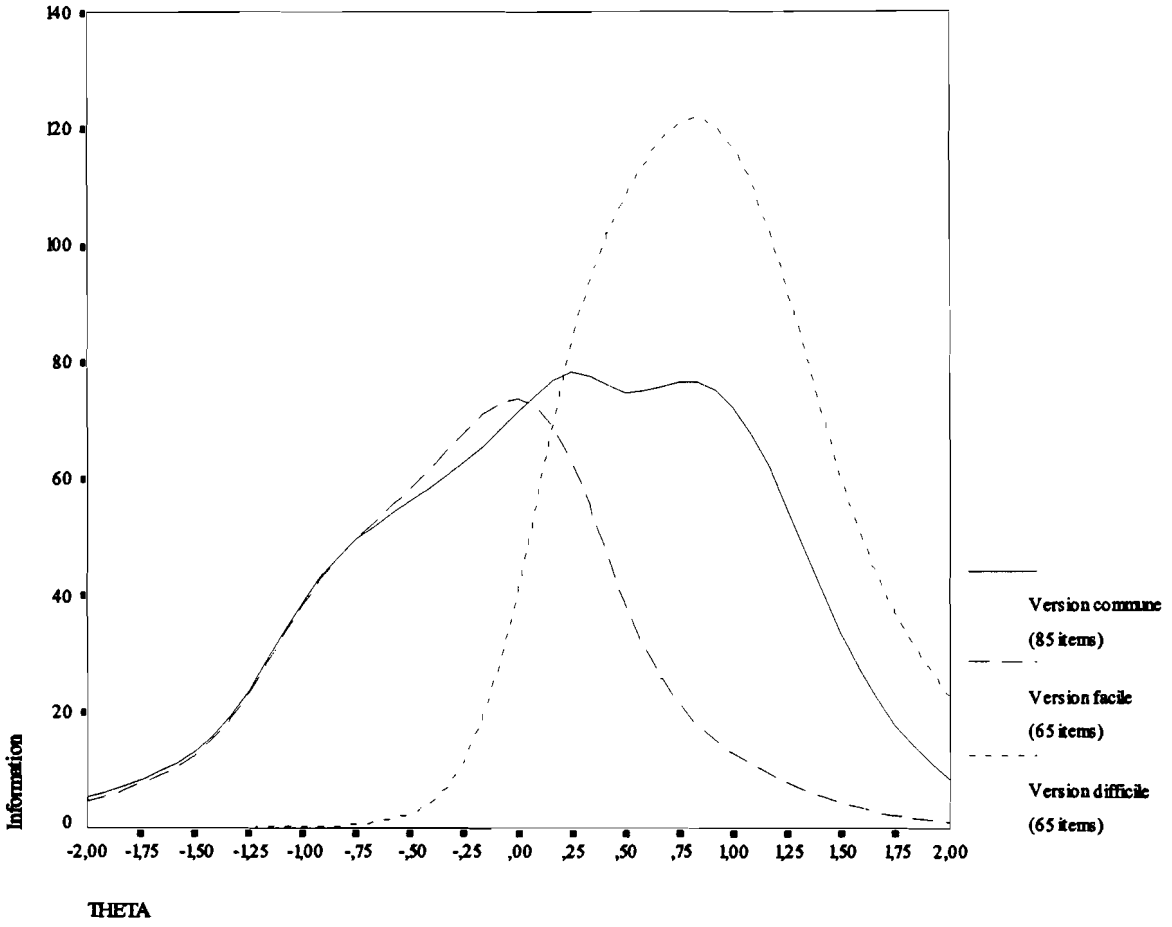

Figure 13
Courbes d'information des trois versions finales

L'aire sous la courbe d'information représente l'information totale du test. En divisant cette information totale par celle apportée par la version expérimentale A, qui contenait 110 items, on obtient un indice de l'efficacité relative. En examinant le tableau 18, on voit ainsi que la version difficile et la version commune, malgré le nombre moins élevé d'items (respectivement 65 et 85 items), apportent presque qu'autant d'information que la version expérimentale. La version facile (65 items) a une efficacité relative plus réduite mais concentre l'information vers les étudiants débutants ou intermédiaires.

