Université de Montréal

Fonctionnement de tâches discrètes et intégrées pour l'évaluation de la lecture en français langue seconde des nouveaux arrivants au Québec.

par
Vincent Folny

Département d'administration et de fondements de l'éducation
Faculté des sciences de l'éducation

Mémoire présenté à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de M.A.
en Mesure et évaluation

Le 22 juin 2009

Université de Montréal
Faculté des sciences de l'éducation

Ce mémoire intitulé :

Fonctionnement de tâches discrètes et intégrées pour l'évaluation de la lecture en français langue seconde des nouveaux arrivants au Québec.

présenté par
Vincent Folny

a été évalué par un jury composé des personnes suivantes :

Monsieur Jean-Guy Blais
Président

Monsieur Michel Laurier
Directeur de recherche

Madame Nathalie Loye
Membre du jury

Je tiens à remercier tous ceux qui ont aidé, de près ou de loin, à la rédaction de ce mémoire. Le MICC, les écoles de langues de l'Université de Montréal et de l'Université du Québec à Montréal (UQAM) ont grandement facilité l'accès à leurs étudiants. Les étudiants, nouveaux arrivants au Québec, se sont livrés avec beaucoup de sérieux et d'intérêt à la passation du test. Ils ont très bien compris les enjeux de la recherche. Enfin, je tiens à remercier l'Université de Montréal pour le soutien qu'elle m'a donné et, tout particulièrement, mon directeur de recherche pour sa patience et sa compréhension.

Résumé

Mots-clés : évaluation en langue seconde, tâche de lecture, difficulté d'un test

Cette recherche s'inscrit dans le cadre de l'évaluation des compétences langagières en français chez des adultes immigrants en vue de leur placement dans des cours de français. Elle porte sur la dimensionnalité, de même que sur la difficulté objective et subjective de tâches discrètes ou intégrées de compréhension écrite, à différents niveaux de maîtrise. Elle propose des analyses de l'estimation de la maîtrise linguistique en fonction de l'appartenance des candidats à des groupes linguistiques distincts.

Pour mener à bien la recherche, un test de six textes et de 30 items a été créé. Il a été administré à 118 immigrants. Ces immigrants suivaient les cours de français proposés par le Ministère de l'immigration et des communautés culturelles du Québec (MICC) dans les écoles de langues de l'Université de Montréal et de l'Université du Québec à Montréal. Après administration, ce test a été soumis à des analyses portant sur la dimensionnalité et la difficulté des tâches discrètes et intégrées ainsi que sur les interactions entre ces tâches et les différents groupes de candidats. Des études plus précises ont été faites sur l'interaction entre le type de tâche, l'appartenance à un groupe linguistique pour des candidats et des items de niveau similaire. Enfin, des analyses ont permis d'étudier la perception de la difficulté des tâches par les candidats.

L'étude, même si elle porte sur un test en rodage, permet de distinguer la dimensionnalité de tâches discrètes de celle de tâches intégrées. Elle permet également de constater les différences de fonctionnement entre ces deux types de tâches. Enfin, elle permet de comprendre l'interprétation de la difficulté par les candidats et, par ricochet, leur vision du test.

In fine, des propositions sont formulées quant à l'opportunité d'utiliser des tâches discrètes et intégrées dans un test de positionnement adaptatif en français langue seconde.

Abstract

Key words: Second language assessment, Reading task, Test difficulty

This research has been conducted within the assessment procedure of the language competence of adult immigrants, for placement purposes in French courses. It relates to the dimensionality as well as the objective and subjective difficulty of discrete or integrated reading tasks at different proficiency levels. Analyses of linguistic proficiency estimates are proposed in relation with candidates' linguistic groups.

In order to conduct this study successfully, a 6-text and 30-item test has been constructed and administered to 118 immigrants. These immigrants were enrolled in French courses offered by the Ministry of Immigration and Cultural Communities (MICC) in language schools at the Université de Montreal and the Université du Québec à Montréal. After the administration, analyses have been made on the dimensionality and difficulty of the discrete and integrated tasks and on the interactions between the tasks and different groups of candidates. More detailed analyses have been made on candidates and items at similar levels. Finally, we were able to study the candidates' perceptions of task difficulty.

Although the study is based on a provisional test, the dimensionality of discrete and integrated tasks has been distinguished. Differences in the way the 2 types of task work have been shown. Finally, the candidates' interpretation of difficulty and therefore, their view of the test, are better understood. *In fine*, proposals are made in regard with the proper use of discrete and integrated tasks in an adaptive placement test in second language.

TABLE DES MATIÈRES

Liste des Tableaux

Liste des figures

Chapitre 1 : Problématique

1.1. Contexte général

1.1.1. L'augmentation de la demande en évaluation linguistique

Depuis quelques années, la demande vis-à-vis des certifications linguistiques semble être à la hausse, notamment, pour les langues les plus diffusées. Plusieurs facteurs sont à l'origine de cette situation. La mondialisation des échanges et du savoir incite les citoyens de nombreux pays à se doter de diplômes de langue ou encore à faire évaluer leurs niveaux linguistiques à l'aide de tests. Dans le monde universitaire, la *Déclaration de l'Association internationale des universités sur l'internationalisation de l'enseignement supérieur* (1998), la *Déclaration de Bologne* (1999) ou encore la *Déclaration de l'Association des Universités et Collèges du Canada sur l'internationalisation et les universités canadiennes* (1995) font, non seulement, la promotion de la diversité culturelle, mais aussi, de l'apprentissage des langues étrangères. Pour la migration, nombre de pays choisissent des tests de langue standardisés. Ils les utilisent soit pour sélectionner les migrants, soit pour positionner ces derniers dans des cours de langue.

Du fait de l'internationalisation des échanges, l'accent est mis sur une plus grande mobilité étudiante. Bien que les deux phénomènes ne soient pas entièrement liés, on assiste de manière concomitante à un accroissement des flux migratoires (M.I.C.C, 2008a) et à une augmentation des demandes de formation (M.I.C.C., 2005 : 13), d'évaluation et de certification linguistiques (M.I.C.C, 2008b : 24).

1.1.2. Un besoin accru de normalisation et de standardisation

Pour répondre à la demande de certification linguistique, plusieurs voies ont été explorées. Dans certains cas, ont été créés des diplômes évaluant un seul niveau, dans d'autres, des tests de langue évaluant différents niveaux sur une échelle. A l'occasion de la sélection des immigrants (notamment au Québec), on a pu charger des agents gouvernementaux de la vérification du niveau linguistique, notamment, par

l'intermédiaire d'entrevues. Le caractère officiel des décisions touchant à l'acceptation d'un étudiant dans une université, dans une école de langue, ou encore, à l'obtention d'un visa migratoire, a amené les sociétés modernes à s'intéresser assidûment aux questions liées à la validation des tests de sélection ou de classement. La validation revêt plusieurs aspects. Elle peut porter sur le processus de conception du test, mais encore sur le déroulement des passations. Quand elle porte sur les processus mis en place pour assurer la qualité optimale de la mesure et à l'évaluation, elle peut être accompagnée d'une validation scientifique extérieure. Les tests peuvent être co-validés par une institution de référence, ou encore, par des équipes de chercheurs reconnus. Dans ce contexte, les utilisations et applications de modèles de mesure, notamment ceux de la théorie de réponse aux items (T.R.I.), ont vu leur importance grandir. Ces modèles ont permis d'améliorer la qualité de la mesure, notamment des tests de compétence en langue étrangère et langue seconde destinés un large public. Pour ce qui est du contenu des tests, divers référentiels de compétences (pour guider l'élaboration de tâches) ont été mis au point et sont à la disposition des concepteurs. Ainsi le Canada a-t-il mis en place des *Standards linguistiques canadiens* (C.C.L.B., 2000, 2002) en anglais et français langues secondes, le Québec des *Niveaux de compétence en français langue seconde pour les immigrants adultes* (M.R.C.I, 1998) et l'Europe un *Cadre européen commun de référence pour les langues (C.E.C.R.)* (Conseil de l'Europe, 2001). Pour ce qui est du Canada, le choix a été fait de privilégier un cadre de référence mixant ceux proposés par Bachman (1990), Bachman et Palmer (1996) mais aussi d'autres sources :

> « Le modèle de compétence communicative des *Standards linguistiques canadiens* est une adaptation et une synthèse ou une « fusion » des modèles suivants : modèle de compétence langagière de Bachman (1990), le modèle de Bachman et Palmer (1996) et un modèle pédagogique de compétence communicative de Celce-Murcia, Dornyei et Thurrell (1995). Tous ces modèles sont des versions actualisées des modèles de compétence communicative classiques de Canale et Swain (1980) et Canale (1983) et tous sont redevables à Hymes (1971) et à son concept de compétence communicative » (Pawlikowska – Smith, 2002: 7).

La rédaction de standards et de niveaux de référence est loin d'être arrivée à son terme. Ainsi certains chercheurs avancent-ils que les référentiels de compétence ne suffisent pas

à assurer une validation optimale. Selon eux, ils ne permettent pas d'atteindre une comparabilité suffisante entre deux tâches d'évaluation, ou encore, entre deux tests prétendant évaluer le même trait. Dans le contexte européen, Cyril Weir (2005a : 297) propose de définir plus en détail les caractéristiques des tâches d'évaluation du *Cadre Européen Commun de référence* (C.E.C.R.) afin d'en assurer la comparabilité. Il propose de construire un cadre de spécification des caractéristiques de ces tâches, de celles des conditions d'administration (contexte) et d'effectuer un va-et-vient entre les échelles de niveau et les observations empiriques[a] :

> "This is encouraging news for language test developers interested in creating a more comprehensive and valid set of descriptors. Eventually if a new version of the CEFR is developed by testers that better defines content by level [...], the coverage of an exam can be profiled against this and then, through calibration and standard setting, the results of content analysis can be checked against psychometric value. The scale's theory-based validity, though of no less importance, may take longer to address. The nature of cognitive and metacognitive processing in spoken and written modes is less overt and susceptible to investigation than context variables, which are more readily amenable to expert scrutiny and empirical investigation.[1]" (Weir, 2005a : 297).

Alderson *et al.* (2006), quant à eux, proposent de revenir sur la rédaction des descripteurs utilisés pour décrire le niveau de compétence linguistique afin d'aider les concepteurs de tests à concevoir des tâches comparables. Pour ce faire, ils identifient différents problèmes. Tout d'abord, il est possible que les descripteurs présentent de sérieux problèmes de rédaction (problème de consistance, de terminologie, de manque de définition, ou encore des « trous » dans les descriptions, etc.). Ensuite, les descripteurs peuvent ne pas avoir été écrits à l'intention des développeurs de test mais à celle des candidats ou encore à celle des institutions utilisant les résultats au test (Alderson, 1991 ; Alderson *et al.* 2006 ; North, 2000). Se pose alors le problème de l'utilisation adéquate des échelles.

Aujourd'hui, la recherche s'oriente vers la rédaction de descripteurs adaptés au construit, décrivant non seulement la compétence linguistique mais aussi les caractéristiques des

[a] Les traductions des citations en anglai s se trouvent toutes en annexe après la bibliographie.

tâches d'évaluation utilisées ou encore l'interaction entre les deux (Alderson *et al.*, 2006 ; Spaan, 2006 ; Weir, 2005a). On cherche également à savoir quelles sont les interactions entre les caractéristiques de la tâche d'évaluation, les composantes constituant cette tâche et les résultats des candidats (Brindley, 1987 ; Carr, 2006 ; Norris *et al.*, 2002).

Si la mise au point de référentiels et de descripteurs de meilleure qualité permettant une meilleure validation est un travail utile, on peut se demander jusqu'où il est nécessaire de décrire les caractéristiques des tâches et activités d'évaluation, jusqu'à quel point on ne risque pas de contraindre (voire restreindre) abusivement, au risque de la stéréotypie, le contenu des examens. Pour le concepteur, il s'agit de trouver un positionnement entre les besoins de la standardisation et la contre-productivité de modèles apportant des contraintes de format trop fortes et visant exclusivement la comparabilité ou encore la précision de la mesure (Puren, 2004 ; Sireci, 1998).

1.1.3. La comparabilité : le cas des enquêtes de lectures internationales

La comparabilité est donc un enjeu du présent et de l'avenir. Elle intervient à des niveaux micro-sociétaux et macro-sociétaux. Elle peut viser des systèmes d'éducation, des approches et méthodes d'évaluation ou, plus spécifiquement, la pertinence des tâches d'évaluation. Pour mieux appréhender son articulation macro-sociétale et micro-sociétale (du contexte global d'une société au contexte particulier de la rédaction des tâches d'évaluation), une enquête, le *International Adult Literacy Survey* (I.A.L.S.) portant sur la littératie servira d'illustration. En particulier, sera évoqué l'incident créé par la France en 1995. Alors que ce pays est sujet à des résultats peu flatteurs, il décide de retirer sa participation à l'enquête, au dernier moment, avant la publication des résultats. Ce retrait serait expliqué par une remise en cause de la validité des tâches d'évaluation. Mais laissons la parole à Bottani et Vrignaud (2004 : 42) qui expliquent comment les tâches de lecture ont été sélectionnées et validées. Ils relatent également la réaction des autorités françaises :

> « Dans l'ensemble, quelque 175 tâches de lecture et d'écriture ont été élaborées pour les tests sur le terrain. De ce nombre, 114 qui se sont avérés valides pour toutes les cultures ont été retenues aux fins de l'évaluation principale. Environ la moitié de ces tâches étaient fondées sur des documents provenant de l'extérieur de l'Amérique du Nord. A la suite de

> la publication des résultats en automne 1995, la DEP [Direction de l'évaluation et de la prospective du Ministère de l'éducation nationale en France] a estimé que ces précautions n'avaient pas été suffisantes et que la composition du test était susceptible de fausser les réponses des interviewés. Le lancement de toute une série d'initiatives visant à apprécier les risques de biais présents dans la méthodologie appliquée jusqu'à 1995 dans les enquêtes scolaires et surtout les conséquences de la transposition aux adultes de cette méthodologie testée auparavant avec les élèves en classe a engendré un important débat scientifique, produit beaucoup de publications sous forme de rapports et d'articles, mais n'est pas parvenu à fournir des preuves irréfutables d'invalidité de ces enquêtes ou à mesurer précisément le degré de distorsion produit par les biais culturels et linguistiques sur les résultats. ».

Comme il est possible de le lire, la réaction de la France se traduit par une remise en cause la comparabilité internationale des tâches. Pour ce pays, le problème est que les tâches utilisées ont été traduites. C'est encore que les méthodologies d'enquête et d'évaluation avaient été conçues pour des adolescents alors qu'on les utilisait pour des adultes.

Aujourd'hui, les appréhensions, vis-à-vis de la comparabilité des évaluations dans un contexte international, semblent avoir été oubliées puisque la France a décidé d'utiliser des tests internationaux, notamment, pour évaluer ses élèves en langue étrangère.

Comme on peut le constater, selon le domaine testé ou encore les sensibilités des uns et des autres, la légitimité des tests internationaux est différemment appréciée. Dans le contexte particulier des tests de langue internationaux, la problématique de la légitimité traduit le choc d'idéologies transnationales et nationales mais aussi des enjeux de pouvoir, de légitimité et de reconnaissance. La question de la concurrence entre les examens des systèmes scolaires publics locaux et ceux issus des « agences » internationales se pose âprement, de même que celle de la pertinence de l'usage des tests internationaux dans des problématiques scolaires locales (Goulier, 2004 ; Zarate, 2004). Et Geneviève Zarate (2004 : 16) de déclarer :

> « Plus exposé que d'autres à l'internationalisation, le champ des langues est contraint de s'adapter à la concurrence internationale. Des innovations, produits d'une industrie des langues, sont à la conquête d'un nouveau marché, celui de l'évaluation des compétences.

> Ce marché, qui ouvre la porte des entreprises et des études à l'étranger, pour peu que l'on ait le minimum requis, tend à être contrôlé par des pays natifs et à affaiblir les tests nationaux (l'Université de Cambridge, l'Alliance française, le Goethe Institut et le WTB ou l'Université de Salamanque en partenariat avec l'Institut Cervantes, regroupés pour une meilleure visibilité européenne) et leurs homologues étrangers validant, par exemple, le *Test of English as a Foreign Language* (TOEFL), le HSK, test de langue chinoise mis au point par la commission d'état pour l'éducation en Chine. Le test de demande d'admission préalable dans le premier cycle universitaire français, dit « test des ambassades », qui reproduit le niveau d'exigence élevé des tâches universitaires est en passe d'être remplacé par un Test de Connaissance du Français, plus souple et plus finement gradué, mieux adapté à l'espace compétitif universitaire. Ces dispositifs doublent généralement la certification universitaire française et, en principe, n'y suppléent pas puisqu'un test n'a pas de valeur diplômante. Mais la visibilité et la crédibilité de ces tests tend à gagner du terrain hors du système éducatif ».

Dans le champ des tests de langue, on s'interroge sur la comparabilité des tests (Alderson *et al.*, 2006). Il faut dire qu'aucune étude (ou recherche), à ce jour, n'a permis véritablement de déterminer si des tests de conceptions, de construits, et de langues différents (même dans le contexte européen) sont comparables et, surtout, jusqu'à quel degré ils le sont (Weir, 2005a). Pour le moment, on « relie » les tests de langue en utilisant des référentiels composés d'échelles de compétences. Rares encore sont les études qui comparent des tests différents conçus pour évaluer une même langue ou des tests traduits en plusieurs langues qui ambitionnent de mesurer une même compétence. Quand on se prête à l'exercice de la comparaison (Alderson *et al.*, 2006), la diversité des caractéristiques des tâches d'évaluation qui sont dites être calquées sur un niveau similaire est troublante. Par conséquent, on peut légitimement douter que ces tests évaluent réellement un même niveau linguistique. Dans le contexte plurilingue européen, s'il y a bien des projets pour construire des banques d'items multilingues (Figueras *et al.*, 2005), il n'existe en revanche aucune instance chargée de vérifier l'équivalence des tests et diplômes produits par ces grandes agences. Ailleurs, lorsque des équivalences sont proposées, elles sont circonstancielles. Elles sont proposées par les agences ou institutions développant un test, ou encore, des instances gouvernementales utilisant un

test pour lequel l'échelle ayant servi à calibrer les tâches d'évaluation est différente de l'échelle nationale ou transnationale utilisée normalement. Dans les pays bilingues ou multilingues, comme au Canada (Ercikan *et al.*, 2004), on tente parfois de proposer des items ou des tâches pour l'évaluation équivalents dans deux ou plusieurs langues. Il est dit alors que les résultats du test ou que les items multilingues correspondent à tels niveaux X et Y d'une échelle A alors que les items ont été calibrés à partir d'une échelle B. Aujourd'hui, il n'existe toujours pas de consensus pour élaborer une échelle « universelle »[b] de la compétence en langue. Mais une telle chose est-elle souhaitable et réalisable ? Les échelles et référentiels de niveau de compétence linguistique sont conçus et utilisés différemment, selon la région géographique, ou encore, la langue apprise. S'il n'y a pas d'homogénéisation générale, quant aux échelles et référentiels utilisés, en revanche, on voit apparaître des phénomènes d'homogénéisation locale. Certaines échelles et référentiels sont très largement utilisés de par le monde, que ce soit au niveau local, régional ou international[c].

La comparabilité des tâches, leur adéquation aux référentiels, aux échelles ou encore leur pertinence pour l'évaluation linguistique ne sont donc pas un vain mot. Elles doivent retenir toute l'attention des chercheurs car il semble que, parfois, on ne prête pas assez attention au phénomène. Si, traditionnellement, on analysait les biais « culturels », depuis quelques années, la recherche s'oriente vers la comparabilité des tests traduits (Bertrand & Blais, 2004 : 258 ; Hambleton *et al.*, 2005). Alors que pour l'analyse des biais, ce qui était mis en avant, c'était la justesse de l'évaluation (Kunnan, 2000 ; Shohamy, 2001), l'adoption récente de la convention sur la diversité culturelle semble pointer vers de nouvelles directions et de nouvelles questions. On se demande, par exemple, si les tests de langue respectent les différences culturelles entre les candidats. Dans le contexte des migrations internationales, on peut se demander jusqu'où les tests doivent respecter et prendre en compte la culture des candidats. Dans le contexte des formations linguistiques proposées aux immigrants (en Australie, au Canada, en France, au Québec,…), la question de savoir jusqu'où doit aller le respect et la prise en compte des identités des uns

[b] Pour un historique des échelles de la mesure de la compétence linguistique, on revoit le lecteur à la lecture de Brian North (2000)

[c] Par exemple, le cadre de référence européen des langues est souvent utilisé dans les pays d'Amérique centrale et du sud.

et des autres et le besoin de cohésion identitaire des sociétés d'accueil est une question cruciale.

Les enjeux liés à des visions dynamiques ou conservatrices de la diversité culturelle ne sont pas de minces enjeux. Ils ne trouvent pas de réponses simples. Ces enjeux, dans le contexte de la mesure et de l'évaluation, incitent le chercheur ou le praticien à se demander si, à l'issue de la calibration d'un test, l'échantillon des personnes est composé d'une seule ou de plusieurs populations (Hambleton, 1989 ; Hambleton, Swaminathan & Rogers, 1991). Le chercheur posera alors la question de la nature des interactions entre le type de tâches utilisées et l'appartenance culturelle des candidats. Une variation de niveau uniquement due à une appartenance culturelle n'aura pas la même signification et le même impact qu'une différence de niveau uniquement due à l'utilisation d'un certain type de tâches. Dans le premier cas, on pourra penser que le test (ou la tâche d'évaluation) défavorise certains candidats. Dans le deuxième cas, on analysera la différence comme étant uniquement due à la difficulté de la tâche.

1.1.4. La comparabilité et la diversité culturelle et linguistique

Le problème posé par la standardisation et la comparabilité des résultats d'évaluation est le fruit de la tension entre d'un côté une uniformité nécessaire et de l'autre la prise en compte d'une certaine variabilité. Alors que la validation peut servir à garantir l'uniformité du contenu ou encore à la fidélité d'un test, la prise en compte de la diversité culturelle et linguistique laisse envisager une certaine variabilité tant au niveau du contenu que de la mesure.

Dans l'évaluation linguistique, il est dit que la population des candidats testés doit être homogène pour que les interprétations des résultats aux tests soient valides. Et Young *et al.* (1996: 25) de dire:

> «The second assumption of IRT is the homogeneity of the population of examinees. That is to say, the only relevant trait causing differences between scores of examinees is their ability on the trait measured by the test[2] ».

Pour pouvoir effectuer des interprétations valides, la tension entre la diversité culturelle et linguistique des candidats et l'exigence d'une mesure adéquate doit être prise en

compte dans la carte d'identité du test, autrement dit, dans le construit. Selon Messick (1989), la validité d'un examen, un concept unifié, correspond à quatre facettes qui forment un tout unifié fondé sur deux pôles : la structure interne du test (relation entre la réponse et la tâche ou l'item) et sa structure externe (relation entre le score au test, d'autres mesures et les variables du contexte). La validité est l'interaction entre d'un côté, les données empiriques du test (les scores) et, de l'autre, l'interprétation des scores et l'usage que l'on fait des tests. Dans les contextes où l'on fait face à des cohortes multiculturelles composées de candidats aux profils divers et variés, il est donc indispensable de définir un construit qui permette de proposer une évaluation uniforme en faisant attention à ne pas favoriser un groupe au détriment d'un autre et de ne pas proposer des interprétations biaisées. Il est également important de proposer des tests qui ne répondent pas uniquement à des exigences de la mesure mais aussi à celles du contenu de l'examen (Sireci, 1998 ; Van der Linden, 2005). Les tests doivent proposer une adéquation entre, d'un côté le contenu de l'examen et, de l'autre, l'usage et l'interprétation que l'on fera des résultats.

Au moment de définir le construit de l'examen, il s'agit de trouver un équilibre entre les problèmes liés à l'homogénéité de la population et ceux liés au contenu testé. Il s'agit aussi de trouver un compromis entre le respect de la diversité culturelle et de la diversité de contenu. En effet, si on ne peut pas sacrifier la diversité du contenu au nom du respect de la diversité culturelle des candidats (dans un tel test, les items seraient créés pour leur « neutralité » face à la diversité culturelle des candidats), on ne peut pas, non plus, sacrifier le respect de la diversité culturelle des candidats au nom de la diversité du contenu (dans ce type de test, les items seraient créés en fonction de la diversité de leur contenu sans se soucier de leurs éventuelles interactions négatives avec la culture des candidats).

1.1.5. Le besoin de métathéorie pour mieux articuler le global et le particulier

Cette préoccupation, pour le respect de la diversité culturelle, la comparabilité et l'absence de biais, mais aussi la qualité de contenu des tests, rejoint celles portant sur l'absence de métathéorie multilingue faisant la synthèse des différentes théories

monolingues et expliquant l'acquisition d'une ou de plusieurs langues dites « secondes » ou encore « étrangères ». Pour la lecture, Fitzgerald (2003), qui évoque le cas de la littératie, rappelle le besoin de voir apparaître une métathéorie multilingue prenant en compte l'articulation entre le général et le particulier, entre le similaire et le semblable. Pour l'heure, seules sont disponibles des études faites par différents pays. On ne dispose pas de métathéories faisant le point sur toutes les recherches entreprises. Fitzgerald (2003) explique que, jusqu'à présent, deux théories concurrentes s'affrontent pour expliquer le développement de la lecture multilingue. La « *General Factor Theory* » affirme que l'apprentissage d'une nouvelle langue n'est pas attaché à une composante de la langue en particulier (phonologie, syntaxe, sémantique…) et non plus à un mode particulier (lire, écrire, écouter, parler) mais que l'on peut apprendre une langue en commençant par n'importe quel mode (parler, écrire, écouter, lire). A l'inverse, l'« *Oral Precedence Theory* » stipule que la compréhension orale d'une nouvelle langue forme la base pour lire et pour écrire cette langue et développer ses compétences. Selon cette théorie, il n'est donc pas possible « d'entrer » dans une langue par l'écrit. Pour Fitzgerald (2003), les conséquences d'une plus grande emphase sur la littératie multilingue seront une meilleure interprétation des phénomènes touchant chacune des langues au travers d'une théorie plus large, une interconnexion des champs de la recherche, une meilleure information pour prendre des décisions politiques et la mise en place d'une métathéorie de la lecture. Ceci étant, cette métathéorie n'aura pas pour ambition de proposer l'uniformité pour tous et toutes les langues mais, au contraire, des pédagogies ou des évaluations diagnostiques (dans un contexte scolaire) et adaptées (Bernhardt, 2003). Cela permettra aussi de créer de nouvelles visions de la recherche et d'amener les chercheurs à voir la lecture sous un nouvel angle. Le vœu de Fitzgerald (2003), mais qui est aussi le vœu de Berg (2003) et de Bernhardt (2003), est celui du respect des diversités culturelle et linguistique. Il est en phase avec les préoccupations contemporaines de nombre de pays, dont le Canada et la province du Québec.

1.1.6. Les caractéristiques des groupes de candidats

Aujourd'hui, si on ne dispose pas d'une métathéorie pour la lecture, on sait en revanche que le profil des populations de candidats et le profil de groupes de candidats font varier le niveau des tâches. Bachman (2002) explique ainsi que la difficulté d'une tâche est fonction des caractéristiques des tâches mais aussi de celles des candidats et de l'interaction entre les deux. Lorsqu'on tente de faire une analyse des facteurs qui peuvent expliquer la difficulté, souvent, le facteur langue maternelle (ou encore « langue 1 ») est retenu. Dans l'évaluation linguistique, on vérifie ainsi que les locuteurs d'une langue A ne soient pas favorisés vis-à-vis des locuteurs d'une langue B. Chen et Henning (1985) ont trouvé que, pour un test de placement en anglais, des items contenant du lexique proche de l'espagnol favorisent les hispanophones au détriment des sinophones. Sasaki (1991), quant à elle, reprend l'étude de Chen et Henning (1985) mais dans des conditions différentes. Si elle trouve encore des différences entre les hispanophones et les sinophones pour la connaissance du lexique, en revanche, elle constate que pour la connaissance des expressions idiomatiques les sinophones sont favorisés. Cette différence s'explique par le fait que les cours proposés aux sinophones ont été axés sur l'apprentissage de ces expressions. Ryan et Bachman (1992) quand à eux, comparent le fonctionnement de deux tests en anglais pour deux groupes de langues différentes. Un premier groupe d'individus possède une langue parmi la famille des langues indo-européennes et un autre groupe possède d'autres types de langue. Des différences de fonctionnement entre ces groupes sont bien trouvées. Enfin, Abbott (2007) propose d'étudier les différences de fonctionnement, non pas seulement en fonction de la langue, mais également, en fonction du type de lecture et de stratégies de lecture utilisés. La chercheure part de l'idée que les locuteurs du mandarin utilisent un modèle de lecture ascendant. Ils lisent d'abord les mots, des segments de phrases pour ensuite passer à la phrase entière et enfin au texte (Fender, 2003 ; Parry, 1996). Les arabophones, eux, lisent d'abord le texte pour s'intéresser ensuite aux aspects discrets du discours.

Les conclusions de toutes ces recherches font apparaître des difficultés d'interprétation. Pour Chen et Henning (1985), les différences trouvées entre les hispanophones et les sinophones sont naturelles. Pour Ryan et Bachman (1992), les différences trouvées entre

groupes linguistiques seraient dues à la culture et à l'éducation. Mais comme le note Abbott (2007), les auteurs ne se positionnent pas sur le fait que cela nuise ou pas à la validité de construit. Pour ce qui est de la recherche de Abbott (2004, 2007), des différences systématiques sont trouvées entre les sinophones et les arabophones. Les items lexicaux et les items demandant à mettre en relation des mots clefs favorisent les sinophones. Les items demandant de repérer des idées principales et de faire des connections entre différentes idées favorisent les arabophones.

Pour ce qui est de l'interprétation des résultats, Chen et Henning (1985) disent que les différences trouvées entre les hispanophones et les sinophones sont réelles. Toutefois, Il faut veiller à ce que les mots proches de l'espagnol ne soient pas plus fréquemment présents dans le test que dans l'usage fait de la langue anglaise. Pour Ryan et Bachman (1992), les différences qu'ils trouvent sont bien dues à la langue mais aussi à des facteurs culturels et éducatifs. Toutefois, comme le note Abbott (2007), les auteurs ne statuent pas sur le fait que cela soit néfaste ou pas pour les candidats. Pour ce qui est de la recherche de Abbott (2007) qui portait sur un test de lecture, l'auteure explique que si on retient un schéma interactif de lecture (Stanovich, 2000), il faut alors concevoir un test avec des items évaluant une lecture « locale » ou encore détaillée et une lecture plus globale (*bottom-up* et *top-down*). Les différences trouvées entre les candidats sinophones et arabophones ne sont pas contraire au construit du test si les items sont représentatifs du contenu des tâches demandant au candidat la mise en œuvre de la compétence que l'on veut inférer.

Le facteur langue maternelle est souvent retenu dans les processus de validation des tests. Les interprétations sont faites en fonction des résultats empiriques mais aussi du construit.

1.1.7. Le renouveau des approches pédagogiques et des approches évaluatives

Le contexte dans lequel sont placés les tests de langue, à savoir, les enjeux liés à l'évaluation linguistique, la prise en compte à la fois de contraintes de contenu mais aussi de la spécificité culturelle des candidats, a amené à un certain renouveau dans les approches pédagogiques et évaluatives. Si le champ de l'évaluation linguistique

certificative est actif, du côté des méthodologies et des approches d'apprentissage des langues mais aussi de l'évaluation mise en place dans un cadre pédagogique, on constate également, depuis quelques années, un certain renouveau. Aussi convient-il de s'y arrêter. Même si les champs de l'évaluation linguistique en contexte d'apprentissage de la langue et hors apprentissage ne sont pas toujours reliés, cela a des conséquences sur les orientations prises dans l'évaluation certificative. Après le succès des approches communicatives des années 1980-1990, les deux nouveautés majeures ont été l'apparition progressive de l'approche par les tâches et la création de l'approche actionnelle proposée dans le *Cadre européen commun de référence pour les langues – apprendre, enseigner, évaluer* (Conseil de l'Europe, 2001). Comme le déclare Christian Puren (2002b), et comme on peut le constater dans le tableau 1.1, l'approche communicative se distingue des autres méthodes par le fait qu'elle met en œuvre la capacité des apprenants à « échanger ponctuellement des informations avec des étrangers » ou encore par le fait qu'elle privilégie des activités sociales consistant à « parler avec / agir sur » et qu'elle propose comme tâches « scolaires » (conçues pour la classe) des jeux de rôles ou encore des actes de paroles (Puren, 2002b).

Tableau 1.1 : aperçu historique de la correspondance entre objectif social de référence et la tâche scolaire de référence dans les différentes approches didactiques de la langue culture (Puren, 2002b : 9)

méthodologie		1. méthodologie traditionnelle	2. méthodologie active	3. approche communicative	4. perspective co-actionnelle co-culturelle
objectif social de référence		compréhension des grands textes de la littérature étrangère	accès à tous documents culturels en langue étrangère	échanges ponctuels avec les étrangers	réalisation commune d'actions sociales
perspective actionnelle	*opération*	traduction	explication	interaction	co-action
	moyen	reproduire	parler sur	parler avec/agir sur	agir avec
perspective culturelle	*type*	universaliste	civilisationnelle	interculturelle	co-culturelle
	orientation	valeurs	connaissances	représentations	conceptions

Évidemment, cette vision défendue par le chercheur est à nuancer. D'une part, elle vise essentiellement l'enseignement en milieu « scolaire ». D'autre part, on ne saurait limiter l'approche communicative, telle que pratiquée dans les salles de classe du monde entier, à une seule grille d'analyse. Il s'agit d'une approche pédagogique, pas d'une méthode. Or, la différence entre une méthode et une approche pédagogique, est que l'approche,

contrairement à la méthode, ne propose pas l'unicité. L'approche propose une pluralité des pratiques pédagogiques à l'intérieur d'un cadre théorique composé de principes généraux. Pour Richards et Rodgers (2001 : 172), l'approche communicative se distingue de la sorte :

> « - Les apprenants apprennent la langue en l'utilisant pour communiquer,
> - une communication authentique et pleine de sens doit être le rôle des activités de classe,
> - l'aisance est une dimension importante de la communication,
> - la communication met en jeu l'intégration de différentes compétences langagières,
> - l'apprentissage est un processus de construction créative et est fait d'essais et d'erreurs ».

Concernant l'apparition de l'approche actionnelle, voici la description qu'en fait Nilsen (2003 : 18), dans sa thèse de doctorat :

> « Une nouvelle approche didactique en langues se profile aujourd'hui, qui est esquissée dans les travaux récents du Conseil de l'Europe (2001), et clarifiée par Puren (2000, 2001a, 2001b, 2002a, 2002b, 2003). Cette approche ne s'oppose pas à l'approche communicative, mais en diffère en plusieurs points. Elle garde la description du « savoir dans la langue » sous forme d'une grille de compétences qui sont : lire, écrire, écouter, parler et communiquer. Mais la communication n'est ici qu'un moyen pour réaliser une tâche (sociale). L'action et l'interaction ne sont plus des fins en elles-mêmes, mais visent la réalisation d'un résultat en langue étrangère. C'est au cours de ses actions et interactions que l'apprenant est en contact avec des documents oraux ou écrits authentiques, c'est avec d'autres « acteurs sociaux », qu'il pratique son expression et qu'il apprend. ».

S'il s'agit d'un bon résumé de ce que l'on peut trouver dans le *Cadre européen commun de référence pour les langues – apprendre, enseigner, évaluer* (Conseil de l'Europe, 2001) , mais aussi, des réflexions méthodologiques menées par Christian Puren, un constat s'impose : bien que l'approche actionnelle semble naître avec la parution du *Cadre européen commun de référence pour les langues* (Conseil de l'Europe, 2001), C.E.C.R., elle n'a pas, à ce jour, donné naissance à beaucoup de publications (autres que celles mentionnées par Nilsen (2003) et Puren (2004)) et à une définition qui permette de la distinguer radicalement des autres approches et méthodes. Puren (2004) consacre un article aux différences existantes entre l'approche par les tâches et l'approche actionnelle. Il déclare que dans les approches communicatives la communication est tout autant

l'orientation de la tâche et son résultat. Le problème est que les tâches ne sont pas uniquement communicatives. Elles peuvent être orientées vers la langue, le processus, la communication, le résultat, le produit, l'action ou encore la procédure. Il faut donc comprendre que le « péché » de l'approche communicative est d'avoir restreint le champ de la réalité aux seules tâches communicatives alors que dans la réalité les tâches sont beaucoup plus variées. Pour l'auteur, ce caractère restreint, mais aussi trop concret, de l'approche communicative la différencie de l'approche actionnelle qui est plus globale et donc plus utile. Et Puren (2004 : 14) de dire : « Plus une idée est concrète, et plus son utilité est réduite à l'environnement particulier dans lequel elle a été élaborée ».

Faut-il donc comprendre que l'approche communicative et l'approche par les tâches sont des cas particuliers de l'approche actionnelle et que ce qui différencie fondamentalement l'approche actionnelle de l'approche communicative, ou encore, l'approche par les tâches, c'est son caractère plus englobant ? Pour Puren (2004), dans la didactique des langues, on a trop souvent cherché à « optimiser les ressources », plus qu'à développer leur adéquation aux apprenants et aux environnements dans lesquels elles s'inscrivent. Si pour Puren, l'approche actionnelle est une approche qui s'adapte à différentes situations, c'est aussi ce que dit le chapitre 2.1 du *Cadre européen commun de référence pour les langues – apprendre, enseigner, évaluer* (Conseil de l'Europe, 2001 : 15). Voici comment ce cadre définit son propre rôle et comment il définit l'approche actionnelle, l'apprenant, les activités langagières, l'action et la tâche :

> « Un Cadre de référence pour l'apprentissage, l'enseignement et l'évaluation des langues vivantes, transparent, cohérent et aussi exhaustif que possible, doit se situer par rapport à une représentation d'ensemble très générale de l'usage et de l'apprentissage des langues. La perspective privilégiée ici est, très généralement aussi, de type actionnel en ce qu'elle considère avant tout l'usager et l'apprenant d'une langue comme des acteurs sociaux ayant à accomplir des tâches (qui ne sont pas seulement langagières) dans des circonstances et un environnement donnés, à l'intérieur d'un domaine d'action particulier. Si les actes de parole se réalisent dans des activités langagières, celles-ci s'inscrivent elles-mêmes à l'intérieur d'actions en contexte social qui seules leur donnent leur pleine signification. Il y a « tâche » dans la mesure où l'action est le fait d'un (ou de plusieurs) sujet(s) qui y mobilise(nt) stratégiquement les compétences dont il(s) dispose(nt) en vue de parvenir à un résultat déterminé. La perspective actionnelle prend donc aussi en compte les ressources cognitives,

> affectives, volitives et l'ensemble des capacités que possède et met en œuvre l'acteur social.».

Dans l'esprit de Puren (2004) et des auteurs du cadre de référence, l'approche actionnelle ne vient donc pas se substituer à l'approche communicative mais vient apporter de nouvelles possibilités. A la lumière des descriptions des méthodes d'enseignement des langues étrangères que font Richards et Rogers (2001), l'approche actionnelle est essentiellement une approche utilisant le concept de tâches, dans le prolongement des approches communicatives, ainsi que le découpage en compétence de l'approche par compétences (Competency-Based language teaching). En effet, elle utilise le concept de tâche et propose de programmer les enseignements à partir de référentiels des composantes de la langue (grammaire, lexique, phonétique, pragmatique). L'évaluation, dans l'approche actionnelle, utilise des grilles mesurant le degré de maîtrise des macro-compétences (compréhension écrite, orale, expression écrite et orale) au travers de la mesure ou de l'appréciation de la maîtrise des composantes de la langue.

Dans la pratique enseignante, même si on aura besoin de plus de recul pour comprendre ce que l'approche actionnelle a changé ou changera, une brève analyse des dernières méthodes publiées en français langue étrangère permet de deviner que le véritable changement porte sur un rééquilibrage entre l'écrit et l'oral (privilégié dans l'approche communicative telle que proposée par les manuels). Là où les approches communicatives utilisaient la communication à partir d'actes de parole, l'approche actionnelle privilégie l'utilisation de tâches demandant la mise en place d'actions (soit en juxtaposant des macro-compétences linguistiques (compétences de réception et de production orale et écrite traitées séparément), soit en proposant une véritable intégration des composantes de la compétence linguistique). Dans l'approche actionnelle, l'apprenant est envisagé comme un véritable utilisateur de la langue et non pas uniquement comme un utilisateur scolaire. L'approche actionnelle, même si elle propose des tâches appartenant à des familles de tâches, qu'elle tend à se détacher comme l'approche communicative du contexte particulier de la communication, est tout de même très ancrée dans la problématique propre à sa sphère géographique, soit celle de l'Europe, du plurilinguisme, de la maîtrise partielle des langues. De ce point de vue, elle prolonge la vision de

l'économie linguistique qui animait déjà les approches communicatives telles que pratiquées en Europe. Pour ce qui est de l'évaluation, le C.E.C.R. est beaucoup moins original et novateur. Le chapitre 9, dédié à l'évaluation, s'attache avant tout à distinguer les différents types d'évaluation. Il permet de comprendre que les différentes possibilités doivent être adaptées au besoin qu'un test ou un dispositif d'évaluation se propose de satisfaire. L'approche retenue est celle de l'évaluation de la performance dans le cadre du concept de « *proficiency* » (maîtrise). Autrement dit, on veut évaluer la compétence linguistique par le biais de tâches, en se gardant de ne pas utiliser des tâches trop contextualisées afin de garder une certaine généralisabilité. Dans la pratique, on constate qu'en Europe, pour le moins, le changement le plus spectaculaire opéré depuis l'apparition du cadre est la généralisation de l'évaluation des quatre macro-compétences pour tous les niveaux. On peut penser que le C.E.C.R. a popularisé le recours systématique aux macro-compétences (qu'elles soient évaluées séparément ou dans des tâches dites « intégrées »).

L'approche par les tâches, quant à elle, est à la fois une approche pédagogique et une perspective d'évaluation aux contours « flexibles » (Norris, 2002 : 338). L'approche par les tâches (Richards & Rodgers, 2001 ; Zanón, 1999), est une approche qui met l'accent sur le processus plus que sur le produit. Elle vise un apprentissage de la langue par le biais de tâches, l'interaction avec les autres apprenants, en fonction d'un but (la réalisation de la tâche complète ou partielle). Contrairement à l'approche par les compétences ou encore l'approche actionnelle, la progression n'est pas faite en fonction d'un référentiel de grammaire ou de vocabulaire ou encore selon une échelle retraçant dans des termes généraux le développement de la compétence langagière, mais par la difficulté intrinsèque des tâches (on verra dans la recension d'écrit que ce concept est loin d'avoir fait la preuve de sa validité). Côté évaluation, sans vouloir, là encore, entrer dans le détail, il est possible de dire que cette approche évalue la performance au travers de la réalisation de la tâche. Les changements perceptibles, par rapport aux approches communicatives, sont que le candidat est évalué au travers d'une tâche dont le but n'est pas uniquement communicatif et que les critères utilisés pour l'évaluation des candidats sont des critères de la « vraie vie ». D'après Norris (2002), l'évaluation par les tâches

n'est pas nouvelle. La notion était déjà présente dans les tests communicatifs et dans l'évaluation de la performance. Les Européens (Puren, 2004 ; Zanón, 1999) donnent pour origine à l'approche par les tâches, les travaux de Nunan (1989) et de Candlin (1987). Si certains précisent que l'évaluation par les tâches est une évaluation scolaire et qu'elle n'a pas pour vocation de remplacer d'autres types d'évaluation (Norris, Brown *et al.*, 2002 ; Zanón, 1999), c'est sans doute parce que, pour les tests, l'évaluation par les tâches est loin d'être facile à mettre en œuvre. C'est aussi parce que, traditionnellement, on oppose l'évaluation de type « *achievement* » (évaluation de l'atteinte des objectifs d'un cours) à l'évaluation « *proficiency* » (visant à placer sur une échelle la performance d'un candidat. Quand l'approche par les tâches est appliquée à la lecture (Long, 2003), elle consiste à proposer des textes et des tâches par ordre de difficulté ou de complexité, comme le proposait déjà en 1978 Widdowson (1981 : 105-108) avec un processus dit « d'approximation graduelle » (on propose un premier compte rendu simple d'un texte à l'apprenant, puis un plus compliqué et enfin le texte original). L'objectif est d'accompagner le développement de la compétence en permettant au lecteur de mettre en œuvre progressivement ses habiletés de lecture et ses connaissances antérieures pour développer sa compétence de lecture. On postule que cet accompagnement permet de faciliter le développement de la compétence parce que le processus d'acquisition est fait de manière intégrée. Les habiletés de lecture et les connaissances linguistiques ne sont pas sollicitées séparément mais dans un ensemble. De même, l'*input* (le texte), parce qu'il n'est pas trop difficile pour le lecteur, permet la mise en œuvre du processus d'acquisition de la compétence de lecture. Aujourd'hui, certains chercheurs s'orientent vers l'établissement d'une taxonomie des tâches (ou encore des textes). *In fine*, le but est de proposer des tâches classées selon leur « complexité » ou difficulté pour permettre aux apprenants d'acquérir du savoir. Mais pour une même tâche, l'enjeu est aussi d'être capable de proposer des sous-tâches qui soient orientées du plus facile vers le plus difficile. Cette exigence, propre à l'approche par les tâches (une tâche simple permet d'acquérir un niveau de compétence qui sera réutilisé pour réussir une tâche d'un niveau plus élevé jusqu'à réalisation complète et sans artefact pédagogique d'une tâche de la « vie réelle »), porte les chercheurs à tenter d'identifier les facteurs qui contribuent à la complexité. Cet intérêt pour la notion de difficulté est accompagné de nouvelles

interrogations et de positionnement vis-à-vis de la notion de « texte authentique ». Long (2003) explique ainsi que si le texte authentique présente un niveau de difficulté trop élevé pour un candidat de niveau faible, voire de niveau moyen, on peut tenter de lui proposer non pas des textes plus faciles mais « explicités ». Le but est de proposer une tâche qui soit dans la « zone proximale de développement » de l'apprenant (Vygotsky, 1985), pour lui laisser l'opportunité d'acquérir du savoir par la tâche et non par le métalangage. Par exemple, plutôt que de remplacer un mot de vocabulaire inconnu par un autre plus fréquent (simplification), on va proposer une paraphrase (explicitation). L'enjeu est de pouvoir placer des tâches appartenant à des familles de tâches sur une échelle de difficulté et ainsi d'obtenir une séquence autorisant l'acquisition de la langue sans autre recours que celui des tâches. D'autres proposent de créer des « sous-tâches » ou tâches partielles. Dans de tels contextes d'apprentissage, on comprend que le texte « purement » authentique ne soit plus aussi prisé que dans le contexte de l'approche communicative. Pourtant, ce point est loin de faire l'unanimité parmi les chercheurs. Ainsi Alderson (2000 : 81) voit-il dans ces modifications le danger de complexifier les textes inutilement :

> « Given the (not very strong) evidence for the impact of linguistic variables, like knowledge of syntax, on first-language reading, test designers should examine carefully the language of questions, rubrics and texts to ensure that they fall within the test population's likely ability range. Although one possible strategy, if texts are found to be too difficult for a given group of learners, is to simplify the texts, not only does this disauthenticate the text, it also risks making the text harder to understand. In addition, an ability to read simplified texts is unlikely to generalise to an ability to read genuine texts. A more appropriate way to adjust for text difficulty might be to develop easier tasks or test questions [3]».

Alors que les partisans de l'approche par les tâches ont souvent étudié les aspects cognitifs des tâches (notamment celle utilisées pour l'évaluation), d'autres s'orientent vers l'étude de leurs caractéristiques. Ce qui est commun à ces deux points de vue, c'est le fait que l'on cherche à comprendre en quoi les tâches sont « adaptées » au niveau de compétence du candidat. On cherche à mettre au point des grilles de descripteurs prenant en compte les caractéristiques des tâches que ces caractéristiques soient, pour les uns,

cognitives, pour les autres, physiques (longueur du texte, complexité syntaxique, vocabulaire, type d'activités à faire à partir du document,...). On a donc besoin de comprendre le fonctionnement des caractéristiques des tâches, de comprendre comment elles peuvent s'adapter au mieux au niveau des candidats et, pour l'évaluation, de permettre une meilleure standardisation ou encore comparabilité.

Enfin, il convient de signaler que les tâches utilisées peuvent être de nature différente (Bachman, 1990, 2002). Dans l'approche dite *Real life*, on utilise des tâches pour l'évaluation ayant des caractéristiques très proches voire similaires à des tâches de la vraie vie. On prétend ne pouvoir faire des inférences sur la compétence des candidats qu'à partir de tâches. En effet, on pense que les caractéristiques des tâches ont une influence sur la compétence du candidat. Dans l'approche dite *interactionnal/ability*, on utilise des tâches qui ne sont pas nécessairement identiques à celles de tâches réelles, mais qui permettent de faire des inférences sur la compétence du candidat. L'utilisation de tâches pas trop spécifiques (ou utilisées dans un contexte trop spécifique) doit permettre d'améliorer les généralisations faites à partir des résultats au test.

Aujourd'hui, que ce soit dans l'approche par les tâches ou dans l'approche actionnelle, de nombreuses voix se lèvent pour informer la communauté des chercheurs et les praticiens de la nécessité d'aller vers plus d'adéquation entre le type d'évaluation utilisé, les besoins et les objectifs à atteindre, et donc une plus grande adaptabilité que ce soit pour les méthodes d'enseignement (Richards & Rodgers, 2001 ; Swan, 2005) ou d'évaluation (Goulier, 2004 ; Puren, 2004 ; Weir, 2005a). Pour l'évaluation, le problème n'est pas tant de classer les outils d'évaluation en fonction de leur appartenance à une pédagogie ou une méthode, sinon de comprendre quels sont leurs avantages et leurs inconvénients et quel est leur degré d'adéquation et de congruence avec les objectifs poursuivis et les usages prévus (Messick, 1989 ; Dassa & Laurier, 2003). Il convient de ne pas oublier que, souvent, l'évaluation n'est pas liée à des enseignements ou à des contenus ou approches d'enseignement. Dans ce cas, elle ne doit pas les prendre comme domaine de référence. Au contraire, elle doit affirmer ses domaines d'ancrage (compétences linguistiques d'un public de migrants ou encore d'un public de professionnels travaillant en langue seconde).

1.1.8. Les tests adaptatifs par ordinateur

Aujourd'hui, il est légitime de penser que les tests « papier-crayon » ne correspondent plus nécessairement aux besoins actuels. Ces tests, dans l'usage que l'on en fait traditionnellement, ne proposent pas d'adapter le contenu de l'examen au candidat pendant la passation. Ils se prêtent assez difficilement à l'adaptabilité. Si les « diplômes » de langue présentent une certaine adaptabilité puisqu'ils certifient des niveaux de compétence linguistique spécifiques, c'est beaucoup moins vrai pour les tests situant les candidats sur une large échelle de compétence. Toutefois, depuis quelques années, on constate que certains tests sont spécifiquement conçus pour évaluer un niveau particulier, notamment le niveau linguistique nécessaire pour entrer dans une université[d]. Si ces tests ne s'adaptent pas au niveau des candidats en cours d'administration, en revanche, ils ont des contenus adaptés à des besoins ou des niveaux spécifiques.

Les tests adaptatifs par ordinateur (mais aussi les tests adaptatifs « papier-crayon ») visent à proposer des items au candidat au plus proche de son niveau. Parce qu'ils utilisent des items pré-calibrés, stockés dans une banque, ils peuvent estimer le niveau de compétence du candidat en temps réel. Ils peuvent lui proposer des items correspondants, entre autres choses, à la dernière estimation de son niveau. Pour choisir les items, des contraintes sont préalablement fixées. Elles peuvent porter sur le contenu du test mais aussi sur la précision de la mesure ou encore selon des règles de passation (règle de départ, de continuation et d'arrêt du test).

Van der Linden (2005) écrit que les tests adaptatifs par ordinateur sont idéaux pour l'estimation du niveau du candidat. En revanche, ils proposent un contenu différent selon le niveau des candidats. Parfois, le critère de sélection du prochain item n'est guidé que par l'atteinte de la précision de la mesure et non par l'atteinte minimale d'un contenu d'examen. Pour éviter une trop grande distorsion dans les versions des tests proposés, il est possible, grâce à des algorithmes, d'astreindre le choix du prochain item à des contraintes autres que celles du niveau de précision de la mesure. Toujours dans ce but, on peut sélectionner une combinaison de n items jusqu'à ce que l'on trouve la meilleure

[d] Le T.O.E.F.L. nouvelle formule propose ainsi des tâches dites intégrées visant essentiellement le niveau requis en anglais pour suivre des cours dans les universités anglophones. En Allemagne, le testDAF n'évalue que les compétences de niveau intermédiaire et avancé. En France, le Test de connaissance du français (TCF) est proposé, pour les épreuves d'expression écrite, dans une version universitaire (TCF-DAP) et est destiné à recruter les candidats aux études en France pour une première année universitaire.

combinaison possible, tant au niveau de la mesure que du contenu. Le but ne consiste donc plus uniquement à trouver le niveau exact du candidat mais également à respecter des contraintes de contenu. Comme le signale Van der Linden (2005), les choix que l'on fera pour l'administration des tests adaptatifs par ordinateur ne sont pas neutres et leurs conséquences non plus. Sireci (1998 : 111) met en garde les chercheurs et les praticiens, contre des tests par ordinateur qui ne privilégieraient que la précision de la mesure au détriment du contenu de l'examen :

> « […] computerized testing and item selection algorithms threaten representation of the content domain if item selection decisions are based solely on statistical indices of item quality (e. g., item difficulty, item discrimination).[4] »

Le problème pour les tests adaptatifs par ordinateur est que leur mise en place nécessite beaucoup de moyens. On ne peut raisonnablement les utiliser que pour des cohortes de candidats très nombreuses. Or, ce type de cohortes, notamment dans le cadre de tests de langue d'envergure internationale, se distingue de par son caractère multiculturel. Les échantillons, qui servent de base pour calibrer les items, sont loin d'être culturellement et linguistiquement homogènes. On peut alors être confronté à des problèmes de validité d'interprétation des résultats (Hambleton, Swaminathan & Rogers, 1991). Brown et Iwashita (1996) trouvent ainsi, comme résultat à leur recherche, que les candidats sont classés différemment selon qu'on les classe avec des items dont le niveau de difficulté est calculé à partir des réponses d'un groupe culturel ou d'un autre.

L'enjeu pour les tests adaptatifs par ordinateur est de trouver un juste milieu entre une évaluation ne portant que sur le niveau du candidat et une évaluation considérant d'autres contraintes comme le contenu. Il s'agit aussi de vérifier que la population que l'on traite présente bien un taux d'homogénéité suffisant (ou que la difficulté des items est comparable pour l'ensemble des candidats). Il s'agit*, in fine,* d'assurer la comparabilité des résultats. Toutefois, les tests adaptatifs par ordinateur devront aussi veiller à ne pas privilégier la complexité (la prise en compte de facteurs ou encore de paramètres trop nombreux), au détriment de « l'interprétabilité » des résultats (ou encore la parcimonie du modèle). C'est pourquoi, il semble utile d'enquêter sur les caractéristiques et les propriétés des tâches d'évaluation avant de les placer dans une banque d'items.

1.1.9. La difficulté objective et la difficulté perçue par les candidats : pour une plus grande adaptabilité des tests

Enfin, une autre facette de l'évaluation peut être prise en considération pour vérifier l'adéquation (ou « l'adaptabilité ») entre les tests proposés et les candidats. La perception de la difficulté d'un test ou plus généralement d'une activité d'évaluation par les candidats ne laisse pas les chercheurs indifférents. Elle suscite de nombreux débats (ce thème sera plus largement développé dans la recension des écrits) parfois rattachés au concept de validité apparente. La validité apparente (« *face validity* », Bachman, 1990 : 285-289), depuis des décennies, provoque des polémiques. Pour certains, elle participe de la validité des tests, pour d'autres, elle n'y participe pas. On peut raisonnablement penser que la prise en compte de la perception des candidats quant à la difficulté des activités d'évaluation est utile. Elle peut permettre de proposer des tests plus adaptés au profil des candidats ou encore de mieux comprendre comment un examen est perçu. Cela peut encore donner l'occasion de mieux informer les candidats quant aux tenants et aux aboutissants d'un test ou des tâches d'évaluation. Il s'agit, en demandant leur opinion aux candidats, de consigner de l'information qualitative, là où les tests ne proposent que des données quantitatives. L'opinion des candidats doit inciter les concepteurs, lorsque c'est possible, à mieux adapter le contenu des tests.

A des fins de clarification, on entend par difficulté objective, la difficulté reflétée par des indices de probabilité ou encore indices issus des modèles de mesure comme le modèle de Rasch. La difficulté perçue, renvoie à la perception qu'ont les candidats d'un test après la passation ou encore à celle des concepteurs de test.

1.2. Contexte particulier

1.2.1. Un test de classement adaptatif par ordinateur du M.I.C.C.

Au Québec, la réalité linguistique dicte des besoins d'apprentissage et d'évaluation de la langue française importants. Alors que la province propose des cours de français aux immigrants allophones, le besoin de positionnement ou encore de certification du niveau de langue sont importants. Pour un immigrant, l'évaluation de son niveau linguistique est

à la fois un problème de reconnaissance de ses compétences vis-à-vis des autres immigrants mais aussi du groupe des québécois francophones.

La présente recherche s'inscrit dans le cadre de la politique de francisation des nouveaux arrivants mise en place par le M.I.C.C. au Québec (Ministère de l'Immigration et des Communautés Culturelles). A ses débuts, elle s'inscrivait dans le cadre d'un projet d'un test adaptatif par ordinateur visant le positionnement des étudiants dans les cours de français langue seconde (Laurier, 2004). Ce test n'ayant pas vu le jour au moment de la mise en place de la recherche, il a fallu que le chercheur procède à des recadrages. Alors que les tâches d'évaluation de la lecture devaient être administrées par ordinateur, il a été décidé de les administrer sur papier-crayon et de respecter la logique adaptative en concevant un test et en sélectionnant des candidats de niveaux comparables. La véritable différence tient donc plus au mode d'administration qu'à des aspects véritablement conceptuels. En effet, dans la version informatisée, les étudiants devaient se voir proposer le test de lecture après estimation de leur niveau. Il n'était pas prévu « d'actualiser » l'estimation du niveau de lecture du candidat, en cours de passation avant la fin du test. De ce point de vue, il n'y a donc pas de différence entre ce qui aurait été une passation sur ordinateur et la passation du test qui a été faite sur papier. La présente recherche a donc pour objet de trouver des réponses à des questions qui se posent dans le testing adaptatif et ce, sans prendre en considération le contexte particulier de la passation du test.

Le test, conçu par le chercheur, répond à une optique de positionnement telle que prévue par le M.I.C.C. et la maquette du test proposée par Laurier (2004). La définition du test de positionnement de Dassa et Laurier (2003) a été privilégiée pour guider la conception des tâches. Il n'a pas pour ambition d'évaluer la maîtrise d'habiletés mais le niveau de développement de la compétence de lecture. Il n'est pas lié à des enjeux critiques pour les candidats. Si Laurier (1993 : 23) et Alderson (2005 : 4-12) expliquent qu'il n'est pas toujours aisé de distinguer les tests de positionnement des tests diagnostiques, pour la présente recherche, la distinction est conservée. La maquette du test prévue par Laurier (2004) traduisait la volonté de créer un test de positionnement et non pas un test diagnostique. Les tâches de lecture devaient être attribuées à chacun des candidats en

fonction d'une estimation préalable de leur niveau. Cette estimation devait être établie à partir des réponses déjà données dans le test. Aucune nouvelle estimation du niveau du candidat n'était prévue avant la fin de l'administration du test de lecture. Les candidats devaient donc répondre à plusieurs questions portant sur plusieurs textes avant que l'on réévalue leur niveau. En aucun cas, on ne cherchait à diagnostiquer des aspects particuliers, des « micro-compétences », des habiletés ou des connaissances.

Le chercheur indiquait clairement que le contenu du test n'était relié à aucun cours (Laurier, 2004). Il avait pour seule ambition de vérifier le niveau général d'usage de la langue française des candidats dans chacune des macro-compétences, telles qu'elles sont proposées dans les *Niveaux de compétence en français langue seconde pour les immigrants adultes* (M.R.C.I., 1998). Ce test devait estimer le niveau de maîtrise de la langue française des candidats, et non diagnostiquer leurs problèmes d'apprentissage. De même, il devait être conçu pour être rapide. Il devait viser l'amélioration de l'efficacité du positionnement du public des immigrants dans les cours de français. Dans de telles conditions, ceux qui seraient tentés d'utiliser ce test de positionnement à des fins diagnostiques feraient une erreur. Étant donné que les cours du M.I.C.C. sont organisés en fonction des niveaux de compétence générale en français, que les groupes sont constitués en moyenne de 18 apprenants, les professeurs n'auraient pas le temps de véritablement exploiter une information diagnostique. Cela dépasserait très certainement la capacité du professeur à traiter une telle somme d'informations. Par ailleurs, les cours étant intensifs, le diagnostic posé en début d'apprentissage perdrait très rapidement de sa valeur.

A l'origine, la mise en place d'un test de positionnement pour les candidats aux cours de français répondait à des besoins exprimés par le M.I.C.C. Elle correspondait à l'accélération de l'apprentissage du français des nouveaux arrivants, au besoin général d'une plus grande adaptation et à l'amélioration des moyens d'évaluation tels qu'exprimés dans le *Plan d'action 2004-2007* (M.I.C.C., 2004 : 68) du M.I.C.C. :

> **« Les objectifs poursuivis**
>
> L'intensification et l'adaptation des programmes de francisation pour les nouveaux arrivants s'inscrivent dans les objectifs gouvernementaux plus larges d'affirmation du français comme langue commune de la vie publique, favorisant de ce fait la rencontre des cultures. La connaissance adéquate du français constitue un catalyseur pour l'intégration des nouveaux arrivants à la société et représente même un préalable à l'égalité en emploi. En effet, la langue constitue un facteur clé d'intégration et de participation puisqu'elle favorise l'accès non seulement au travail, mais aussi à la vie sociale et à la vie culturelle du Québec. Le partage du français comme langue commune des Québécois de toutes origines contribue aussi à l'enrichissement du patrimoine commun. Les cours de français seront d'autant plus efficaces que l'accès aux services sera rapide, que les contenus de formation seront plus poussés et que les services seront mieux adaptés aux différentes clientèles. […] Pour assurer la pérennité du fait français au Québec, les objectifs à atteindre sont les suivants :
>
> - Accélérer l'apprentissage du français,
> - Accroître la francisation en adaptant les services aux besoins.
>
> **[…]**
>
> **Objectif 9 Accroître la francisation en adaptant les services aux besoins**
>
> Mesure 9.1
>
> - Améliorer les stratégies pédagogiques, les moyens d'évaluation et les services pour mieux répondre aux besoins d'intégration en emploi et au profil des clientèles […] ».

Même si le projet de test de positionnement n'était pas directement mentionné dans le *Plan d'action 2004-2007* (M.I.C.C., 2004), il répondait, non seulement, à la problématique de l'adaptabilité, mais encore, au souci de l'accélération de l'apprentissage du français. Le problème identifié était donc celui de l'uniformité des procédures et, finalement, de l'efficacité globale du processus de francisation.

1.2.2. Présentation de la recherche : principes et choix

Pour cette recherche, le choix a été fait de travailler sur des tâches de lecture en français langue seconde pour un public d'adultes immigrants. Un test composé de tâches de différentes natures a été conçu pour étudier leur comparabilité et l'opportunité de les utiliser dans un test de positionnement adaptatif. Le but était de comprendre les

différences de fonctionnement entre des tâches de lecture en français langue seconde, de niveaux différents, mais aussi des caractéristiques différentes. Le test, qui a été conçu pour l'occasion, est composé de trois tâches dites « discrètes » (composées chacune d'un texte et de cinq questions) et de trois autres tâches dites « intégrées » (réunissant trois textes concaténés et cinq questions chacune). Toutes ces tâches ont été développées en vue d'une utilisation avec un groupe d'immigrants multilingues, ayant un niveau linguistique comparable. A l'aide d'un questionnaire, on a cherché également à savoir quelle perception les candidats avaient de la difficulté des tâches.

1.2.3. Problème général, pertinence scientifique et sociale de la recherche

Le problème général auquel s'intéresse cette recherche est de savoir s'il est possible de proposer des tâches pour évaluer la lecture en français langue seconde plus variées et adaptées à des candidats de langues maternelles différentes, en vue d'une utilisation éventuelle dans un test adaptatif par ordinateur, visant le positionnement « d'apprenants émigrés », dans des cours de français. Du point de vue scientifique, l'objectif consiste à formuler un diagnostic sur l'interaction entre différentes tâches et différents groupes de candidats pour définir des directions à emprunter ou non dans les tests adaptatifs par ordinateur visant le positionnement dans des cours de langue. L'objectif social est de fournir une analyse permettant de donner des directions quant à l'usage de divers types de tâches de lecture pour proposer une évaluation, à la fois, diversifiée quant à son contenu et juste pour des candidats dont la « multiculturalité » peut faire craindre l'apparition de biais culturel. *In fine*, il s'agit de proposer des outils ou plus modestement des analyses permettant de faciliter l'apprentissage du français.

1.2.4. Problème spécifique et questions de recherche

« Traditionnellement », dans les tests de langue, et, plus spécifiquement, pour l'évaluation de la compréhension écrite, il est plus fréquent d'utiliser des tâches discrètes que des tâches intégrées. Ceci étant, récemment, on a vu apparaître des tests et autres diplômes proposant des tâches dites « intégrées ». Ces tâches utilisent souvent la compréhension écrite (ou orale) comme « *input* » (intrant). Elles ont pour « *output* »

(extrant) l'expression orale ou encore l'expression écrite. Par ailleurs, du fait de la multiplication des sources d'information, de l'Internet, les tâches de lecture se font de plus en plus à partir de plusieurs documents (Goldman, 1997 : 358). Le « travail » du lecteur consiste alors à intégrer l'information à partir de plusieurs sources ou encore de faire le lien entre différents documents. Or, il est rare que les tests actuels utilisent des tâches « intégrées » pour évaluation de la compétence de lecture seule, sans que l'évaluation ne se fasse par la mise en œuvre de compétences de production. Si l'utilisation de textes multiples était présente dans le projet T.O.E.F.L. 2000 (Enright *et al.* 2000 : 5), le mode de réponse choisi était l'expression écrite ou orale. Ce type de tâches intégrées, très spécifiques et demandant la correction d'une production (orale ou écrite), nécessite des moyens financiers et du temps. Ces tâches ne sont donc pas adaptées aux tests visant le positionnement de cohortes de candidats dans des cours de niveaux différents. Utiliser des tâches de lecture intégrées ne faisant pas appel à des corrections trop coûteuses et permettant d'évaluer une compétence de lecture générale est une avenue qui demande à être explorée. C'est d'autant plus vrai si l'on travaille avec un test visant le positionnement, sans visée diagnostique, mais dont une meilleure validité de contenu est recherchée.

Si l'on veut statuer sur l'opportunité d'utiliser des tâches intégrées de compréhension écrite dans le cadre d'un test de positionnement adaptatif par ordinateur pour des cohortes de candidats multiculturels, il faut proposer des tâches qui permettent d'étudier les qualités des tâches discrètes et intégrées. Il convient de savoir si ces deux types de tâches évaluent une compétence générale de compréhension écrite ou bien si elles évaluent une dimension secondaire. Pour les tâches intégrées, il se peut que les corrélations entre les questions liées à un passage soient plus élevés ce qui reflète une dimension supplémentaire reliées à chacun des passages. Dans les tâches discrètes, il est possible de penser que l'ensemble des questions constitue une banque de questions relativement interchangeables qui se distinguent essentiellement par leur niveau de difficulté.

Il faut alors tenter de découvrir s'il existe (pour des candidats de niveau comparable) des interactions entre le type de tâche et les groupes particuliers de candidats. Il convient de comprendre si ces interactions sont les mêmes pour tous les types de tâches et pour tous les groupes de candidats. Analyser la perception des candidats, quant à la difficulté des

tâches, permettra de comprendre, d'une part, quelle perception ils en ont et, d'autre part, en quoi cette perception peut différer. Comprendre ces interactions (et les décrire) devrait permettre de proposer des tâches et des tests en adéquation avec les préoccupations liées au contenu des tests et au respect de la diversité culturelle des candidats.

1.3. Questions de recherche

L'objectif général de la recherche est d'étudier le fonctionnement de tâches discrètes et intégrées de trois points de vue distincts :

- la dimensionnalité,
- la difficulté objective,
- la difficulté subjective.

On pose comme hypothèse que la proximité linguistique permet d'expliquer les variations.

Les questions de recherche sont les suivantes :

- **Question 1** : Les épreuves composées de tâches de lecture discrètes et celles composées de tâches plus intégrées sont-elles unidimensionnelles ?

Les sous-questions qui correspondent à cette question sont les suivantes :

- Peut-on utiliser toutes les tâches pour les placer sur une échelle de mesure unique ?
- Y a-t-il une différence de dimensionnalité entre les tâches de lecture discrètes et intégrées ?

- **Question 2 :** Pour des candidats de niveau de compétence équivalent, des tâches pour l'évaluation discrètes et intégrées de niveaux différents (mais suffisamment proche du niveau des candidats), présentent-elles un niveau de difficulté homogène pour tous les candidats ?

Les sous-questions qui correspondent à cette question sont les suivantes :

a) Lequel des deux types de tâches, soit les tâches discrètes et les tâches intégrées, est plus difficile (ou plus facile) pour des candidats de même niveau ?
b) Comment le classement des candidats varie-t-il selon que l'on calibre leurs compétences avec l'ensemble des tâches ou selon chaque type de tâches ?
c) Y a-t-il des interactions entre le type de tâche et la famille linguistique à laquelle appartiennent les candidats ?

- **Question 3 :** Comment la difficulté des tâches et des différents types de tâches est-elle perçue par l'ensemble des candidats, et par les candidats regroupés selon leur appartenance à une famille linguistique ?

Les sous-questions qui correspondent à cette question sont les suivantes :

a) Lequel des deux types de tâches, soit les tâches discrètes et les tâches intégrées, est perçu comme plus difficile (ou plus facile) pour des candidats de même niveau ?
b) Les perceptions de la difficulté des tâches varient-elles en fonction de la famille linguistique à laquelle appartiennent les candidats ?

Chapitre 2 : Recension des écrits

La recension des écrits débutera par une présentation de la définition de la compétence langagière puis de la compétence de compréhension écrite. Pour ce qui est de la compétence de lecture, les thèmes de l'unidimensionnalité, du construit, des différentes visions de la difficulté et plus spécifiquement de la difficulté des tâches d'évaluation seront abordés. Enfin, une présentation de la distinction entre les tâches discrètes et intégrées sera proposée au lecteur. Après avoir évoqué les aspects liés au contenu des tests, l'emphase sera mise sur la présentation des modèles qui peuvent être utilisés pour mesurer la compréhension écrite. En particulier, seront présentés les modèles de la théorie classique et de la théorie de réponse aux items, leurs propriétés, leur fonctionnement et l'utilisation que l'on peut en faire. Pour finir, une synthèse des points abordés précédemment sera proposée afin de faire le lien avec les aspects liés aux tests adaptatifs par ordinateur. Cette partie de la recension des écrits traitera plus spécifiquement de l'évaluation de la compréhension écrite dans le contexte d'une évaluation adaptative. Après la présentation des caractéristiques des tâches et des textes, la recension s'achèvera par une facette de l'évaluation, touchant plus directement les caractéristiques des candidats, celle de la perception subjective de la difficulté des tâches. Cet aspect est présenté à la fin de la recension des écrits car, s'il est important, il ne participe pas à la validité du test ou à la définition du construit.

2.1 L'évaluation d'un objet multi-facettes, la compétence langagière

Depuis des années, dans l'évaluation linguistique, les chercheurs prennent en compte les interactions entre les caractéristiques des candidats, les tâches d'évaluation, ou encore, les conditions de passation (Bachman, 1990). Pour comprendre l'origine de cette évolution, il convient de tracer un bref historique de l'évaluation de la compétence langagière et des méthodes utilisées. En 1961, Lado et Caroll définissent un cadre qui distingue quatre macro-compétences (compréhension écrite, orale et expression écrite, orale) et des composantes du langage (grammaire, vocabulaire, phonologie et graphologie). Cependant, comme le signale Bachman (1990), ce cadre ne permet pas de comprendre si

les compétences sont autre chose qu'une simple manifestation des composantes du savoir. Bachman (1990) propose donc une nouvelle vue de la compétence langagière, cette fois-ci, une vue communicative. Le résultat de recherches empiriques menées par le chercheur montre qu'il y a trois composantes qui constituent la compétence communicative :

- la compétence langagière qui est un ensemble de savoirs qui sont utilisés dans la communication via la langue,

- la compétence stratégique qui est une capacité mentale permettant d'utiliser les composantes de la compétence langagière dans une vision communicative,

- les mécanismes psychophysiologiques qui sont des processus neurologiques et psychologiques impliqués dans l'exécution de la langue.

En 1996, Bachman et Palmer retiennent l'hypothèse selon laquelle la compétence langagière se divise en deux facteurs principaux :

- la compétence organisationnelle (compétences grammaticale et textuelle) et,

- la compétence pragmatique (compétences illocutoire et sociolinguistique).

L'importance de l'usage est donc affirmée. Bachman et Palmer (1996) relient cet usage à des facteurs affectifs. Les orientations prises, entre autres, par Bachman et Palmer (1996) et McNamara (1996) mettent en avant la performance. Toutefois, si les premiers auteurs insistent pour garder la notion de construit et de trait latent, McNamara (1996), lui, prend soin à distinguer des degrés de performance. Pour lui, la notion de performance peut être placée sur un continuum et être comprise de différentes manières. Dans un sens fort, la performance se réfère à l'exécution de tâches qui correspondent ou s'approchent des tâches « de la vraie vie ». La performance est jugée avec des critères se voulant représentatifs du monde réel. Ici, des facteurs non-linguistiques sont pris en compte. Dans un sens faible, la performance est focalisée sur la production langagière. On donne une tâche au candidat qui devra « représenter » une tâche de la vraie vie. Cependant, le but le l'évaluation n'est pas d'évaluer la capacité à réussir la tâche comme dans la vraie vie, mais la maîtrise de la langue seconde (on postule que la compétence est latente et qu'elle n'est pas directement observable). Pour l'auteur, distinguer « performance » au sens faible et

« performance » au sens fort est un bon moyen pour éviter d'affirmer que l'on évalue une compétence de la vraie vie quand on évalue une simple compétence langagière :

> « Performance tests in the weak sense seem to have had such convincing face validity that they have managed to convince language testers themselves that they were more than tests of language proficiency[5] » (McNamara, 1996: 44).

Comme cela a été signalé dans le premier chapitre, dans la mouvance de l'évaluation de la performance, depuis environ 10 ou 15 ans, un nouveau type d'évaluation est apparu, l'évaluation par les tâches. On pourrait qualifier ce nouveau type d'évaluation de cas particulier de l'évaluation de la performance (Bachman & Palmer, 1996 ; McNamara, 1996). Dans ce type d'évaluation, ce à quoi on s'intéresse, c'est avant tout à l' « output » (la production, le résultat). Ce nouveau type d'évaluation peut avoir différentes orientations. Pour les uns (Bachman & Palmer, 1996 ; McNamara, 1996 ; Nunan, 2004), la compétence langagière peut-être évaluée en macro-compétences (compréhension / expression orales et écrites) mais aussi en tâches. Pour les autres, elle doit être évaluée avec des tâches qui, avant tout, requièrent le recours au langage pour leur accomplissement et qui, enfin, sont identiques à celles que les gens font dans leur vie quotidienne (Norris *et al*., 1998). Long et Norris (2000) disent que l'approche par les tâches, dans cette optique, prend la tâche, elle-même, comme l'unité fondamentale d'analyse qui motive la sélection de l'item, la construction des instruments pour l'évaluation et l'évaluation de la performance à la tâche. Brown *et al.* (2002) précisent que la performance à ces tâches est évaluée selon des critères et des niveaux de critères du monde réel. Si cette approche ne constitue pas la seule nouveauté des dernières années (on a déjà évoqué le cas de l'approche actionnelle), en revanche, c'est celle qui a véritablement suscité des débats concernant l'évaluation.

2.2 Une définition de la compétence de compréhension écrite en langue seconde

Avant même de définir l'évaluation de la compréhension écrite en langue seconde, il convient de définir ce qu'est la compétence de compréhension écrite en langue seconde en général. Si, depuis des années, les chercheurs tentent de lui donner une définition

satisfaisante, il faut bien avouer que la tâche est ardue et que cette compétence est encore l'objet de nombreuses recherches.

2.2.1 Différence entre langue maternelle et langue seconde

Alderson (1999, 2000) explique que depuis longtemps, on s'interroge pour savoir si la compréhension écrite en langue seconde est un problème de lecture ou bien plus directement un problème de langue. Bernhardt et Kamil (1995) montrent que le lien entre les aptitudes de lecture en langue maternelle et langue seconde n'est pas aussi fort qu'on peut le croire. Seulement 20% de la variance de la lecture en langue seconde est due aux compétences de lecture en langue maternelle. 30% de la variance est due à la connaissance de la langue seconde, mais surtout 50% de la variance reste inexpliquée. Cela représente encore une proportion importante. Bernhardt (2003) explique que l'on doit continuer à examiner la question de la variance résiduelle. Selon Koda (2005), la compétence de lecture en langue maternelle est le fruit de l'interaction intégrative de l'information présente dans les textes et les connaissances antérieures du lecteur. La compréhension se produit quand le lecteur extrait et intègre plusieurs informations du texte et les combine avec ce qu'il sait déjà. La compétence de lecture en langue seconde se distingue par le fait qu'on commence à la travailler à tout âge, que l'on peut avoir eu l'expérience de la lecture. Enfin, l'absence de savoirs linguistiques en langue seconde explique en partie l'impossibilité de lire en langue seconde. Selon Grabe (1999), ce qui distingue la compétence de compréhension écrite en langue maternelle et en langue seconde, c'est :

- la connaissance du vocabulaire (pour comprendre un texte, il faut comprendre 95% du vocabulaire),
- l'impact des textes authentiques sur les lecteurs (on ne sait pas si un texte authentique va réellement motiver la lecture),
- la conscience de la langue,
- l'aisance et la rapidité,
- le savoir culturel et,

- enfin, l'atteinte d'un « niveau seuil » (un niveau minimal en langue pour pouvoir lire le texte).

Selon Koda (2005), il est non seulement nécessaire de prendre en compte la distance inter-linguistique mais il est aussi nécessaire de prendre en compte la particularité des lecteurs. Leurs niveaux de lecture varient selon leurs compétences de lecture en langue maternelle, selon leurs niveaux linguistiques en langue seconde, et la manière d'apprendre à lire en langue maternelle et en langue seconde. On ne peut donc pas affirmer que le transfert est identique, pour toutes les compétences, pour tous les lecteurs. Si cette approche proposée par Koda permet d'étudier les différences entre les individus et les groupes linguistiques, elle semble plus opérante pour des évaluations privilégiant le diagnostic que celles privilégiant le positionnement ou la certification du niveau de compétence des lecteurs.

Si la distance interlinguistique est un concept difficile à manier, la distance intralinguistique, autrement dit entre une même langue, ou encore des variétés d'une même langue, elle l'est tout autant. C'est un concept multidimensionnel complexe à analyser. Éloy (2004a : 397) en propose une description qui laisse peu de place à la parcimonie et, *mutatis mutandis*, en empruntant le vocabulaire de Blais (1987 : 14), au « pouvoir explicatif » dont on a besoin pour un test de lecture :

> « L'appréhension des « distances » entre les langues ainsi conçues demande un outil multidimensionnel, faisant leur place aux dimensions suivantes.
>
> - degrés de différences phonologiques, morphosyntaxiques, lexicales (et, au plan pratique, la possibilité de maîtriser ces différences par des règles de conversion ; ces différences seront appréciées tant dans l'oral vernaculaire que dans l'écrit standardisé ; une question difficile reste celle de la pondération des différents niveaux d'analyse ;
>
> - possibilités d'intercompréhension des locuteurs. Il s'agit là d'un problème psycholinguistique autant que linguistique, qui met en jeu la psychologie (dispositions à comprendre et à se faire comprendre) et la microsociolinguistique (variabilité selon les situations), ainsi que les habitudes de contact des individus et des groupes ;
>
> - degrés d'institutionnalisation des langues : degrés de standardisation des pratiques quotidiennes, poids idéologique des standards, importance de l'enseignement ;

> - degrés de vitalité ethnolinguistique, incluant l'appréciation de la « conscience linguistique » ou conscience de l'individuation, qui résultent en pratiques « oppositives » (Éloy, 2004b), et des habitudes de mixage. »

Si dans le concept de « compréhension écrite en langue seconde », la notion de « compréhension écrite » ne parvient pas à faire l'unanimité, celle de « langue seconde » n'y parvient pas non plus. Ainsi Cummins (1991) remet-il en cause la notion de « langue seconde ». Il déclare que dans le cas des langues majoritaires et minoritaires, il est possible que des enfants apprennent à lire d'abord dans la langue seconde puis dans la langue maternelle. Ces enfants peuvent encore ne pas lire du tout dans la langue maternelle. Dans le cas des personnes bilingues, souvent la distinction « langue 1 » et « langue 2 » est loin d'être aussi simple qu'il y paraît. Pour eux, l'apprentissage du « français langue seconde » est peut-être en réalité un apprentissage en français langue « troisième ». Par ailleurs, les bilingues peuvent, à la fois, posséder une langue proche du français et une autre fortement éloignée. Dans ce cas, la question se posera de savoir si cette personne appartient à un groupe de langue proche du français, à un groupe éloigné ou bien à un groupe distinct. Enfin, dans un monde où les cultures sont de plus en plus au contact des autres, il n'est pas rare que des apprenants ou des candidats à un test soient en possession de compétences partielles dans diverses langues. Sans doute que ces personnes développent des compétences transversales consistant à apprendre diverses langues, diverses langues « secondes » et, ce, de manière simultanée.

Il faut donc faire attention à ce que le concept de « langue seconde » soit réellement uniforme pour tous les candidats. Une vision trop restrictive de la variable « langue seconde » pourrait amener à conclure que la langue maternelle influe sur le niveau linguistique, là où on aurait des biais culturels, des biais de scolarité, de sexe, etc. Pour le chercheur, il s'agit surtout de comprendre s'il est possible d'opérer des distinctions utiles et pertinentes dans la population des candidats à partir de l'indicateur « langue seconde ».

2.2.2 Le paradigme cognitiviste et constructiviste de la compétence de lecture

Alors que le statut des langues dites « premières » et « secondes » soulève encore bien des questions et que l'on sait que cette distinction n'explique qu'une partie de la variance de compétence de lecture, ce qui semble plus consensuel, c'est la vue constructiviste de la compétence de compréhension écrite (Alderson, 2000 ; Grabe, 1999 ; Koda, 2005). Dans cette perspective, les compétences de lectures sont inscrites dans un processus. Le lecteur inscrit ses compétences et ses stratégies de lecture dans des processus qui lui permettent d'atteindre son but de lecture et ainsi de construire « sa » compréhension des textes (Goldman, 1997 : 373). Selon Grabe (1999), lire est un processus rapide et stratégique. Cela requiert une activité interactive, se fait en fonction d'un but. Lire demande, également, des savoirs linguistiques et la connaissance du monde pour un sujet donné. Selon l'auteur, la capacité à faire des inférences à partir d'un texte, est un bon moyen de distinguer les bons des mauvais lecteurs. D'aucuns (dont Alderson, 2000) opinent que l'évaluation de la compréhension écrite passe par la prise en compte du processus. La remarque est pertinente pour peu que le type d'évaluation envisagé et les objectifs fixés s'y prêtent.