L'indice de fidélité est égal à la variance de la population divisé par la somme de la variance de la population et par le carré de l'erreur moyenne. Contrairement au coefficient alpha qui représente une borne inférieure à la fidélité du test, cet indice est une estimation de sa valeur véritable. On voit que cet indice se compare au coefficient alpha que nous avions calculé pour les versions expérimentales.

	Information totale	Efficacité relative	Indice de fidélité
Version facile	115,0	0,63	0,96
Version difficile	166,0	0,91	0,96
Version commune	175,8	0,96	0,97

Tableau 18
Efficacité des trois versions finales

5.2 - Les scores de césure

Puisque le souhait de la majorité des répondants au sondage était d'en arriver à un instrument qui uniformiserait les niveaux à travers la province, nous avons cherché à établir des scores de césure. S'est alors posé la question du critère pour déterminer où commence et où finit chaque niveau. Au départ, nous avons dû postuler, que la majorité des étudiants, malgré les écarts entre les établissements et les difficultés de classement, recevait un enseignement approprié. Ne pas admettre ce postulat aurait signifié que nous mettions sérieusement en doute l'efficacité de l'ensemble du programme d'enseignement de l'anglais au collégial. D'autre part, nous nous étions prémunis contre des écarts trop importants dans l'interprétation des niveaux en limitant l'échantillon à des étudiants qui avait été classés avec le TCALS, l'outil de classement le plus répandu, aussi imparfait soit-il. Par la suite, nous avons regroupé les étudiants en fonction du niveau déclaré et produit des diagrammes en boîte de façon à visualiser la distribution des étudiants par niveau en fonction de l'estimation selon le modèle de Rasch. Cette estimation, rappelons-le, est comparable à celle que fournit le cumul des bonnes réponses, mais elle permet de tenir compte du fait que le score provenait de trois versions expérimentales qui ne comprenaient pas les mêmes items. La figure 14 illustre la répartition des étudiants. La ligne foncée au centre de chaque boîte représente la médiane. Comme les cotés inférieurs et supérieurs de chaque boîte correspondent respectivement au début du deuxième quartile et à la fin du troisième quartile, la moitié des sujets sont inclus dans la boîte. L'ensemble des sujets est compris entre les deux traits horizontaux à l'extérieur de la boîte, à l'exception des cas aberrants qui sont identifiés par de petits cercles ou des astérisques.