2.2.3 Les modèles de lecture

Si la lecture est vue par tous comme un processus rapide, interactif, fait en fonction d'un but (Alderson, 2000 ; Carr, 2006 ; Grabe, 1999 ; Koda, 2005), le consensus s'arrête ici. Ainsi, pour ce qui est des modèles de lecture (essentiellement, les modèles « *bottom-up* », « *top-down* », interactif et interactif compensatoire) les opinions sont–elles moins consensuelles. Alderson (2000 : 16-20) explique que l'approche « *bottom-up* » décrit le processus de la lecture comme une série d'événements concaténés qui va du déchiffrage jusqu'à la compréhension. Ici, le lecteur est passif, fait toujours la même chose, suit toujours la même séquence. Dans l'approche « top-down », les lecteurs activent un schéma de connaissances existant pertinent. Ils placent les informations récoltées dans le texte à l'intérieur de ce schéma. Plus le degré de pertinence du schéma est grand, plus la compréhension est élevée. Selon Alderson (2000 : 18), ni les modèles « *bottom-up* », ni

les modèles « *top-down* » ne permettent de décrire correctement le processus de lecture. Il préfère le modèle interactif de Grabe où chaque composante peut interagir avec une autre composante d'un individu et où les interactions entre les composantes compensent les déficits. Dans ce modèle, le processus est parallèle plus que concaténé. Un dernier modèle, dit « interactif compensatoire » de Stanovich (2000), stipule que le degré d'interaction entre les composantes dépend du déficit de connaissance des composantes d'un individu et les interactions entre les composantes pour compenser les déficits. L'existence de ce dernier modèle doit alerter les chercheurs sur la nécessaire attitude de prudence dont on doit faire preuve lorsque l'on parle de la compréhension de lecture en langue seconde. Il semble risqué de vouloir en faire uniquement un processus mécanique. Sans doute est-il bon de se souvenir que les interprétations d'un texte varieront selon les individus (Shohamy, 2001).

2.2.4 Les compétences de lecture

Dans l'espoir de décrire le processus dans lequel s'engage le lecteur, et parallèlement à l'étude des modèles précédemment cités, la recherche s'est intéressée aux compétences de lecture (autrement appelées micro-compétences). Selon Alderson (2000 : 10-11), il y a eu deux manières principales de penser le « découpage » de la compétence de compréhension écrite au cours de l'histoire. En 1968, Davis effectue une analyse factorielle des réponses aux questions de tests de lecture et analyse les « facteurs » pour en déduire des micro-compétences[a], et sous-compétences. Ce découpage en compétences permet d'isoler les compétences de compréhension écrite à tester, de diagnostiquer les problèmes de lecture et donc de proposer des outils apparemment utiles pour la construction des tests. En 1978, Munby propose une taxonomie théorique. Il lance alors l'idée d'un découpage en compétences qui permettait d'établir des axes transversaux entre les compétences. En 1993, Lumley critique cette classification qui, pour lui, est plus la vision d'un chercheur dans son laboratoire qu'un cadre fondé sur une étude empirique. Il explique que l'on peut demander à des experts de classer les items en fonction de la compétence utilisée pour y répondre. Après des conversations fréquentes entre les

[a] Les micro-compétences, dans ce travail, sont définies comme étant des éléments constitutifs des macro-compétences (soit les compréhensions orale et écrite, les productions orale et écrite). Lire un texte pour s'informer, lire pour faire des inférences, lire pour s'orienter,…, sont des micro-compétences de la compétence de compréhension écrite.

experts, on peut arriver à de hauts niveaux d'accord entre experts. Le problème, d'après Alderson (1990, 2000) et Alderson et Lukmani (1989), c'est que, bien souvent, quand on demande à des experts d'identifier la compétence qui est évaluée par un item (une question, une activité d'évaluation), l'accord n'est pas facile à trouver. Pour Lumley (1993), qui est plus optimiste sur la capacité des experts à repérer des compétences de lecture dans des items, ce genre d'accord reflète aussi le fait que ces experts aient pour projet initial de trouver un consensus et que bien souvent, ils le trouvent !

Selon Alderson (2000 : 11), une partie du processus de lecture met en jeu l'usage variable et simultané de différentes et supposées « habiletés ». Cependant, la distinction entre habiletés faciles et difficiles ne semble pas justifiée. Elle n'est repérable ni par des experts, ni par des instruments statistiques. L'analyse des compétences et son utilisation doit donc se faire uniquement en cas de besoin réel (pour un besoin de formation spécifique, pour une population spécifique, pour un test diagnostique). Si on suppose que de telles compétences existent, elles interagiront en fonction du but de lecture poursuivi, des caractéristiques de la tâche et des particularités personnelles des candidats.

2.3 L'évaluation de la compétence de compréhension écrite

2.3.1 Une évaluation unidimensionnelle

La tentative de découpage de la compétence de lecture en micro-compétences, correspond à deux visions du développement de la compétence linguistique. Soit l'acquisition des compétences est vue comme étant le fruit d'un développement naturel, issue d'une « taxonomie naturelle », soit ce découpage est perçu comme étant empirique et la taxonomie des micro-compétences varie en fonction des contextes, des apprenants et des activités d'évaluation. Faire le pari de l'une ou de l'autre de ces options n'est pas sans conséquences. Si certains chercheurs, plutôt cognitivistes, de l'approche par les tâches, ont fait le choix d'un ordre naturel des tâches hiérarchisées selon leur « complexité cognitive » (Norris *et al*, 2002 ; Skehan, 1996 ; Robinson, 2001), d'autres ont fait le choix de préférer l'ordre empirique établi à partir de la difficulté (Alderson, 1999 ; Bachman, 2002, Bachman & Palmer, 1996 ; Brindley & Slatyer, 2002 ; Carr, 2006 ; Weir, 2005a). Cette obsession récurrente dans le champ de l'évaluation linguistique du

« découpage » de la compétence langagière fait sans doute oublier aux chercheurs qu'avant de proposer des taxonomies, il convient d'opérer des distinctions. D'une part, il faut définir le type d'évaluation que l'on vise pour savoir si un découpage en compétences est pertinent ou pas (par exemple, pour une évaluation diagnostique). D'autre part, il est utile de savoir si les éléments placés dans une taxonomie de compétences appartiennent tous à une même classe d'objet ou bien s'il s'agit d'objets différents. De ce point de vue, la recherche sur la dimensionnalité semble capable d'apporter des réponses intéressantes.

Depuis quelques années, on s'interroge pour savoir ce qui, dans la compétence de lecture et plus généralement langagière, est unidimensionnel. Pour ce qui est de compétence langagière, Bachman (1990) et Bachman et Palmer (1996) ont montré qu'il ne s'agit pas d'une compétence unidimensionnelle. En général, pour l'évaluation de la lecture, par le truchement de tests, on vise l'unidimensionnalité (cela permettant, entre autres choses, une interprétation plus aisée de la mesure). Au moment de la création du test, les chercheurs ou développeurs de test doivent faire la démonstration de la dimensionnalité de leur(s) test(s).

Ceci étant, le débat entourant la notion d'unidimensionnalité dans les tests de langue est loin d'être clos. Si Reckase (1990) distingue l'unidimensionnalité psychologique de l'unidimensionnalité psychométrique, Henning (1992), pour les tests de langue, reprend cette distinction. Les définitions données par Henning étant plus opérantes que celles données par Reckase, ce sont celles qui seront proposées ici. Alors que Dassa et Laurier (2003 : 110) déclarent que la dimensionnalité psychologique est fondée sur les analyses des experts de contenu et les habiletés cognitives et la dimensionnalité psychométrique est fondée sur les analyses psychométriques, voici *in extenso* les définitions proposées par Henning (1992 : 2-3) :

> « Psychological unidimensionality in a test implies that the test scores are intended to be interpreted as reflective of the extent of the presence of some known unitary psychological construct or trait. […] Thus, inferences may be drawn about the extent of the presence of the intended construct from item performance. ».
>
> « Psychometric unidimensionality is similar to psychological unidimensionality in that both refer to the capacity of a test to measure some primary dimension or trait, but psychometric

> unidimensionality differs from psychological unidimensionality in several important aspects that I intend to demonstrate in this study. First of all, psychometric unidimensionality can be present when the test measures a variety of correlated underlying psychological dimensions. Furthermore, psychometric unidimensionality can be present even when there is no explicit interpretation possible of the primary dimension said to be measured, apart from the operational definition provided by the items themselves[6] ».

L'auteur explique que les unidimensionnalités psychologique et psychométrique sont indépendantes. Lorsqu'on élabore un test, il faut donc opérer des choix qui définissent un construit unidimensionnel ou multidimensionnel. En effet, on ne peut pas se fier uniquement à l'unidimensionnalité psychométrique car il n'est pas impossible que les mesures indiquent de l'unidimensionnalité pour les données, là où, en réalité, ce que l'on veut mesurer est multidimensionnel. Il est possible également de trouver de la multidimensionnalité là où ce que l'on veut mesurer est unidimensionnel (Henning, 1992 ; McNamara, 1996 ; Reckase, 1990). Lorsque l'on évalue une compétence possiblement unidimensionnelle, il faut se référer à la littérature scientifique sur le sujet (recherches antérieures). Ensuite, il faut vérifier si le construit du test mesure bien une seule compétence ou plusieurs. Si les mesures contredisent l'unidimensionnalité, on peut penser que certains items du test mesurent d'autres dimensions. Il est encore possible que l'ensemble des items mesure des dimensions différentes et, donc, que le construit ou la compétence que l'on mesure ne soient pas ceux ou celles que l'on cherchait à mesurer. Enfin, il est possible que les items soient mal écrits, qu'ils contiennent des erreurs (McNamara, 1991, 1996). Pour la mesure de la compétence des personnes, trois phénomènes peuvent apparaître lorsque l'on pense avoir créé un construit unidimensionnel et que l'on mesure de la « multidimensionnalité » (McNamara, 1991, 1996) :

1) La performance à un item particulier n'était pas indicative de la compétence du candidat en général, et doit être le résultat de facteurs non pertinents comme la fatigue.
2) La population des candidats est hétérogène.
3) Il y a des trous dans le savoir testé par le test chez les candidats.

D'un point de vue plus conceptuel, Reckase (1990) démontre qu'un test multidimensionnel d'un point de vue psychologique (ou encore édumétrique, autrement

dit, pour les spécialistes du domaine) peut être unidimensionnel d'un point de vue psychométrique. Il dégage deux cas. Dans un premier cas, le test est bien composé de plusieurs compétences (ou composantes de la compétence) n'appartenant pas à une seule dimension. Cependant, les composantes de la compétence appartenant à des dimensions différentes étant toutes orientées dans un même sens (les vecteurs associés au niveau de maîtrise des habiletés, savoir-faire ou connaissances sont dirigés dans un même sens), les analyses statistiques amènent à conclure que la compétence est unidimensionelle. Dans un deuxième cas, la différence de difficulté entre les items appartenant à deux dimensions différentes étant si importante, il devient impossible de distinguer la multidimensionnalité. A l'issue des analyses statistiques, on conclut alors à l'unidimensionnalité. Reckase (1990 : 28) conclut sur ce point que d'un point de vue opérationnel, il est important que les tests mesurent toujours la même combinaison de composantes de la compétence. Il faut également que les composantes de la compétence aient bien la même orientation pour que les scores soient interprétables, pour que les inférences faites sur la compétence à partir des résultats soient valides et, *in fine*, pour que les résultats soient suffisament généralisables. C'est une condition indispensable pour que la compétence soit unidimensionnelle d'un point de vue statistique.

Face à cette interprétation de deux types d'unidimensionnalité proposée par Reckase (1990) et reprise par Henning (1992), on trouve un autre point de vue. Blais et Laurier (1995a, 1995b) après avoir étudié la dimensionnalité d'un test de positionnement en utilisant plusieurs méthodes, disent que l'on peut voir la dimensionnalité d'un test comme un continuum. Selon les auteurs, les tests ne sont jamais purement unidimensionnels. Ce qui importe, c'est d'avoir un test qui ne s'éloigne pas trop de l'unidimensionnalité. S'il s'agit d'un aspect essentiel dans la recherche de la validité du construit, l'unidimensionnalité « pure » n'est pas une condition *sine qua non* :

> « Unidimensionality is not a yes/no issue; it is rather a matter of degree considering the purpose of the test. To what extend does a departure from unidimensionality rule out the use of a test in a given situation? In fact, we know that a language test is never fully unidimensional. Moreover, unidimensionality concerns should not force test developers to restrict the nature and the range of tasks whenever validity would entail diversity and complexity. Dimensionality analysis as part of the construct validity study of a test is long-

> term process, in search of an adequate but perfectible construct.[7] » (Blais & Laurier, 1995a : 88).

L'objectif de la recherche de l'unidimensionnalité dans un test consiste à permettre une interprétation claire des résultats plutôt que de savoir avec justesse quel est le profil « réel » ou encore « exact » du candidat. Et Blais (1987 : 14) de dire :

> « En fait, même si à la limite toutes les situations de mesure sont multidimensionnelles, on considère que la multidimensionnalité n'est pas souhaitable parce qu'il est ardu d'observer et d'identifier les dimensions en présence et de saisir l'influence spécifique de chacune des dimensions et de leurs interactions, ce qui ne permet pas une interprétation simple et compréhensible des résultats. Comme l'idéal de l'unidimensionnalité « pure » est difficile à réaliser, on souhaite surtout que le rendement observé soit principalement influencé et explicable par une dimension dominante ; en un sens c'est le pouvoir explicatif de la mesure obtenue qui est recherché plutôt que sa justesse. ».

On retrouve la même idée chez Embretson et Reise (2000 : 309) pour qui l'unidimensionnalité permet une interprétation moins ambiguë des scores. Ils rajoutent que lorsque l'on est en présence de plusieurs dimensions on doit alors utiliser des sous-échelles de scores ou encore des modèles unidimensionnels :

> « Generally speaking, the more strictly unidimensional the scale items, the less ambiguous the interpretation of the resulting raw scale scores and corrections for attenuation are legitimate (Smith, 1996). Also, the application of unidimensional IRT measurement models is more valid and reasonable. If item responses are influenced by two or more common factors, the researcher may wish to consider the creation and scoring of subscales or the application of multidimensional IRT models (Reckase, 1997)[8] ».

Enfin, certains chercheurs n'acceptent pas l'idée d'évaluer une compétence unidimensionnelle. Ils penchent plutôt vers une interprétation multidimensionnelle ou encore « pluridimensionnelle » de la compétence (Springer, 2002). Le problème, avec ce type de position, est qu'elle est focalisée sur certains types d'évaluation particuliers (diagnostique, contextualisée). Du côté européen, certains chercheurs orientés vers l'évaluation multidimensionnelle ou encore « pluridimensionnelle » disent privilégier des

approches « communicatives » ou « actionnelles ». Or, cette orientation multidimensionnelle semble renoncer à la notion d'intégration telle que définie par Widdowson (1981). Elle signe un retour au « découpage » de la compétence (Springer, par exemple, emploie le terme de « gabarit »). Elle privilégie l'évaluation d'étapes de l'acquisition de la compétence, au risque de pratiquer une évaluation de type « *achievement* » (évaluation de rendement), là où on a besoin d'une évaluation de type « *proficiency* » (évaluation de la maîtrise), au risque d'évaluer des composantes de la compétence là où on a besoin d'un test demandant une évaluation globale de la compétence langagière. Par ailleurs, le concept de compétence suppose une certaine unidimensionnalité dans la mesure où la compétence est un attribut que se développe dans le temps et qu'on peut reconnaître certaines manifestations qui correspondent à divers degré de développement. S'il en en était autrement, la compétence ne serait pas un « vecteur orienté » (vers la compétence de l'expert) et son développement irait dans tous les sens.

Pour certains, la confusion semble régner quant aux liens entre la compétence communicative, les problèmes liés à la mesure et les différents types d'évaluation. Et ainsi Springer (2002 : 61) de dire :

> « Évaluation et apprentissage en langue étrangère sont intimement liés dans la mesure où on a toujours tenté de mesurer d'une manière aussi juste que possible les performances des apprenants. Il existe en effet de nombreux tests de langues qui permettent de mesurer tel ou tel aspect langagier. Le testing en langue s'est petit à petit constitué en discipline rigoureuse répondant à une méthodologie stricte et à un cahier des charges précis. Cette approche psychométrique continue à marquer les recherches anglo-saxonnes. Cependant, la révolution communicative a fini par ébranler le socle épistémologique du testing, pourtant solide. L'idée d'évaluer de manière qualitative la compétence en langue commence à faire son chemin. »

2.3.2 Le construit

Quelle que soit la position que l'on adopte face à l'unidimensionnalité avant de concevoir un test, il convient de définir un construit. Voici la définition que donne Bond et Fox (2001 : 229) du construit : « A single latent trait characteristic, attribute, or dimension

assumed to be underlying a set of item. ». Si Chapelle (1999 : 156) rappelle que selon Messick (1981), le construit permet « d'interpréter la performance au test utilement et significativement », elle dit aussi que la nouveauté apportée par Bachman (1990) et Bachman et Palmer (1996) est de proposer un construit en fonction d'un contexte d'utilisation langagière (Chapelle, 1999 : 161). Cette définition du construit inclut, à la fois, la compétence cognitive et le domaine dans lequel la compétence est pertinente (par exemple, lire un texte académique pour un test d'entrée dans une université). Ce construit, qui demande une définition du trait latent et du contexte, est appelée « interactionnel » (Messick, 1981, 1989). Il a beaucoup influencé l'évaluation de la performance (McNamara, 1996 ; Douglas, 2000). Chapelle (1999 : 154) rappelle qu'il y a deux types d'évaluation : celle qui s'attache à la compétence et celle qui s'attache à la performance. L'évaluation qui s'attache à la compétence ne peut se faire qu'au travers d'une performance. Cette performance est le résultat de la mise en œuvre, non seulement de la compétence, mais aussi d'habiletés sous-jacentes. Si ces habiletés sont utilisées pour la performance, ce qui est visé, c'est uniquement l'évaluation de la compétence, et rien d'autre. Cela permet d'établir des inférences au sujet d'une compétence non-observable, pour ensuite faire des prédictions sur la performance dans des situations de non-test. Dans la logique de l'évaluation par la performance, on n'observe pas la compétence directement mais au travers de la performance au test pour faire des inférences au sujet de la performance dans une situation de non-test. Ici, la performance doit être le plus proche possible de la situation de non-test. Le problème sera alors de trouver un test qui soit le plus proche possible de la situation de non-test.

L'attachement à la notion de la performance, dans une situation de non-test, mais aussi à celle de trait latent a, sans doute, amené les concepteurs de test à proposer des activités intégrées. La notion de « tâche intégrée » renvoie à une activité demandant au candidat d'utiliser l'oral et l'écrit ou encore la compréhension et la production pour accomplir la tâche. Une tâche intégrée est aussi une tâche qui demande au candidat de réutiliser le matériel linguistique déjà présenté, de « l'intégrer » pour l'aider à accomplir une tâche complète où plusieurs documents ou supports sont proposés à la lecture ou à l'écoute. Alderson (2000 : 26-27) déclare que l'on a une très petite idée de l'effet que peut avoir l'intégration ou l'isolement des tâches et que la recherche est nécessaire. Le problème est

de savoir comment évaluer la compréhension à partir d'une production ou encore de savoir comment la production est altérée par la compréhension. Selon l'auteur, lorsque l'on veut un construit avec des tâches intégrées ou discrètes, cela dépend du type d'évaluation que l'on veut mettre en place et du type de résultats que l'on veut obtenir. Par exemple, un test proposant des tâches intégrées mélangeant les macro-compétences pourra correspondre à une évaluation diagnostique et un autre test proposant une évaluation évaluant les macro-compétences séparément pourra correspondre à une évaluation pour fins de sélection, autorisant une image claire du niveau de lecture (Alderson, 2000 : 30).

2.3.3 La notion de difficulté : deux visions différentes

La mise au point d'un test d'évaluation linguistique, en fonction d'un construit défini préalablement, vise à placer sur une échelle commune des items et des personnes selon leur niveau de difficulté ou de compétence. L'unidimensionnalité d'un test étant plus une question de degré que de présence ou d'absence d'unidimensionnalité (Blais & Laurier, 1995a, 1995b), il semble utile de s'interroger quant à la nature, la « composition » et les « mécanismes » de la difficulté des items. Est-elle « intrinsèque », dans le sens où les items d'un test auraient une difficulté indépendante des caractéristiques de l'échantillon testé ? Est-elle « locale », provisoire, empirique, dans le sens où elle dépendrait des caractéristiques des items, des personnes et de l'interaction entre les deux ? Aujourd'hui, dans l'évaluation en langue seconde, on a principalement deux visions de la difficulté qui s'affrontent (Bachman, 2002) :

1) *Celle de l'approche par les tâches* : dans cette approche, on identifie les différentes caractéristiques des tâches dont on pense qu'elles affectent la difficulté d'une tâche donnée (Norris *et al.* 2002 ; Elder *et al.,* 2002 ; Brindley & Slatyer, 2002). La difficulté est perçue comme étant due aux caractéristiques de la tâche elle-même, à sa « nature ».

2) *Celle des tâches ciblées sur l'usage langagier. (Task language use)* : ici, on tente d'identifier les caractéristiques des tâches qui sont considérées comme étant indépendantes de la compétence du candidat pour pouvoir étudier les liens entre ces caractéristiques et des indicateurs de difficulté empiriques (Bachman & Palmer, 1996).

La difficulté n'est pas perçue comme étant uniquement liée aux caractéristiques de la tâche mais également à celles des candidats.

Dans l'approche par les tâches (Brown *et al.*, 2002), on ne veut pas seulement faire des interprétations au sujet de la compétence d'un individu pour accomplir une tâche particulière. On veut aussi savoir comment la performance à une tâche peut aider à faire des prédictions sur sa capacité à effectuer une autre « tâche-cible » similaire. La difficulté est perçue comme étant le résultat de la complexité cognitive des tâches, les particularités du contexte de la performance et un ensemble de variables des caractéristiques des tâches. Dans cette approche, on tente de proposer un cadre de référence avec une description de la complexité des tâches afin de pouvoir généraliser les résultats. Pour ce faire, on utilise les variables proposées par Skehan (1996) : complexité du code, complexité cognitive et exigences communicatives. On pense pouvoir faire des prédictions de la difficulté à partir de ce schéma et surtout, on pense que la complexité cognitive des tâches permet de faire varier l'output, autrement dit, la production du candidat. Si la recherche est parvenue à montrer que l'output varie en fonction de la tâche, le problème est que le lien avec la difficulté n'est pas établi[b]. Le cadre de Skehan (1996) est celui qui a été retenu par Norris *et al.* (2002 : 72-82) pour construire leur théorie T.B.P.A. (Task Based Performance Assessment).

Ce concept de la prédictibilité de la difficulté par la complexité des tâches n'a pas encore fait la preuve de sa pertinence. Il n'est toujours pas opérationnel. Dunkel (1999) rapporte que le cadre conceptuel, qui avait été conçu pour concevoir un test de Hausa, n'a pas permis de prédire la difficulté des items. Iwashita, McNamara et Elder (2001) ainsi que Elder, Iwashita et McNamara (2002), échouent, eux aussi, dans leur tentative d'isoler les variables de la difficulté. Selon les auteurs, il est donc difficile de prévoir la difficulté à partir du cadre de Skehan (1996). Brindley et Slatyer (2002), quant à eux, arrivent à la conclusion qu'il n'est pas possible de prévoir la difficulté d'une tâche avec un facteur général de difficulté comme on le fait avec des items individuels.

Une critique pertinente du concept de prédiction de la difficulté vient de Bachman (2002) selon lequel, il y a deux problèmes avec l'idée de prédire le niveau de difficulté des tâches. Premièrement, il y a une confusion entre l'effet de la compétence du candidat et

[b] Par exemple, Nunan (2004 :85-92), dans un sous-chapitre qu'il intitule « task difficulty » utilise les termes de « difficulté » et de « complexité » ce qui peut prêter à confusion.

l'effet de la tâche. On introduit la notion de « difficulté de la tâche » comme déterminant de la performance au test. Or, la difficulté est une interaction entre le candidat et la tâche. Les difficultés cognitive et communicative vont varier en fonction des candidats. La difficulté de la tâche ne peut donc pas être un facteur isolé.

On a donc deux positions qui se dégagent : une première qui affirme qu'il est difficile d'avoir des tâches comparables d'une administration d'examen à l'autre (surtout quand on change de population), une deuxième qui affirme pouvoir prévoir la difficulté, d'en comprendre les mécanismes pour proposer des tâches équivalentes aux candidats.

Selon Bachman (2002 : 458), le problème de toutes ces recherches (portant sur les tâches ciblées sur l'utilisation langagière -*T.L.U.*- et les tâches de la vraie vie), c'est qu'elles donnent des résultats difficiles à prévoir. Autrement dit, il est hasardeux de tenter de prédire la difficulté de tâches complexes et authentiques. Selon lui, on doit considérer les tâches comme des ensembles de caractéristiques plus que comme des ensembles holistiques. Il y a trois facteurs qui influencent la performance au test : les caractéristiques de la tâche, les attributs du candidat et les interactions entre les deux

Nunan et Keobke (1995), quant à eux, proposent une synthèse des travaux de Brown et Yule (1983), Nunan (1989, 1995), Anderson et Lynch (1988) portant sur la difficulté de la tâche d'évaluation. Ce travail, très documenté, fait ressortir des oppositions permettant d'expliquer la facilité ou la difficulté de la tâche selon trois facettes : le candidat, la tâche et le texte. Cet instrument permet de porter des diagnostics sur la difficulté d'une tâche mesurée empiriquement.

2.3.4 La difficulté des tâches de lecture, un concept « multi-facettes »

Dans l'évaluation de la compréhension écrite en langue seconde, on retrouve cette distinction entre, d'un côté, ceux qui soutiennent la prédictibilité de la difficulté des tâches et, de l'autre, ceux qui s'opposent à cette idée. Depuis que Munby (1978) a proposé une classification des habiletés de lecture, on s'interroge pour savoir comment, on peut, non seulement, classer les tâches de lecture selon leur difficulté, mais encore, faire des prédictions au sujet de cette difficulté. Lumley (1993), cherchant à classer les habiletés de lecture en fonction de leur difficulté, propose de comparer la difficulté

évaluée par des juges à la difficulté mesurée avec un test mettant en œuvre ces habiletés. Si les résultats semblent plus ou moins probants, la question qui se pose, alors, est surtout de savoir comment définir la difficulté.

Selon Koda (2005 : 230-231), plusieurs perspectives sont actuellement proposées pour l'évaluation de la compréhension écrite. Une première distingue trois niveaux hiérarchiques (du plus facile au plus difficile) : le déchiffrage, la construction du sens et l'utilisation combinée des connaissances antérieures avec des informations contenues dans le texte. Une seconde (cognitive) distingue des degrés de compréhension selon que le texte est explicite ou implicite. Une troisième perspective, que l'on doit à Carver (1990, 1997, 2000), privilégie le but du lecteur. L'exigence de lecture augmente selon le but du lecteur (repérage, compréhension de base, apprentissage ou mémorisation du contenu informatif). Selon Carver (2000), ces différents buts produisent des différences de rapidité de lecture. Cette perspective « développementale » met en avant la différence de degré de compréhension qu'il y a entre « apprendre à lire » et « lire pour apprendre ». Elle est intéressante car elle montre les limites de la lecture évaluée sous l'angle de la « compréhension écrite ». Ici, il ne s'agit plus de mesurer la compréhension, mais la compréhension pour faire quelque chose, ce qui, on en conviendra, est une activité cognitive différente. Carver (1997) distingue ainsi la compréhension de base qu'il appelle « *rauding* » de l'activité qui consiste à lire pour apprendre, « *read to learn* » (Enright *et al.*, 2000). Alors que Enright *et al.*, (2000) en commençant leur recherche, pensent qu'il est possible de placer les activités d'évaluation de la compréhension écrite selon une « hiérarchie naturelle » conçue à partir de quatre buts du lecteur, soit « lire pour trouver de l'information », « lire pour comprendre », « lire pour apprendre » et « lire pour intégrer », leur conclusion est quelque peu différente. Le but du lecteur devient un facteur affectant la difficulté parmi d'autres. Un appel est lancé pour effectuer de plus amples recherches :

> « **A Difficulty Continuum**. To some extent the four reader purposes form a kind of difficulty continuum. To be sure, easy tasks could be designed for reading to learn or reading to integrate information, and difficult tasks asking examinees to find discrete information or read for general comprehension could be designed by manipulating task and linguistic/syntactic variables. Still, we are more likely to find reading to learn and reading to integrate information across texts associated with more challenging academic tasks and

> to require more sophisticated processing abilities than reading to find discrete information or reading for general comprehension. The implication is that, as the reading purpose changes from reading to find information to reading for basic comprehension to reading to learn and reading to integrate information across texts, more reading is required and more efficient strategies are necessary. We believe, therefore, that reader purpose itself may be one of the variables that can contribute to task difficulty when combined with appropriate texts and tasks. It will be important to explore this as part of the research agenda.[9] » (Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt & Schedl, 2000: 6-7).

En 2005, Trites et McGroarty, à l'aide d'une analyse discriminante, trouvent que la compétence de lecteurs non-natifs (de niveau intermédiaire dans l'échantillon étudié) pour des tâches mesurant la compétence à lire pour apprendre et d'autres mesurant la compétence à lire pour intégrer (mesurée à l'aide de la lecture de deux textes et d'un exercice de synthèse utilisant la production écrite) ne peut pas être prédite par une mesure de leur compréhension de base. Un tiers de ceux qui avaient un niveau élevé pour la compréhension de base n'ont plus qu'un niveau intermédiaire quand on mesure leur compétence à lire pour apprendre et pour intégrer. Les auteurs concluent que ce nouveau type de tâche évalue autre chose que ce qu'évalue la compréhension de base :

> « Considering results from both the non-native speakers and native speakers, we conclude that the new tasks did access something different from basic academic English proficiency had been achieved[10] ». (Trites & McGroarty, 2005 : 198)

Pour ce qui est de la difficulté des tâches, encore une fois, les résultats ne montrent pas une hiérarchie des tâches :

> « We had hoped to find clear evidence of a hierarchy suggesting that Reading to Learn was demonstrably more difficult than basic comprehension and Reading to integrate demonstrably more difficult than Reading to learn, but results did not yield an obvious hierarchy [11] » (Trites & McGroarty, 2005: 198).

Déjà en 1999, Alderson (1999 : 53-54) déclare que les recherches empiriques montrent qu'il n'y a pas une simple correspondance entre les items, les personnes, la compétence et la difficulté. Le niveau de difficulté dépend de différents facteurs. Ailleurs, Alderson

(2000 : 103-109) ajoute que les facteurs qui affectent la difficulté de la lecture des textes des items sont la familiarité avec le contenu, la présence du texte au moment de répondre et la longueur du texte. Selon McNamara (1996), la difficulté de la tâche de lecture dérive, naturellement, du but de lecture et varie considérablement d'un texte à l'autre. Au jugé de ce dernier, la difficulté de la tâche est définie en considérant la combinaison des caractéristiques du texte et le processus mis en œuvre pour effectuer la tâche. Les variables du processus de lecture incluent :

- le type d'informations que le lecteur est censé identifier selon un degré d'abstraction,
- le type de correspondances requises entre le document et la tâche,
- la possibilité d'extraire de l'information avec le texte.

Parmi les facteurs qui influencent la difficulté des items de compréhension écrite, on peut encore trouver la langue de rédaction des items, soit la langue maternelle ou encore la langue cible (Alderson, 2000 ; Koda, 2005). Lorsque les amorces des items sont rédigées dans la langue maternelle du lecteur, les items sont plus faciles. Les items sont plus difficiles lorsque les amorces sont rédigées dans la langue cible[c]. On a donc bien deux types de tâches différentes. Le type de questions posées sur les textes (sur un passage ou sur l'ensemble du texte) a aussi un impact sur la difficulté. D'après Pearson et Johnson (1978) et Garcia (1991), les questions sont plus ou moins difficiles selon qu'elles sont :

- textuellement explicites (la question et la réponse sont des paraphrases ou tirées d'une seule phrase dans le texte),
- textuellement implicites (la réponse est à trouver dans plusieurs phrases, elle n'est jamais dans une seule phrase, il faut combiner divers passages du texte pour pouvoir répondre aux questions),
- basées sur les connaissances antérieures du lecteur (intégration du savoir du candidat à la réponse).

Les conditions de passation semblent, sous certaines conditions, avoir un impact important sur la difficulté, autrement dit sur le nombre de bonnes réponses données (Davey & Lasasso, 1984). Ainsi, lors d'une passation d'examen, quand les candidats sont autorisés à effectuer des retours en arrière, il n'y a pas de différence entre des réponses

[c] Dans le cas du français langue seconde, en français.

choisies (Q.C.M.) et des réponses construites. Mais quand ils ne sont pas autorisés à réviser leurs réponses, les Q.C.M. sont plus « faciles » que les réponses construites. Ce qui est intéressant, c'est que c'est vrai pour les items explicites et implicites. Il faut donc comprendre que les Q.C.M. sont plus faciles que les réponses construites sous certaines conditions. Il convient donc être prudent avec les problèmes liés à l'administration des examens. Cette prudence doit être d'autant plus grande que la recherche a encore besoin d'analyser les interactions entre le texte, les caractéristiques des tâches d'évaluation et les caractéristiques des candidats (Alderson, 2000 ; Bachman, 2002 ; Carr, 2006).

D'autres approches, empiriques, ne s'intéressent pas aux types de questions, d'activités ou de tâches proposées aux candidats pour étudier la difficulté, mais à la lisibilité des textes. Au Québec, le logiciel SATO (Daoust, 1996) permet de calculer un « indice Gunning » de la difficulté des textes. Voici comment cet indice de lisibilité est calculé :

> « La formule utilisée pour calculer l'indice de Gunning est la suivante :
> (longueur-moyenne-des-phrases + pourcentage-de-mots-longs) x 0.4
> où longueur-moyenne-des-phrases est : nombre-de-mots / nombre-de-phrases.
> et pourcentage-de-mots-longs est :
> (nombre-de-mots-de-9-lettres-ou-plus / nombre de mots) x 100 » (Daoust, 1996 : 114).

Stenner et Stone (2004) proposent une autre solution, baptisée « lexile ». Alors que dans le modèle de Rasch, les deux facettes évaluées sont celles de la compétence des candidats et la difficulté des questions, les chercheurs proposent de mesurer la compréhension du texte en faisant la différence entre la compétence du lecteur et la lisibilité du texte au moyen d'une équation algébrique combinant la mesure de la fréquence des mots et la longueur des phrases tirés d'un texte (Stenner, 1996; Linacre, 1999a). La compétence du lecteur et la lisibilité du texte sont positionnées sur une échelle continue arithmétique ayant pour unité le lexile (si pour le thermomètre Celsius, l'unité équivaut à 1/100 de la distance entre le point de congélation et le point d'ébullition, le lexile, lui, est le 1/1000 de la distance entre des textes faciles tirés de textes élémentaires (« *primers* ») et difficiles (« *encyclopaedias* »)[d]). La compréhension d'un texte par un lecteur est une fonction de la différence entre la compétence du lecteur et la lisibilité du texte. Cette

[d] Pour plus de détails, le lecteur intéressé pourra consulter Stenner (1996 : 13-15)) et Linacre (1999a).

solution amène à reconnaître que les généralisations au sujet des performances des lecteurs peuvent être indépendantes des textes (on mesure alors la compétence du lecteur) ou dépendante du texte (on mesure alors la compréhension). Stenner et Stone (2004) expliquent qu'un lecteur avec un haut niveau de compétence de lecture pourra avoir une très mauvaise compréhension d'un texte (relativement difficile) d'une décision de la Cour Suprême. À l'inverse, un lecteur avec une compétence de lecture relativement faible pourra faire un bon résumé d'un livre pour enfant.

Alderson (2000 : 73-74), quant à lui, met les chercheurs en garde contre une vue trop simpliste de la difficulté dont pourrait éventuellement découler un usage irraisonné des indices de lisibilité. Selon l'auteur, la difficulté ne peut pas être définie dans des termes absolus. Il trouve préférable pour le concepteur de test de choisir un ensemble de textes qui soient ceux que le lecteur cible pourrait être amené à lire :

> « However, readability formulae give only crude measure of text difficulty, and are rarely suitable for second-or foreign language readers [...]. Given the range of variables that affect text difficulty – topic, syntactic complexity, cohesion, coherence, vocabulary and readability – language testers should beware a simplistic approach to language difficulty when selecting tests [...]. In many circumstances, text difficulty will not be definable in absolute terms, and instead testers will prefer to identify a range of authentic texts that might have to be read in the test-taker language use situation.[12] ».

Linacre (1999a) propose la création d'un test adaptatif de compréhension écrite en utilisant le lexile comme paramètre de difficulté pour les items. Les items seraient donc rangés dans une banque de données en fonction de leur valeur en lexile et non pas en fonction l'indice « traditionnel » de difficulté. Pour définir le niveau de compétence du candidat estimé, lui aussi, en lexile), il propose d'utiliser comme règle un taux de réussite de 70 à 80%. Ainsi, le premier item proposé au candidat aura 200 lexiles de moins que l'estimation initiale de son niveau. On s'attend alors à ce que le candidat ait une probabilité de 75% de réussite. Selon l'auteur, cette règle, permettra de distinguer les lecteurs compétents pour des niveaux de lecture calculés en lexile. L'idée centrale de cette proposition est de centrer l'activité sur la difficulté des textes plutôt que sur la difficulté des items.

Pour intéressantes que soient ces solutions, elles ne doivent pas faire oublier que, comme le propose Bachman (2002), la difficulté pour le candidat est fonction de trois facteurs : les caractéristiques de la tâche (texte et activité -ou encore but de lecture-), celles du candidat et l'interaction entre les deux. Pour la lecture, Nunan (2004 : 171) cite Brindley (1987) qui considère que la personne, la tâche et le texte interagissent pour déterminer la difficulté. La difficulté de l'activité d'évaluation varie donc en fonction des caractéristiques du candidat, du texte mais aussi de l'activité à effectuer. Pour la lecture, cette affirmation semble être confirmée par les recherches les plus récentes (Carr, 2006 ; Jang & Roussos, 2007 ; Gorin, 2005 ; Grabe *et al.*, 2000 ; Trites & McGroarty, 2005)

Le schéma de la figure 2.1 permet de prendre en considération non seulement la lisibilité du texte et la compétence de lecture mais encore les caractéristiques de l'activité de lecture (ce que le candidat doit faire et le but de lecture). Si comprendre un discours présidentiel est plus difficile que de comprendre un livre pour enfant, lire un article de presse et démontrer sa compréhension par des réponses à des Q.C.M. ou une réponse écrite peut ne pas présenter la même difficulté pour tous les candidats.

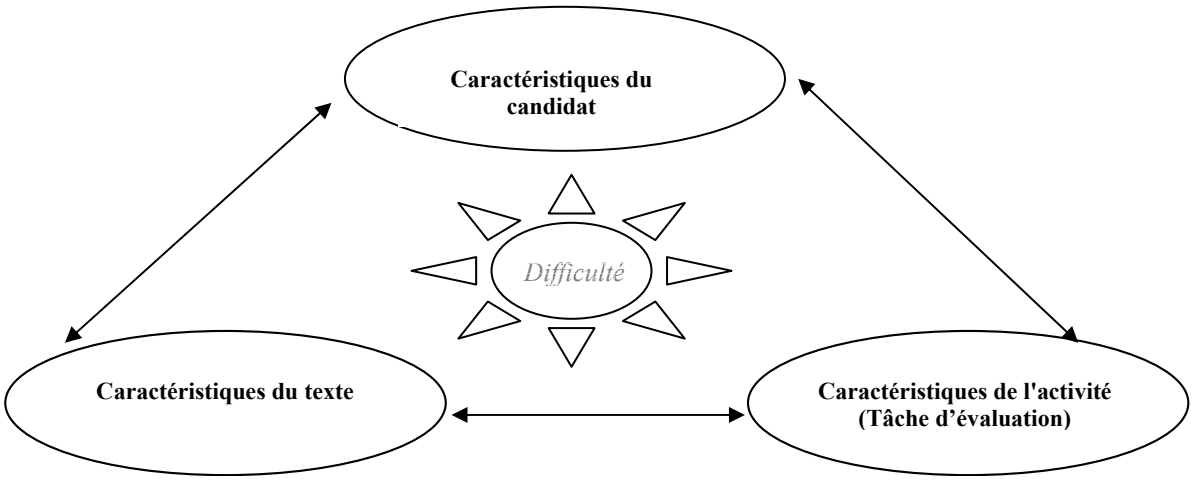


Figure 2.1 : la notion de difficulté pour une tâche de lecture dans un test de langue

Pour conclure sur ce point, les propos de Nathan Carr (2006) traduisent parfaitement ce qui semble constituer un consensus parmi les chercheurs (opérant dans le champ de l'évaluation des langues secondes) quant aux relations qui peuvent exister entre les différentes composantes de la lecture et leurs interactions :

> "Most researchers agree that reading is rapid, purposeful, and interactive (Grabe, 1999; Alderson, 2000). Specifically, this interactivity occurs at two levels. First, regardless of the exact components or levels that they posit, models of the reading process depict these components or levels as interacting with each other. Second and more relevant to this study, reading is interactive in that the reader's background knowledge and other attributes interact with the content of the text. Given that reading involves interaction between readers and texts, it logically follows that characteristics of both the reader and the text will affect the reading process (Alderson, 2000)[13]" (Carr, 2006: 271).

2.3.5 Les caractéristiques des tâches d'évaluation

2.3.5.1 Les tâches d'évaluation discrètes et intégrées

Après s'être intéressé aux recherches menées sur les caractéristiques des textes (Daoust, 1996, Stenner, 1996 ; Stenner & Stone, 2004), il convient de s'intéresser aux caractéristiques des tâches d'évaluation et à la littérature portant sur le domaine. La présentation qui sera faite commencera par des tâches « discrètes » (tâches ne demandant pas de production) pour s'achever par des tâches « intégrées » (tâches demandant une production). Les tâches ne sont pas présentées par ordre de difficulté. Toutes peuvent être proposées à différents niveaux. Les travaux de Koda (2005) et Alderson (2000), entre autres, seront utilisés pour établir les profils des différents types de tâches, leurs avantages et leurs inconvénients. Seront présentées celles qui sont les plus illustratives du continuum entre tâches discrètes et tâches intégrées.

Mais avant d'entrer plus dans le détail des activités d'évaluation qui peuvent être mises en place dans l'évaluation de la lecture en langue seconde, il convient de définir ce que l'on entend par tâches d'évaluation discrètes (parfois appelées « indépendantes » dans la littérature) et intégrées. Pour ce faire, on se référera aux travaux de Widdowson (1981) et de Savignon (1983), Enright *et al.* (2000) et Trites & McGroarty (2005). Si les travaux de Widdowson sont plus orientés vers la pédagogie des langues secondes, la position qu'il

adopte vis-à-vis de l'intégration est intéressante et pertinente pour la présente recherche. Widdowson (1981 : 27) défend l'idée selon laquelle on doit associer la langue étrangère à son emploi. Il opère une distinction entre l'usage (renvoyant aux aspects du système linguistique) et l'emploi de la langue (renvoyant à la communication). L'auteur ne voit pas le processus de lecture comme un simple décodage mais bien comme un acte par lequel le lecteur doit interpréter le matériel linguistique pour lui donner du sens, et, ce, au fur et à mesure qu'il lit. En ce sens, lire n'est pas une somme d'actes de lecture isolés mais bien d'actes « intégrés » reposant sur la capacité à interpréter. Et Widdowson (1981 : 84) de dire :

> « La capacité d'interpréter sur laquelle repose la lecture efficace met donc en jeu l'actualisation de la valeur propositionnelle et illocutionnaire par référence avec ce qui a précédé et l'anticipation de la valeur propositionnelle et rhétorique de ce qui va suivre. ».

Ce qui est intéressant dans cette vision de l'intégration, c'est qu'on ne dénonce pas en soit le découpage de la matière linguistique, sinon son découpage lorsqu'il vise à produire des éléments indépendants les uns des autres. À ce sujet, l'auteur déclare que les pédagogies traditionnelles ont un principe de base qui « semble être la ségrégation plutôt que l'intégration : « diviser pour régner ». » (Widdowson, 1981: 163). Le problème de l'authenticité des textes, n'est pas seulement posé en termes de « discours non-écrits à des fins pédagogiques » mais bien de discours dépendants d'un contexte, voire d'un co-texte, inscrits dans une « unité de communication » :

> « D'ordinaire, nous n'appréhendons pas le discours sous forme de passages à lire indépendants mais sous forme d'unités rhétoriques complètes : essais, articles, lettres, reportages journalistiques, etc. En outre, ces unités renvoient à notre propre réalité sociale et psychologique. » (Widdowson 1981 : 93).

La vision de l'authenticité de Widdowson définit l'authenticité d'un texte non pas uniquement par son usage mais aussi par son emploi (on retrouve la même vision chez Bachman, 1990). Autrement dit, l'authenticité du texte ne se limite pas uniquement au fait qu'il n'ait pas été écrit ou modifié pour un apprenant, mais à la manière dont le lecteur-apprenant devra lire ce texte. La dépendance au contexte permet d'établir la différence entre les tâches indépendantes (ou encore discrètes) et intégrées (ou encore dépendantes).

Le principe d'intégration défendu par l'auteur laisse à penser que les activités pédagogiques et d'évaluation peuvent être plus ou moins intégrées, selon un contexte, plus ou moins élargi ou restreint. Le degré d'intégration se situe donc entre deux pôles, soit celui de l'analyse d'éléments discrets du discours (un mot, une phrase) hors contexte et celui de l'interprétation d'un document écrit dans un contexte large et un usage social donné (par exemple, lire des formulaires administratifs universitaires). Pour le concepteur de test, il s'agit donc de situer les tâches d'évaluation sur ce continuum en fonction du construit qu'il aura choisi, mais aussi en fonction des interprétations requises des résultats.

Savignon (1983), quant à elle, analyse des tâches non pas pour l'apprentissage mais bien pour l'évaluation. Pour son analyse, elle reprend à Carroll (1972) les termes d'« intégré » et de « discret ». Voici la définition qu'elle propose des tâches discrètes[e] :

> « A discrete-point task is one that focuses on an isolated bit of language, typically surface features of phonology, morphology, syntax, or lexicon [...] In their purest form, discrete-point item include but one channel (oral or written) and one direction (receptive or productive); that is, they test "separate" skills of listening, reading, speaking, and writing.[14] » (Savignon, 1983 : 249).

Si Savignon (1983) ne donne pas de définition pour les tâches intégrées, Alderson (2000 : 207) précise que les tâches discrètes évaluent une chose à la fois, alors que les tâches intégrées permettent une évaluation plus holistique dans laquelle les différentes composantes sont en interdépendance :

> « In discrete-point approaches, the intention is to test one « thing » at a time, in integrative approaches, test designers aim to gain a much more general idea of how well students read. In the latter case, this may be because we recognise that the whole is more than the sum of the part[15] ».

Toutefois, il convient de ne pas donner de définitions trop strictes aux deux types de tâches. Dans les faits, il est très peu probable d'avoir des tâches intégrées et discrètes

[e] L'auteur ne donne pas de définition des tâches intégrées. On peut cependant facilement la deviner derrière la définition qu'elle donne des tâches discrètes.

« pures ». Ce qui est intéressant, c'est que les modes de réponse discrets peuvent être utilisés aussi bien avec les tâches discrètes qu'avec les tâches intégrées, alors que les modes de réponses plus globaux sont plus facilement utilisables avec des tâches intégrées.

Savignon (1983) précise que la distinction tâches intégrées / tâches discrètes n'est pas toujours facile à faire. Il faut analyser les tâches selon deux axes, un axe vertical (la tâche, soit le type de document utilisé et la nature de l'activité) et un axe horizontal (le mode de réponse) qui correspondent à deux continuums : discret-global, discret-intégré. Ainsi peut-on à partir du même enregistrement d'une conversation demander aux candidats de noter les verbes entendus au présent (tâche et mode de réponse discrets) ou bien de prouver en répondant à des Q.C.M. qu'ils ont bien compris la situation de communication (tâche plus intégrée avec un mode de réponse « discret »). L'auteur invite le lecteur à faire attention au fait que dire d'une tâche qu'elle est « discrète » ne signifie pas qu'elle ne soit pas communicative. Les tâches « intégrées » ne sont pas non plus « de fait » ou par nature communicatives.

Pour les tâches d'évaluation de la compréhension écrite, Enright *et al.* (2000) et Trites & McGroarty (2005) partent des travaux de Perfetti (1997) et Goldman (1997) pour définir les tâches intégrées et distinguer différents types de tâches discrètes. Perfetti (1997) et Goldman (1997) ont étudié la compétence qui consiste à « lire de multiples textes » ou encore l'intertextualité. Le constat fait par Perfetti (1997 : 339) est que ni le modèle *bottom-up* ni le modèle *top-down* ne sont suffisants pour expliquer les mécanismes de la compréhension d'un texte. Selon Perfetti (1997), l'interaction entre les deux modèles joue un rôle important. Le lecteur a besoin de comprendre à la fois les phrases mais aussi les relations qui les unissent, ou encore, dans quel co-texte elles s'inscrivent. Le modèle qu'il propose pour la compréhension des textes multiples est appelé « *Document Model* ». Le lecteur doit comprendre, à la fois, les relations intertextes, mais aussi, la situation en général (Perfetti, 1997 : 346) pour avoir un *Document Model* complet, ou encore pour comprendre pleinement les textes. Le lecteur peut saisir la situation en général (les auteurs de deux textes sont en désaccord) mais ne pas avoir établi de relations intertextes (le lecteur ignore au sujet de quoi les auteurs ne sont pas d'accord). Le lecteur peut aussi comprendre la relation intertexte (un premier auteur dit une chose, un second, autre

chose) sans avoir établi de lien de situation (le lecteur ignore si les auteurs sont en désaccord). Ce qui est intéressant, c'est que selon Perfetti (1997), cette manière d'intégrer l'information est la même que celle que l'on utilise dans des textes indépendants et la compréhension écrite en général, quoiqu'elle soit plus complexe. Et Perfetti (1997: 351) de dire :

> « From my assessment of individual differences, we can anticipate that individuals will differ in their adaptive use of multiple text environments, just as they differ in the ability to comprehend single texts.[16] ».

Selon Enright, Grabe *et al.* (2000 : 6) et Trites et McGroarty (2005 : 175), les tâches de lecture se distinguent en fonction du but du lecteur. Celles consistant à lire pour intégrer de l'information (dans cette recherche « tâches intégrées ») sont des tâches dans lesquelles le but du lecteur est de faire le lien entre des informations contenues dans plusieurs textes et la matière textuelle. Dans ces tâches, les lecteurs doivent générer leur propre vison des relations entre les textes (Trites & McGroarty, 2005 : 176). Concrètement, le candidat doit prouver qu'il a bien « intégré » les informations contenues dans les différents textes. Il doit montrer sa capacité à repérer les structures textuelles, à retenir des informations et à les retransmettre soit dans un mode oral, soit dans un mode écrit.

Les autres activités de compréhension écrite qui ne portent que sur un seul texte sont dites « indépendantes » (dans cette recherche « discrètes »). Celles qui consistent à évaluer la compétence de lecture avec le but de « lire pour apprendre » portent sur la capacité à se souvenir et à classer l'information contenue dans un seul texte. L'évaluation se fait à partir de résumés ou encore de Q.C.M. Pour évaluer la compétence de lecture, lorsque l'on a pour but de « lire pour comprendre », on pose des questions de compréhension sur les phrases et, non pas sur l'ensemble du texte. Enfin, pour évaluer la compétence de lecture utilisée pour « lire pour trouver de l'information », des questions de repérage peuvent être proposées. Comme on le voit, il est possible de placer ces quatre types de tâche définis selon le but du lecteur sur un continuum discret-intégré.

En conclusion sur ce point, et avant d'entreprendre une description plus formelle des activités d'évaluation les plus couramment utilisées, la synthèse des travaux d'Alderson

(2000), d'Enright *et al.* (2000), de Goldman (1997), de Trites & McGroarty (2005), de Perfetti (1997), de Savignon (1983) et Widdowson (1981) révèle que les tâches intégrées sont des tâches dépendantes non seulement d'un co-texte (qui constitue un contexte restreint) mais également d'un contexte. Le lecteur, pour réussir ce type de tâche, doit se fixer comme objectif de faire le lien entre les informations intra-textuelles et les différents textes. Il doit comprendre quelle est la relation entre les textes. Pour la création d'un test, il convient de définir le degré d'intégration qui convient aux tâches d'évaluation que l'on veut utiliser mais surtout au type de construit que l'on veut mettre en place (Alderson, 2000 : 30). Pour la lecture, le choix du type de texte, du degré d'indépendance ou de dépendance (intégration) des tâches doit être relié à l'authenticité correspondant au construit choisi pour le test. Le choix d'un mode de réponse peu intégratif ne signifie pas que l'on renonce à l'aspect communicatif ou encore interactionnel d'une tâche. Ce choix peut découler de contraintes imposées par des aspects liés à la mesure ou, plus simplement, à des problèmes liés à l'administration du test, ou à un principe de faisabilité. Enfin, Lee (2006 : 134) cite Lewkowictz (1997) selon lequel l'avantage des tâches intégrées est de proposer des tâches qui ne désavantagent pas les candidats selon leurs connaissances antérieures (la tâche intégrée fournit le contexte ou le co-texte). Ces tâches ont également l'avantage d'être plus proches des tâches de la vraie vie.

2.3.5.2 Typologie sommaire de quelques activités d'évaluation

2.3.5.2.1 Les questions à choix multiple

Cette partie de la recension des écrits n'a pas pour ambition de décrire de manière exhaustive les activités d'évaluation. De même, il ne s'agit pas de proposer les meilleures techniques d'évaluation, puisqu'en réalité, il n'y a pas de meilleures techniques sinon des techniques adéquates, efficaces, efficientes ou pas (Alderson 2000 : 203 ; Laveault & Grégoire, 2002 : 35). Il s'agit de présenter des tâches emblématiques qui contrairement à celles présentes dans le nouveau T.O.E.F.L. n'ont pas pour principe de prendre en compte le but du lecteur. Seront d'abord présentées les Q.C.M., le résumé, le rappel libre et enfin les tâches de la vraie vie.

Pour les activités bâties à partir de questions, les travaux d'Alderson (2000), Haladyna, Downing et Rodriguez (2002), Gorin (2005), Haladyna (2004), Widdowson (1981), Millman et Greene (1989) seront utilisés. Certains ouvrages cités étant en priorité consacrés à l'évaluation dans le cadre de la classe, ne seront conservées que les informations pertinentes pour l'évaluation dans le cadre des tests et autres examens. Pour une revue plus complète des activités d'évaluation (vrai / faux, questions ouvertes, exercices d'appariement, tests de closure...), le lecteur intéressé est renvoyé aux écrits d'Alderson (2000, chapitre 7 ; 2005, chapitre 15), Koda (2005, chapitre 11), Haladyna (2004, chapitres 3 et 4) et Weir (2005b, chapitre 8).

En 2002, Haladyna, Downing et Rodriguez proposent une synthèse de 31 ouvrages traitant des directives et des conseils à suivre pour la rédaction des items à choix multiples. Leurs résultats mettent en avant l'existence d'une certaine « stéréotypie » dans les règles utilisées pour l'écriture des questions et plus généralement des questions à choix multiple (Q.C.M.). Ils montrent aussi que certaines des règles de rédaction le plus souvent préconisées par les manuels de rédaction des Q.C.M. n'ont pas été validées par la recherche. Voici quelques-uns uns des résultats de leur méta-analyse, et ce que la recherche dit aujourd'hui. Premièrement, l'amorce d'une Q.C.M. (partie de l'item précédent les leurres[f] et suivant la consigne donnée au candidat) peut amener à deux types d'activités : soit on répond à une question, soit on répond à une phrase que l'on doit compléter. Les recherches qui ont été menées sur l'usage de ces deux types d'amorces sont quelque peu contradictoires. Si toutes les recherches recensées (Crehan & Haladyna, 1991 ; Eisley, 1990 ; Raschor & Gray, 1996) s'accordent à dire que lorsqu'on utilise l'une ou l'autre des options, il n'y a pas de différences de la valeur de la discrimination (indice mesurant la capacité à distinguer les candidats ayant un haut niveau de compétence de ceux ayant un bas niveau), en revanche, une étude trouve une augmentation de la difficulté lorsque des questions sont posées. Haladyna (2004) conclut que, si les concepteurs de test peuvent utiliser les deux formats, sans doute est-il plus utile de poser des questions. Elles permettent de placer l'idée principale dans l'amorce de l'item plutôt que dans les leurres. Pour ce qui est de la formulation de l'item, le

[f] Dans la littérature en français, le terme anglais « distractor » est aussi traduit « distracteur ». Certains préfèrent ce terme à la traduction « leurre » (Morissette, 1996 :46).

vocabulaire utilisé doit être à la portée des candidats. La négation, lorsqu'elle est utilisée dans l'amorce produit des résultats contradictoires en terme de difficulté et de discrimination de l'item. Les auteurs de l'étude préconisent, le cas échéant, de les mettre en caractère gras pour éviter les confusions. Pour ce qui est du nombre de leurres par item, là encore, la littérature est contradictoire. Parfois, en réduisant le nombre de leurres, on augmente la difficulté, parfois, on la fait diminuer, on augmente la discrimination ou on la fait diminuer. Trois choix de réponse par item permettent d'atteindre une meilleure optimisation des ressources et de la qualité des items. C'est tout au moins l'idée maîtresse qui se dégage d'une méta-analyse de 80 ans faite par Rodriguez (2005). Si l'on choisit d'écrire des items à quatre choix de réponse, il faut savoir que des recherches empiriques ont démontré qu'il est difficile d'écrire quatre choix de réponse plausibles, que cela prend du temps à concevoir. Cela allonge le temps de réponse des candidats et, par conséquent, limite le nombre de questions auxquelles ils peuvent répondre. Toutefois, l'avantage des Q.C.M. à quatre choix de réponse est de permettre de diminuer les réponses données au hasard (de 33% pour les Q.C.M. à trois choix de réponse à 25% pour les Q.C.M. à quatre de choix de réponse). Cependant, encore une fois, cela n'est vrai que si le concepteur du test réussi à écrire trois leurres de bonne qualité (Haladyna, 2004 ; Rodriguez, 2005). Pour conclure sur ce point, sans doute faut-il retenir que, s'il n'y a pas de liens entre le nombre de leurres efficaces et la difficulté, en revanche, le nombre de leurres efficaces a un lien avec de meilleurs indices de discrimination (Haladyna, 2004). Si l'écriture d'items avec quatre choix de réponse permet de réduire les chances de réussite à l'item au hasard (Haladyna, 2004), réduire les items de quatre choix de réponse à trois réduit légèrement la difficulté, augmente légèrement la discrimination et augmente légèrement la fidélité (Rodriguez, 2005: 10).

Les aspects concernant l'ordre dans lequel apparaissent les items dans le test, quant à eux, semblent faire l'unanimité parmi les chercheurs. Tous préconisent de les proposer par ordre « logique ». Pour ce qui est de l'ordre des choix de réponse, lorsqu'ils sont placés au hasard, cela permet d'augmenter les indices de discrimination. On peut en déduire que le hasard n'est pas bénéfique aux candidats les plus faibles (d'où l'augmentation de l'indice de discrimination). L'homogénéité du contenu des choix de réponse et des structures grammaticales utilisées est recommandée par l'ensemble de la littérature.

Toutefois, cette affirmation manque de preuves empiriques. On n'a pas la preuve de son effet sur la difficulté et la discrimination. En ce qui concerne les choix de réponse « autre réponse » ou « aucune de ces réponses », les recherches consultées trouvent toutes qu'ils augmentent la difficulté. Pour ce qui est de la discrimination, certaines études trouvent qu'ils l'augmentent ou la font baisser. Étant donné l'interaction entre la difficulté de la question et ces leurres, il est préférable d'en limiter l'usage, voire de le supprimer. Pour ce qui est du leurre « toutes les réponses », la littérature conseille de ne pas l'utiliser. Entre autres choses, il fait baisser les indices de fidélité des tests. Enfin, pour la présentation des items, notamment dans les livrets destinés aux candidats, certains auteurs préconisent une présentation verticale (texte au dessus des questions), d'autres, une présentation horizontale (questions et texte placés en vis-à-vis). Selon Haladyna (2004), parce que la littérature ne permet pas de trancher la question, il est possible d'utiliser les deux formats.

Quand bien même les Q.C.M. sont fréquemment utilisées, notamment pour l'évaluation de la compréhension écrite, elles sont loin de faire l'unanimité parmi la communauté des scientifiques ou encore des praticiens. Selon certains, les Q.C.M. sont une activité séparée de la lecture. Alderson (1999 : 59) signale que Hudson (1996 : 19) a trouvé que le processus qui implique de sélectionner un choix de réponses dans une Q.C.M. est très différent de celui qui consiste à lire un texte puis à répondre à des questions ouvertes ou encore de celui qui consiste à rédiger un résumé à partir d'un texte. Par ailleurs, on peut augmenter son score en se préparant. Il est possible d'ajouter à cela qu'on n'enseigne pas à répondre aux Q.C.M. dans tous les pays. On peut apprendre à éliminer les leurres improbables ou à utiliser uniquement la logique pour répondre (on peut, dans un mauvais item, éliminer des choix de réponse sans lire le texte). De même, les candidats peuvent être amenés à choisir des réponses auxquelles ils n'auraient jamais pensé sans la présence des leurres. Comme le mentionne Haladyna (2004 : 98), Katz et Lautenschlager (1999), en donnant des items avec ou sans le texte correspondant, ont montré que certains individus peuvent trouver les bonnes réponses avec ou sans le texte :

> « They found that some students could perform because of their out-of-school experience and testwiseness, thus casting some doubt on the capacity of MC formats for measuring reading comprehension. The problem with measuring reading comprehension is not with

> the format used but with writing items that are truly passage dependent yet can be independently answered[17] ».

L'idée d'items fonctionnant différemment selon des groupes définis d'individus est présente ailleurs dans la littérature. Ainsi, Gorin (2005 : 368), dans une étude de l'évaluation de la compréhension écrite en anglais pour des natifs anglophones avec des Q.C.M., trouve des résultats qui ne concordent pas toujours avec la littérature scientifique. Alors que des manipulations sont faites sur les textes et les Q.C.M. pour en augmenter ou en diminuer la difficulté et la discrimination, elle ne trouve pas de résultats significatifs. Elle conclut que le problème vient de la spécificité de la population qu'elle a testée. Il semble donc encore difficile de généraliser le « savoir » acquis quant à l'écriture des items pour en manipuler la difficulté dans toutes les situations d'évaluation et avec toutes les populations. La vision de la difficulté de Bachman (2002) et Brindley (1987), qui résulte de la rencontre entre les caractéristiques des tâches d'évaluation, de celles de la population et l'interaction entre les deux, semble pouvoir être généralisée à l'évaluation faite avec des Q.C.M.

Si, comme il est possible de le comprendre, les Q.C.M. soient loin de convenir à toutes les situations d'évaluation, il semble important de ne pas les condamner trop rapidement. Ainsi, dans le champ de l'évaluation linguistique, Freedle et Kostin (1993, 1996, 1999) ont-ils démontré que leur difficulté est principalement reliée aux caractéristiques des textes et à l'interaction entre les textes et les items. Freedle (1997 : 400), à partir d'une analyse corrélationnelle, arrive à prédire tout de même 60 % de la variabilité de la difficulté d'items en compréhension écrite. En ce qui concerne leur complexité cognitive (c'est un grief souvent évoqué contre leur validité), Haladyna (2004 : 58) signale qu'Hibbison (1991) a montré qu'elles peuvent évaluer des aspects cognitifs plus complexes qu'il n'y paraît. Parfois, selon l'auteur, la « pauvreté » cognitive des Q.C.M. est plus due à l'absence de « talent » du rédacteur qu'à la possibilité réelle d'écrire de tels items. Cette remarque est importante, notamment, lorsque l'on réfléchit à la gestion d'une banque d'items et aux coûts qui lui sont associés. L'interaction entre la qualité des items (et donc le talent du rédacteur) et la qualité d'une banque d'items est très forte (Ariel, Van der Linden &Veldkamp, 2006). Dès lors, si on veut une banque d'items permettant

de proposer de nombreuses versions d'un même test, il paraît difficile de faire l'économie de la formation des rédacteurs aux « conseils » de rédaction pour avoir des items satisfaisants, tant au niveau du contenu que de leurs qualités psychométriques. Comme le signalent Brown *et al.* (2002 : 4) les Q.C.M., comme tous les autres types de questions, ont des avantages et des inconvénients. S'ils ont l'avantage d'être faciles à administrer, à corriger, leur inconvénient est d'être relativement difficile à concevoir et de ne pas évaluer le langage en production. Toutefois, selon les auteurs, ils permettent d'évaluer correctement les compétences de réception en lecture et en compréhension auditive. Alors que souvent on affirme que les questions construites demandant de la production sont meilleures que les Q.C.M., Brown, Hudson, *et al.* (2002 : 4) mettent en garde le chercheur et le praticien contre des positions trop tranchées. Si les Q.C.M peuvent donner l'occasion au candidat de trouver la réponse au hasard, dans une réponse construite ce dernier peut faire croire qu'il a un niveau qui n'est pas le sien :

> « Unfortunately, they also have the disadvantage that bluffing is possible (i.e., a student who is clever can construct a linguistically sophisticated and well-organized response that does not actually accomplish the task at hand, bet gets partial or even full credit), administration and scoring are time-consuming, and scoring is both difficult and somewhat subjective[18] » (Brown, Hudson, *et al.*, 2002: 4).

Pour la compréhension écrite, l'usage des Q.C.M. pourrait, entre autres, servir à augmenter la qualité psychométrique de la banque d'items et à baisser le coût de conception du test sans pour autant abandonner l'idée d'évaluer essentiellement la compréhension écrite. Le concepteur aura avantage à prendre en compte les conseils d'écriture énoncés par la recherche et à faire preuve parfois d'imagination. Au moment de la rédaction des items, il ne s'agit pas de viser le contrôle de la difficulté des items, mais d'améliorer leur fonctionnement, leur discrimination et d'éliminer les erreurs de mesure dues aux problèmes de rédaction. Il s'agit aussi de les utiliser quand l'économie du test le demande.

2.3.5.2.2 Le recueil d'information

Autre activité, plus intégrative, le recueil d'information demande au candidat d'extraire de l'information d'un texte ou de divers documents (graphiques, figures, entrevues…) pour les placer, entre autres possibilités, dans des tableaux. Le problème, avec ce type d'activité, c'est qu'on prend le risque de tester les capacités du candidat à manipuler des chiffres ou encore des statistiques (Alderson, 2000 : 248). Il faut donc veiller à ce que ce type d'activité soit en accord avec le construit du test.

2.3.5.2.3 Le résumé

Le résumé, bien que parfois perçu, à tort, comme un exercice scolaire est une activité relativement proche des activités de la vraie vie. Il présente une exigence de faisabilité ou encore de fidélité beaucoup plus importante que pour les autres activités d'évaluation. Il fait appel à des corrections subjectives que l'on a souvent beaucoup de peine à standardiser. Ainsi est-il ardu de savoir ce que l'on doit corriger : doit-on compter le nombre d'idées ou encore évaluer selon des critères ou indicateurs inclus dans une grille d'évaluation ? Une certitude, c'est que pour atteindre un minimum d'efficacité, il faut évaluer le résumé à plusieurs personnes pour voir quelles sont les idées principales et secondaires retenues. On pourra alors retenir les idées qui présentent de 75 à 100 % d'accords inter-juges (Alderson, 2000 : 233).

2.3.5.2.4 Le rappel libre

Quoiqu'il n'y ait pas en absolu de « meilleure » technique d'évaluation (Alderson, 2000 : 203), de l'avis de Koda (2005) mais aussi d'Alderson (2000), la technique qui semble être la plus efficace pour évaluer la compréhension écrite, est celle du « rappel libre ». Cette activité consiste à demander au candidat de lire un texte de le mettre le côté et, ensuite, d'écrire (ou de dire) tout ce dont il se souvient. Si c'est de la compréhension pure, c'est aussi très long à concevoir et à corriger. On va également avoir du mal à décider de ce que sont les idées principales et les idées secondaires. Une technique très similaire est celle qui est utilisée, aujourd'hui, dans certaines des tâches proposées dans le

T.O.E.F.L., nouvelle version, évaluant la compétence à lire en fonction d'un but (cf. le point 2.3.5.3. Les tâches d'évaluation discrètes et intégrées), soit lire pour apprendre et lire pour intégrer. Dans ces activités, on laisse quatre minutes au candidat après avoir lu le ou les textes puis on lui enlève le texte avant de l'inviter à en faire la présentation à l'oral ou à l'écrit (Trites & McGroarty, 2005 : 180-181). Le problème de ce type de tâche, c'est qu'elles posent de sérieux problème de généralisabilité. Pour les nouvelles tâches utilisées pour le T.O.E.F.L, Lee (2006 : 162) explique que si les tâches ont une difficulté moyenne identique, elles ne sont pas uniformément difficiles pour tous les candidats. Pour assurer une mesure optimale, il faut augmenter le nombre de tâches utilisées (ce qui est plus efficace que d'augmenter le nombre de correcteurs).

2.3.5.2.5 Les tâches de la vraie vie

Pour finir, les tâches de la vraie vie consistent à évaluer une tâche de lecture que le candidat aura à faire dans la vraie vie, selon des critères, eux aussi, de la vie réelle. Ici, le concepteur doit se demander ce que le candidat fera dans la vraie vie avec certains types de documents et comment il sera évalué lors de l'exécution de cette tâche. Là encore, l'usage de tâches de la vraie vie pose de sérieux problème de généralisabilité. Si on utilise de telles tâches, il faut veiller à ce qu'elles soient en adéquation avec ce que l'on veut évaluer et que l'on ne fasse pas de généralisations abusives des résultats (cf. pour de plus amples détails, le point déjà abordé dans la recension des écrits, 2.3.6).

2.3.5.3 Le continuum des tâches évaluant le processus ou le produit

Si la généralisabilité peut servir à différencier les tâches discrètes des tâches intégrées, il est encore possible de distinguer les deux types de tâches, selon qu'elles évaluent le processus (évaluation du « chemin » parcouru par le candidat pour arriver au résultat) ou le produit (prise en compte, exclusive, du résultat). Traditionnellement, les activités discrètes (ou encore indépendantes) sont réputées pour privilégier le produit, les activités intégrées le processus et le produit. Les limites de l'approche produit sont la variabilité des « résultats » auxquels arrivent les candidats (les lecteurs parfois interprètent les textes différemment sans que ces interprétations soient bonnes ou mauvaises) et la méthode

utilisée pour la mesure, les lecteurs n'ayant pas tous la même mémoire de ce qu'ils ont lu (Alderson, 2000 : 5). L'approche processus, elle, met en avant l'interaction entre le lecteur et le texte (voir, déchiffrer, penser au sujet de la lecture,…). Pour mettre en place une telle évaluation, les concepteurs identifieront des micro-compétences et des stratégies de lecture que le lecteur devra mettre en œuvre pour arriver au résultat escompté. Le problème de cette deuxième approche, c'est qu'il est difficile de prévoir quelles sont les micro-compétences ou les stratégies spécifiques que le candidat utilisera. Parfois, lorsque les candidats utilisent la micro-compétence prévue, ils répondent mal et, d'autres fois, ils utilisent des micro-compétences non-prévues et répondent correctement (Alderson, 2000 : 304-305 ; Li, 1992). Si Mokhtari et Reichard (2004) montrent que la conscience des habiletés méta-cognitives de deux groupes aux langues et cultures différentes mais de même niveau de lecture, est quasiment identique pour les deux groupes, en revanche, les stratégies de lecture les plus citées et les moins citées par les deux groupes ne sont pas les mêmes. On peut donc avoir un niveau de lecture identique, avec un même niveau de conscience des habiletés méta-cognitives mais ne pas déclarer utiliser les mêmes stratégies. La variabilité des stratégies semble donc être présente à tous les niveaux que ce soit dans les stratégies utilisées par les lecteurs ou encore la conscience qu'ils en ont. Ces résultats vont dans le sens de ceux trouvés pour la lecture en langue maternelle, résultats qui stipulent que les meilleurs lecteurs sont ceux qui utilisent le plus de stratégies sont ceux qui sont capables de retenir le plus d'informations à partir d'un texte (Goldman, 1997 : 360-361). Ces résultats pourraient être généralisés à la compétence de compréhension en général, les évaluations de la compréhension orale et auditive étant fortement corrélées (Perfetti, 1997 : 343).

Pour ce qui est de l'évaluation et notamment de l'évaluation de la compétence de compréhension écrite, évaluer le processus est un vrai défi (Alderson, 2000 : 5). Cela étant dit, évaluer le processus est surtout utile pour l'évaluation diagnostique (Alderson, 2000 : 306). Pour les tests à enjeux critiques mais aussi les tests de positionnement, l'approche produit a sa raison d'être. Le problème n'est pas tant de savoir comment le lecteur lit, mais plutôt ce qu'il lit, comprend, ou encore, ce qu'il peut faire à partir de ses lectures.

2.4 Les modèles de mesure

La partie de la recension des écrits sur les tâches de lecture aura permis de constater que les évolutions récentes vont vers une plus grande diversité des tâches utilisées. S'il est possible de diversifier le type de tâches, sans doute cela vient-il en partie d'une demande du public ou encore des institutions consommatrices d'évaluation linguistique. Aujourd'hui, l'usage de l'outil informatique pour les passations d'examen semble ouvrir la voie à d'éventuelles réponses aux demandes de diversification des tâches d'évaluation. L'ordinateur permet de renouveler le type de tâches administrées parce qu'il facilite la mise en œuvre de certains aspects de l'administration des tests. Il autorise, non seulement, l'évaluation des tâches sur différents supports (écrits, oraux, authentiques, fabriqués), mais aussi, un stockage immédiat des données issues des passations, ou encore, une adaptabilité du niveau des items au niveau du candidat (l'estimation du niveau du candidat se faisant instantanément par la machine). L'ordinateur a donc l'immense intérêt de faciliter le recours aux modèles de mesure.

Ceci dit, comme le signale Dooey (2008), Green et Maycock (2004 : 4) mettent en garde les concepteurs de test. Ce n'est pas parce que les choses deviennent possibles sur le plan technique qu'elles sont valables. L'usage de modèles de mesure peut ainsi nuire à la variété du type de tâche utilisé dans un test. Si l'ordinateur a permis l'usage de nouveaux modèles de mesure (ceux de la théorie de réponses aux items), encore faut-il démontrer leur pertinence. Aujourd'hui, certains concepteurs de tests favorisent l'usage de la théorie de réponse aux items (T.R.I), même s'ils ne délaissent pas totalement la théorie classique des items. Mais pourquoi la T.R.I. a-t-elle un tel succès ? En quoi semble-t-elle mieux répondre aux besoins de diversification des tâches d'évaluation ? Quels sont ses avantages du point de vue de la qualité de la mesure ?

2.4.1 Théorie classique des items, différences avec la théorie des réponses aux items (T.R.I .)

La théorie classique des items ne sera pas présentée dans le détail. Le lecteur intéressé pourra lire, entres autres références, les ouvrages fondateurs de Gulliksen (1950), de Lord et Novick (1968) ou encore celui en français de Laveault et Grégoire (2002). Le but, ici,

est surtout de présenter quelques unes des propriétés du modèle, comme ont pu le faire Bertrand et Blais (2004), Laveault et Grégoire (2002). Il s'agit encore de différencier de la théorie classique des items de la théorie de réponse aux items (T.R.I.), comme l'ont fait Embretson et Reise (2000).

Fondamentalement, la théorie classique s'attache à obtenir un score observé qui soit le plus proche possible du score vrai des candidats. Le but est donc d'obtenir une mesure libre d'erreur de mesure. Par ailleurs, l'erreur de mesure doit être due au hasard (Bachman, 2004 : 157). Voici la définition que donnent Bertrand et Blais (2004 : 40) du score vrai (libre d'erreur de mesure) dans cette théorie :

> « Le score vrai d'un individu à un test donné [est] la moyenne des scores observés obtenus lorsque le même test est administré à cet individu un très grand nombre de fois (un nombre indéterminé de fois ! ».

Voici les propriétés[g] du modèle telles que décrites par Bertrand et Blais (2004) :

- la moyenne des erreurs de mesure pour un individu à qui on a administré un test un très grand nombre de fois est nulle,
- Il doit y avoir une corrélation nulle entre les erreurs de mesure et les scores vrais d'un grand groupe d'individus,
- la corrélation entre les erreurs de mesure de deux tests différents administrés à une même population d'individus doit également être nulle,
- deux formes d'un même test sont parallèles si la variance des erreurs de mesure des deux formes est la même,
- la variance des scores observés d'un test (variance totale) est égale à la somme de la variance des scores vrais et celle des erreurs de mesure.

L'avantage de cette théorie est peut-être sa simplicité mais surtout qu'elle s'accommode des statistiques descriptives habituelles. Toutefois, il faut bien dire que le problème de cette théorie c'est qu'elle ne permet pas d'obtenir des paramètres d'items invariants (l'estimation de la difficulté des items et de la compétence des personnes dépend des

[g] Pour information, Laveault et Grégoire (2002 :106- 108) parlent de « postulat » et non de propriétés. Nous préférons l'emploi du terme propriété.

items et des personnes). Il ne peut donc pas réellement proposer des tests comparables. Bertrand et Blais (2004 : 108) font remarquer à ce sujet que la théorie classique manque de « réalisme ». Il est clair que la théorie classique s'inscrit dans une perspective déterministe où l'on décrit les données alors que la T.R.I. prétend modéliser les données. Cela permet d'identifier les tâches qui ne se conforment pas au modèle (*misfit*).

Embretson et Reise (2000, chapitre 2) opposent la théorie classique des items et la théorie des réponses aux items à partir de « règles anciennes et nouvelles ». Voici ce que déclarent les auteurs.

- Dans la théorie classique, l'erreur de mesure dépend de l'échantillon. Elle est constante alors que, dans la T.R.I., elle dépend de la population en général et qu'elle varie selon le score obtenu.
- Dans la T.R.I., il n'est pas obligatoire d'avoir un test long pour augmenter la fidélité. Lorsque l'on a des items correspondant au niveau des candidats, l'erreur de mesure pourra être moindre avec peu d'items, alors que, dans la théorie classique, il faut des tests parallèles pour pouvoir comparer les résultats entre les candidats. Dans la T.R.I., on veut avoir des tests évaluant des niveaux de difficulté appropriés à chacun des candidats.
- La difficulté, dans la théorie classique, est la proportion de personnes ayant réussi. La discrimination est la corrélation entre le résultat à l'item et le résultat total, soit la corrélation bisérielle (ou point-bisérielle). Ces deux statistiques varient selon l'échantillon utilisé et surtout si l'échantillon utilisé n'est pas représentatif de la population cible. Dans la T.R.I., ces statistiques ne sont pas censées dépendre de l'échantillon des personnes ou des items[h].
- Enfin, dans la théorie classique, une analyse factorielle menée à partir d'items à choix dichotomiques produit des artefacts plutôt que des facteurs (le problème étant que les facteurs se constituent à partir d'items de même niveau de difficulté). Selon certains chercheurs, avec la T.R.I., l'analyse factorielle faite à partir de toutes les données brutes liées à l'item est plus performante (« *Full information factor analysis* ») (Embretson & Reise, 2000: 36-37; Bock, Gibbons & Muraki, 1998).

[h] Bien que les auteurs parlent de « statistiques », il est certainement plus approprié de parler d'indices. Dans la T.R.I., avec le modèle à deux paramètres, on obtient un para mètre de discrimination.

Dans le contexte d'un besoin d'une diversification des tâches d'évaluation, d'une utilisation des items aux paramètres invariants, d'une évaluation, parfois, à valeur plus nationale que locale (ou encore pour les tests dits internationaux, plus internationale que nationale), et (dans certains cas) d'une meilleure adaptabilité des tests, il est clair que la T.R.I. est mieux armée que la théorie classique pour répondre aux défis actuels. Cependant, dans certaines situations, notamment dans les contextes locaux, le coût d'utilisation de la T.R.I. et sa complexité ne sont peut-être pas nécessaires et la théorie classique se montre encore utile. Par exemple, la théorie classique peut être utilisée pour certaines étapes de la calibration des items. Sans doute faut-il voir la T.R.I et la théorie classique comme des instruments complémentaires plutôt que rivaux. Toutes deux répondent à des besoins et usages spécifiques.

2.4.2 Les spécificités des différents modèles de la théorie des réponses aux items (TRI)

Lorsque le praticien (ou encore le chercheur) fait le choix d'utiliser un modèle de mesure de la T.R.I., dites encore « théorie du trait latent » (Birnbaum, 1968), en fonction d'un construit inscrit dans une perspective unidimensionnelle, deux options ou « paradigmes » (Andrich, 2002, 2004) s'offrent à lui ainsi que plusieurs modèles ou encore familles de modèle. Tout d'abord, il doit décider s'il veut que les données de sa recherche correspondent au modèle de mesure utilisé ou s'il souhaite que ce modèle corresponde aux données. Comme le précise Andrich (2002, 2004), cela demande au chercheur d'effectuer un positionnement « paradigmatique » le situant dans une utilisation particulière du modèle à un paramètre.

Lorsque le choix est fait d'utiliser un modèle à un paramètre dans lequel on veut que les données soient ajustées au modèle, on peut choisir de travailler avec un des modèles de la famille de Rasch. S'il existe plusieurs modèles à un paramètre, visiblement ils sont un peu trop rapidement regroupés sous l'étiquette « modèle de Rasch ». Ce qui distingue essentiellement ces modèles, c'est le choix que l'on fait d'avoir un modèle qui colle aux données ou encore le choix que l'on fait d'avoir des données qui collent au modèle (pour plus de détails sur les différences entre les types de modèle à un paramètre, le lecteur intéressé pourra notamment lire Bertrand et Blais (2004: 128) et *Rasch Measurement*

Transactions[i] (2005). Ainsi les modèles à un paramètre non classés dans la famille de Rasch sont utilisés quand ils sont adaptés à la nature des données et qu'il n'y a pas lieu d'utiliser des modèles correspondant mieux aux données. Si on préfère utiliser un modèle qui soit ajusté aux données, qui en épouse les formes, dans ce cas, on utilisera (selon les données que l'on veut décrire) des modèles à un, deux, trois ou quatre paramètres, soit dans l'ordre, la difficulté, la discrimination, la pseudo-chance (« *guessing* ») et l'inattention ou l'étourderie (« *carelessness* »). Dans ce paradigme, on choisira le modèle en fonction de la nature des données (Choi & Bachman, 1992 : 74). On procèdera alors par élimination:

> « Model selection can be aided by an investigation of the principal assumptions underlying the popular unidimensional item response models. Two assumptions common to all these models are that the data are unidimensional and the test administration was not speeded. An additional assumption of the two-parameter model is that guessing is minimal; a further assumption of the one-parameter model is that all item discrimination indices are equal.[19] » (Hambleton, Swaminathan, & Rogers, 1991 : 55)

Concrètement, voici le type de raisonnement qui peut être tenu dans ce type de paradigme :

> « The assumption of no guessing is most plausible with free-response items, but it often can be met approximately with multiple-choice items when a test is not too difficult for the examinees[20]. » (Hambleton, Swaminathan, & Rogers, 1991 : 55)

En ce qui concerne les modèles de la famille de Rasch, selon Andrich (2002, 2004) et Linacre (2006b) mais aussi Luecht (1999 : 198), leur utilisation permet de découvrir des anomalies dans les données. On peut alors étudier ces anomalies même si parfois leur analyse est ardue :

> « However, in constructing data to fit the model, there is potential for the model to disclose anomalies in the data, anomalies that must be understood, but for which the model provides no substantive answer-merely clues as to where to look[21] » (Andrich, 2002: 332).

[i] Il n'est pas fait mention de l'auteur dans l'article publié sur le site internet de la revue.