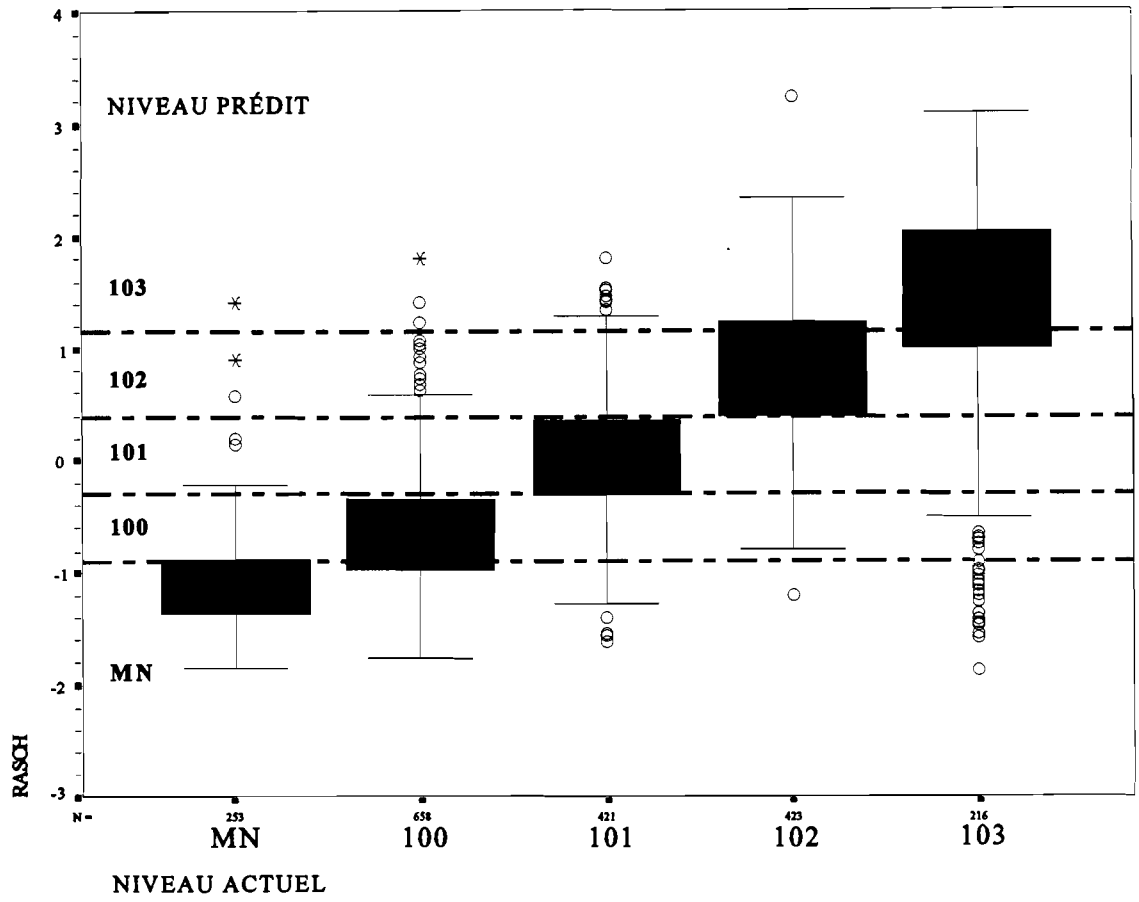

Figure 14
Distribution des sujets selon les niveaux déclarés

Nous avons par la suite traversé le diagramme de lignes horizontales, correspondant aux scores de césure du nouveau test, de façon à placer le maximum de sujets dans leur niveau d'origine. L'opération était relativement aisée puisqu'à l'exception d'un léger chevauchement entre les deux niveaux les plus avancés (102 et 103), la majorité des sujets (55%) n'ont pas changé de niveau. Le tableau 19 permet de voir comment se distribuent les sujets par rapport aux niveaux déclarés et aux niveaux prédits par le nouveau test. Les totaux marginaux montrent que la proportion des sujets classés à chacun des niveaux a légèrement changé. En adoptant les seuils proposés, les étudiants se répartiraient plus également d'un niveau à l'autre puisque globalement une centaine d'étudiants classés au niveau 100 passeraient au niveau inférieur (mise à niveau) et qu'une autre centaine passeraient du niveau 102 au niveau le plus avancé. On peut voir par ailleurs, grâce au tableau 19, que les différences de plus d'un niveau sont relativement rares à l'exception de 24 étudiants, classés au niveau 104 par leur établissement, et que le résultat à

l'épreuve expérimentale placerait dans un cours de mise à niveau. Deux explications sont possibles : soit que ces étudiants ont été surclassés par leur établissement, soit qu'ils n'ont pas répondu au test expérimental comme ils auraient dû le faire.

		Niveaux déclarés					
		MM	100	101	102	103	Total
Niveaux prédits	MN	181	177	13	1	24	396 20,1%
	100	63	362	110	16	13	564 28,6%
	101	6	96	201	90	7	400 20,3%
	102	2	17	79	187	19	304 15,4%
	103	1	6	18	129	153	307 15,6%
	Total	253 12,8%	658 33,4%	421 21,4%	423 21,5%	216 11,0%	1971 100,0%

Tableau 19
Répartition des sujets dans les niveaux prédits par rapport aux niveaux déclarés