Dans le cadre du test de compréhension écrite du T.O.E.F.L., Choi et Bachman (1992 : 63) montrent que le modèle de Rasch permet de découvrir beaucoup plus d'items présentant des problèmes d'ajustement au modèle que les modèles à deux ou trois paramètres.

Andrich (2002 : 351) ajoute que lorsque le modèle indique que les données ne sont pas ajustées au modèle, cela signifie que l'on est face à un problème de mesure. Essentiellement, les données ne permettent pas d'avoir des conditions suffisantes pour la mesure, pour assurer son invariance, soit la stabilité de l'estimation du niveau de difficulté de l'item et de la compétence des personnes :

> « Instead of simply describing data, the Rasch models provide an opportunity to understand data by the exposure of anomalies, which is the prime function of measurement in physical science research (Kuhn 1961/1977). The reason that the Rasch model can be used in this way is that the case for the model is, not that it describes any data, but that it formalises conditions for invariance, which lead to properties of measurement. Thus when a data deviate from a Rasch model, it deviates from the requirement of measurement. To consider that when there is a mismatch between the data and the model it might be a problem with the data rather than the model, is in itself a considerable perceptual shift from the traditional perspective on the data-model relationship[22] » (Andrich, 2002: 351).

Dans ce paradigme, si on ne change pas de modèle, on « supprime » les données qui provoquent des anomalies. Ces suppressions interviennent lorsqu'elles permettent d'améliorer la mesure. Cette démarche pose souvent des problèmes à ceux pour qui les données sont « sacrées » et « sont la vérité ». Or, dans le modèle de Rasch, c'est le trait latent qui « est la vérité » et non pas les données. Il importe donc de ne pas accepter de données amenant à une mesure s'écartant trop du trait latent :

> « Statisticians can find it difficult to adjust to Rasch methodology. They tend to believe that the data points tell the truth and that it is the task of statisticians to find models, which explain them, and to find the latent variables, which underlie them. Rasch methodology takes an opposite position. It says that the latent variable is the truth, and when that latent variable is expressed in linear terms, it is the Rasch model that is necessary and sufficient to describe it. Consequently those data points, which do not accord with the Rasch model,

> are giving a distorted picture of the latent variable. They may be telling us very important things, e.g., "the students were disinterested", "the scoring key was wrong" - but those do not pertain to the central variable[23] » (Linacre, 2006a : 5).

En admettant qu'un chercheur choisisse d'utiliser un modèle de la famille de Rasch pour mesurer la compétence de lecture dans une optique de mesure « unidimensionnelle » ou tout au moins généralisable à une population plus ou moins spécifique, il devra opérer son choix en fonction du type de tâches d'évaluation qu'il utilisera ou encore à l'analyse qu'il voudra faire des données. S'il évalue la compétence des candidats à l'aide de questions à score dichotomique (ex, réussi / échoué, question à choix multiple ou vrai/faux, etc.) il pourra utiliser le modèle dit « modèle de Rasch classique ». S'il veut utiliser des échelles d'appréciation (échelles de Likert) ou encore des scores partiels (par exemple, un score de 15 point sur 20), il pourra utiliser soit le modèle d'Andrich, appelé encore « *rating scale* » (Andrich, 1978a, 1978b) soit le modèle de Masters, dit aussi « *partial credit* » (Wright & Masters, 1982). Enfin, s'il veut étudier chacun des aspects d'une situation d'évaluation (item, compétence, tâche, évaluateur,...), autrement dit les « facettes » de l'évaluation, il pourra utiliser le modèle étendu de Rasch, soit le modèle multi-facettes développé par Linacre (1989).

D'un point de vue conceptuel, tous les modèles de Rasch proposent une relation mathématique entre la compétence et la difficulté, autrement dit, une relation entre deux facettes ou plus (comme pour le modèle multi-facettes). Ce sont des modèles de mesure probabilistes. La calibration des items et des personnes permet de produire des estimations de la compétence des personnes et de la difficulté des items, et de les placer sur une échelle commune. Lorsqu'un item a un niveau comparable à celui d'une personne, cette personne a 50% de chance de réussir ou d'échouer à l'item (Bond & Fox, 2001). La difficulté et la compétence sont placées sur une échelle arithmétique commune, celle des *logits*. Cette échelle possède des propriétés additives et soustractives. Lorsque l'échelle est centrée sur la difficulté des items, la valeur 0 de la difficulté des items, qui est aussi la moyenne de l'échelle, est une valeur arbitraire. De même, le *logit* n'a en soit aucune valeur et la valeur qu'il prend dépend de l'échantillon et des items utilisés (Bond & Fox, 2001). Les modèles de Rasch permettent encore d'étudier les configurations de

réponses pour voir comment elles renforcent la configuration générale ou bien la contredisent («*fit* »). Cela permet, non seulement, de détecter les items qui fonctionnent correctement de ceux qui fonctionnent mal, mais aussi, les configurations de réponses suspectes des candidats. Il s'agit d'un avantage non-négligeable.

Le modèle « *rating scale* » permet d'analyser les items qui utilisent des échelles d'appréciation, comme les échelles de Likert (Bond & Fox, 2001). Ce modèle ne peut être utilisé que si les items du test possèdent tous le même nombre de catégories de réponse (par exemple, « 1, 2 ,3 ,4 et 5 » ou encore, « jamais, souvent, toujours »). Fox et Bond (2001 : 160), suite à la recension des écrits qu'ils ont faite, expliquent que le nombre de catégories utilisées (par exemple, « jamais », « parfois », « toujours ») doit être défini empiriquement et non pas d'une position purement théorique :

> « It is therefore the job of the test developer to determine empirically the optimal number of response categories every time a new rating scale is developed or when an existing rating scale is developed or when an existing rating scale is used with a new population. Thus, the analyst must discover empirically, rather than assert, the optimal number of rating scale categories for measuring a given construct[24] » (Lopez, 1996).

McNamara (1996 : 255), quant à lui, explique que le modèle "*rating scale*" peut aussi permettre de comprendre comment les évaluateurs interprètent chaque point de score brut sur l'échelle de mesure et quelle est la consistance des interprétations de chaque « étape » (« *step* »). Si dans la littérature, certains comme Masters (1982) et McNamara (1996) parlent de « steps » (Molenaar, 1983) ou encore d'intervalles (Linacre, 2006b), d'autres parlent de « seuil » (« *threshold* », Andrich, 2002) ou encore « d'intersections » (Linacre, 2005). Selon Andrich (2002 : 349), avec le modèle Rasch, il est plus approprié de parler de « seuils » (« *threshold* »), car, selon lui, l'idée de distance, véhiculée par le terme « *step* » n'est pas appropriée au modèle de Rasch. Dans ce modèle, ce qui importe ce n'est pas tant la distance entre les intersections mais l'ordre des catégories. Encore une fois, le but du modèle de Rasch, n'est pas de décrire les données mais de vérifier si les données récoltées avec les différents dispositifs d'évaluation, sont aptes à produire des estimations de la difficulté des items et de la compétence des candidats stables et un score plus juste (Kubinker, 2005 : 381). Dès lors, si les catégories ne sont pas bien ordonnées

(par exemple, les catégories « jamais, souvent, toujours » sont placées sur l'échelle de *logits* dans l'ordre « jamais, toujours, souvent »), on n'a pas un problème d'interprétation, mais plus directement un problème avec les données.

Le "*partial credit model*" permet, lui aussi, de mesurer la différence d'interprétation des graduations de l'échelle par les évaluateurs. Toutefois, il le fait de manière plus raffinée puisqu'il permet l'analyse de la structure des intersections pour des items n'ayant pas le même nombre de catégories. Lorsque l'on utilise le « *partial credit* », on peut avoir à la fois des items à score dichotomique (vrai /faux, question à choix multiple) et des items polytomiques (ex : score sur une échelle allant de 0 à 3) (Bond & Fox, 2001 : 89). Ce modèle permet également de savoir comment les évaluateurs interprètent et utilisent les scores aux tests (McNamara, 1996 : 256) comme dans le modèle d'Andrich.

Le modèle multi-facettes, quant à lui, est une extension du modèle classique de Rasch à deux facettes, soit la compétence du candidat et la difficulté de l'item (Bachman, 2004 ; Bachman, Lynch & Mason, 1995 ; Fox & Bond, 2001 ; McNamara, 1996). Il permet de situer les deux facettes du modèle de base, sur une échelle de *logits*, en fonction de facettes supplémentaires (par exemple, le candidat, le type de tâche, l'évaluateur, les conditions de la passation…). Selon Bond et Fox (2001 : 230), le concept de facette renvoie aux aspects touchant les conditions de passation. Avec le modèle multi-facettes, il est possible de calculer le niveau des candidats et des tâches, au moyen d'un processus itératif qui a pour but de trouver la meilleure correspondance entre les configurations des scores observés et les prédictions du modèle, mais cette fois-ci en fonction de facettes (McNamara, 1996 : 134). On peut (avec les statistiques de l'ajustement des données au modèle, « *infit* » et « *outfit* »[j]) vérifier le fonctionnement des facettes ou détecter un dysfonctionnement. Par exemple, il est possible de savoir si un évaluateur n'est pas constant dans ses évaluations (« *misfit* ») ou encore si une facette présente moins de variabilité que celle que l'on attendait (« *overfit* »). Selon Bachman, Lynch et Mason (1995 : 255) et Bachman (2004 : 148-149), les avantages du modèle multi-facettes sont les suivants :

[j] Ces statistiques seront présentées plus en détail dans la partie sur la méthodologie.

- fournir des informations sur une tâche, une personne, un évaluateur,
- faire des pondérations quand on utilise le *logit* à la place du score brut,
- découvrir des biais (estimation du degré avec lequel des combinaisons de personnes et des différentes conditions sur les facettes de mesure sont plus ou moins inconsistantes ou quand il y a un biais),
- former les évaluateurs.

Enfin, Bond et Fox (2001 : 110) ajoutent que, par rapport aux analyses par variance l'avantage du modèle multi-facettes est qu'il permet non seulement de vérifier la consistance du classement des candidats par les différents évaluateurs, mais aussi, et surtout, de pouvoir comparer la sévérité ou la générosité des évaluateurs. Cette donnée peut être prise en compte pour le calcul de la probabilité d'une bonne réponse :

> « The probability of any correct response is a function of the ability of the person and the difficulty of the item, with allowance made for the severity of the rater, and for which particular form of the test was taken (i.e., probability = function of (ability-difficulty-rater-test))[25] » (Bond & Fox, 2001: 110).

2.4.3 Le fonctionnement différentiel (F.D.I.) et les biais

Quels que soient les modèles utilisés dans la théorie des réponses aux items, ceux qui les utilisent, en général, le font parce qu'ils veulent une mesure non dépendante de la distribution d'une population (et donc obtenir des paramètres d'item invariants), ou encore, parce qu'ils ont besoin de créer des versions différentes d'un même test à partir d'une banque d'items (et donc avoir des versions similaires d'un même test sans que ces versions soient strictement parallèles). À un niveau plus général, et surtout pour les tests visant un public large, multiculturel et multilingue, les concepteurs de tests veulent que les mesures qu'ils proposent soient justes pour tous. Ces mesures ne doivent pas favoriser ou défavoriser un sous ou des sous-groupes de manière systématique. Si la théorie classique répond mal à ce genre de demandes, la T.R.I. est plus en mesure de fournir des réponses adéquates.

Dans une optique unidimensionnelle, après avoir choisi un modèle de mesure de la T.R.I., quand les données ont été calibrées, il est possible d'utiliser les données du test pour enquêter sur les fonctionnements différentiels et les biais de personne ou d'item. Le but

de ce type d'analyse est de vérifier la qualité de la calibration des items mais aussi de la validité du test. Hambleton, Swaminathan et Rogers (1991 : 109) distinguent le fonctionnement différentiel d'item (F.D.I.) du biais de la manière suivante : selon eux, le biais évoque une conclusion (par exemple, « les items discriminent les hispanophones »), alors que le F.D.I. désigne la preuve empirique d'un fonctionnement différent d'un groupe à l'autre sans établir de causalité. Il est important de noter que les F.D.I. ne sont pas systématiquement synonymes de biais. Le F.D.I. est une condition nécessaire mais pas suffisante pour le biais d'item (Bertrand & Blais, 2004 : 285 ; Shimizu & Zumbo, 2005 : 113).

Selon Shimizu et Zumbo (2005 : 113), l'analyse F.D.I. est une procédure statistique qui vise à déterminer si les items du test sont appropriés pour mesurer le savoir de plusieurs sous-groupes de candidats. Le postulat qui soutient les F.D.I. est que les candidats qui ont une compétence similaire doivent avoir des résultats similaires pour des items individuels du test. Ces résultats ne doivent pas être affectés par le bagage culturel du candidat, ses appartenances culturelles, sa langue ou autre. Si des items particuliers fonctionnent différemment pour des groupes spécifiques de candidats, ils doivent refléter un biais qui n'est pas relié au domaine devant être testé. Dans les publications, on trouve cinq buts lorsque l'on effectue une analyse des F.D.I. On peut :

- vouloir assurer la justesse et l'équité dans les tests (selon la langue, la communauté ethnique ou le genre),
- fournir une preuve en cas de litige (quand un candidat dépose une plainte),
- enquêter pour savoir si les items changent en termes de difficulté et de discrimination à travers le temps,
- utiliser les F.D.I. pour faire des comparaisons de groupes et des mesures de violation des règles psychométriques comme méthode d'explication de la différence entre les groupes
- ou encore comprendre le processus de réponse aux items.

Depuis quelques années, on est passé à une nouvelle génération d'étude du F.D.I. :

> « Today, […] in addition to matters of bias, DIF technology is used to help answer a variety of basic research and applied measurement questions wherein one wants to compare item performance between or among groups when taking into account the ability

> distribution. At this point, applications of DIF have more in common with the research methodology aligned with analysis of covariance (ANCOVA) or attribute-by-treatment-interaction (ATI) than test bias per se.[26] » (Shimizu & Zumbo, 2005: 112).

L'étude des F.D.I. a de multiples applications possibles. Entre autres, elle permet de travailler sur la traduction des tests ou encore l'adaptation interculturelle. Shimizu et Zumbo (2005 : 112) ajoutent que beaucoup d'applications de F.D.I. sont mises en place parce que les précédentes études de différence de groupe comparaient les différences de moyenne de performance des groupes sans prendre en compte le continuum de compétence sous-jacent. S'il est « à la mode » de critiquer les analyses F.D.I. pour ne pas expliquer la raison des différences à la performance, il faut se rappeler que dans les analyses F.D.I. la raison (ou causalité) est secondaire. On cherche avant tout à garantir l'adéquation des inférences faites à partir des scores au test et donc à réduire les biais de test affectant les sous-groupes de candidats.

2.5 L'évaluation de la compétence de compréhension écrite dans les tests adaptatifs par ordinateur

L'émergence de l'étude des fonctionnements différentiels, si elle relativement récente, montre l'intérêt grandissant des utilisateurs pour une plus grande justice et justesse des tests. Phelps (2005), en dirigeant l'élaboration d'un ouvrage collectif pour la défense des tests standardisés, se fait l'écho du besoin d'une évaluation impartiale. Sans vouloir entrer dans les considérations philosophiques, le problème posé au concepteur de test contemporain est de répondre à une demande d'évaluation complexe. Cette demande oscille entre le général et le particulier. Les résultats individuels doivent être comparables et être en adéquation avec ceux de l'ensemble d'une population. Les tests doivent souvent permettre d'évaluer des échantillons de population de candidats ayant des niveaux de compétence très étendus, sans que le test ne soit, ni trop long, ni trop cher. Il doit aussi fournir une estimation suffisamment précise du niveau de chaque candidat, sans biais ou fonctionnement différentiel. Parmi les différents types de tests existant, les tests adaptatifs par ordinateur sont ceux qui idéalement semble le mieux répondre à plusieurs de ces défis.

2.5.1 L'évaluation par ordinateur

Avant d'expliquer ce que sont les tests adaptatifs par ordinateur, il convient de se poser la question de la légitimité de l'évaluation par ordinateur. En 2006, le choix de l'ordinateur (et cela est valable pour tous les tests informatisés) semble susciter chez les candidats de moins en moins de craintes quant à la légitimité du mode d'administration et des résultats. Déjà, il y a une dizaine d'années, Laurier (1993: 215-216) a montré que le stress des candidats passant un test de langue en français langue seconde dans un environnement informatique, est le même que dans un environnement « papier-crayon ». Aujourd'hui, dans nombre de pays, les activités quotidiennes de lecture peuvent être réalisées soit sur support papier, soit sur des écrans d'ordinateurs. Par conséquent, il semble légitime de l'utiliser pour évaluer un niveau général de compétence de compréhension écrite en langue seconde (Linacre, 1999a). Pour le concepteur de test, l'ordinateur permet un usage facilité de modèles d'évaluation de plus en plus sophistiqués. Il peut encore aider à la constitution de bases de données constituées à partir des réponses des candidats et permettre une analyse rapide (Alderson, 1999 : 68). Dans le cas plus particulier du testing adaptatif, l'ordinateur permet de traiter les réponses des candidats *in situ*, d'avoir un degré d'interactivité homme-machine plus important que dans le cas des tests papier-crayon passés sur ordinateur (Laurier, 1993: 37). L'ordinateur, de par sa puissance de calcul, permet de faciliter la mise en œuvre de l'adaptabilité qui vise une meilleure précision de l'estimation du niveau des candidats ou bien un contrôle du contenu du test. Selon Embretson et Reise (2000 : 268), l'utilisation du testing adaptatif avec l'ordinateur présente les avantages suivants : l'ordinateur peut continuellement contrôler les qualités psychométriques des items, on peut enregistrer le temps de réaction du candidat (si cela est utile) et on peut utiliser de nombreux gabarit d'items dans un même test. Ce dernier aspect n'est pas inutile, au vu du contexte actuel et de l'émergence de nouveaux formats d'items ou de type de tâches (Alderson, 2005 ; Cumming *et al.*, 2006 ; Trites & Groarty, 2005) et l'opportunité octroyée par certains modèles de la T.R.I. d'utiliser des items de formats différents.

2.5.2 Les particularités du testing adaptatif

Si on tente de donner une définition de ce qu'est un test adaptatif, on peut dire pour commencer que ce sont des tests qui utilisent des items pré-calibrés et stockés dans une banque (Wainer & Mislevy, 2000 : 78) et qui sélectionnent les items proposés à chaque candidat en fonction de règles préétablies. Ces tests estiment, au moment même de l'administration des items, le niveau du candidat. Ils utilisent l'estimation de la compétence des personnes (préalablement établie ou récemment calculée) pour choisir des items en fonction d'un niveau de difficulté (plus ou moins proche de l'estimation du niveau du candidat). Chaque test adaptatif, comme les autres types de tests, suit un algorithme, autrement dit, un ensemble de règles spécifiant les règles de la passation, soit le commencement du test, son déroulement et sa fin. Cependant, dans le cas des tests adaptatifs, les algorithmes utilisés sont beaucoup plus complexes et de nature adaptative (Thissen & Mislevy, 2000). Dans un tel test, le concepteur doit donc décider de :

- la façon dont sera commencé le test : estimation initiale du niveau de compétence du candidat… ;
- comment les items proposés aux candidats seront choisis au cours de la passation : estimation du niveau des candidats après chaque réponse donnée, choix des items les plus informatifs (autrement dit, les items permettant de réduire au maximum l'écart entre l'estimation du niveau du candidat et son vrai niveau), choix des items en fonction de contraintes de contenu, ou encore en fonction de contraintes concernant la passation du test, choix mixtes… ;
- comment et quand l'arrêter : en fonction de la correction de l'estimation du niveau, du nombre d'items administrés, du temps écoulé, du contenu minimal que le test doit évaluer… (Laurier, 1993 ; Thissen & Mislevy, 2000 ; Van der Linden, 2005).

Les tests adaptatifs font généralement appel à un modèle de réponse aux items. Cependant, ils n'utilisent pas tous le même modèle. De la même façon, on ne peut pas réduire les tests adaptatifs à des types d'activités d'évaluation ou encore à des designs particuliers. Par exemple, on peut faire le choix d'utiliser des tests à plusieurs « étages » (Laurier, 1993, 2004) pour évaluer des compétences différentes à différents moments du

déroulement du test (par exemple, compréhension écrite puis expression écrite,…). On peut aussi choisir des algorithmes différents en fonction du but poursuivi par le test. Van der Linden (2005 : 284) explique que lorsque l'on utilise des algorithmes séquentiels (estimation de la compétence du candidat effectuée après chaque item administré et prévision faite pour le prochain item à administrer), on ne se préoccupe que du maximum d'information, autrement dit, de la précision de la mesure de la compétence du candidat sans se préoccuper du contenu. L'information, notée I(θ), qui est l'inverse du carré de l'erreur-type de mesure associée à l'estimation de la compétence (Bertrand & Blais, 2004: 143) (ou encore, la valeur attendue de l'inverse de la variance de l'erreur de mesure associée à l'estimation de la compétence) a des propriétés additives et varie en fonction du niveau de compétence du candidat, de difficulté et de discrimination de l'item. La quantité d'information de chaque item dépend de la valeur de ses paramètres (Wainer & Mislevy, 2000: 73-75). Dans une optique séquentielle, la règle de sélection de l'item suivant ne répond qu'à la contrainte d'avoir un item suivant le plus informatif possible, délaissant quelque peu le contenu. Au contraire, lorsque l'on utilise un algorithme simultané, on ne sélectionne pas des items individuels mais des séries de combinaisons de n items. On choisit la meilleure combinaison possible d'items restant à passer en fonction de contraintes définies avant la passation.

Pour construire un test adaptatif, il faut donc choisir un construit, circonscrire un domaine et un contenu d'évaluation, concevoir des items respectant les contraintes liées au domaine et au contenu, choisir un modèle de réponse aux items, choisir un design et des contraintes adaptés au construit. Ce modèle doit permettre de faire des inférences sur la compétence des candidats et d'interpréter les scores des candidats. On devra aussi prendre en compte les possibilités et les limites posées par des instruments techniques ou encore, les ressources allouées au test.

Dans une perspective historique, on peut dire que la T.R.I. a fourni les fondements théoriques pour un certain type de tests adaptatifs par ordinateur qui s'adaptent principalement sur la base du paramètre de difficulté de l'item (Bunderson, Inouye & Olsen, 1989 : 381). En 1960, Rasch crée un modèle de réponse aux items qui permet de calculer des paramètres pour des items particuliers et non plus pour un test dans sa totalité. La création de ce modèle a permis le classement des items du plus facile au plus

difficile, et de construire des tests où la place de l'item n'importe pas. Elle a également permis d'utiliser un item dans deux tests différents pour mesurer la même compétence. Il est évident que la calibration et le calcul des paramètres de tels items ont permis une « administration adaptative ».

Toutefois, comme le signalent Dorans (2000 : 153), Embretson et Reise (2000 : 265), il faut être très attentif à l'administration du test et au respect de certaines règles (ou certains postulats) pour assurer une comparabilité suffisante entre les versions parallèles du test. Dans la mesure où les candidats passent des tests parallèles, les items proposés ne sont pas toujours les mêmes. Cela peut représenter une menace à la validité de contenu du test et jeter des doutes sur le fait que les candidats ont bien passé des tests parallèles (Thissen & Mislevy, 2000 : 120-121). Par ailleurs, les items n'apparaissant pas dans le même ordre, des problèmes dits d'indépendance locale peuvent surgir. Ces phénomènes (parfois dits « effets de contexte »), apparaissent lorsque la performance observée pour un item influence la performance à un autre item ou lorsque le contexte, dans lequel est placé l'item a un effet sur lui (Bunderson, Inouye & Olsen, 1989 : 386; Wainer & Kiely, 1987: 187; Wainer & Mislevy, 2000: 90). Si les candidats ne reçoivent pas les items dans le même ordre, l'ordre de passation, potentiellement différent pour chaque candidat, peut affecter les paramètres des items (Wainer & Kiely, 1997 : 187). Kingston et Dorans (1984) trouvent ainsi que la difficulté des items de compréhension écrite augmente légèrement lorsque les items sont proposés en fin de passation.

Autre difficulté liée au type d'administration mis en œuvre dans les tests adaptatifs : on doit veiller à ce que les items ne soient pas systématiquement (ou trop souvent) proposés aux candidats. Pour veiller à la bonne « exposition » des items, des règles devront être mises en place (Thissen & Mislevy, 2000 : 119-120). Si les items sont trop souvent proposés, le risque est d'avoir des candidats qui décrivent, voire, qui dévoilent, la nature des items à d'autres candidats (Wainer & Eignor, 2000 : 274-276). Cela peut provoquer une entorse à la dimensionnalité du test et *in fine* à la validité de l'interprétation des résultats.

Dans un tel contexte, pour construire des tests équivalents, il faut que les items qui les composent évaluent une même compétence unidimensionnelle (par exemple, la compétence langagière). Il faut aussi qu'ils mesurent le même construit (par exemple,

l'évaluation de la compréhension écrite dans le cadre d'un test de positionnement pour des apprenants de français du M.I.C.C. au Québec), des niveaux de compétence ciblés ou encore un large spectre de compétence. Le défi sera de construire une banque d'items ayant chacun une valeur de discrimination élevée. En effet, plus l'item discrimine, plus il apporte d'information sur le niveau du candidat. Lorsque les candidats auront obtenu la même quantité d'information que les autres, ils auront aussi reçu un test équivalent en termes de précision de la mesure :

> « For an equal-precise measure, items must be selected so that the test information curve is relatively flat across the trait range. This means that a set of highly discriminating items with a broad range of difficulty parameters must be identified[27] » (Embretson & Reise, 2000: 270).

Toutefois, un usage exclusif de la fonction d'information (et donc d'items discriminant fortement) est à éviter. Si on n'utilise que la fonction d'information, on risque d'avoir des problèmes de présentation des items et *in fine* de contenu. Deux items, s'ils discriminent tous deux fortement, n'en mesurent pas pour autant le même contenu (Raîche, 2004 : 331). Si on sélectionne les items dans le but d'avoir une information maximale pour tous les candidats et d'obtenir une quantité d'information équivalente indépendamment du niveau des candidats (Thissen, 2000 : 166-167), on risque d'avoir des résultats avec une validité de contenu très limitée (Thissen & Mislevy, 2000 : 110-111; Van Der Linden, 2005 : 284; Wainer *et al.*, 2000 : 238). Enfin, certains contenus ou certaines habiletés sont plus difficiles à mesurer avec précision ce qui ne signifie pas qu'il faille les éliminer pour autant. Privilégier l'optimisation de la mesure (avoir pour seul objectif de réduire l'erreur de mesure) peut donc se faire au détriment de la pertinence de contenu. Pour un meilleur résultat, il faut à la fois viser l'amélioration de la mesure mais aussi celle du contenu (on visera alors une meilleure quantité d'information sans que cela altère le contenu du test et une meilleure diversité de contenu en gardant une information suffisante). Pour les tests adaptatifs, Laurier (1999 : 129) explique que s'il est préférable d'utiliser des modèles à trois paramètres pour des tests utilisant des questions à choix multiples (et des modèles à deux paramètres quand on ne rencontre pas de problèmes de

pseudo-chance), le problème est que les items les plus discriminants seront plus souvent proposés, posant alors des problèmes de surexposition.

Afin d'éviter les phénomènes de dépendance locale, ou encore, les effets de contexte, on peut utiliser des minitests (Laurier, 1999 : 129 ; Thissen & Mislevy, 2000 : 125-128 ; Wainer & Kiely, 1987 ; Wainer *et al*, 2000 : 238). On proposera ainsi des items apparaissant dans un ordre de difficulté ou encore une sélection d'item à l'intérieur du minitest en fonction du niveau de compétence du candidat (Eignor, 1999 : 176). Toutefois, l'utilisation de ces minitests n'est pas sans poser de problèmes. Comme le signale Eignor (1999 : 176), les items inclus dans un minitest ne sont pas forcément tous du niveau du candidat. Le minitest (pris dans sa totalité) ne permet que d'obtenir une difficulté moyenne. Cela occulte le fait que certains items ne soient pas adaptés au niveau des candidats ou encore qu'ils ne soient pas pertinents. Par conséquent, on obtiendra une information plus ou moins utile, plus ou moins importante. Toutefois, même s'il s'agit d'un avantage indirect, les minitests parce qu'ils sont composés d'items qui cernent avec plus ou moins de précision le niveau des candidats, peuvent permettre de détecter des patrons de réponses anormaux, ce que l'on fait encore relativement difficilement avec des tests adaptatifs quand bien même la recherche est active dans le domaine (Blais & Raîche, 2005).

Autre avantage des minitests, ils permettent de proposer des textes plus longs avec plusieurs questions et, ainsi, d'éviter de proposer uniquement des items dans lesquels un seul paragraphe est associé à une seule question (Alderson, 1999 : 65). Néanmoins, Yen (1993) affirme que proposer plusieurs questions pour un texte crée de la dépendance locale. C'est donc parce qu'on traite l'item comme un « macro-item » avec un score polytomique, que la dépendance locale peut être moindre, voire disparaître (artificiellement). Afin de respecter l'esprit adaptatif du test, on pourra proposer des minitests composés de questions sélectionnées en fonction du niveau des candidats (Hambleton & Xing, 2006). Cependant, un autre problème affecte l'utilisation des minitests dans le testing adaptatif. Il n'est possible d'éviter les problèmes de surexposition que si beaucoup de minitests sont conçus (Eignor, 1999 : 176). Dès lors, les coûts de conception du test peuvent devenir beaucoup plus importants.

Enfin, les minitests ayant un nombre d'items prédéfini, l'estimation de la compétence se fait parfois avec un nombre d'items insuffisant. En créant des minitests, on réduit considérablement le nombre d'items qui entrent dans l'estimation de la compétence. Pour atteindre une précision suffisante, il faudrait administrer plusieurs minitests. On serait alors confronté à des épreuves interminables ! Si les minitests sont composés d'un nombre fixe items, il peut s'avérer très difficile pour différentes raisons d'administrer un nouveau minitest.

2.5.3 Validité apparente de la lecture avec les tests adaptatifs par ordinateur

Pour finir, après avoir traité des caractéristiques des tâches (texte et activité), il faut évoquer certaines des caractéristiques des candidats. Comme l'expliquent Bachman (1990, 2002) et Bachman et Palmer (1996), certaines de ces caractéristiques, leur interaction avec les caractéristiques des tâches ont (ou peuvent avoir) une influence importante sur la difficulté réelle. Pour ce qui est de la perception de la difficulté (objet de la troisième question de recherche), il n'y a pas de consensus. Certains, à l'instar de Robinson (2001 : 49), prétendent que la perception de la difficulté est en partie reliée à la complexité cognitive de la tâche. Si les candidats perçoivent qu'une tâche comme étant plus complexe, ce n'est pas pour autant qu'ils en ont une mauvaise image. Elder, Iwashita et McNamara (2002) étudient les liens entre la complexité de la tâche, au sens de Skehan (1996), et la difficulté perçue par le candidat. Les données (questionnaires donnés aux étudiants et résultats aux tests) montrent qu'il existe bien des relations significatives entre la perception de la difficulté et la performance. Cependant, ces relations ne sont pas constantes au travers des tâches et les corrélations sont faibles. Elder, Iwashita et McNamara (2002 : 361-362) concluent qu'ils ne peuvent pas soutenir l'hypothèse de la relation entre la perception de la difficulté et la complexité de la tâche (ce qui contredit Robinson, 2001 ; Skehan & Foster, 1997 ; 1999) ou, dit autrement, entre la perception de la difficulté et la difficulté réelle de la tâche. Nunan et Keobke (1995 : 7), eux aussi, avaient déjà trouvé que la difficulté perçue par les candidats ne permet pas de faire de prédictions sur la capacité réelle du candidat à effectuer ou non la tâche. Pour eux, ce que l'on ignore, mais que l'on gagnerait à connaître, c'est l'impact que peut avoir la

perception de la difficulté de la tâche sur les probabilités de réussir ou d'échouer à la tâche d'évaluation :

> « The implications of students' inability to properly perceive task difficulty hinge on the question of what each student brings to a task in term of effort, and whether this effort is tempered by perception. More simply put, if the question is harder, does the student try harder?[28] » (Nunan et Keobke, 1995 : 7).

Enfin, à la simple évocation de la perception du test par le candidat, certains avertissent des dangers liés à la validité apparente (Bachman, 1990 : 285). Alderson, Claphan et Wall (1995 : 172-173), donnent deux acceptions au concept. Dans son acception négative, elle signifie que le test n'a aucune validité, et que sa validité n'est que de surface, non soutenue scientifiquement. Dans son acception positive, elle traduit l'acceptabilité du test par les candidats. Si la difficulté perçue n'est pas un bon indicateur de la difficulté réelle, elle permet toutefois de comprendre comment les candidats se représentent la difficulté des tâches et du test. Il semble donc légitime de considérer ces représentations comme un indice de l'acceptabilité du test par les candidats.

Chapitre 3 : Cadre conceptuel

Dans ce bref chapitre, une présentation sera faite des points abordés dans la recension des écrits et qui seront retenus pour le cadre conceptuel. Il va de soi que les points déjà abordés dans la recension des écrits ne subiront plus de développement dans cette partie. Le lecteur est donc invité à se référer aux informations déjà contenues dans la recension des écrits pour de plus amples détails. Ce cadre est aussi conçu pour définir le construit du test de lecture rédigé par le chercheur.

3.1. La vision de la compétence langagière et la compétence de lecture en langue seconde

Pour l'évaluation de la compétence langagière, le modèle retenu est celui de l'évaluation multi-facettes proposé par Bachman (1990), Bachman et Palmer (1996). La compétence langagière est incluse dans une compétence communicative plus vaste regroupant la compétence langagière, la compétence stratégique et les mécanismes psychophysiologiques. La compétence langagière se divise en deux facteurs principaux, la compétence organisationnelle et la compétence pragmatique. La compétence de lecture dépend des caractéristiques des lecteurs, des tâches (questions et textes) (Brindley, 1987 ; Carr, 2006). Le niveau de lecture en langue seconde peut être expliqué par les compétences de lecture en langue maternelle et la connaissance de la langue seconde (Bernhardt & Kamil, 1995). Il existe bien un seuil minimal en langue seconde pour pouvoir lire des textes (Alderson, 2000). Ainsi une connaissance insuffisante du vocabulaire ne permet pas de comprendre les textes (Grabe, 1999). Étant donné que 50% de la variance de la compétence de lecture n'est toujours pas expliqué, il est possible de croire que cette variance puisse être expliquée par des facteurs affectifs ou encore par des facteurs tels que la langue maternelle ou par les caractéristiques du candidat lui-même (Koda, 2005). Enfin, la lecture est bien un processus et un produit (Alderson, 2000).

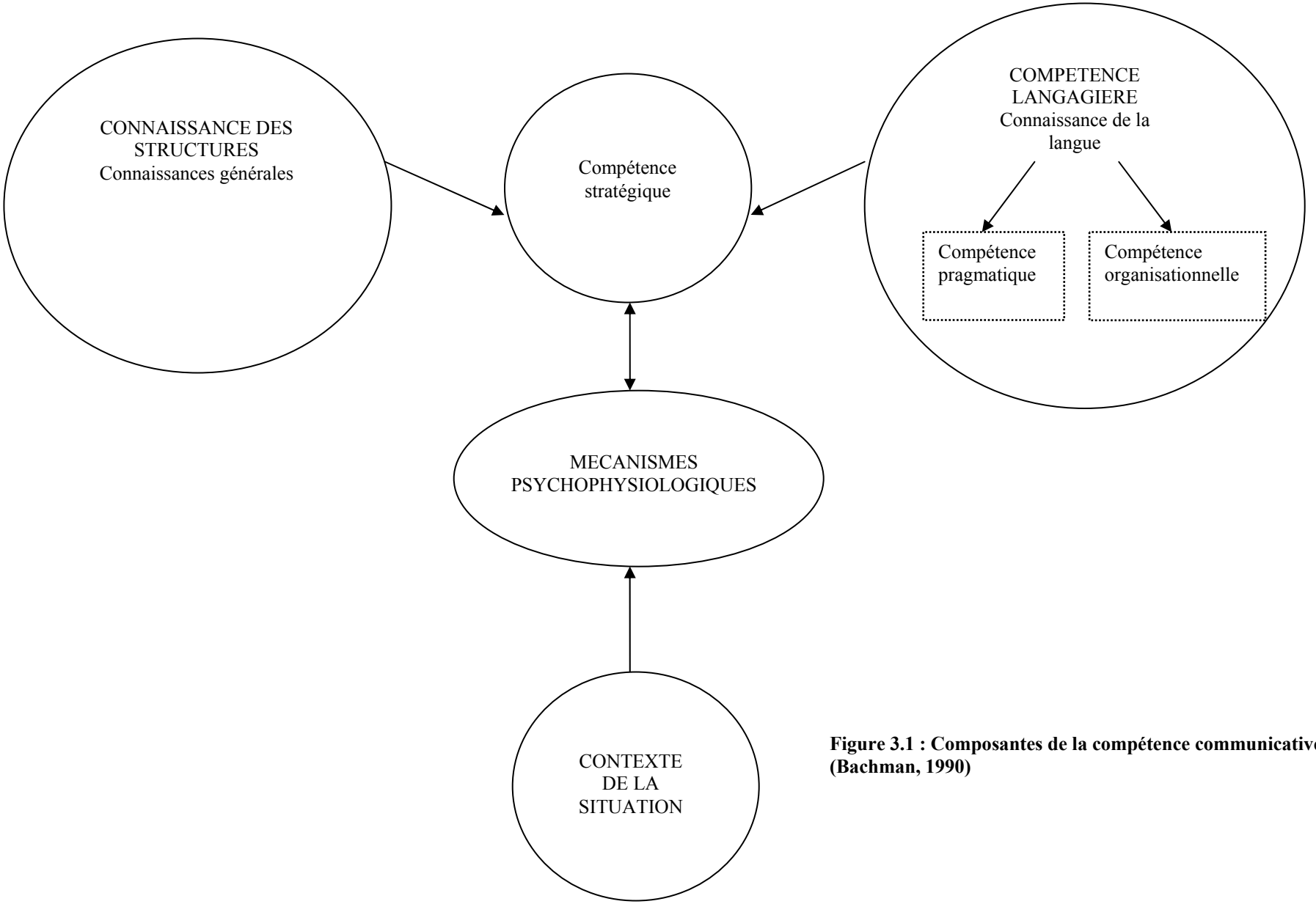


Figure 3.1 : Composantes de la compétence communicative (Bachman, 1990)

3.2. Les caractéristiques des tâches d'évaluation

Dans le cadre d'un test de positionnement, portant sur des cohortes fort nombreuses, on utilisera des tâches évaluant le produit et non pas le processus ou encore les étapes de la réalisation d'une tâche (Dassa & Laurier, 2003 : 118). Il est important de considérer le but du lecteur lors de la conception des tâches, dans la mesure où ce but va orienter l'interprétation ou encore la représentation que le lecteur se fera du texte (Goldman, 1997 : 366). Ce ou ces buts sont importants et doivent être pris en compte pour le construit d'un test de lecture (Alderson, 2000 : 123). Dans la présente recherche, les buts de lecture sont les suivants : lire des textes indépendants pour comprendre et apprendre, lire des textes intégrés, interdépendants en faisant des liens entre les textes pour comprendre et pour apprendre.

Pour les tâches d'évaluation, aucune hypothèse de taxonomie cognitive ou d'ordre naturel et universel de difficulté des tâches ne sera retenue. La difficulté doit être mesurée empiriquement. Elle est le fruit de la rencontre entre les caractéristiques de la tâche (texte et activité), les attributs du candidat et les interactions entre les deux (Bachman, 2002). Ceci dit, il est possible de tenter un diagnostic de la difficulté en fonction des caractéristiques de la tâches, du texte et du candidat, tels que décrits par Nunan et Keobke (1995). En revanche, la complexité (cognitive) des tâches n'est qu'un facteur, parmi d'autres, qui joue sur la difficulté. Elle ne constitue pas l'essence même de la difficulté. Par conséquent, dans cette recherche, l'idée de prédire la difficulté d'une tâche en fonction de sa seule « complexité cognitive » ou en fonction des habiletés de lecture utilisées par les candidats n'est pas retenue.

Pour guider la conception d'un test, il est utile d'utiliser des types de tâches décrits dans des référentiels comme les *Standards linguistiques canadiens* (C.C.L.B., 2002) ou encore les *Niveaux de compétence en français langue seconde pour les immigrants adultes (M.R.C.I., 1998*). La forte généralisabilité et la concision des descriptions holistiques des *Niveaux de compétence en français langue seconde pour les immigrants adultes (M.R.C.I., 1998*) en font un outil de choix pour un test de positionnement ayant pour visée d'évaluer un niveau linguistique général, sans aucune visée diagnostique particulière.

3.3. Différents types de tests et d'évaluation

Pour la recherche, les tâches d'évaluation sont conçues pour mesurer la compétence de lecture des candidats au travers de la performance sur une échelle de niveau de compétence unique et continue (McNamara, 1996). L'évaluation de la performance permettra d'inférer la compétence de lecture (qui est conçue comme étant un trait latent non-observable directement). Les tâches devront encore autoriser une généralisation des interprétations des résultats au test à des situations de non-test. Afin de pouvoir se livrer à des interprétations valides, ces tâches doivent s'inscrire dans un même domaine de référence. Elles doivent encore respecter le postulat de l'unidimensionnalité. Autrement dit, elles doivent toutes évaluer une même habilité, la compétence de lecture en français langue seconde.

L'évaluation qui sera mise en œuvre est celle de la maîtrise (« *proficiency* ») (Laurier, 1993 : 23 ; North, 2000 : 41-54). Par conséquent, il s'agit d'évaluer une compétence non reliée à l'organisation d'un cours ou de séquences pédagogiques. Il s'agit de proposer une évaluation de l'usage de la langue en dehors de tout contexte particulier (Ingram, 1985 : 220). Le but est d'obtenir des résultats dont les interprétations sont facilement généralisables à un ensemble de situations. L'objectif n'étant pas de diagnostiquer finement la compétence de compréhension écrite, il ne sera pas procédé à l'évaluation d'habiletés de lecture. Quoique la distinction entre de test de positionnement et test diagnostique (Alderson, 2005 ; Laurier, 1993 : 23) ne soit pas toujours facile à établir, le type d'évaluation, qui sera mis en œuvre dans la recherche, aura pour principale fonction de sélectionner les candidats à l'aide de tâches différentes (discrètes et intégrées) mais conçues pour évaluer une compétence de lecture générale, vue comme un ensemble holistique. Il ne s'agit donc pas de viser l'utilisation par le lecteur d'éléments discrets (stratégies ou habiletés) mais des éléments plus englobants (inférences, type de tâches, buts de lecture).Cette vue unitaire de la compétence de lecture est choisie parce qu'elle est appropriée pour un test de positionnement.

3.4. Le type de textes, format et mode de réponse

Les tâches de lecture comme le propose Savignon (1983) peuvent être plus ou moins intégratives, selon le type de texte proposé (une ligne, un paragraphe, plusieurs paragraphes, un texte, plusieurs textes), et, le mode de réponse choisi (Q.C.M., réponse construite, portfolio). Pour le concepteur, il s'agit donc de placer les tâches d'évaluation sur un continuum discret / intégré et ce, à la fois pour le type de texte et le type de document utilisés. Le choix du degré d'intégration doit être fait en fonction du type d'évaluation que l'on veut mettre en place (Dassa & Laurier, 2004) mais aussi des buts de lecture que le lecteur devra avoir pour effectuer la tâche, ou encore, en fonction de contraintes de faisabilité. Pour un test de positionnement visant l'évaluation rapide de candidats sur une échelle de difficulté, il semble opportun d'utiliser des Q.C.M. à quatre choix de réponse (en prenant en compte les savoir du domaine pour leur rédaction). Ce format d'item peut être appliqué à des documents discrets ou intégrés, à des textes aussi bien longs que courts. Il permet l'usage de textes intégrés avec un mode de réponse discret. Sa simplicité permet de respecter le besoin de résultats immédiats et peu onéreux qui est le propre des tests de positionnement. Dans un test de lecture adaptatif, il est utile de proposer des questions d'inférence (mais pas exclusivement) puisque cela permet d'avoir des items qui sollicitent à la fois le « modèle » du texte (ou encore l'intertexte) et la « situation » (Grabe, 1999 : 20). Des items évaluant différents aspects de l'intégration de l'information textuelle (Grabe, 1999 : 20) autorisent une évaluation holistique de la compétence de lecture. Les items demandant au lecteur d'inférer sont d'autant plus intéressants que leur « profil inférenciel » varie en fonction du but du lecteur (Grabe, 1999 : 18). Dans l'optique d'un test de placement adaptatif, ce type d'item permet d'évaluer autre chose que des éléments discrets. Par conséquent, il permet d'éviter une évaluation de type diagnostique, tout en gardant une certaine variété de contenu. Le type d'inférence variant en fonction du but du lecteur, les items devraient différer selon que les tâches sont discrètes ou intégrées. Une telle orientation permet de penser que les items du test devraient être unidimensionnels. Pour un test visant une éventuelle utilisation dans un test adaptatif, c'est un aspect crucial, l'unidimensionnalité étant indispensable aux

tests adaptatifs (Bunderson, Inouye & Olsen, 1989 : 382) mais aussi aux tests visant le positionnement des candidats sur une échelle de niveau.

Pour la présente recherche, l'utilisation d'items rédigés pour être indépendants, qui demandent au lecteur de faire des inférences, permet de penser que l'on obtiendra des items répondant, à la fois, à des contraintes de variété de contenu et de précision de la mesure.

3.5. Le modèle de réponse aux items

Pour choisir un modèle de réponses aux items, il y a plusieurs options. Andrich (2002) explique que la famille des modèles de Rasch s'inscrit dans un paradigme qui n'est pas celui des modèles utilisant plusieurs paramètres d'item. On peut choisir d'opter pour un des modèles de la famille de Rasch, soit un modèle visant l'ajustement des données au modèle. On peut encore faire le choix d'appliquer des modèles à un, deux ou trois paramètres pour ajuster le modèle aux données (Choi & Bachman, 1992). Pour cette recherche, le choix a été fait de privilégier le modèle de Rasch parce qu'il est centré sur le paramètre de la difficulté et qu'il permet de détecter les items et les personnes qui ne correspondent pas au modèle. L'échantillon étant de taille modeste, il n'aurait pas été possible d'utiliser des modèles à plusieurs paramètres.

Par ailleurs, la recherche n'ambitionne pas de décrire le fonctionnement de tous les paramètres des items. Elle ambitionne de décrire le fonctionnement « global » des items, leur ajustement au modèle (Choi & Bachman, 1992 : 63 constatent que le modèle de Rasch permet de « détecter » plus d'items qui ne sont pas ajustés au modèle que les autres modèles) ou encore les problèmes de fonctionnements différentiels. Pour ce qui est de l'ajustement des données au modèle, au besoin, et en fonction de la disponibilité des données, des items et des personnes seront éliminés de l'échantillon. Le test est en « rodage », « au banc d'essai ». Il s'agit d'en vérifier les qualités avant une éventuelle application à grande échelle. Cet aspect devra être pris en compte dans les orientations méthodologiques touchant notamment à la calibration.

3.6. La perception de la difficulté par les candidats

La perception de la difficulté des tâches par les candidats ne peut servir à décider de la validité de ces tâches. En revanche, il est possible de s'en servir d'indice de l'acceptabilité, ou encore d'indicateur quant à la représentation que les candidats se font de la hiérarchie de la difficulté des tâches et des textes discrets et intégrés. La variabilité de la compétence de lecture n'étant pas complètement expliquée, il n'est pas inutile de se pencher sur ces aspects affectifs (Nunan & Keobke, 1995). La comparaison de la mesure empirique et objective de la difficulté et sa perception par les candidats peut s'avérer utile, surtout, pour un test destiné à un public multiculturel.

Chapitre 4 : Méthodologie

Pour rappel, les questions de recherche ont pour objectif d'analyser la dimensionnalité des tâches discrètes et intégrées tant du point de vue de la dimensionnalité psychométrique que psychologique. A partir d'un ensemble d'items, appartenant à deux types de tâches, conçu pour être unidimensionnel du point de vue psychométrique mais dont la dimensionnalité psychologique pose question (pour les concepteurs du test, mais, surtout, pour les candidats), on veut tester des hypothèses permettant de statuer sur la dimensionnalité. Dans un premier temps, il s'agit d'étudier la dimensionnalité des items pour l'ensemble de l'échantillon et, dans un deuxième temps, pour les différents groupes linguistiques. Pour ce faire, les interactions entre des candidats et des items de niveaux de compétence et de difficulté similaires seront étudiées. Enfin, on vérifiera si la dimensionnalité psychologique des tâches est la même pour les concepteurs de ces tâches et pour les candidats et quel est le lien avec la dimensionnalité psychométrique. Là aussi, on tentera de savoir si l'appartenance à un groupe linguistique a un impact sur la difficulté perçue.

4.1 Nature de la recherche

La recherche est une recherche exploratoire au sens que lui donne Van der Maren (1996 : 192). Elle a donc pour but de :

> « Générer des hypothèses, c'est-à-dire d'examiner un ensemble de données afin de découvrir quelles relations peuvent être observées, quelles structures peuvent y être construites ».

Un test, composé de deux sous-tests, a été créé (annexes 1 et 3). Le premier sous-test est composé de trois tâches discrètes (trois textes indépendants et cinq questions à choix multiple par texte), le second d'une tâche intégrée (trois textes concaténés, dépendants, avec cinq questions à choix multiple pour chaque texte). Ces deux sous-tests ont un total de 30 items, soit 15 pour chaque sous-test.

Le test a été administré à un échantillon de 118 individus. Utilisant un modèle de Rasch, un échantillon de 118 individus pour 30 items dichotomiques fournit suffisamment de

données pour l'analyse, le minimum étant d'environ une centaine de personnes pour 20 items à réponses dichotomiques (Wright & Stone, 1979). Cette contrainte a été préférée à la simulation de données. Comme cela a déjà été expliqué dans la problématique (point 1.2.1.), quoique prévu pour éventuellement fonctionner comme test adaptatif par ordinateur, le test a été administré selon une modalité conventionnelle (papier-crayon). L'objectif est donc de répondre à des questions qui se posent dans le cadre de la mise en place d'un test adaptatif. Les sujets de l'échantillon ont été choisis en fonction de leur niveau. Les tâches et les items ont été conçus à partir d'une estimation de ce niveau. Avant la passation, les candidats ont dû remplir un questionnaire de renseignements sociodémographiques. À l'issue de la passation, ils ont été invités à répondre à des questions portant sur leur perception de la difficulté du test. Pour les deux sous-tests, les *Niveaux de compétence en français langue seconde pour les immigrants adultes* (M.R.C.I., 1998) et les *Standards linguistiques canadiens français langue seconde* (C.C.L.B., 2002) ont servi de référentiels pour la conception des tâches et la sélection des textes.

4.2 L'échantillon et la cueillette de données

4.2.1. Provenance des candidats : choix et justification

Les sujets sont tous des volontaires, inscrits dans les cours de français financés par le M.I.C.C et dispensés par les écoles de langue de l'Université de Montréal et de l'Université du Québec à Montréal. Ils sont inscrits dans ce que l'on nomme le « Bloc 3 », soit le cours FIA-300-3, selon l'appellation du M.I.C.C. Au moment de la cueillette des données, les sujets avaient suivi les cours du Bloc 3 depuis une semaine. Le choix de cette population a été dicté par le besoin de tester des candidats ayant un niveau comparable. Il a été fait à partir d'informations données au chercheur par un agent du M.I.C.C.[1], en décembre 2005. Selon ces informations, les personnes commençant les cours du Bloc 3 avaient, le plus souvent, un niveau estimé en compréhension écrite

[1] Nous tenons à remercier Louis Kelly. Son aide et les informations qu'il nous a données ont permis de mener à bien la recherche. En effet, il a été possible de cerner au plus près, et ce dès la conception du test, le niveau des candidats.

équivalent au niveau 5 des *Niveaux de compétence en français langue seconde pour les immigrants adultes* (M.R.C.I., 1998).

Dans un premier temps, il a été décidé de retenir les élèves du M.I.C.C. inscrits dans les universités à Montréal. Cela permettait, d'une part, d'avoir un public homogène quant au niveau d'études (les étudiants envoyés dans les universités sont des étudiants avec un haut niveau de scolarisation) et, d'autre part, de profiter d'une concentration des candidats de niveau 5 plus grande que dans les autres institutions dispensant les cours de français aux immigrants (CEGEP,…). Le nombre de candidats à l'Université de Montréal n'ayant pas été pas suffisant, une deuxième collecte de données a eu lieu une semaine plus tard à l'école de langue de l'Université du Québec à Montréal.

4.2.2. Les passations

Comme cela était prévu dans le devis de recherche, les passations ont commencé à l'Université de Montréal en février 2006. Une rencontre a eu lieu avec la coordinatrice pédagogique en charge des cours M.I.C.C. Le test lui a été présenté à cette occasion. On a alors décidé de le proposer aux quatre groupes des élèves des cours du bloc 3. Une semaine avant la passation, un document a été envoyé aux professeurs pour leur expliquer les principes de la recherche (cf. annexe 2). Le jour de la passation, les professeurs étaient invités à participer à une séance d'information afin que le chercheur puisse leur expliquer le déroulement de la passation, les aspects éthiques et répondre aux éventuelles questions. Le chercheur est passé dans chacune des classes pour expliquer aux candidats le but de la recherche. Il était présent pendant toute la durée du test. Au total, 55 élèves ont participé aux passations du test sur un total d'environ 80 élèves.

À l'U.Q.A.M., là encore, une rencontre a eu lieu avec la responsable pédagogique des cours du M.I.C.C. À cette occasion, une lecture du test lui a été proposée. La décision a été prise de proposer le test aux quatre groupes du bloc 3 sur deux jours et de laisser le chercheur se charger des passations. Avant chacune des deux séances, le chercheur a été présenté par la responsable pédagogique. Au total, 63 élèves ont passé le test. La majorité des élèves d'un groupe n'a pas assisté à la passation du test, leur professeure étant souffrante le matin de la passation.

Le test a été proposé en dehors des cours réguliers, à l'occasion des activités annexes prévues l'après-midi et, ce, pour tous les candidats (soit 118). Dans les deux universités, les passations se sont déroulées normalement. Les professeures et les coordinatrices pédagogiques ont été très coopératives. A quelques exceptions près, les candidats étaient intéressés par l'idée de participer à une recherche, de passer un test de lecture et de recevoir une note à l'issue de ce test. Après envoi des résultats par courriel, deux candidats ont voulu avoir des précisions quant à l'interprétation de leur score. Une réponse leur a alors été adressée.

4.3 Caractéristiques et fonctionnement du matériel

4.3.1. Fonctionnement général des deux sous-tests

Les deux sous-tests ont été passés l'un à la suite de l'autre. Les candidats n'étaient pas informés qu'ils passaient deux sous-tests différents. La durée prévue de l'examen était d'une heure à laquelle il fallait ajouter 15 minutes pour le questionnaire. Dans la pratique, certains candidats sont sortis après 35 minutes et la plupart avant les 75 minutes prévues initialement.

Afin de vérifier les effets de fatigue, le sous-test 1 et le sous-test 2 ont été administrés soit en première position, soit en deuxième. Autrement dit, les candidats ont eu soit une version du test commençant par la question 1, soit par la question 15. La répartition s'est faite au hasard, les tests étaient distribués aux candidats en alternance dans une version puis dans l'autre[2].

Les trois tâches discrètes et les trois tâches intégrées ont été conçues pour avoir un niveau et un contenu comparables. Toutefois, le concepteur ne prétendait pas concevoir des tâches ayant exactement le niveau 4, 5 ou 6 des *Niveaux de compétence en français langue seconde pour les immigrants adultes* (M.R.C.I., 1998) , mais bien de proposer des tâches ayant un niveau de difficulté semblable et qui correspondent au niveau de compétence des sujets de l'échantillon sélectionné (autrement dit que les candidats aient des items correspondant à leur niveau de compétence en français).

[2] Candidat 1 (version 1), candidat 2 (version 2), candidat 3 (version 1), candidat 4 (version 2).

4.3.1.1. *Le sous-test 1*

Le sous-test 1 (annexe 1 et 3) est un test composé de trois tâches de trois niveaux de difficulté différents. Pour la rédaction des tâches, on s'est inspiré des descriptions des niveaux 4, 5 et 6 des *Niveaux de compétence en français langue seconde pour les immigrants adultes* (M.R.C.I, 1998) en compréhension écrite. Chaque tâche est composée d'un texte auquel sont associés cinq items. Ces items sont des questions à choix multiple (à correction dichotomique et quatre choix de réponse). Les documents ont, eux aussi, été choisis selon les descriptions présentes dans les *Niveaux de compétence en français langue seconde pour les immigrants adultes* (M.R.C.I., 1998) en tenant également compte des *Standards linguistiques canadiens français langue seconde* (C.C.L.B., 2002). Chaque document est indépendant des autres documents, tant au niveau du contenu que du thème. Les questions, elles aussi, étaient prévues pour être indépendantes. : répondre à une question ne doit pas permettre de mieux répondre aux autres.

4.3.1.2. *Le sous-test 2*

Ce sous-test est composé, comme le premier, de trois tâches d'évaluation, pour un total de 15 items (annexes 1 et 3). Cependant, dans ce sous-test, les textes sont concaténés ou encore dépendants. Le texte de la tâche 3 a un lien (situation et intertexte) avec les textes des tâches 1 et 2. Le texte de la tâche 2 a un lien avec le texte de la tâche 1. Il faut donc lire le premier texte pour mieux comprendre le contexte d'écriture du second texte, les deux premiers textes pour mieux comprendre celui du troisième. Comme dans le « sous-test 1 », pour la rédaction des tâches, le concepteur s'est inspiré des descriptions disponibles des niveaux 4, 5 et 6 des *Niveaux de compétence en français langue seconde pour les immigrants adultes* (M.R.C.I., 1998).

4.3.2. Choix des documents

La conception du test s'est faite à partir d'une sélection de textes. Ces textes ont été sélectionnés à partir de sites internet québécois et de la base de données « Eureka.ca » regroupant des articles de la presse francophone canadienne et étrangère. Les articles des

tâches discrètes et intégrées ont été choisis afin de présenter une certaine similitude quant au contenu linguistique, discursif et au type de texte. Les paires de texte 1 et 4, 2 et 5, 3 et 6 sont des textes de longueur et de nature comparables. Le tableau de l'annexe 4 permet de se faire une idée des ressemblances et des différences entre les textes.

4.3.3. Révision des textes

Afin de calibrer les textes, une révision d'expert à été mis en place (à l'aide d'un expert, le directeur de recherche). Pour une meilleure homogénéité des niveaux et des longueurs de texte, plusieurs modifications ont été faites. Des parties de textes ont été enlevées, d'autres modifiées ou encore simplifiées. Toutes les modifications qui ont été faites, l'ont été afin de produire des textes comparables en termes de complexité (longueur des phrases, vocabulaire et complexité linguistique) et de longueur (nombre de mots). Aucune modification, pouvant modifier le type de texte (ou encore la « *situation* » au sens de Perfetti, 1997) n'a été faite. Il convient ici de préciser que ces modifications correspondent à la notion d'interaction authentique telle que définie par Bachman (1990 : 317) ou encore Long (2003). La nature du texte et le but que se fixait le scripteur au moment de la rédaction ont été respectés. La relation entre le lecteur et le scripteur du texte est restée la même. Le ton employé par le scripteur a été conservé afin de conserver « l'homogénéité et la couleur » du texte[3].

Enfin, pour la longueur des textes, on a veillé à trouver un compromis entre la longueur telle qu'elle est décrite dans les référentiels et la contrainte temporelle (une heure de passation). Il fallait tenir compte de la longueur cumulée des six textes – au total 7137 mots.

4.3.4. Choix et calibration des items

Parce que l'échantillon des candidats risquait de ne pas être assez important pour mener la recherche à son terme (la collecte de données ne pouvait pas être étalée sur plus d'une session de cours), avant même les passations, il a été décidé de ne pas organiser de pré-

[3] Habituellement, les textes dans les forums de discussion ne se démarquent pas par leur cohérence. Ceux qui apparaissent sur le site Internet de l'hebdomadaire « voir.ca » n'échappent pas à la règle. Il ne semble donc pas utile de porter une attention trop importante à « l'homogénéité et la couleur du texte ». Si le choix avait été fait de travailler avec des textes littéraires, il en eût été tout autrement.

test. Par conséquent, la rédaction des items devait être rigoureuse, respecter les règles de rédaction telles que décrites dans le cadre conceptuel. Les questions ont été rédigées pour évaluer deux buts de lecture :

- lire des textes indépendants pour comprendre et apprendre,
- lire des textes intégrés, interdépendants en faisant des liens entre les textes, pour comprendre et pour apprendre.

Le principe d'authenticité interactionnelle de Bachman (1990) a été retenu. Dès lors, et autant que peut ce faire, les questions ont été rédigées pour correspondre à celles que se poserait un lecteur face aux documents retenus. Toutefois, il serait très prétentieux d'avancer que les 30 questions du test final sont celles que les candidats se poseraient après avoir lu les textes pour la première fois. Pour aboutir à un tel résultat, il eût fallu sonder les candidats pour statuer sur la vraisemblance des questions et leur représentativité. Enfin, comme cela a été expliqué dans le cadre conceptuel, le choix a été fait de privilégier des questions demandant au lecteur d'inférer. En général, ces questions, lorsqu'elles sont correctement rédigées, permettent de bien distinguer les mauvais lecteurs des bons lecteurs (Grabe, 1999 : 20-21). Toutefois, afin d'offrir plus de diversité de contenu, d'autres types de questions ont été proposés. Ainsi le test comporte-t-il des questions de repérage.

Après une première rédaction, les questions ont été proposées au directeur de recherche. Certaines ont été directement validées. D'autres ont été directement éliminées. Après modifications, une deuxième version du test a été proposée au directeur de recherche. Ce dernier a proposé des modifications pour certaines questions. Ces propositions prises en compte, par la suite, d'autres sont encore intervenues. Les questions ont alors été soigneusement lues et analysées plusieurs fois avant d'arrêter une version finale. Le travail itératif de rédaction et de révision des textes jusqu'à la version définitive de l'examen a duré environ trois à quatre semaines avant de parvenir à l'ultime version (annexe 1).

4.3.5. Les questionnaires sociodémographiques et de perception de la difficulté

Bien que les candidats ne soient pas des francophones d'origine, toutes les questions sociodémographiques ont été rédigées en français. Le niveau des candidats permettait une telle chose. Ce questionnaire était à compléter avant la passation, après avoir lu et signé le formulaire de consentement (*cf.* annexe 5). Il visait à récolter des informations permettant d'éclairer l'analyse des résultats au test. Les informations récoltées sont les suivantes :

- âge ;
- sexe ;
- niveau d'étude,
- langue maternelle ;
- langue(s) parlée(s) dans la rue, au travail, à la maison ;
- nombre de mois et d'années écoulées depuis l'arrivée au Québec.

Dans le deuxième questionnaire sur la perception de la difficulté des tâches, complétée après la passation, il était demandé au candidat de s'exprimer sur ses perceptions de la difficulté pour chaque type de tâche, pour chacune des tâches et pour chacun des textes. Au total, le candidat devait donc répondre à 14 énoncés, distribués de la manière suivante :

- deux énoncés portant sur la difficulté (générale) de l'ensemble des tâches discrètes (A, B, C) et de l'ensemble des tâches intégrées (D, E, F) ;
- six énoncés portant sur la difficulté des questions liées à chacun des textes ;
- six énoncés portant sur la difficulté de chacun des textes.

Pour ces énoncés, des échelles de Likert graduées de 1 à 5 (ordre ascendant de difficulté) ont été utilisées. Comme pour les textes et les questions, ces deux questionnaires ont été validés auprès du directeur de recherche.

4.4 Aspects éthiques, confidentialité et transmission des résultats

Les candidats étaient tous des volontaires. Ils ont été invités à signer un formulaire de consentement éclairé avant de participer au test (annexe 5). Comme cela avait été recommandé au chercheur par le comité d'éthique de l'Université de Montréal, les principaux points apparaissant sur le formulaire de consentement ont été expliqués oralement aux candidats avant qu'ils ne signent ledit formulaire. Une séance de questions sur le sujet à été proposée aux candidats avant signature. Le test n'a été distribué qu'après signature du formulaire de consentement. Les candidats ont été informés qu'en participant à la recherche, ils pouvaient contribuer au développement des connaissances et à l'amélioration des cours de français offerts aux nouveaux arrivants au Québec. Leur participation pouvait également leur permettre de faire évaluer leur niveau de compréhension écrite en français. S'ils le souhaitaient, ils pouvaient donner une adresse courriel pour que les résultats leur soient communiqués. Les candidats ayant indiqué une adresse lisible et fonctionnelle ont reçu leurs résultats, soit leur score brut. Après envoi des courriels, deux candidats ont voulu avoir des précisions quant à l'interprétation de leur score. Une réponse leur a alors été adressée. Enfin, chacun des candidats a été informé qu'il pourrait avoir à répondre à des questions d'un niveau supérieur ou encore inférieur au sien. Ils ont également été informés que la note qu'ils recevraient pour ce test, ne serait pas prise en compte pour l'accréditation du cours. Après la passation, les informations contenues dans les copies d'examen et dans les questionnaires ont été rendues anonymes. Chaque candidat s'est vu attribuer un numéro d'identification (de 0 à 118).

4.5 La méthode d'analyse des données

Afin de choisir une méthode d'analyse qui permette de répondre aux questions de recherche, d'autres recherches ont été considérées (Bachman, Lynch, & Mason, 1995 ; Du, Wright, & Brown 1996 ; Eckes, 2005 ; Elder, 1996 ; Elder, McNamara, & Congdon, 2003 ; Iwashita, McNamara, & Elder, 2001 ; Jones, 2005 ; McNamara, 1996 ; Kondo-Brown, 2002).

4.5.1. Calibration des items

Pour la calibration et une partie de l'analyse des données, le logiciel WINSTEPS (Linacre, 2005) a été utilisé. WINSTEPS permet (entre autres choses et outre la calibration des items) d'analyser les résultats à un test avec le modèle de Rasch et de transformer le score brut des candidats en un score calculé en *logits*. Avec WINSTEPS, il est possible d'avoir des informations sur la qualité de la mesure. Lorsque celle-ci est suffisante, il est possible d'utiliser le score obtenu en *logits* pour faire des inférences au sujet d'une compétence non-observable. Il devient alors possible de généraliser les interprétations des résultats au test. L'utilisation de ce logiciel permet donc d'inscrire la recherche dans le cadre conceptuel qui a été retenu et ainsi de pouvoir répondre aux questions de recherche.

Une fois les paramètres du modèle estimés, ce logiciel permet d'obtenir des informations sur la dimensionnalité, les fonctionnements différentiels d'item et de personne. Il autorise encore l'étude de l'interaction entre des groupes de personnes et des groupes d'items.

Des analyses complémentaires ont été faites à l'aide du logiciel SPSS (version 11, étudiant).

4.5.2. Méthode d'analyse pour répondre à la première question de recherche

Pour répondre à la première question de recherche (sur la dimensionalité psychométrique), le logiciel TESTFACT 4 (du Toit, 2003) a été utilisé. Cette partie de l'analyse a pour but de déterminer, d'une part, si les tâches discrètes et intégrées sont

unidimensionnelles et, d'autre part, jusqu'à quel degré le type de tâche (ou éventuellement les textes) peuvent expliquer la variabilité des données. L'analyse en composantes principales proposée par le logiciel WINSTEPS (Linacre, 2005) n'a pas été utilisée, car elle ne correspondait pas aux données de la recherche et surtout ne propose pas une correction pour le « *guessing* », comme le logiciel TESTFACT peut le faire (du Toit, 2003 : 582). De plus, TESTFACT propose un traitement des données (commande « *BIFACTOR* ») qui convient à l'analyse d'items associés à des minitests :

> « The BIFACTOR command is used to request full information estimation of loadings on a general factor in the presence of item-group factors[29] ». (du Toit, 2003: 804).

La commande *BIFACTOR* de TESTFACT utilise, pour mener une analyse factorielle, la méthode de l'information complète (Bock, Gibbons, & Muraki, 1988 ; du Toit, 2003 : 585). Cette méthode a pour avantage de ne pas « reposer sur l'analyse de la matrice des corrélations des items » (comme dans les analyses factorielles classiques) et «utilise plutôt les fréquences pour chaque patron de réponse observé et modélise le tout avec une fréquence multinomiale » (Bertrand & Blais, 2004 : 212). La commande *BIFACTOR* permet encore de mener une analyse factorielle confirmatoire, avec un facteur principal et des facteurs liés à des groupes d'items. Voici la description proposée par du Toit (2003: 410) ;

> « TESTFACT includes yet another full-information method that provides an important form of confirmatory item factor analysis called 'bifactor' analysis. The factor pattern in bifactor analysis consists of a general factor on which all items have some loading, plus any number of so-called 'group factors' to which non-overlapping subsets of items, assigned by the user, are assumed to belong. The subsets typically represent small numbers of items that pertain to a common stem such as a reading passage or problem-solving exercise.[30] ».

Traditionnellement, cette commande est utilisée pour les minitests de compréhension écrite ou de résolution de problème. Néanmoins, il est possible de l'utiliser pour étudier le regroupement de tous les items autour d'un facteur général et les regroupements d'items autour de facteurs spécifiques. du Toit (2003 : 809-811) prend ainsi pour exemple

un test qui a été conçu à partir de questions de chimie, biologie et physique et analysée à l'aide de ladite commande.

La présente recherche vérifie quel modèle explique le mieux les données parmi les modèles suivants :

1) Un facteur général et six facteurs spécifiques, un pour chaque groupe de cinq items associés à chaque texte.
2) Un facteur général et quatre facteurs spécifiques, trois pour les trois groupes de cinq items portant chacun sur un texte indépendant et un regroupant les 15 items des tâches intégrées.
3) Un facteur général et deux facteurs spécifiques, soit un pour les 15 items des tâches discrètes et 15 pour les tâches intégrées.
4) Un facteur général.

L'objectif est d'étudier la dimensionnalité du test et l'importance des textes ou du type d'item dans l'explication de la variabilité.

Enfin, la correction pour le paramètre de pseudo-chance sera utilisée afin d'améliorer l'analyse. Les Q.C.M., qui ont été utilisées, comportant quatre choix de réponse, ce paramètre sera fixé à 0,25 pour tous les items. De même, les réponses manquantes ne seront pas prises en compte. Par conséquent, elles ne seront pas considérées comme étant de mauvaises réponses. Le but est de ne pas augmenter indûment le niveau de difficulté des items qui se trouvent à la fin du test.

4.5.3. Méthode d'analyse pour répondre à la deuxième question de recherche

Pour répondre à cette question de recherche (sur la difficulté objective par rapport aux groupes linguistiques), seuls des candidats de niveaux comparables ont été retenus (la procédure utilisée pour la répartition des candidats et des items en groupes de niveau est expliquée en annexe 7). Pour chacun des groupes de niveau, la différence de difficulté (de chacune des tâches mais aussi de chaque ensemble de tâche, dans un premier temps avec tous les items du test puis avec les items au plus proche du niveau des candidats) est étudiée. Les « interactions » proposées dans WINSTEPS ont été mises à contribution pour détecter les fonctionnements différentiels. Enfin, l'analyse a été dupliquée mais cette fois avec des items calibrés séparément pour les tâches discrètes et intégrées. Pour

finir, les différences de classement entre les candidats ont été étudiées, pour l'ensemble des candidats et pour les candidats répartis par groupe linguistique selon que l'on prend en compte l'ensemble des tâches, les tâches discrètes ou les tâches intégrées pour établir le classement. Sur la base de l'estimation du niveau de compétence des candidats des rangs leur ont été assignés.

4.5.4. Méthode d'analyse pour répondre à la troisième question de recherche

Pour cette dernière question (sur la perception de la difficulté subjective par rapport aux groupes linguistiques), là encore, le logiciel WINSTEPS a été utilisé, mais cette fois-ci, avec le modèle de la famille de Rasch « *rating scale* » (convenant au traitement des données compilées lors de l'usage d'échelle de likert). La difficulté perçue des textes, des items associés à ces textes et des deux ensembles de tâches (discrètes et intégrées) a été calculée. Pour chaque type de tâche, est étudié le niveau de difficulté perçue par les groupes de langues maternelles différentes. Le portrait des différences de perception entre les groupes est alors dressé.

Enfin, est établi un portrait croisé entre les résultats obtenus au test pour les tâches discrètes et intégrées et le niveau de difficulté perçue pour les textes, les items et les deux ensemble de tâches, pour l'ensemble des candidats et pour chacun des trois groupes de candidats.

4.5.5. Choix méthodologiques pour la recherche

4.5.5.1. La saisie des données

Deux étapes ont été nécessaires à la saisie des données. Après avoir supprimé toute mention relative à l'identité des candidats dans les données, le premier codage effectué par le chercheur a été vérifié avec l'aide d'une seconde personne. Les configurations de réponses pour le test et le questionnaire sociodémographique de tous les candidats ont ainsi été vérifiées. Cette démarche a permis de constater que très peu d'erreurs avaient été commises.

4.5.5.2. Démarche opérationnelle de la calibration pour vérifier l'ajustement entre les données et le modèle

Avant même de procéder à la calibration des items, il a été décidé d'effectuer une recension des écrits. L'ajustement, entre les données et le modèle, est un aspect incontournable quand un modèle de mesure est utilisé. Il demande au chercheur une bonne compréhension de ses tenants et aboutissants. De la qualité de la calibration dépend la qualité de l'analyse. La revue de littérature qui a été faite n'avait pas pour ambition d'être exhaustive. Elle a été conçue comme une synthèse des différents positionnements des chercheurs et un argumentaire pour justifier l'utilisation des statistiques de l'ajustement des données au modèle pour la recherche. Cet argumentaire n'a pas pour ambition d'affirmer « la » vérité ou « la bonne manière » d'ajuster les données au modèle mais de dégager des principes opérationnels.

Comme suite de la recension des écrits, le lecteur trouvera ci-dessous la démarche (en quatre étapes) et les principes qui ont été retenus pour calibrer les données. Il convient de préciser que, concernant les items, il s'agissait surtout, pour le chercheur, de savoir s'ils avaient les qualités minimales requises pour la calibration. Comme cela a déjà été évoqué, l'échantillon de candidats n'était pas assez étoffé pour procéder à une pré-calibration des items. Il a donc été décidé de procéder à une validation d'expert.

L'objectif de la calibration n'était pas d'avoir des items permettant d'avoir un test immédiatement utilisable, mais bien de profiter de l'expérimentation pour tester des hypothèses sans avoir la « pression » de produire un instrument de mesure parfaitement fiable et calibré. Il va de soi, qu'il ne s'agit pas non plus de tester des hypothèses à partir de données non valides, ce qui n'aurait aucun intérêt.

Le test de compréhension écrite étant composé de 30 questions réparties dans six textes, trois portant sur des tâches intégrées et trois sur des tâches discrètes, il est très difficile d'envisager de supprimer un item, à moins d'un énorme dysfonctionnement. À ce sujet, Wilson (2005 : 132), déclare que quand un item n'est pas de très bonne qualité, le chercheur doit décider s'il peut ou non le garder. Dans le cadre de la présente recherche, il fallait donc faire un choix entre avoir un test avec un nombre d'items suffisant mais de qualité inégale et avoir un test avec des items de bonne qualité mais en nombre insuffisant.

Première étape de l'ajustement

Tout d'abord l'enquête est dirigée vers la corrélation point-bisérielle (appelée « *point-mesure* » dans WINSTEPS, puisqu'elle est calculée à partir du score en *logits*). Tous les items ou personnes avec une corrélation négative sont éliminés de l'échantillon. Une telle corrélation indique un dysfonctionnement grave, une mesure qui va dans le sens contraire des mesures des autres items ou candidats.

Deuxième étape

Les corrélations *point-bisérielles* de moins de 0,3 sont examinées attentivement. On interprète ces corrélations en fonction du niveau de la personne ou encore de l'item. La décision est alors prise de garder les items ou les personnes en fonction des analyses de l'ajustement des données au modèle.

Troisième étape

Pour l'ajustement des données au modèle, tout d'abord, les statistiques *infit* sont examinées. L'*infit*, autrement dit, l'indice d'ajustement interne, est plus sensible que l'*outfit*. C'est une statistique (basée sur le chi-carré) qui est plus sensible aux patrons de réponses non-attendues des personnes pour des items qui sont proches de leur niveau (ou des patrons de réponses aux items inattendues pour des personnes qui sont proches du niveau de ces items). Pour la recherche, sont utilisées les valeurs des carrés moyens. Est pris en considération le fait que les valeurs des carrés moyens sont sensibles à la taille de l'échantillon et que plus l'échantillon est grand, plus leur variabilité est faible (Smith, Schumacker, & Bush, 1998 ; Wang & Chen, 2005 : 402). Malgré la recommandation de Smith, Schumacker et Bush (1998) d'utiliser les transformations standardisées des carrés moyens pour éviter les problèmes de taille d'échantillon, les carrés moyens sont souvent plus faciles à interpréter. Dans notre études, les carrés moyens sont privilégiés pour l'interprétation car l'hypothèse nulle posée par la recherche est de savoir si les données

s'ajustent au modèle suffisamment et non pas parfaitement (Linacre, 2002 : 878). Les valeurs des statistiques *t* standardisées sont utilisées lorsque les valeurs des carrés moyens ne sont pas comprises dans l'intervalle décrit, ici, au point 5. La statistique *infit* est considérée comme étant plus importante que la statistique *outfit* puisque le propos de la recherche est de mener l'enquête pour des items proches du niveau des candidats et des personnes proches du niveau des items. L'*outfit*, autrement dit l'indice d'ajustement externe, est une statistique (basée sur le chi-carré) sensible aux valeurs extrêmes. Elle permet de détecter des patrons de réponses non-attendues des personnes face à des items qui sont relativement difficiles ou faciles pour ces personnes (ou des patrons de réponses non-attendues des personnes qui ont un niveau nettement supérieur ou inférieur pour aux items). Pour la recherche, l'*outfit* est consulté lorsque l'item présentera des problèmes d'*infit*.

Quatrième étape

Le cas échéant, les patrons de réponses « suspects » des candidats sont examinés afin d'expliquer le sens des statistiques *infit* et *outfit*. Toutefois, une attention particulière a été portée pour ne pas proposer d'interprétations causales trop abusives. On s'efforce de juger du gain (ou la perte) entraîné par la suppression des réponses données par un candidat (sauf cas extrême, les items ne peuvent pas être supprimés) en relation avec les questions de recherche. Les suppressions sont validées chaque fois qu'elles améliorent la qualité de l'échantillon.

Pour les items suspects, on procède à une analyse des leurres *post hoc*. Sont mises en parallèle les informations fournies par les statistiques *infit* et *outfit* et l'analyse du fonctionnement des leurres.

Intervalles des valeurs pour les carrées moyens et les t standardisés des statistiques *infit* et *outfit*

Pour un test de lecture en « rodage » (n'ayant pu procéder qu'à une révision d'expert, si les questions ne sont pas parfaites, elles sont, pour le moins, adéquates), calibré avec le

modèle de base de Rasch et le logiciel WINSTEPS, l'intervalle des valeurs de carrés moyens de l'*infit* (et de l'*outfit*, le cas échéant) est de [0,75;1,3] (Bond & Fox, 2001 : 179). L'intervalle de valeur utilisé pour les valeurs standardisées est de [-2 ; 2]. Comme le signalent Wang et Chen (2005 : 386), le choix de l'intervalle fixe [0,75;1,3] pour les carrés moyens de l'*infit* et de l'*outfit*, avec un échantillon de taille modeste et 30 items, est un choix conservateur. Même s'il n'y a pas vraiment de limites absolues pour les valeurs *infit* et *outfit* (Wilson, 2005 : 129 ; Wang & Chen, 2005 : 402-403), l'intervalle a été choisi afin de détecter les items et les personnes ne s'ajustant pas au modèle et de garder les items et les personnes même s'ils ne s'ajustent pas parfaitement au modèle.

Chapitre 5 : Présentation des résultats liés aux questions de recherche

Afin faciliter la lecture des résultats, les analyses préliminaires (qui portent sur la calibration des items et les analyses afférentes) ont été placées en annexe (annexe 10). Un bref résumé de ces analyses est présenté ci-dessous.

A l'issue des différentes calibrations, les 30 items composant le test ont été conservés. 113 personnes composent l'échantillon. La consistance interne est correcte. En effet, l'alpha de Cronbach a une valeur de 0,82 pour les personnes et de 0,78 pour les items. L'examen des corrélations entre l'ensemble du test et ses différentes parties (items intégrés, items discrets, items pairs, items impairs) montre que la corrélation la plus basse est celle entre les items des tâches discrètes et intégrées (r=0,614). La compétence des candidats mesurée avec l'ensemble des items n'est pas significativement différente selon que les candidats sont des hommes ou des femmes. La variable « âge » ne permet pas de faire de prédiction au sujet des scores des candidats. Aucun effet de fatigue n'a été trouvé. En revanche, l'appartenance à un groupe linguistique explique 34 % des différences de moyennes entre les personnes. Il faut ajouter à cela que les performances des Sinophones et des Slaves aux tâches discrètes et aux tâches intégrées ont très peu de lien entre elles. En effet, il est difficile de prédire la performance des candidats à partir de l'un ou l'autre type de tâche pour ces deux groupes linguistiques.

Les calibrations séparées des items des tâches discrètes et intégrées ne montrent pas de différences notables d'estimation de la difficulté des items avec la calibration de l'ensemble des items. Que l'estimation de la compétence de lecture soit faite à partir de l'ensemble des items du test, uniquement les items des tâches discrètes ou uniquement les items des tâches intégrées, le classement des items ne change pas. Lorsque la compétence des candidats est estimée avec l'ensemble des items, les moyennes calculées à partir des items discrets et des items intégrés sont sensiblement identiques.

Aucun groupe linguistique ne présente de problème d'ajustement des données par rapport au modèle. Que la compétence de lecture soit estimée avec l'ensemble des items, uniquement les items des tâches discrètes ou uniquement les items des tâches intégrées, le

classement des groupes linguistiques ne change pas. Par ordre décroissant de compétence, les groupes linguistiques sont classés de la manière suivante : groupe des langues latines, groupe des langues slaves, groupes des multilingues et groupe des Sinophones. Pour ce qui est des moyennes des groupes selon le type d'items, à l'exception des Sinophones, tous les autres groupes ont de meilleurs résultats aux items des tâches discrètes qu'aux items des tâches intégrées. En effet, pour les Sinophones, les moyennes aux items des tâches intégrées et discrètes sont identiques.

Concernant le classement des candidats, selon que la compétence est estimée à partir de l'ensemble des items, les items des tâches discrètes, les items des tâches intégrées, les items pairs et les items impairs, la corrélation la plus faible est celle trouvée entre la compétence estimée à partir des items des tâches discrètes et intégrées (r=0,606). Enfin, les classements des Sinophones et du groupe des langues latines, établis selon la différence des rangs obtenus avec les items des tâches discrètes et intégrées, sont significativement différents.

Comme il est possible de le constater, les analyses préliminaires font écho aux questions de recherche. En effet, si la corrélation entre la compétence des candidats calculée à partir des tâches discrètes et celle calculée à partir des tâches intégrées est moyenne et est la plus basse de toutes, elle pose la question de la dimensionnalité. S'il existe des différences de moyennes significatives entre les groupes linguistiques, il est légitime de de s'interroger sur la persistance de ces différences lorsque les candidats ont un niveau similaire. Enfin, si l'appartenance à un groupe de candidat explique 34 % de la variabilité, on peut se demander comment cela se reflète dans la perception de la difficulté du test par les candidats.