En utilisant l'échelle de *logits*, que fournit la calibration selon le modèle de Rasch, il devenait possible de convertir les seuils entre les niveaux de façon à obtenir des scores de césure pour chacune des trois versions. Le tableau 20 présente les fourchettes que nous avons initialement établies pour attribuer un niveau à partir du score obtenu à l'une ou l'autre des versions. Il importe de préciser que ces scores de césure ont été fixés à partir de l'ensemble de l'échantillon et que leur application dans un établissement en particulier peut mener à une répartition entre les niveaux fort différente de celle à laquelle cet établissement est habitué. Cela est susceptible de se produire dans les établissements où le haut degré de bilinguisme de la population conduit à un enrichissement des contenus de cours et inversement dans les établissements où une faible exposition de la population à la langue anglaise risque de provoquer la disparition des cours plus avancés. Il n'est pas exclu également que des disparités s'observent dans les établissements qui procèdent actuellement à leur classement à l'aide d'outils qui sont fort différents du TCALS (en se basant sur les notes de Secondaire V, par exemple).

	Mise à niveau	Anglais 100	Anglais 101	Anglais 102	Anglais 103
Version facile	0-33	34-42	43-50	51-57	58-65
Version difficile	0-15	16-23	24-33	34-44	45-65
Version commune	0-37	38-48	49-59	60-99	70-85

Tableau 20
Scores de césure pour chacune des versions : première proposition

Afin de nous assurer du bon fonctionnement du test dans son ensemble, nous avons procédé à une mise à l'essai de l'instrument en l'utilisant la version commune pour réaliser le classement d'une cohorte de 774 étudiants à la session d'hiver 1997, au Cegep André-Laurendeau. Nous avons pu ainsi vérifier des aspects tels que la qualité sonore, la clarté des consignes, le temps imparti, etc. L'opération visait également à nous assurer de la pertinence et de la justesse des scores de césure. La moyenne de la cohorte se situait à 62,11 (sur 85) avec un écart-type de 15,94. Il faut reconnaître que de tels résultats ont causé un certain émoi. Il est en effet apparu que le test était trop facile et qu'en conséquence, par rapport aux années précédentes, une proportion plus grande d'étudiants se retrouvait aux niveaux plus avancés. La figure 15 rend compte de la différence entre le degré de difficulté que représente les différents items de la version commune (chacun correspond à un point du graphe) pour la cohorte d'André-Laurendeau par rapport à l'ensemble de l'échantillon d'expérimentation. Nous avons pu faire cette comparaison en fixant l'échelle d'habileté et en recalibrant les items. Le décalage de l'ensemble (environ 0,5 sur l'échelle des *logits* montre clairement que dans l'ensemble les items sont plus faciles pour les étudiants d'André-Laurendeau. Il faut souligner que bien qu'il ne soit pas localisé dans une zone géographique à forte densité anglophone, le Cégep André-Laurendeau attire une clientèle qui provient principalement de l'Île de Montréal.

Il apparaissait nécessaire, à la lumière des résultats au Cégep André-Laurendeau de revoir les scores de césure. Dans l'espoir d'en arriver à une distribution plus convenable pour les établissements susceptibles de se trouver confrontés au même problème, c'est-à-dire pour la majorité des collèges de la région montréalaise, nous avons cherché à remanier les scores de césure. Globalement, il s'agissait d'abaisser les étudiants d'un niveau, sauf pour ce qui est des très avancés (pas plus de 15%) qui devaient rester au niveau 103. De plus, une contrainte supplémentaire, à savoir la disparition possible des cours de mise à niveau, nous incitait à restreindre le nombre de candidats classés dans ce cours (pas plus de 10%). En comparant différents scénarios, nous en sommes venus à une deuxième échelle (tableau 21).