5.1 Première question de recherche : l'unidimensionnalité

5.1.1 Introduction

Tout d'abord, sera étudiée la dimensionnalité de l'ensemble des tâches, puis, pour chacun des deux types de tâches. Pour l'ensemble des tâches, une approche confirmatoire ayant été retenue, la fonction « *bifactor* » de Testfact a été utilisée. Pour l'ensemble du test, ont été testées :

- une première solution avec un facteur général et deux facteurs correspondant aux 15 items des tâches discrètes et aux 15 items des tâches intégrées ;
- une deuxième solution avec un facteur général et six facteurs regroupant les items de chaque texte ;
- une solution avec un facteur général, un autre pour tous les items des tâches intégrées et trois regroupant les items de chacun des tests discrets ;
- et enfin une solution avec un facteur général.

Pour les tâches discrètes et intégrées calibrées séparément, trois solutions ont été étudiées :

- une première solution avec un facteur général et trois facteurs regroupant les items des trois textes discrets ;
- une deuxième solution avec un facteur général et un facteur spécifique ;
- une troisième solution avec un facteur général.

5.1.2 Unidimensionnalité pour l'ensemble du test

Des trois solutions appliquées à l'ensemble des items (tableau 5.1), celle qui permet d'expliquer le plus de variance est celle qui opère une distinction entre les 15 items des tâches intégrées et les 15 items des tâches discrètes (40% de variance expliquée). Cette solution regroupe sur un facteur général 32,4% de la variance, 3,2% pour les items des tâches discrètes et 4,5% pour les items des tâches intégrées. Cette solution est aussi celle qui permet d'obtenir la meilleure corrélation entre les deux groupes d'items (r=0,673). La deuxième solution, pour les six textes, explique presque autant de variance (37,5%). Elle regroupe un peu plus de variance sur le facteur général (33,2%). Toutefois, chacun des six facteurs (textes), pris isolément, explique très peu de variance. La troisième solution ne permet d'expliquer que 31,8% de la variance et, enfin, la solution avec un seul facteur général n'explique que 24,1 % de la variance. L'analyse des corrélations de ces deux solutions « six textes » et « trois textes discrets et un ensemble de textes intégrés » ne permet pas de dégager des schémas facilement interprétables.

La première solution (avec un facteur général, un facteur pour les tâches intégrées et un facteur pour les tâches discrètes) présente donc l'avantage d'expliquer plus de variance

que les autres solutions. Elle présente encore l'avantage d'être plus économe que les solutions avec « six textes » et « trois textes discrets et un ensemble de textes intégrés ». En effet, avec moins de facteurs, elle explique plus de variance. Cela est remarquable car généralement plus on ajoute de facteurs et plus on explique de variance.

Tableau 5.1 : variance expliquée par les différentes solutions appliquées à l'ensemble des items du test

Solution 15 items discrets et 15 items intégrés	Solution six textes	Solution trois textes discrets, un ensemble de textes intégrés	Solution avec un facteur général
GENERAL 0 32.3968 ITEM GROUP 1 3.2358 ITEM GROUP 2 4.4772 UNIQUENESS 59.8902 SUBTEST SCORE CORRELATIONS DISCRETE / INTEGREE 0.673	GENERAL 0 33.1640 ITEM GROUP 1 0.8677 ITEM GROUP 2 0.2130 ITEM GROUP 3 0.7644 ITEM GROUP 4 0.0348 ITEM GROUP 5 0.9913 ITEM GROUP 6 1.4019 UNIQUENESS 62.5629 SUBTEST SCORE CORRELATIONS T1 T2 T3 T4 T5 T2 0.443 T3 0.536 0.436 T4 0.270 0.296 0.463 T5 0.408 0.421 0.491 0.294 T6 0.447 0.414 0.449 0.271 0.293	GENERAL 0 24.1351 ITEM GROUP 1 1.0630 ITEM GROUP 2 0.0549 ITEM GROUP 3 1.1344 ITEM GROUP 4 5.3448 UNIQUENESS 68.2679 SUBTEST SCORE CORRELATIONS T1 T2 T3 T2 0.443 T3 0.536 0.436 INT 0.516 0.538 0.665	GENERAL 0 24.4506 ITEM GROUP 1 0.0000 UNIQUENESS 75.5494

Pour la solution avec six textes, les items des tâches intégrées sont moins corrélés entre eux que les items des tâches discrètes. Pour la troisième solution (« trois textes discrets et un ensemble de textes intégrés »), les items intégrés ont des corrélations plus fortes avec chacun des trois textes discrets que les textes discrets entre eux (toutefois, on doit se souvenir que les tâches discrètes ont cinq items chacune alors que l'ensemble des textes intégrés est composé de 15 items).

L'analyse de la solution avec les 15 items discrets et les 15 items intégrés montre que tous les items n'ont pas la même qualité (tableau 5.2). Si tous les items sont liés au facteur général, cinq items des tâches discrètes (A4, B3, B4, C4 et C5) ont des corrélations négatives avec le facteur spécifique et six items des tâches intégrées ont également des corrélations négatives (D5, E4, E5, F1, F2 et F3). Lorsque l'on prend en compte le facteur général et le facteur spécifique (« *communality* », carré du facteur général + carré du facteur spécifique), là encore, le comportement des items varie

beaucoup. Dans cette solution, les items qui s'écartent le plus du modèle sont les items D5, E5 et F5.

Tableau 5.2 : analyse des résultats de la solution « *bifactor* » pour les items discrets et intégrés

ITEM	GROUP	DIFFICULTY	COMMUNALITY	GENERAL	SPECIFIC
1 A1	1	-0.7375	0.1135	0.3336	0.0463
2 A2	1	-0.2053	0.5553	0.7370	0.1101
3 A3	1	0.9440	0.5655	0.5643	0.4971
4 A4	1	-0.3427	0.2394	0.4880	-0.0361
5 A5	1	0.4682	0.1437	0.3180	0.2062
6 B1	1	-0.5291	0.5498	0.7341	0.1041
7 B2	1	-0.3936	0.2102	0.4581	0.0183
8 B3	1	0.5373	0.4335	0.6163	-0.2316
9 B4	1	-0.1726	0.3268	0.5666	-0.0760
10 B5	1	-0.0371	0.3285	0.5704	0.0558
11 C1	1	1.7384	0.9648	0.8064	0.5609
12 C2	1	-0.2573	0.1405	0.2628	0.2673
13 C3	1	-0.6569	0.3141	0.5217	0.2047
14 C4	1	-0.3424	0.2116	0.4599	-0.0075
15 C5	1	0.3606	0.5766	0.6423	-0.4049
16 D1	2	0.0142	0.1081	0.3171	0.0868
17 D2	2	0.0775	0.5813	0.6126	0.4540
18 D3	2	-0.1253	0.3620	0.5947	0.0915
19 D4	2	1.1048	0.8833	0.6496	0.6792
20 D5	2	0.2920	0.0538	0.2044	-0.1098
21 E1	2	-0.8032	0.3822	0.5811	0.2110
22 E2	2	-0.2574	0.2006	0.4346	0.1084
23 E3	2	2.0253	0.9842	0.9570	0.2616
24 E4	2	-0.4356	0.4281	0.6466	-0.1000
25 E5	2	0.6112	0.0737	0.2338	-0.1381
26 F1	2	0.8964	0.6901	0.7663	-0.3208
27 F2	2	-0.3157	0.7352	0.6677	-0.5379
28 F3	2	-0.0287	0.2007	0.3191	-0.3144
29 F4	2	0.2601	0.6248	0.7887	0.0527
30 F5	2	-0.1314	0.0511	0.2259	0.0111

5.1.3 Unidimensionnalité pour les taches discrètes et pour les tâches intégrées

5.1.3.1 *Unidimensionnalité pour les tâches discrètes*

Tableau 5.3 : variance expliquée par les différentes solutions appliquées pour les items des tâches discrètes calibrés à partir de l'ensemble des items du test

Solution avec trois textes

GENERAL	0	32.0549
ITEM GROUP	1	0.8377
ITEM GROUP	2	2.3599
ITEM GROUP	3	5.3104
UNIQUENESS		59.4371

Solution avec facteur général et un ensemble de textes

GENERAL	0	28.6292
ITEM GROUP	1	11.1029
UNIQUENESS		60.2679

Solution avec un facteur général

GENERAL	0	30.2891
ITEM GROUP	1	0.0000
UNIQUENESS		69.7109

Les deux premières solutions expliquent presque autant de variance avec un léger avantage pour la solution avec les trois textes (40,5% de variance expliquée) (tableau 5.3). Le facteur général regroupe plus de variance pour la solution à trois textes (32%) que pour la solution à un ensemble de textes (28,5%). Pour la solution avec trois textes, la variance expliquée par les facteurs spécifiques est de 8,5% au total. La solution avec un facteur général est celle qui explique le moins de variance. Toutefois, comme on peut le constater, la variance expliquée pour ce facteur général est assez proche de celle expliquée dans les deux autres solutions (30,3 % contre 32% et 28%).

L'analyse des items discrets dans la solution « un facteur général » montre que tous les items ont des corrélations positives avec le facteur général (tableau 5.4).

Tableau 5.4 : analyse des résultats de la solution « *bifactor* » (un facteur général) pour les items discrets

	ITEM	GROUP	DIFFICULTY	COMMUNALITY	GENERAL	SPECIFIC
1	A1	0	-0.8782	0.0743	0.2725	0.0000
2	A2	0	-0.1291	0.8938	0.9454	0.0000
3	A3	0	0.7161	0.3788	0.6155	0.0000
4	A4	0	-0.4253	0.1472	0.3837	0.0000
5	A5	0	0.2419	0.0792	0.2814	0.0000
6	B1	0	-0.5510	0.3376	0.5811	0.0000
7	B2	0	-0.4940	0.1629	0.4036	0.0000
8	B3	0	0.3336	0.2461	0.4961	0.0000
9	B4	0	-0.3689	0.0925	0.3042	0.0000
10	B5	0	-0.0866	0.5264	0.7255	0.0000
11	C1	0	1.4920	0.8090	0.8994	0.0000
12	C2	0	-0.3390	0.1146	0.3385	0.0000
13	C3	0	-0.6574	0.4746	0.6889	0.0000
14	C4	0	-0.4778	0.0936	0.3060	0.0000
15	C5	0	0.2648	0.1128	0.3358	0.0000

5.1.3.2 **Unidimensionnalité des tâches intégrées**

Pour les items des tâches intégrées, la solution avec un ensemble de textes est de loin la meilleure (tableau 5.5). Elle permet d'expliquer 44% de la variance. La solution à trois textes permet d'expliquer 31% de la variance sur un facteur général, 0,5 % pour le groupe d'items du premier texte, 3,5% pour celui du deuxième et 2% pour celui du troisième. La solution avec un facteur général est celle qui explique le moins de variance avec 29%. Toutefois, la variance expliquée par le facteur principal (29%) est presque identique à celle expliquée dans la solution avec les trois textes (31%).

Tableau 5.5 : variance expliquée par les différentes solutions appliquées pour les items intégrés calibrés à partir de l'ensemble des items du test

Solution avec trois textes			Solution avec facteur général et un ensemble de textes			Solution avec un facteur général		
GENERAL	0	31.3598	GENERAL	0	36.2730	GENERAL	0	29.0588
ITEM GROUP	1	0.5909	ITEM GROUP	1	7.9032	ITEM GROUP	1	0.0000
ITEM GROUP	2	3.2899	UNIQUENESS		55.8238	UNIQUENESS		70.9412
ITEM GROUP	3	1.8212						
UNIQUENESS		62.9383						

Si on analyse de plus près les résultats de la solution avec un ensemble de 15 items (tableau 5.6), on constate que l'item E3 a une corrélation négative avec le facteur général. Dans l'analyse menée à partir de tous les items, l'item était également fortement corrélé au facteur général mais sa corrélation était fortement positive (0,95). L'item F5 n'a rien de commun avec le facteur général (« *communality* »)

Tableau 5.6 : analyse des résultats de la solution « *bifactor* » pour les items intégrés

	ITEM	GROUP	DIFFICULTY	COMMUNALITY	GENERAL	SPECIFIC
1	D1	0	-0.0150	0.1079	0.3285	0.0000
2	D2	0	0.0212	0.5610	0.7490	0.0000
3	D3	0	-0.1331	0.3196	0.5654	0.0000
4	D4	0	1.0400	0.7429	0.8619	0.0000
5	D5	0	0.2369	0.0521	0.2283	0.0000
6	E1	0	-0.7670	0.3252	0.5703	0.0000
7	E2	0	-0.3054	0.2317	0.4814	0.0000
8	E3	0	2.4398	0.7285	**-0.8535**	0.0000
9	E4	0	-0.4528	0.4129	0.6426	0.0000
10	E5	0	0.4396	0.1112	0.3335	0.0000
11	F1	0	0.6464	0.1446	0.3803	0.0000
12	F2	0	-0.3651	0.1810	0.4254	0.0000
13	F3	0	-0.0515	0.0709	0.2663	0.0000
14	F4	0	0.1867	0.3522	0.5935	0.0000
15	F5	0	-0.1762	0.0170	0.1305	0.0000

5.2 Analyse des résultats liés à la deuxième question de recherche

Pour mémoire, la seconde question de recherche est la suivante : « Pour des candidats à niveau de compétence équivalent, des tâches pour l'évaluation discrètes et intégrées de niveaux différents (mais suffisamment proche du niveau des candidats), présentent-elles un niveau de difficulté homogène pour tous les candidats ? ».

Pour répondre à cette question de recherche, les variations de niveau pour les candidats de niveau intermédiaire ou encore de « niveau 2 » (cf. annexe 10 -figure A.10.13- et annexe 7), selon les items auxquels ils répondent (ensemble de tous les items, tous les items des tâches discrètes ou bien intégrées, uniquement les items du même niveau que celui des candidats (items de « niveau 2 », cf. annexe 10 -figure A.10.13- et annexe 7), ont été étudiées. Enfin, l'étude a porté sur le classement des personnes en fonction du type de tâche. Les changements de classement des candidats de « niveau 2 » attribuables au type de tâche ont aussi été étudiés.

5.2.1 Différence des résultats des candidats du « niveau 2 » aux tâches discrètes et intégrées calibrées avec l'ensemble des items ou bien séparément

Dans un premier temps, et pour bien distinguer les niveaux d'analyse, la différence de niveau est étudiée avec les items calibrés avec l'ensemble des items du test, puis avec les items des tâches discrètes et intégrées calibrés isolément.

Lorsque le niveau des items et celui des candidats sont obtenus à partir de l'ensemble des items, WINSTEPS, nous signale une seule différence significative, pour des candidats de « niveau 2 », entre l'estimation du niveau de compétence aux tâches discrètes et intégrées (tableau 5.7). Ce candidat, le numéro 13, qui a reçu la version 2 du test (tâches intégrées puis tâches discrètes), a mieux réussi que prévu les tâches intégrées que les tâches discrètes. Toutefois un autre candidat, le numéro 83 est à la limite du seuil de signification. Celui-ci qui a reçu la version 1 du test (tâches discrètes puis intégrées) a mieux réussi que prévu les tâches discrètes que les tâches intégrées. Enfin, le candidat

117 de niveau 3 (à la limite supérieure du niveau 2) (annexe 10 -figure A.10.13-), a mieux réussi que prévu les items des tâches intégrées que des tâches discrètes.

Tableau 5.7 : différences significatives entre l'estimation de la compétence des candidats du « niveau 2 » pour les tâches discrètes et intégrées calibrées conjointement

```
INPUT18 personnes, 30 items  MEASURED: 49 personnes, 30 items, 2 CATS    3.60.1
-------------------------------------------------------------------------------
DPF class specification is: DPF=$S4E12
+-------------------------------------------------------------------------------------+
| item        DPF   DPF   item        DPF   DPF      DPF    JOINT                personne
|
| CLASS     MEASURE S.E.  CLASS     MEASURE S.E.  CONTRAST  S.E.   t  d.f. Prob. Number
|
|-------------------------------------------------------------------------------------|
| Discrètes  -1.81   .83  Intégrées    .78   .76     -2.59  1.13 -2.30  16 .0350     13 |
| Discrètes   1.20   .71  Intégrées   -.65   .58      1.85   .92  2.01  27 .0544     83 |
| Discrètes   -.80   .67  Intégrées   2.14   .83     -2.94  1.06 -2.76  23 .0110    117 |
+-------------------------------------------------------------------------------------+
```

L'étude des corrélations de Pearson des estimations du niveau de compétence des candidats obtenues à partir de l'ensemble des items du test ou bien uniquement avec ceux des tâches discrètes et intégrées (tableau 5.8) pour les personnes du groupe « niveau 2 », nous donne les indications suivantes :

- La corrélation de la compétence des candidats de « niveau 2 » estimée à partir des items des tâches discrètes et intégrées est proche de zéro (-0,11). Elle est la plus basse des corrélations. Toutefois, alors que toutes les autres corrélations sont significatives, celle-ci ne l'est pas (p=0,44).

- Les corrélations entre la compétence des candidats de « niveau 2 » estimée à partir des items de « niveaux 2 » (cf. annexe 10 -figure A.10.13-) et tous les items est moyenne (0,58) et relativement basse avec les tâches discrètes et intégrées.

Tableau 5.8 : corrélations de Pearson entre les différentes versions du test (calibration 30 items et calibrations séparées des 15 items des tâches discrètes, des 15 items des tâches intégrées et des 16 items de « niveau 2 ») pour les candidats de « niveau 2 »

Correlations

		Estimation habileté à partir des items des tâches discrètes	Estimation habileté à partir des tâches intégrées	Estimation habileté à partir de tous les items	Estimation habileté à partir de tous les items de niveau 2
Estimation habileté à partir des items des tâches discrètes	Pearson Correlation	1,00	-,11	,68**	,42**
	Sig. (2-tailed)		,44	,00	,00
	N	48,00	48,00	48,00	48,00
Estimation habileté à partir des tâches intégrées	Pearson Correlation	-,11	1,00	,56**	,34*
	Sig. (2-tailed)	,44		,00	,02
	N	48,00	49,00	49,00	49,00
Estimation habileté à partir de tous les items	Pearson Correlation	,68**	,56**	1,00	,58**
	Sig. (2-tailed)	,00	,00		,00
	N	48,00	49,00	49,00	49,00
Estimation habileté à partir de tous les items de niveau 2	Pearson Correlation	,42**	,34*	,58**	1,00
	Sig. (2-tailed)	,00	,02	,00	
	N	48,00	49,00	49,00	49,00

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

L'examen des moyennes de chacun des groupes linguistiques du « niveau 2 »[1] aux tâches discrètes et intégrées calibrées séparément (tableau 5.9 ; figure 5.1) permet de constater que les sinophones obtiennent une moyenne presque identique pour les deux types de tâches.

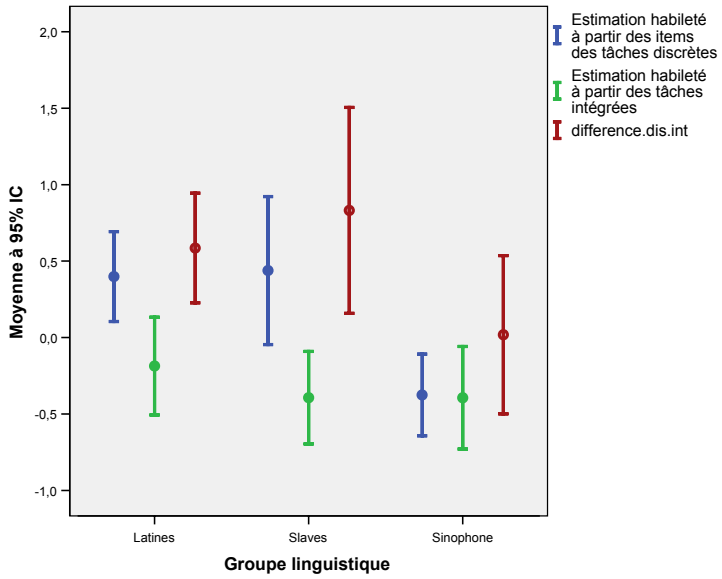
Tableau 5.9 : moyennes de l'estimation de la compétence des candidats de « niveau 2 » pour les 15 items des tâches discrètes et les 15 items intégrées calibrés séparément et moyenne de la différence de ces deux estimations

Report

GRLING		Difference estimation habileté avec tâches discrètes et intégrées	Estimation habileté à partir des items des tâches discrètes	Estimation habileté à partir des tâches intégrées
latines	Moyenne	,59	,40	-,19
	N	17	17	17
slaves	Moyenne	,83	,44	-,39
	N	11	11	11
sinophones	Moyenne	,02	-,38	-,39
	N	17	17	17
Total	Moyenne	,43	,12	-,32
	N	45	45	45

[1] Le groupe des « multilingues » comprenant quatre personnes, il n'a pas été retenu pour cette partie de l'analyse.

Figure 5.1 : moyenne à 95 % de l'estimation de l'habileté à partir des 15items des tâches discrètes, des 15 items des tâches intégrées, et la moyenne de la différence entre ces deux types de tâches pour les candidats de « niveau 2 »



La différence de leurs moyennes aux tâches discrètes et aux tâches intégrées n'est que de 0,2 *logit*. Ce résultat contraste avec celui des autres groupes pour lesquels la différence est importante (0,59 *logit* pour le groupe des langues latines), et significativement différente pour le groupe des Slaves (0,83 *logit*). Par ailleurs, les sinophones du « niveau 2 », par rapport aux deux autres groupes linguistiques de « niveau 2 », s'écartent plus de la moyenne obtenue aux tâches discrètes qu'aux tâches intégrées (écart significatif avec les autres groupes).

Les moyennes atteintes par les candidats des trois groupes linguistiques (langues latines, slaves et sinophones) ne diffèrent pas significativement (figure 5.1) quand elles sont calculées à partir de l'ensemble des items ou encore à partir de tous les items intégrés. Par contre, celles calculées à partir de tous les items des tâches discrètes sont presque significativement différentes. Les sinophones de « niveau 2 » sont moins performants que les locuteurs de langues latines et slaves de « niveau 2 » lorsqu'ils répondent aux items des tâches discrètes. Toutefois, les moyennes de la différence de compétence obtenue avec les items de l'ensemble du test issus des tâches discrètes et ceux tâches intégrées ne diffèrent pas selon le groupe linguistique.

5.2.2 Fonctionnement des items de « niveau 2 » des tâches discrètes et intégrées, calibrés séparément pour les personnes des groupes linguistiques de « niveau 2 »

Dans cette partie de l'analyse, la calibration des items a été effectuée à partir des seuls candidats de « niveau 2 » (49 candidats) et des seuls items de « niveau 2 » (16 items). Il s'agit de vérifier le fonctionnement des items des tâches discrètes et intégrées lorsque les items et les personnes ont un niveau, soit de difficulté, soit de compétence, comparable. Les moyennes (avec des candidats de « niveau 2 » et des items de « niveau 2 ») de chacun des groupes linguistiques sont très proches les unes des autres (tableau 5.10, figure 5.2).

Tableau 5.10 : moyenne des groupes linguistiques de « niveau 2 » aux tâches linguistiques de niveaux 2 » et moyennes des items des tâches discrètes et intégrées de « niveau 2 »

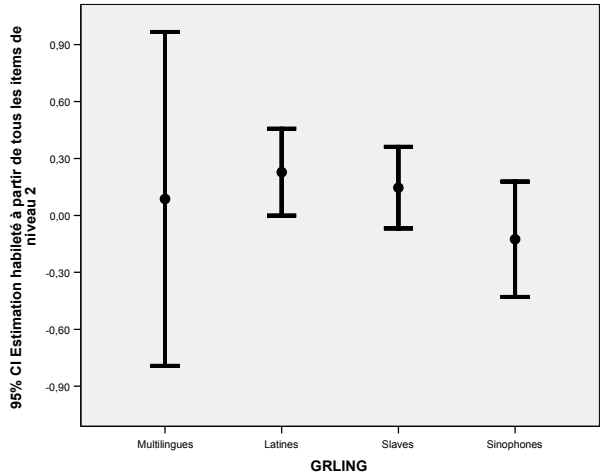
INPUT: 118 personnes 30 items MEASURED: 49 personnes 16

person COUNT	MEAN MEASURE	S.E. MEAN	OBSERVED S.D.	MEDIAN SEPARATION	REAL	CODE
49	.09	.07	.49	.01	.00	*
4	.12	.31	.53	.14	.00	A
17	-.11	.14	.57	-.26	.22	C
17	.23	.11	.43	.26	.00	L
11	.16	.09	.29	.26	.00	S

item COUNT	MEAN MEASURE	S.E. MEAN	OBSERVED S.D.	MEDIAN SEPARATION	REAL	CODE
16	.00	.10	.39	.09	.77	***
7	-.06	.17	.42	.06	.91	Discrètes
9	.04	.13	.37	.11	.60	Intégrées

Légende :
A= multilingues,
C= sinophones,
L= langues latines ;
S= langues slaves

Figure 5.2 : moyennes à 95% des groupes linguistiques de « niveau 2 » calculées à partir des 16 items de « niveau 2 »



Du côté des items, on constate la même chose. La moyenne de la difficulté des items des tâches intégrées et des tâches discrètes est similaire. Afin de vérifier la moyenne atteinte par chacun des groupes linguistiques de « niveau 2 » avec les items des tâches discrètes et les items des tâches intégrées de « niveau 2 », il n'est pas possible de modéliser les

Figure 5.3 : moyennes avec un intervalle de confiance à 95 % pour les candidats, par groupes de langues composés des candidats de « niveau 2 » et des items de « niveau 2 » issus des tâches discrètes

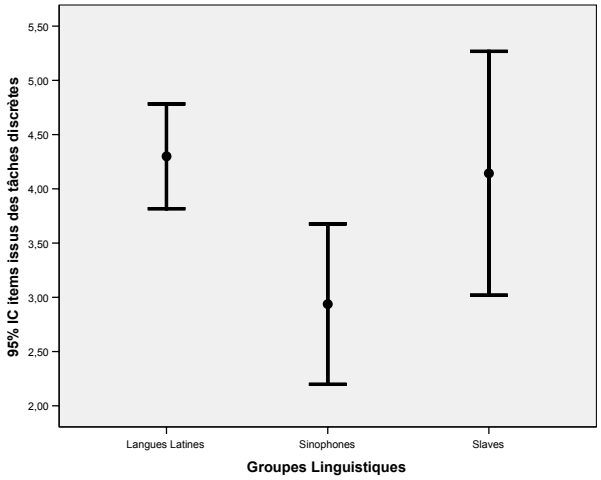


Figure 5.4 : moyennes avec un intervalle de confiance à 95 % pour les candidats, par groupes de langues composés des candidats de « niveau 2 » et des items de « niveau 2 » issus des tâches intégrées

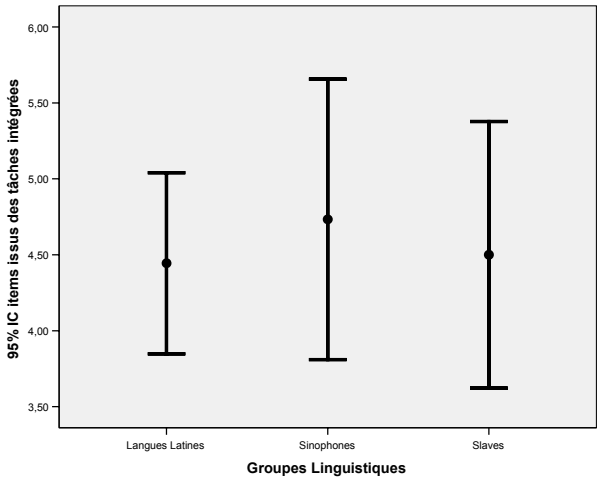


Tableau 5.11 : mesure d'association entre la variable groupe linguistique composés de candidats de « niveau 2 » et items discrets issus des tâches de « niveau 2 »

Mesure D'association

	Eta	Eta carré
discrets * GR.LING	,483	,234

données avec le modèle de Rasch. En effet, il y a seulement sept items issus des tâches discrètes et neuf issus des tâches intégrées. Afin de procéder à des analyses, il a été décidé d'utiliser les scores bruts (le nombre de bonnes réponses aux items) des candidats de « niveau 2 » ayant répondu aux items de « niveau 2 ». Les items ayant déjà été calibrés avec l'ensemble les 118 candidats et des 30 items, on sait qu'ils ont une mesure relativement comparable. Le groupe des « multilingues », composé de quatre éléments n'a pas été retenu pour l'analyse.

L'étude des scores bruts nous apprend (figures 5.3 et 5.4) que les moyennes atteintes par l'ensemble des groupes linguistiques de « niveau 2 » sont similaires pour l'ensemble des candidats pour les items issus des tâches intégrées. En revanche, pour les tâches discrètes, la moyenne atteinte par les sinophones de « niveau 2 » est significativement plus basse que celle atteinte par les autres groupes de « niveau 2 ».

L'examen de la mesure d'association entre la variable « score brut aux sept items discrets de niveau 2 » et la variable « groupe linguistique de niveau 2 » (tableau 5.11) nous apprend que l'appartenance à un groupe linguistique pour des candidats de « niveau 2 » ayant répondu à des items issus des tâches discrètes de « niveau 2 » permet d'expliquer 23 % de la variance.

En résumé, on remarque que lorsqu'on calibre ensemble les 16 items de « niveau 2 » avec des candidats de « niveau 2 », la moyenne calculée en *logits* des items des tâches discrètes et des tâches intégrées est similaire ainsi que les moyennes des groupes linguistiques (tableau 6.10). En revanche, lorsqu'on analyse les scores bruts des candidats de « niveau 2 » aux items de « niveau 2 » selon le type de tâches, on constate une différence de moyenne significative entre les sinophones et le groupe des langues latines. L'appartenance à un groupe linguistique dans ce dernier cas explique une bonne partie de la variance (tableau 5.11).

5.2.3 Différences de classement pour les candidats de « niveau 2 »

Pour étudier le classement des candidats, la corrélation de rang de Spearman ainsi que des tests non-paramétriques ont été utilisés. Ensuite, la distribution des candidats de « niveau 2 » (pour l'ensemble du test) dans les niveaux 1, 2 et 3 pour les tâches discrètes et intégrées a été étudiée.

L'examen des corrélations de Spearman (tableau 5.12) entre le test complet (30 items calibrés avec 118 personnes), les deux sous-tests calibrés séparément (15 items et 118 personnes) et l'ensemble des items de « niveau 2 » calibrés séparément (49 personnes de « niveau 2 » et 16 items de « niveau 2 ») révèle que les deux sous-tests (items discrets et items intégrés) sont corrélés significativement au test complet. La corrélation entre le test complet et les tâches discrètes (r=0,68, sig.=0,00) est plus forte que celle entre le test complet et les tâches intégrées (r= 0,54, sig.=0,000). Par ailleurs, la corrélation de rang entre l'estimation de la compétence calculée à partir des items des tâches discrètes et intégrées n'est pas significative (sig.=0,39). Pour ce qui est de l'estimation de la compétence des candidats de « niveau 2 » calculée à partir des items de « niveau 2 », les corrélations sont moyennes avec toutes les autres estimations. La corrélation avec les items des tâches intégrées est la plus faible (r=0,34).

Tableau 5.12 : corrélation de rangs (Spearman) entre le test complet et les deux sous-tests calibrés séparément, pour les candidats de « niveau 2 » (compétence estimée à partir des scores calculés en *logits* pour les 118 candidats)

Correlations

			Items des tâches discrètes	Items des tâches intégrées	Tous items	Tous les items de niveau 2
Spearman's rho	Items des tâches discrètes	Correlation Coefficient	1,00	-,12	,68**	,46**
		Sig. (2-tailed)	.	,39	,00	,00
		N	49	49	49	49
	Items des tâches intégrées	Correlation Coefficient	-,12	1,00	,54**	,34*
		Sig. (2-tailed)	,39	.	,00	,02
		N	49	49	49	49
	Tous items	Correlation Coefficient	,68**	,54**	1,00	,58**
		Sig. (2-tailed)	,00	,00	.	,00
		N	49	49	49	49
	Tous les items de niveau 2	Correlation Coefficient	,46**	,34*	,58**	1,00
		Sig. (2-tailed)	,00	,02	,00	.
		N	49	49	49	49

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Un test non-paramétrique de Wilcoxon (tableau 5.13), pour échantillon pairé, a été fait pour vérifier si le classement des candidats de « niveau 2 »[2] varie selon qu'ils répondent aux 15 items des tâches discrètes ou bien aux 15 items des tâches intégrées (calibrés séparément). La réponse à cette question est positive (z=-2,513 ; sig=0,012).

Tableau 5.13 : test de Wilcoxon pour tester la différence de classement pour les candidats de « niveau 2 » aux tâches discrètes et intégrées calibrées séparément

Ranks

		N	Mean Rank	Sum of Ranks
Estimation habileté à partir des tâches intégrées - Estimation habileté à partir des items des tâches discrètes	Negative Ranks	31[a]	26,87	833,00
	Positive Ranks	17[b]	20,18	343,00
	Ties	0[c]		
	Total	48		

a. Estimation habileté à partir des tâches intégrées < Estimation habileté à partir des items des tâches discrètes

b. Estimation habileté à partir des tâches intégrées > Estimation habileté à partir des items des tâches discrètes

c. Estimation habileté à partir des tâches intégrées = Estimation habileté à partir des items des tâches discrètes

Test Statistics[b]

	Estimation habileté à partir des tâches intégrées - Estimation habileté à partir des items des tâches discrètes
Z	-2,513[a]
Asymp. Sig. (2-tailed)	,012

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Afin de vérifier si la différence de classement des candidats de « niveaux 2 » avec les items des tâches discrètes et des tâches intégrées est due à l'appartenance à un groupe linguistique, un test de Kruskal-Wallis sur la variable de la différence de score pour les

[2] Dans ce test, on vérifie l'hypothèse nulle selon laquelle la somme des rangs de différences positives est égale à la somme des rangs négatifs

candidats de « niveaux 2 » aux 15 items des tâches discrètes et aux 15 items des tâches intégrées a été administré (tableau 5.14). Ce test des moyennes des rangs n'est pas significatif (chi-carré=5,386 ; dll.=2 ; sig.=0,068).

Tableau 5.14 : test de kruskal-Wallis sur la variable de la différence des scores aux tâches discrètes et intégrées pour trois groupes linguistiques de « niveau 2 » estimation de la compétence faite avec les 30 items

Ranks

	GRLING	N	Mean Rank
difference.dis.int	latines	17	25,18
	Slaves	11	28,32
	sinophone	17	17,38
	Total	45	

Test Statistics[a,b]

	difference.dis.int
Chi-Square	5,386
df	2
Asymp. Sig.	,068

a. Kruskal Wallis Test

b. Grouping Variable: GRLING

Enfin, deux tests de Kruskal-Wallis (tableau 5.15) ont été faits pour tester l'égalité des moyennes des rangs entre les différents groupes linguistiques de « niveaux 2 » pour les tâches discrètes et intégrées calibrées séparément. Alors que la moyenne des rangs pour les tâches intégrées ne varie pas significativement entre les groupes linguistiques (chi-carré=0,791, dll.=2, sig.=0,673) la moyenne des rangs pour les tâches discrètes varie significativement (chi-carré=14,078, dll.=2, sig.=0,001).

Tableau 5.15 : tests de Kruskall-Wallis pour tester la différence de la moyenne des rangs pour les tâches intégrées et discrètes calibrées séparément pour les différents groupes linguistiques de « niveau 2 »

Ranks

	GRLING	N	Mean Rank
Niveau des candidats pour les tâches intégrées	Latines	17	25,18
	Slaves	11	21,18
	Sinophones	17	22,00
	Total	45	
Niveau des candidats pour les tâches discrètes	Latines	17	28,65
	Slaves	11	28,68
	Sinophones	17	13,68
	Total	45	

Test Statistics[a,b]

	Niveau des candidats pour les tâches intégrées	Niveau des candidats pour les tâches discrètes
Chi-Square	,791	14,078
df	2	2
Asymp. Sig.	,673	,001

a. Kruskal Wallis Test

b. Grouping Variable: GRLING

Le groupe des sinophones est celui qui présente la plus grande variation pour la moyenne de rang (μ=22 pour les tâches intégrées et μ=13,68 pour les tâches discrètes). Le test de la médiane (test de rangs sur deux échantillon indépendants) lui aussi est significatif pour les tâches discrètes (Chi-carré= 11,876, dll=2, sig.=0,003) (tableau 5.16). Le groupe des sinophones a donc un profil différent des autres groupes linguistiques. Alors que ce

groupe a presque autant de candidats au-dessus et au-dessous de la médiane des rangs, pour les tâches intégrées, une seule personne est au-dessus de la médiane et 16 sont en dessous pour les tâches discrètes.

Tableau 5.16 : résultats du test de la médiane pour tester la différence de la médiane des rangs pour les tâches intégrées et discrètes calibrées séparément pour les différents groupes linguistiques de « niveau 2 »

Frequencies

		GRLING		
		Latines	Slaves	Sinophones
Niveau des candidats pour les tâches discrètes	> Median	10	6	1
	<= Median	7	5	16
Niveau des candidats pour les tâches intégrées	> Median	10	4	8
	<= Median	7	7	9

Test Statistics[c]

	Niveau des candidats pour les tâches discrètes	Niveau des candidats pour les tâches intégrées
N	45	45
Median	-,4025	-,3461
Chi-Square	11,876[a]	1,385[b]
df	2	2
Asymp. Sig.	,003	,500

a. 1 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 4,2.

b. 0 cells (,0%) have expected frequencies less than 5. The minimum expected cell frequency is 5,4.

c. Grouping Variable: GRLING

Tableau 5.17 : tests de Kruskall-Wallis pour tester la différence de la moyenne des rangs pour les groupes linguistiques composés de personnes de « niveau 2 » pour une estimation de la compétence faite à partir des 16 tâches de « niveau 2 »

Ranks

	groupe.linguistique	N	Mean Rank
Tous les items de niveau 2	Langues latines	17	26,79
	Sinophones	17	18,12
	Langues slaves	11	24,68
	Total	45	

Test Statistics[a,b]

	Tous les items de niveau 2
Chi-Square	4,018
df	2
Asymp. Sig.	,134

a. Kruskal Wallis Test

b. Grouping Variable: groupe.linguistique

Avec les 16 items de « niveau 2 », le test de Kruskal-Wallis entre la variable dépendante de l'estimation de la compétence linguistique et la variable indépendante des groupes linguistiques de « niveau 2 » (tableau 5.17) n'est pas significatif. L'appartenance à un groupe linguistique pour les candidats de « niveau 2 » n'a donc pas de lien avec le classement des candidats lorsque l'ensemble des items des tâches discrètes et intégrées de « niveau 2 » est utilisé pour l'évaluation des candidats.

Pour finir, le classement des candidats de « niveau 2 » a été étudié (tableau 5.18). Pour ces analyses, l'estimation de la compétence des candidats aux différents types de tâches a

été calibrée à partir des 30 items. L'objectif était de considérer les éventuelles erreurs de classement. Les trois niveaux qui avaient été définis dans les analyses préliminaires ont été utilisés (annexe 10 -figure A.10.13-, annexe 7).

Tableau 5.18. : comparaison du classement des candidats de « niveau 2 » pour l'ensemble du test aux tâches discrètes et intégrées

Count

		Classement avec les tâches discrètes			
		1,00	2,00	3,00	Total
Classement avec les items des tâches intégrées	1,00	0	12	2	14
	2,00	3	22	8	33
	3,00	1	0	0	1
Total		4	34	10	48

Tableau 5.19 : comparaison du classement des candidats lorsque l'estimation de la compétence des candidats est faite à partir de l'ensemble des items ou bien à partir des items des tâches discrètes et des tâches intégrées

Count

		Classement avec les tâches discrètes			
		1,00	2,00	3,00	Total
Classement avec les trente items	1,00	11	2	0	13
	2,00	4	34	10	48
	3,00	0	8	44	52
Total		15	44	54	113

Count

		Classement avec les items des tâches intégrées			
		1,00	2,00	3,00	Total
Classement avec les trente items	1,00	10	3	0	13
	2,00	14	33	1	48
	3,00	0	14	38	52
Total		24	50	39	113

Lorsque la compétence des candidats est estimée à partir des 15 items des tâches discrètes et 15 items des tâches intégrées calibrés séparément (tableau 5.18, tableau 5.19), 33 candidats conservent le « niveau 2 » estimé avec l'ensemble des items. En revanche, de nombreux candidats obtiennent un classement différent, selon que leur compétence est estimée avec les tâches discrètes ou intégrées. Avec les tâches discrètes, calibrées séparément, quatre candidats de « niveau 2 » obtiennent le « niveau 1 » et dix le « niveau 3 ». Avec les tâches intégrées, calibrées séparément, 14 obtiennent un « niveau 1 » et un candidat le « niveau 3 ». Pour les candidats de « niveau 2 », il y a presque autant de candidats qui obtiennent un classement différent avec les tâches intégrées et les tâches discrètes. En revanche, la tendance est au « sous-classement » avec les tâches intégrées et au « sur-classement » avec les tâches intégrées.

A présent, si on examine, le classement des candidats de « niveau 2 » aux tâches discrètes et intégrées (tableau 5.20), on constate que 22 candidats ont un « niveau 2 », pour

l'ensemble des items calibrés conjointement, et les items des deux types de tâches calibrés séparément. Seuls trois candidats présentent une différence de deux niveaux (« niveau 1 » et « niveau 3 »). 11 candidats présentent une différence de classement d'un niveau.

Tableau 5.20. : classement des candidats de « niveau 2 » pour les tâches discrètes et intégrées en fonction du groupe linguistique

Classement avec les items des tâches intégrées * Classement avec les tâches discrètes * gr. linguistiques

Count

gr.linguistique			Classement avec les tâches discrètes 1,00	2,00	3,00	Total
Langues latines	Classement avec les items des tâches intégrées	1,00		4	1	5
		2,00		8	4	12
	Total			12	5	17
Sinophones	Classement avec les items des tâches intégrées	1,00	0	6		6
		2,00	2	7		9
		3,00	1	0		1
	Total		3	13		16
Langues slaves	Classement avec les items des tâches intégrées	1,00		1	1	2
		2,00		6	3	9
	Total			7	4	11

Différence de classement pour les candidats de "niveau 2"

pour les deux types de tâches* gr.linguistique

Count

		Groupe linguistique Langues latines	Sinophones	Langues slaves	Total
Différence des rangs	-2,00	0	1	0	1
	-1,00	0	2	0	3
	,00	8	7	6	22
	1,00	8	6	4	20
	2,00	1	0	1	2
Total		17	16	11	48

En prenant en compte la variable « groupe linguistique » (tableau 5.20) pour analyser ces résultats, différents phénomènes apparaissent :

- La moitié des personnes du groupe des langues latines et slaves de « niveau 2 », conservent leur « niveau 2 », que l'estimation de leur niveau de compétence soit faite avec les 15 items des tâches discrètes ou les 15 items des intégrées. L'autre moitié obtient un, voire, deux niveaux de plus pour les tâches discrètes que pour les tâches intégrées.