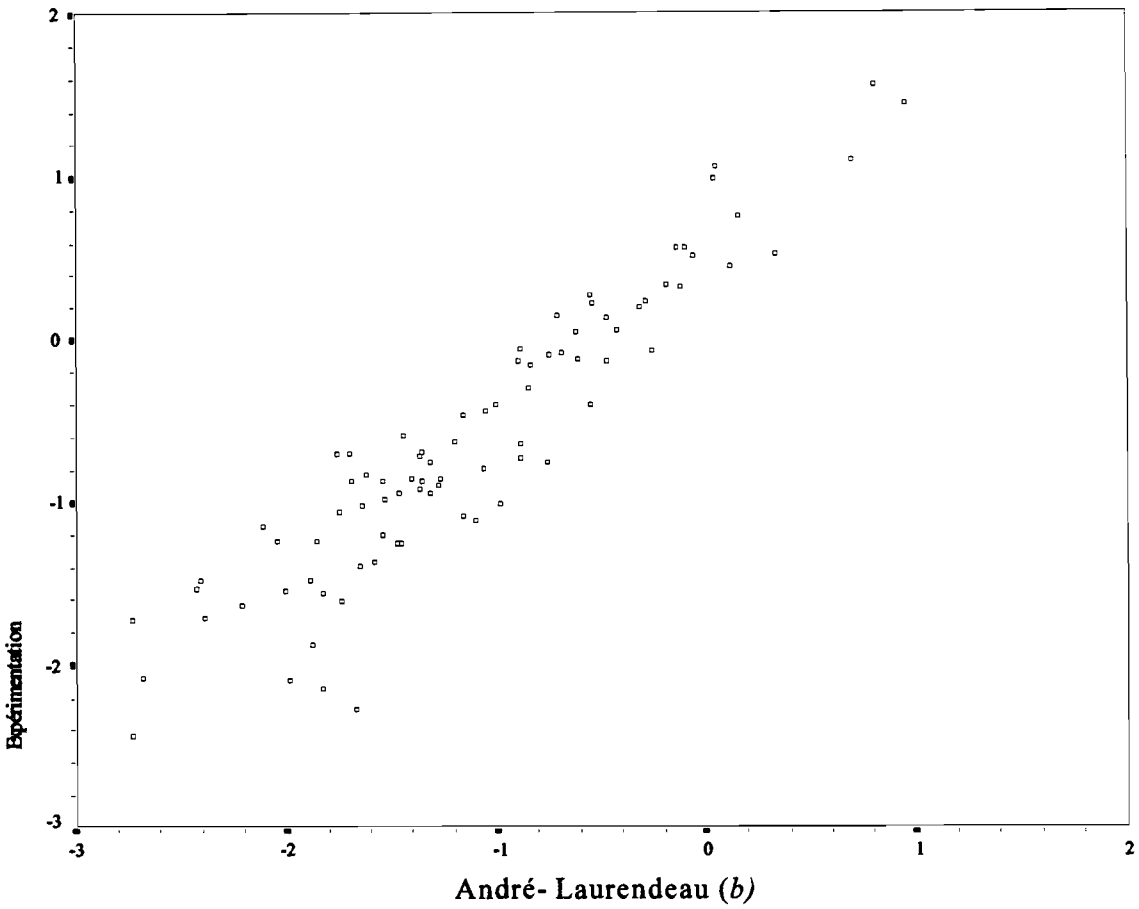

Figure 15
Difficulté des items à André-Laurendeau par rapport à l'échantillon expérimental

	Mise à niveau	**Anglais 100**	**Anglais 101**	**Anglais 102**	**Anglais 103**
Version facile	0-34	35-51	52-58	59-62	63-65
Version difficile	0-16	17-34	35-47	48-56	57-65
Version commune	0-38	39-60	61-72	73-79	80-85

Tableau 21
Scores de césure pour chacune des versions : deuxième proposition

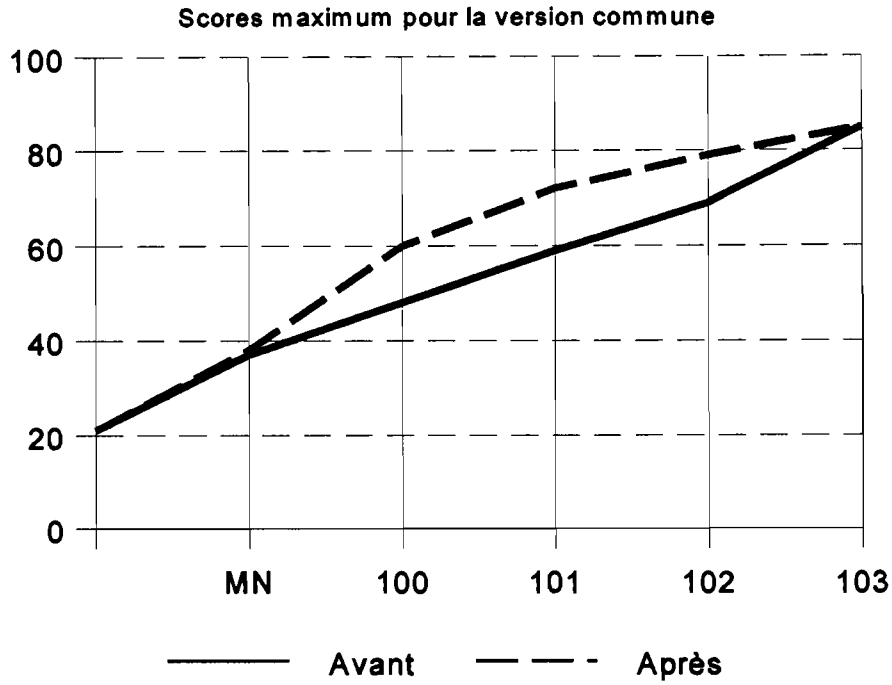

Figure 16
Modification des scores de césure après la mise à l'essai

Le diagramme de la figure 16 permet de comparer la première proposition des scores de césure avec la deuxième. On peut voir que pour la version commune du test, le seuil de passage du cours de mise à niveau au cours 100 n'a que légèrement augmenté (un point) ce qui, dans le cas du Cégep André-Laurendeau se traduit par moins de 10% d'inscriptions dans le cours de mise à niveau. Les seuils de passage entre les cours 100 et 101, 101 et 102, de même que 102 et 103 ont été considérablement haussés (respectivement de 13, 12 et 10 points), ce qui laisse tout de même un nombre suffisants d'étudiants pour assurer la viabilité du cours 103.

Lors d'une autre mise à l'essai, à plus petite échelle en région, nous avons cependant constaté que l'application de la deuxième échelle risquait, dans les établissements qui se trouvent dans des régions où l'exposition à l'anglais est réduite, de provoquer une concentration de la clientèle étudiante dans les niveaux inférieurs au point de mettre en péril l'existence des cours avancés. De plus, un plus grand nombre d'étudiants se retrouvait dans le cours de mise à niveau à un moment où, de surcroît, le statut de ce cours était remis en question. Nous avons donc suggéré d'appliquer la première échelle en indiquant que celle-ci était provisoire.

La situation présente montre de façon évidente que l'application de scores de césure uniformes pose des problèmes majeurs qui sont reliés aux disparités entre les clientèles. Si tous

s'entendent sur l'importance d'adopter des normes communes, plusieurs en redoutent les effets. La solution dépasse les objectifs du présent projet. De fait, la solution optimale du point de vue métrologique (la première proposition) risque de ne pas rallier la majorité des établissements. Dans ces circonstances, il nous apparaît que l'établissement de scores de césure définitifs ne pourra se faire qu'après des discussions avec tous les intervenants sur la pertinence d'uniformiser non seulement les scores de césure d'un test, mais aussi d'homogénéiser les contenus de cours. Un comité réunissant des représentants de divers établissements a d'ailleurs été créé avec le mandat d'étudier la question des scores de césure. Au cours de l'hiver 1998, de nouveaux seuils ont été proposés. C'est à l'usage qu'on pourra voir si les scores de césure proposés satisfont la majorité des établissements.

Conclusion

Les analyses que nous avons rapportées dans le présent document ont servi à composer un ensemble de test qui a été distribué à travers la province. Chaque collège a reçu une copie du matériel qui inclut, pour chacune des versions, le questionnaire et un modèle de feuille de réponses accompagné d'un corrigé. À ce matériel, est joint un document d'information destiné aux utilisateurs et un protocole qui établit les conditions qui régissent la reproduction et l'utilisation du matériel. Étant donné les coûts associés à l'élaboration d'un instrument de classement et les conséquences irrémédiables d'une diffusion inconsidérée du matériel de test, il est impérieux que les établissements n'utilisent le test que pour leurs propres besoins et ne le communiquent pas à d'autres organisations ou individus. Plus encore, il est essentiel que les utilisateurs n'aient recours à ce test que pour les fins pour lesquelles il a été conçu, soit le classement des cégépiens en anglais langue seconde. Les responsables du test dans les établissements s'engagent, en signant le protocole, à se conformer à des règles strictes d'utilisation.

Sans que le test ne soit imposé de quelque façon, plusieurs établissements y ont déjà eu recours pour procéder au classement de leurs étudiants. Leurs commentaires sont élogieux en ce qui trait au contenu, à la facilité d'utilisation et à la présentation matérielle (tant pour les document écrits qu'enregistrés). La question de la détermination des scores de césure risque de poser encore problème mais, comme nous l'avons souligné, la solution déborde l'élaboration de l'instrument puisqu'elle implique une réflexion sur l'ensemble du programme en anglais langue seconde. Nous comptons sur le comité mis sur pied pour poursuivre la réflexion entreprise et réévaluer les seuils convenus.

Malgré un échéancier très serré, nous avons suivi toutes les étapes d'élaboration d'un instrument de mesure, n'hésitant pas à recourir à des techniques raffinées pour produire un outil innovateur et d'une qualité métrologique supérieure. Cependant, la validation d'un instrument de mesure est un processus continu qui se complète avec l'accumulation des données et qui doit tenir compte de l'évolution des contextes d'utilisation. En ce sens, il faudra prévoir un suivi à l'élaboration. Le suivi devrait également comprendre la mise au point de versions ultérieures, les précautions prises ne garantissant jamais tout à fait que les items ne seront jamais divulgués. Le plan d'élaboration que nous avons adopté permet d'ailleurs la production d'items qui pourront enrichir la banque en s'arrimant à l'échelle existante. Parmi les versions possibles, on devrait considérer sérieusement la possibilité d'une version informatisée qui appliquerait les principes de testing adaptatif. Il en résulterait un test plus court et individualisé. Au plan administratif, outre les avantages d'un traitement automatique des résultats (par exemple pour la production des listes de classe), on pourrait alors songer à une administration à distance qui tirerait profit des ressources de l'internet.

Références

1. WALL Diane., CLAPHAM Caroline. et ALDERSON Charles J., "Evaluating a Placement Test". *Language Testing,* vol.11, no3, 1995, pp 23-56.

2. BACHMAN Lyle .F. et PALMER Adrian S., *Language Testing in Practice*. Oxford : Oxford University Press, 1996.

3. LUSSIER Denise et TURNER Carolyn E., *L'évaluation en didactique des langues*. Montréal : Centre éducatif et culturel, coll. Le point sur...,1995.

4. BARSALOU FROIO Lydia, *Needs Assessment and Formative Evaluation in the Development of a College Level English Language Placement Test* (mémoire de M.A. non publié). Montréal : Concordia University. 1997.

5. LAURIER Michel, *L'informatisation d'un test de classement en langue seconde*. Québec: CIRAL, Université Laval, Faculté des Lettres, 1994.

6. OLLER John W. Jr, *Language Test at School*. Londres: Longman, 1979.

7. FROIO Lydia et PEARO Charles, *English Second Language Language Level Profiles*. Document interne disponible auprès des auteurs, 1996.

8. BLAIS Jean-Guy et LAURIER Michel, "The Dimensionality of a Placement Test from Several Analytical Perspectives". *Language Testing*, vol 12, no 11, 1995, pp 72-98.

9. BAKER Rosemary, *Classical Test Theory and Item Response Theory in Test Analysis*. Lancaster:: Language Testing Update, Special report No 2, 1997.

10. HENNING Grant, *A Guide to Language Testing : Development, Evaluation, Research*. Cambridge, MA : Newbury House.