- La moitié des personnes du groupe des sinophones de « niveau 2 », conserve son « niveau 2 » que l'estimation du niveau de compétence soit faite avec les 15 items des tâches discrètes ou les 15 items des intégrées. Six obtiennent un niveau de plus pour les tâches discrètes que pour les tâches intégrées. Trois obtiennent un niveau de plus pour les tâches intégrées que pour les tâches discrètes.

5.3 Analyse des résultats liés à la troisième question de recherche

5.3.1 Calibration des items et des personnes

Pour la calibration des personnes et des items du questionnaire portant sur la perception de la difficulté par les candidats, l'échantillon de départ qui a été utilisé est celui des 113 candidats issus de la calibration du test composé de 30 items. La première calibration est le résultat de 32 itérations menées avec le logiciel WINSTEPS (annexe 9). Parmi les 113 personnes conservées, 14 présentent trop de réponses manquantes pour être conservées pour la calibration des données. Une personne a un score extrême. Le nombre de personnes conservées est donc de 98, sans le score extrême, et 99 avec le score extrême. Pour les personnes, la mesure maximale est de 4,70 et la mesure minimale de –3,32 *logits*. Pour ce qui est des items, le maximum de 1,46 et le minimum de -1,27. L'indice de fidélité KR-20 des personnes (« *reliability* ») signalé par WINSTEPS est de 0,95. Pour les items, le logiciel signale une valeur de 0,90.

Les corrélations point-bisérielles (tableau 5.21) des items présentent des valeurs satisfaisantes. Les statistiques de l'ajustement des données par rapport au modèle (*infit* et *outfit*) se situent dans les intervalles [0,6 ; 1,4], intervalles généralement utilisés pour calibrer les données dans un modèle « *rating scale* » (Bond & Fox, 2001 : 179). Aucune valeur standardisée n'a de valeur au-delà de l'intervalle –2, +2.

Tableau 5.21 : ajustement des données par rapport au modèle pour les items du questionnaire de perception subjective de la difficulté

INPUT: 118 personnes, 14 questions MEASURED: 99 personnes, 14 questions, 5 CATS

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	question
4	300	95	-.15	.17	1.27	1.8	1.25	1.6	A .67	52.6	58.5	DQB
5	279	96	.51	.16	1.12	.8	1.11	.8	B .66	60.4	56.8	DTC
8	328	94	-.99	.17	1.07	.5	1.09	.7	C .65	67.0	58.3	IQD
13	253	84	.18	.17	1.09	.6	1.08	.6	D .66	58.3	57.5	TD3
6	314	95	-.53	.17	1.08	.6	1.07	.6	E .65	53.7	58.7	DQC
11	304	96	-.12	.16	1.05	.4	1.05	.4	F .77	57.3	58.5	ITF
7	283	97	.50	.16	1.03	.3	1.03	.2	G .74	55.7	56.7	ITD
10	319	93	-.84	.17	1.01	.1	1.03	.2	g .72	53.8	58.6	IQE
9	286	96	.33	.16	1.02	.2	1.02	.2	f .81	52.1	57.3	ITE
2	271	92	.35	.17	.92	-.5	.92	-.5	e .73	54.3	57.2	DQA
1	244	96	1.46	.16	.89	-.8	.90	-.7	d .77	62.5	57.3	DTA
12	335	93	-1.27	.17	.75	-1.8	.87	-.9	c .76	62.4	58.2	IQF
14	275	83	-.62	.18	.78	-1.5	.80	-1.3	b .75	69.9	59.0	TI3
3	251	95	1.19	.16	.76	-1.8	.76	-1.8	a .82	65.3	56.7	DTB
MEAN	288.7	93.2	.00	.17	.99	-.1	1.00	.0		58.9	57.8	
S.D.	28.0	4.2	.77	.00	.15	1.0	.13	.9		5.5	.8	

Pour ce qui est des personnes, 16 présentent des valeurs *infit* au-delà de 1,4, mais aucune ne dépasse 2. Ôter la totalité ou bien quelques-unes unes de ces personnes de l'échantillon ne permet pas d'améliorer la qualité de la mesure. Il a donc été décidé de conserver l'échantillon de 99 personnes.

Pour ce qui est des catégories utilisées pour le test, au nombre de cinq (1, 2, 3, 4, 5), leur distribution est normale. Il y a plus de dix observations par catégorie. La valeur calculée en *logits* augmente entre chaque catégorie (tableau 5.22). Ces informations permettent de dire que les catégories fonctionnent correctement (Bond & Fox, 2001 : 162). Par ailleurs, l'intervalle entre chaque catégorie est de plus de 1,4 et de moins de 5 *logits* (figure 5.5) ce qui indique un bon fonctionnement des catégories (Bond & Fox, 2001 : 163 ; Linacre, 1999b).

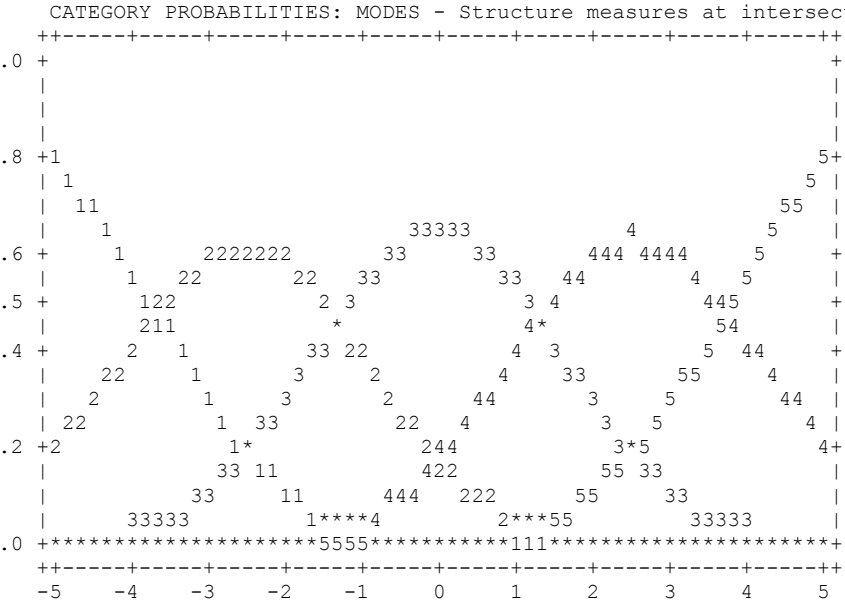
Tableau 5.22 : catégories de réponses pour le test portant sur la perception de la difficulté par les candidats

INPUT: 118 personnes, 14 questions MEASURED: 99 personnes, 14 questions, 5 CATS

SUMMARY OF CATEGORY STRUCTURE. Model="R"

CATEGORY LABEL	SCORE	OBSERVED COUNT	%	OBSVD AVRGE	SAMPLE EXPECT	INFIT MNSQ	OUTFIT MNSQ	STRUCTURE CALIBRATN	CATEGORY MEASURE	
1	1	52	4	-2.79	-2.63	.90	.91	NONE	(-4.84)	1
2	2	277	20	-1.33	-1.35	.93	.93	-3.67	-2.53	2
3	3	554	40	.07	.06	.95	.95	-1.34	-.02	3
4	4	336	24	1.56	1.55	1.11	1.15	1.29	2.53	4
5	5	86	6	3.11	3.22	1.01	1.01	3.72	(4.88)	5
MISSING		67	5	1.73						

Figure 5.5 : intersection des catégories pour la perception de la difficulté par les candidats



5.3.2 Analyse de la perception de la difficulté des tâches, des questions et des textes pour l'ensemble des candidats

Pour rappel, à l'issue de la passation du test, les candidats devaient répondre à un questionnaire (annexe 1) sur leur perception de la difficulté des tâches, des questions et des textes. Ils devaient répondre à des items destinés à mesurer leur perception de la difficulté des tâches discrètes et intégrées dans leur ensemble, puis à leur perception de la difficulté de chacun des textes et des cinq items associés à chacune des six tâches. Pour mesurer la difficulté subjective, il leur était proposé des échelles de Likert à cinq échelons. La difficulté subjective a été étudiée pour l'ensemble des candidats puis pour chacun des groupes de langues. Enfin, la difficulté subjective des candidats de « niveau 2 » a été analysée.

Dans l'ensemble (tableau 5.23 et 5.24), les candidats pensent que les tâches discrètes (variable TD3) sont plus faciles que les tâches intégrées (variable TI3). La différence est de près d'un *logit*. Ils pensent encore que les textes (variables DTA, DTB, DTC, ITD, ITE, ITF) sont plus faciles que les questions (variables DQA, DQB, DQC, IQD, IQD, IQF).

Tableau 5.23 : Classement de la difficulté perçue pour l'ensemble des candidats et par groupe linguistique

Classement croissant de la difficulté perçue[3]	Tous	Langues latines	Sinophones	Langues Slaves	Candidats « Niveau 2 »
Classement par type de tâche					
Tâches discrètes (TD3)	7	7	4	5	5
Tâches intégrées (TI3)	11	11	12	9	11
Classement par texte					
Texte A (discret, DTA)	1	1	1	1	1
Texte B (discret, DTB)	2	2	5	2	2
Texte C (discret, DTC)	3	5	3	6	3
Texte D (intégré, ITD)	4	4	8	4	6
Texte E (intégré, ITE)	6	3	9	3	9
Texte F (intégré, ITF)	8	9	10	7	8
Classement par questions associées au texte					
Questions A (discret, DQA)	5	6	2	8	4
Questions B (discret, DQB)	9	8	6	13	7
Questions C (discret, DQC)	10	10	7	11	10
Questions D (intégré, IQD)	13	13	11	14	12
Questions E (intégré, IQE)	12	12	13	10	13
Questions F (intégré, IQF)	14	14	14	12	14

[3] Le rang 1 correspond donc au plus facile et le rang 14 au plus difficile.

Ainsi les textes A, B, C et D sont-ils perçus comme étant les « items » les plus faciles (variables DTA, DTB, DTC, ITD) et les questions portant sur les textes E, D, F comme les plus difficiles (variables IQE, IQD, IQF). Selon les candidats, les textes les plus faciles sont les textes des tâches discrètes puis ceux des tâches intégrées. L'ordre de difficulté pour les textes est le suivant : A, B, C, D, E et F. Seules les questions associées à la tâche A sont perçues comme étant plus faciles que les textes E et F.

Pour les questions, celles des tâches discrètes (variables DQA, DQB et DQC) sont perçues comme étant les plus faciles et celles des tâches intégrées comme étant les plus difficiles. L'ordre de difficulté pour les questions est le suivant : A, B, C, E, D, F. Alors que le texte D est perçu comme étant plus facile que le texte E, les questions du texte E sont perçues comme étant plus faciles que celles du texte D. Toutefois, les questions les plus difficiles pour les candidats sont bien les questions du texte F.

Tableau 5.24 : perception de la difficulté des questions et des textes par l'ensemble des candidats

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	question
1	244	96	1.46	.16	.89	-.8	.90	-.7	.77	62.5	57.3	DTA
3	251	95	1.19	.16	.76	-1.8	.76	-1.8	.82	65.3	56.7	DTB
5	279	96	.51	.16	1.12	.8	1.11	.8	.66	60.4	56.8	DTC
7	283	97	.50	.16	1.03	.3	1.03	.2	.74	55.7	56.7	ITD
2	271	92	.35	.17	.92	-.5	.92	-.5	.73	54.3	57.2	DQA
9	286	96	.33	.16	1.02	.2	1.02	.2	.81	52.1	57.3	ITE
13	253	84	.18	.17	1.09	.6	1.08	.6	.66	58.3	57.5	TD3
11	304	96	-.12	.16	1.05	.4	1.05	.4	.77	57.3	58.5	ITF
4	300	95	-.15	.17	1.27	1.8	1.25	1.6	.67	52.6	58.5	DQB
6	314	95	-.53	.17	1.08	.6	1.07	.6	.65	53.7	58.7	DQC
14	275	83	-.62	.18	.78	-1.5	.80	-1.3	.75	69.9	59.0	TI3
10	319	93	-.84	.17	1.01	.1	1.03	.2	.72	53.8	58.6	IQE
8	328	94	-.99	.17	1.07	.5	1.09	.7	.65	67.0	58.3	IQD
12	335	93	-1.27	.17	.75	-1.8	.87	-.9	.76	62.4	58.2	IQF
MEAN	288.7	93.2	.00	.17	.99	-.1	1.00	.0		58.9	57.8	
S.D.	28.0	4.2	.77	.00	.15	1.0	.13	.9		5.5	.8	

Légende :
Première lettre : D= discrète, I= intégrée, T=tâche
Deuxième lettre : T= texte, Q= question, I, D= discrète, I= intégrée
Troisième lettre : A, B, C, D, E, F correspondent à chacune des tâches portant la même lettre. 3 signifie ensemble des trois tâches.

5.3.3 Analyse de la perception de la difficulté par les groupes de candidats

Le groupe des langues slaves étant affecté par un sujet atyique, pour cette partie de l'analyse sont pris en compte les 97 candidats non-extrêmes.

Tableau 5.25 : perception de la difficulté par chacun des groupes de candidats

EXTREME AND NON-EXTREME SCORES

person COUNT	MEAN MEASURE	S.E. MEAN	OBSERVED S.D.	MEDIAN SEPARATION	REAL	CODE
98	.32	.19	1.85	.07	3.14	*
9	.82	.75	2.13	1.03	3.69	A
21	.88	.36	1.59	1.24	2.84	C
52	-.31	.21	1.52	-.45	2.71	L
16	1.31	.56	2.16	1.03	2.95	S

UMEAN=0 USCALE=1

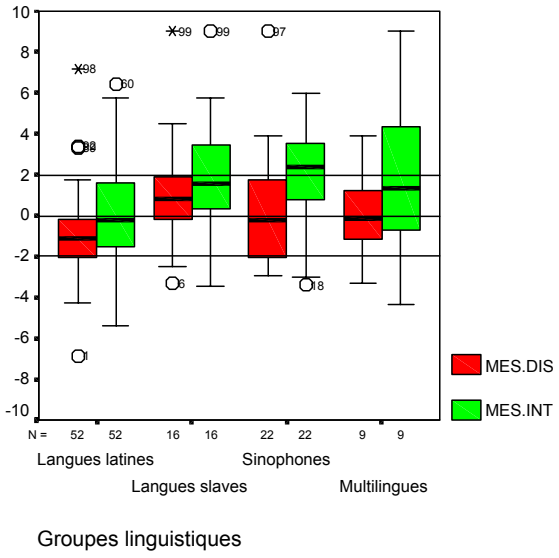
NON-EXTREME SCORES ONLY

person COUNT	MEAN MEASURE	S.E. MEAN	OBSERVED S.D.	MEDIAN SEPARATION	REAL	CODE
97	.24	.17	1.69	-.03	3.03	*
9	.82	.75	2.13	1.03	3.69	A
21	.88	.36	1.59	1.24	2.84	C
52	-.31	.21	1.52	-.45	2.71	L
15	.88	.38	1.44	1.03	2.50	S

UMEAN=0 USCALE=1

Légende = A= multilingue, C= sinophones, L= langues latines, S= langues slaves.

Figure 5.6 : boîtes à moustache de la difficulté perçue par les candidats (calculée en logits) pour les tâches discrètes et intégrées calibrées séparément



Les groupes pour lesquels le test, dans son ensemble, a été le plus difficile (tableau 5.25) sont le groupe des slaves et celui des sinophones, suivis des multilingues et, enfin, des langues latines. Le groupe des langues latines est donc celui qui a pensé que le test était le plus facile.

Afin de vérifier si la difficulté perçue (calculée en logits) par le groupe des langues latines diffère de celle perçue par les autres candidats, il a été procédé à un test ANOVA (tableau 5.26). Il ressort de

ce test qu'effectivement la perception de la difficulté des items par le groupe des langues latines varie significativement de celles de l'ensemble des autres candidats.

Comme il est possible de le constater dans la figure 5.6, pour l'ensemble des groupes linguistiques, les tâches intégrées sont perçues comme étant plus difficiles que les tâches discrètes. L'écart le plus marqué entre la médiane de la difficulté des tâches discrètes et intégrées est celui du groupe des sinophones (différence de 2,5 *logits*).

5.3.4 Difficulté perçue par le groupe des langues latines

Pour le groupe des langues latines, dans l'ensemble, les textes sont perçus comme étant plus faciles que les questions. Les textes des tâches discrètes ne sont pas systématiquement perçus comme étant plus difficiles que ceux des tâches intégrées (tableau 5.27).

Tableau 5.26 : test sur la perception de la difficulté du test par le groupe des langues latines et les autres groupes linguistiques (variable perception de la difficulté calculée en logits et bariable groupe linguistique recodée)

Avec tous les candidats

Descriptives

Estimated personne Measure: UMEAN=.00 USCALE=1.00

	N	Mean	Std. Deviation	Std. Error
langues latines	52	-,3052	1,53331	,21263
autres groupes	46	1,0196	1,94915	,28739
Total	98	,3167	1,85479	,18736

Test d'homogénéité des variances

Estimated personne Measure: UMEAN=.00 USCALE=1.00

Levene Statistic	df1	df2	Sig.
1,374	1	96	,244

ANOVA

Estimated personne Measure: UMEAN=.00 USCALE=1.00

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	42,837	1	42,837	14,138	,000
Within Groups	290,866	96	3,030		
Total	333,703	97			

Avec les candidats non-extrêmes

Descriptives

Estimated personne Measure: UMEAN=.00 USCALE=1.00

	N	Mean
langues latines	52	-,3052
autres groupes	45	,8705
Total	97	,2403

ANOVA

Estimated personne Measure: UMEAN=.00 USCALE=1.00

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	33,347	1	33,347	12,937	,001
Within Groups	244,885	95	2,578		
Total	278,232	96			

En effet, les derniers textes des tâches discrètes et intégrées (C et F) sont perçus comme étant les plus difficiles. Le texte D, premier texte des tâches intégrées est perçu comme étant plus difficile que le texte E. L'ordre croissant de difficulté perçue pour les textes est donc le suivant : A, B, E, D, C, F.

Pour les questions portant sur les textes, ceux des tâches discrètes sont perçues comme étant plus faciles que celles des tâches intégrées. Selon ces candidats, les questions de la tâche D sont plus difficiles que les questions de la tâche E. L'ordre croissant de difficulté pour les questions est donc le suivant : A, B, C, E, D, F.

Tableau 5.27 : difficulté des textes et des tâches pour les candidats du groupe des langues latines

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	question
1	124	52	1.34	.23	.65	-2.0	.66	-2.0	.81	67.3	57.5	DTA
3	124	52	1.34	.23	.85	-.8	.85	-.7	.78	57.7	57.5	DTB
9	135	51	.62	.23	.76	-1.2	.75	-1.3	.83	68.6	57.8	ITE
7	141	52	.49	.22	1.03	.2	1.02	.2	.66	59.6	58.0	ITD
5	142	52	.44	.22	1.33	1.6	1.32	1.5	.49	51.9	58.1	DTC
2	149	52	.10	.22	.79	-1.0	.79	-1.0	.65	65.4	58.8	DQA
13	126	45	.02	.24	1.27	1.2	1.25	1.1	.40	55.6	59.0	TD3
4	153	52	-.09	.22	1.17	.9	1.17	.8	.61	55.8	59.1	DQB
11	152	51	-.17	.22	1.14	.7	1.12	.6	.79	51.0	59.2	ITF
6	163	52	-.57	.22	1.22	1.1	1.23	1.1	.56	46.2	57.6	DQC
14	138	45	-.64	.23	.74	-1.2	.73	-1.3	.71	66.7	57.7	TI3
10	161	50	-.77	.22	1.00	.1	.98	.0	.65	48.0	57.1	IQE
8	168	51	-.91	.22	1.09	.5	1.18	.9	.54	58.8	55.7	IQD
12	172	50	-1.23	.22	.77	-1.2	.90	-.4	.76	50.0	54.8	IQF
MEAN	146.3	50.5	.00	.22	.99	-.1	1.00	.0		57.3	57.7	
S.D.	15.5	2.4	.76	.01	.22	1.1	.21	1.1		7.2	1.2	

Pour le groupe des langues latines, le classement par ordre de difficulté des textes et des questions portant sur ces textes est identique pour les questions et les textes des tâches A, B et F. Il diffère quelque peu pour les tâches C, D, E. Il y a, toutefois, une certaine similitude entre ces deux classements.

5.3.5 Difficulté perçue par le groupe des langues slaves

Pour le groupe des langues slaves, l'ordre croissant de la perception de la difficulté des textes est A, B, E, D, C, F (tableau 5.28). Comme pour le groupe des langues latines, le texte D est considéré comme étant plus difficile que le texte E. Le texte C est considéré comme étant plus difficile que les textes E et D. Il semble donc que l'interprétation pour

les Slaves est que les deux premiers textes des tâches discrètes sont les plus faciles. Suivent les deux premiers textes des tâches intégrées (avec le deuxième texte plus difficile que le premier). Enfin, les derniers textes des tâches discrètes (C) et intégrées (F) sont perçus comme étant les plus difficiles. Le dernier texte des tâches discrètes est perçu comme étant plus facile que celui des tâches intégrées. Pour les slaves, les textes des tâches intégrées ne sont donc pas systématiquement plus difficiles que ceux des tâches discrètes.

Pour les items portant sur les textes, l'interprétation est différente de celle des autres groupes. L'ordre croissant de difficulté est A, E, C, F, B, D. Cet ordre n'est pas relié au type de tâche ou encore à la difficulté des textes (les questions des textes C, E et F sont considérées comme étant plus faciles que les textes sur lesquels elles portent et les questions des textes B et D plus difficiles que les textes). Seuls les rangs donnés pour les questions de la tâche A correspondent au rang donné pour le texte. Pour le groupe des langues slaves, dans l'ensemble, l'ordre de difficulté des questions et des textes n'est pas le même. Toutefois, les questions associées aux textes sont perçues comme étant plus difficiles que les textes eux-mêmes.

Tableau 5.28 : difficulté des textes et des tâches pour les candidats du groupe des langues slaves

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	question
1	35	14	2.62	.47	.83	-.3	.81	-.4	.87	71.4	63.4	DTA
3	38	13	1.34	.48	.88	-.2	.93	-.1	.87	69.2	59.7	DTB
9	42	14	1.10	.47	1.01	.2	.99	.1	.87	50.0	60.1	ITE
7	45	14	.45	.47	.71	-.7	.71	-.7	.88	71.4	62.3	ITD
13	37	11	.25	.52	1.08	.3	1.07	.3	.85	54.5	61.9	TD3
5	46	14	.23	.47	.90	-.1	.87	-.2	.86	64.3	63.6	DTC
11	46	14	.23	.47	.78	-.5	.80	-.4	.87	78.6	63.6	ITF
2	35	11	.22	.53	1.12	.4	1.12	.4	.82	54.5	63.3	DQA
14	39	11	-.30	.53	.66	-.8	.66	-.8	.82	81.8	64.2	TI3
10	45	13	-.42	.49	1.30	.8	1.28	.8	.76	46.2	62.4	IQE
6	48	13	-1.12	.49	.62	-1.1	.61	-1.1	.79	69.2	60.3	DQC
12	48	13	-1.12	.49	.41	-1.9	.41	-1.9	.85	84.6	60.3	IQF
4	49	13	-1.36	.49	2.24	2.6	2.20	2.6	.50	30.8	59.8	DQB
8	52	13	-2.10	.50	1.24	.7	1.27	.8	.57	53.8	62.0	IQD
MEAN	43.2	12.9	.00	.49	.98	.0	.98	.0		62.9	61.9	
S.D.	5.3	1.1	1.17	.02	.42	1.0	.41	1.0		14.6	1.5	

5.3.6 Difficulté perçue par le groupe des sinophones

Pour les sinophones (tableau 5.29), les textes, perçus comme étant les plus faciles sont ceux des tâches discrètes, les plus difficiles, ceux des tâches intégrées. L'ordre de difficulté subjective est le suivant : A, C, B, D, E, F. Le texte C est ressenti comme étant plus facile que le texte B. En revanche, les questions portant sur le texte C sont ressenties comme étant plus difficiles que celles du texte B.

Pour ce qui est des questions portant sur chacun des textes, l'ordre de la difficulté subjective est sensiblement le même que pour les textes : les questions portant sur les trois tâches discrètes sont ressenties comme étant les plus faciles, les questions portant sur les trois tâches intégrées comme étant les plus difficiles. L'ordre de difficulté ressenti est le suivant : A, B, C, D, E, F.

Pour les sinophones, la différence de difficulté perçue entre les tâches intégrées et les tâches discrètes est relativement importante ; plus de 1,5 *logits* séparent les deux mesures. Enfin, pour les sinophones, les items sont, en général, perçus comme étant plus difficiles que les textes.

Tableau 5.29 : difficulté des textes et des tâches pour les candidats du groupe des sinophones

ENTRY NUMBER	RAW SCORE	COUNT	MODEL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	question
1	53	19	1.43	.36	1.45	1.4	1.48	1.5	.76	57.9	55.3	DTA
2	57	19	.91	.36	.89	-.2	.89	-.3	.82	47.4	55.7	DQA
5	62	20	.68	.36	1.12	.5	1.21	.7	.72	60.0	55.6	DTC
13	57	18	.65	.38	.71	-.9	.78	-.6	.84	66.7	56.1	TD3
3	63	20	.56	.36	.45	-2.2	.43	-2.2	.87	80.0	55.9	DTB
4	64	20	.43	.36	.77	-.7	.75	-.8	.81	70.0	56.5	DQB
6	66	20	.16	.37	.84	-.4	.86	-.4	.76	70.0	58.8	DQC
7	66	20	.16	.37	1.19	.7	1.06	.3	.74	50.0	58.8	ITD
9	70	20	-.39	.38	1.51	1.5	1.60	1.6	.67	55.0	62.8	ITE
11	71	20	-.54	.38	1.17	.6	1.19	.7	.64	60.0	63.5	ITF
8	72	20	-.69	.39	.70	-.9	.62	-1.2	.76	80.0	63.8	IQD
14	67	18	-.91	.42	1.07	.3	1.34	1.0	.48	55.6	65.0	TI3
10	75	20	-1.15	.40	.98	.0	1.12	.5	.66	55.0	64.7	IQE
12	76	20	-1.31	.40	.80	-.5	1.04	.2	.67	70.0	64.2	IQF
MEAN	65.6	19.6	.00	.38	.97	-.1	1.03	.1		62.7	59.8	
S.D.	6.6	.7	.81	.02	.29	.9	.32	1.0		9.9	3.8	

5.3.7 Difficulté perçue par les candidats du « niveau 2 »

Pour les candidats de « niveau 2 » (tableau 5.30), les trois textes des tâches discrètes ont été perçus comme étant les plus faciles (DTA, DTB, DTC). Les items portant sur les trois tâches intégrées ont été perçus comme étant les plus difficiles (IQD, IQE, IQF). L'ordre de difficulté des textes et le suivant : A, B, C, D, F, E. Le texte E a donc été perçu comme étant plus difficile que le texte F.

Tableau 5.30 : difficulté des textes, des questions et des tâches pour les candidats du « niveau 2 »

INPUT: 118 personnes 14 items MEASURED: 44 personnes 14 items 5 CATS 3.63.2

personne: REAL SEP.: 2.49 REL.: .86 ... item: REAL SEP.: 2.67 REL.: .88

item STATISTICS: MEASURE ORDER

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	item
1	115	41	1.41	.26	1.17	.8	1.21	1.0	.68	56.1	58.9	DTA
3	118	40	1.01	.26	.66	-1.6	.66	-1.6	.78	75.0	57.9	DTB
5	125	41	.75	.25	1.12	.6	1.09	.5	.58	65.9	58.7	DTC
2	113	37	.64	.27	.91	-.3	.91	-.3	.77	54.1	58.7	DQA
13	111	35	.43	.27	.98	.0	.96	-.1	.65	57.1	58.3	TD3
7	135	42	.34	.25	1.00	.1	1.00	.1	.68	54.8	58.3	ITD
4	133	40	.00	.26	1.21	1.0	1.20	.9	.63	55.0	58.3	DQB
11	143	42	-.16	.25	1.11	.6	1.09	.5	.70	59.5	58.4	ITF
9	140	41	-.21	.25	1.21	1.0	1.18	.9	.71	56.1	58.5	ITE
6	139	40	-.40	.26	1.13	.7	1.12	.6	.58	52.5	58.4	DQC
14	124	34	-.85	.28	.68	-1.4	.66	-1.6	.75	79.4	58.1	TI3
8	146	40	-.85	.26	.85	-.7	.93	-.3	.65	70.0	58.2	IQD
10	144	39	-.99	.26	1.12	.6	1.12	.6	.64	53.8	58.0	IQE
12	150	40	-1.12	.26	.66	-1.8	.79	-1.0	.72	65.0	57.8	IQF
MEAN	131.1	39.4	.00	.26	.99	.0	.99	.0		61.0	58.3	
S.D.	12.8	2.4	.77	.01	.20	.9	.18	.8		8.3	.3	

Pour ce qui est des questions, l'ordre est le suivant : A, B, C, D, E, F. La différence de perception de la difficulté entre les tâches discrètes et les tâches intégrées (TD3 et TI3) est assez importante puisqu'elle est de plus de 1,20 *logits*.

Pour les candidats du « niveau 2 », les textes sont donc perçus comme étant plus faciles que les questions et les tâches discrètes plus faciles que les tâches intégrées. Toutefois, les textes ne sont pas perçus comme étant systématiquement plus faciles que les questions.

Chapitre 6 : Discussion

6.1 La dimensionnalité des tâches discrètes et des tâches intégrées

L'étude sur la dimensionnalité consistait à apporter des réponses quant à la nature de la dimensionnalité des tâches de lecture discrètes et intégrées. Les sous-questions posées étaient celles d'un possible positionnement sur une échelle de mesure unique et des différences de dimensionnalité entre les tâches de lecture discrètes et intégrées. Tout d'abord, pour ce qui est de la consistance interne, globalement, elle est satisfaisante. L'alpha de Cronbach (KR-20) pour les personnes (tableau A.10.1) (calibration 30 items) a une valeur de 0,82. L'étude des corrélations (qui a été faite à partir des scores calibrés obtenus par les candidats aux 30 items, aux 15 items pairs, aux 15 impairs et aux 15 items des tâches discrètes et aux 15 items des tâches intégrées) a montré que ces corrélations sont relativement élevées (tableau A.10.12). De plus, les items des tâches discrètes et intégrées, calibrés séparément, ont des corrélations assez élevées avec les 30 items du test calibrés conjointement (les deux corrélations ont une valeur de 0,89). Pour ce qui est de la corrélation entre les items des tâches discrètes et intégrées calibrées séparément, elle est modérée puisqu'elle n'est que de 0,614 (tableau A.10.11). Ces résultats montrent que la fidélité des items des tâches discrètes et intégrées semble être acceptable. La mesure de la compétence de lecture en langue française par les tâches discrètes et par les tâches intégrées produit des résultats relativement consistants. Tous ces résultats laissent à penser que les deux types de tâches, d'une part, permettent d'obtenir des résultats assez fidèles et d'autre part, participent d'une même dimension (la compétence de lecture en français). En effet, si tous les indices de fidélité (les alphas de Cronbach calculés pour le test avec 30 items, ceux calculés pour les 15 items des tâches discrètes et 15 items des tâches intégrées) avaient été faibles, ont aurait pu penser qu'ils étaient affectés par la présence d'items qui ne covariaient pas avec les autres items.

Pour ce qui est de l'étude de la dimensionnalité, proprement dite, le pourcentage de variance expliquée par l'analyse factorielle menée avec TESTFACT n'est pas très élevé. En effet, 40 % de la variance sont expliqués dans la meilleure des solutions, soit avec un

facteur général et deux facteurs spécifiques : un premier pour les 15 items des tâches discrètes et, un second pour les 15 items des tâches intégrées (tableau 5.1). Toutefois, on doit se souvenir que Bernhardt et Kamil (1995), dans leur étude, trouvent que 50% de la variance pour la lecture en langue seconde reste inexpliqués. Autrement dit, 50% du processus de lecture reste inexpliqué et n'est ni relié au niveau en langue seconde ou encore au niveau de lecture en langue maternelle.

Si on reprend l'idée de Blais et Laurier (1995a et b), selon laquelle il existe un continuum pour la dimensionnalité, on peut penser que le test qui a été développé ne présente pas une unidimensionnalité très forte. Toutefois, dans le cadre d'un test de positionnement, cela pourrait s'avérer être suffisant (d'autant plus qu'une version améliorée du test permettrait sans doute d'augmenter le pourcentage de variance expliquée). Pour ce qui est des facteurs spécifiques, le groupe des items discrets explique environ 3% de la variance et le groupe des items intégrés 4%. Ces pourcentages, s'ils ne sont pas négligeables, n'en restent pas moins relativement faibles. Par ailleurs, si le facteur général seul regroupe 24% de la variance, lorsqu'il est associé aux deux facteurs spécifiques (relativement faibles), il explique 32% de la variance. Ces données font penser au cas d'unidimensionnalité psychométrique mentionné par Reckase (1990) dans lequel deux compétences corrélées ont des vecteurs dirigés vers la compétence de l'expert.

Lorsque l'on analyse séparément les items des tâches intégrées et des tâches discrètes (tableau 5.3 et 5.5), avec la solution à un seul facteur général, la saturation sur le facteur général est supérieure à celle observée lorsqu'on analyse l'ensemble des 30 items (environ 30% contre 24%). Dans les solutions avec un facteur général et des facteurs spécifiques, pour les tâches discrètes, la saturation sur le facteur général est meilleure lorsque les items sont regroupés sur trois facteurs spécifiques correspondant aux trois textes. Pour les tâches intégrées, la saturation sur le facteur général est meilleure (36%) lorsque tous les items sont rassemblés sur un seul facteur spécifique (tableau 5.5). Pour les deux types de tâches étudiés séparément, la solution la meilleure est celle où un facteur général et des facteurs spécifiques sont associés.

Le type de tâche semble donc bien définir des dimensionnalités spécifiques. D'une part, lorsque les items sont calibrés tous ensemble, la meilleure solution est celle qui distingue un facteur général et deux facteurs spécifiques (un pour les tâches discrètes et un pour les

tâches intégrées). D'autre part, lorsque les items des tâches discrètes et intégrées sont analysés séparément, la meilleure solution pour les tâches discrètes est celle d'un facteur général et de trois facteurs spécifiques pour chacun des textes et la meilleure solution pour les tâches intégrées est celle d'un facteur général et d'un facteur spécifique regroupant toutes les tâches.

Ce résultat semble aller dans le sens des travaux de Widdowson (1981) sur les notions d'authenticité et d'intégration et ceux de Perfetti (1997) et Goldman (1997) sur les notions de « *document models* » et de « *situation* » (lien général entre deux textes). Il semble bien que le type de lecture qui doit être faite par le lecteur influe sur sa compétence. L'estimation du niveau de compétence à lire en langue étrangère est affectée selon que le lecteur doit lire des textes reliés entre eux ou pas. Il y a une différence sans doute due aux liens entre les textes. Lire trois textes sans lien de contenu, et, lire trois textes ayant des liens quant à leur contenu, a un effet sur la dimensionnalité de la tâche, et, sur la manière dont les facteurs spécifiques fonctionnent. Il est possible de rapprocher ces résultats de ceux obtenus par Jang et Roussos (2007 : 12) qui ont étudié la dimensionnalité du T.O.E.L.F. Ils trouvent que les items de compréhension écrite, liés à des minitests, présentent une multidimensionnalité faible à modérée. Ils expliquent ce résultat par le fait que les sous-dimensions du test doivent être fortement corrélées, mais plus intéressant, ils trouvent que les sous-dimensions ne sont pas uniquement dues à la présence de minitests, mais aux micro-compétences de lecture :

> « These results constitute strong evidence for identifiable and substantively meaningful reading comprehension dimensionality structure beyond that of testlet effects and with important measurement implications.[31] » (Jang et Roussos, 2007: 16).

Ces résultats semblent montrer que la nature des tâches et des items proposés aux candidats influence bien le résultat. La présence de dimensions sous-jacentes n'est pas uniquement due à la présence des minitests. Si Bachman (2002) explique la difficulté objective par la relation entre les caractéristiques des tâches, celles des candidats et le lien existant entre les tâches et les candidats, il est probable que cette relation influe également sur la dimensionnalité du test.

Pour ce qui est des caractéristiques des candidats, l'appartenance à un groupe linguistique permet de trouver des différences de moyennes significatives (tableau A.10.21 et A.10.22) lorsque le niveau de compétence des candidats est calculé avec les 30 items du test, soit tous les items des tâches discrètes et intégrées. Cependant, ce résultat est peut-être attribuable à une différence réelle de moyenne due au niveau de compétence de chacun des groupes et non pas à une différence attribuable à l'appartenance à un groupe en particulier. Rien dans les données ne permet de conclure de manière certaine que l'appartenance à un groupe linguistique conditionne entièrement une différence de moyenne. Toutefois, on peut émettre l'hypothèse selon laquelle le groupe des sinophones ne fonctionne pas de la même manière que les autres groupes.

La prédiction de l'estimation du niveau de compétence à partir des 15 items des tâches intégrées et des 15 items des tâches discrètes, calibrées séparément, est différente selon que l'on appartient au groupe des langues latines, des sinophones ou des slaves (figure A.10.3). Si pour les personnes du groupe des langues latines, l'estimation du niveau de compétence calculé à partir des items des tâches discrètes est un prédicteur moyen du niveau de compétence aux tâches intégrées (Rcarré = 0,37), pour les deux autres groupes, cette prédiction est nulle (Rcarré= 0,01 et Rcarré=0,02). Avec plus de sujets, on aurait pu vérifier si la structure factorielle trouvée pour l'ensemble des candidats se maintient pour les sinophones.

6.2 La difficulté des tâches intégrées et des tâches discrètes de « niveau 2 »

Pour rappel, cette deuxième question de recherche ambitionnait d'apporter des réponses aux questions concernant le niveau de difficulté empirique des tâches discrètes et intégrées. L'étude a porté sur la difficulté des items associés aux deux types de tâches, au classement des candidats selon le type de tâches utilisé et enfin aux éventuelles interactions entre le type de tâche et l'appartenance des candidats à une famille linguistique.

L'étude des items en fonction du type de tâche pour les candidats de « niveau 2 » nous apprend qu'il y a une corrélation nulle, non significative, entre les tâches discrètes et les tâches intégrées (tableau 5.9). Cette corrélation entre les items des tâches discrètes et

intégrées pour les candidats de « niveau 2 » est la seule qui pose problème puisque toutes les autres corrélations sont significatives. Comme cela a pu être observé lors de l'analyse, si les moyennes des groupes linguistiques de « niveau 2 » aux items de « niveau 2 »des tâches discrètes sont identiques (figure 5.3), il n'en est rien pour celles atteintes pour les items des tâches intégrées (figure 5.4). Par ailleurs, les sinophones de « niveau 2 » atteignent une moyenne identique aux 15 items des tâches intégrées et aux 15 items des tâches discrètes (tableau 5.9). Cela va dans le sens inverse des autres groupes (figure 5.1). Toutefois, il convient de prendre ces résultats avec prudence car cette partie de l'étude a porté sur 49 candidats et 15 items par type de tâches.

Pour ce qui est des moyennes des candidats de « niveau 2 » aux 15 items de chacun des deux types de tâches, la différence de moyenne est moins importante pour les sinophones que pour les autres groupes linguistiques (tableau 5.8 ; figure 5.1) elle n'est que de 0,02 *logit* pour ce groupe. Toutefois, cette différence de moyenne entre les groupes linguistiques n'est pas significativement différente. Lorsque l'on ne garde que les candidats de « niveau 2 » et les 16 tâches de « niveau 2 », on ne constate pas, non plus, de différence de moyenne entre les items des tâches intégrées et discrètes significatives (figure 5.2). En revanche, l'étude a montré que les moyennes du groupe des sinophones et du groupe des langues latines pour les items des tâches discrètes de « niveau 2 » sont significativement différentes (figure 5.3). Alors que les sinophones de « niveau 2 » sont aussi compétents que les autres groupes de candidats pour les items de « niveau 2 » des tâches intégrées, ils sont moins compétents que les candidats de langue latine lorsqu'ils répondent à aux items de « niveau 2 » des tâches discrètes.

En résumé, ces résultats semblent indiquer qu'il peut exister des interactions selon le type de tâche lorsque les candidats répondent à des questions correspondant à leur niveau ou encore pour des candidats se trouvant au milieu de l'échantillon des 113 individus ayant servi pour l'étude.

Pour ce qui est du classement des candidats de « niveau 2 », l'étude des corrélations de rang indique que le lien entre les rangs obtenus avec l'estimation de la compétence avec les 30 items et les 15 items des tâches discrètes est le plus fort, même si sa valeur est moyenne (tableau 5.12). La corrélation la plus basse est obtenue entre les 15 items des tâches intégrées et les 16 items des items de « niveau 2 ». La corrélation de rang entre les

15 items des tâches discrètes et les 15 items des tâches intégrées n'est pas significative. Toutefois, ces résultats, là encore, sont à interpréter avec prudence car il aurait fallu beaucoup plus d'items pour obtenir des corrélations facilement interprétables.

Un test de Wilcoxon fait entre l'estimation du niveau de compétence des candidats de « niveau 2 », nous apprend que les rangs obtenus à partir des 15 items des tâches discrètes et des 15 items des tâches intégrées diffèrent de manière significative pour l'ensemble des candidats (tableau 5.13). Pour ce qui est des groupes linguistiques, la moyenne des rangs est significativement différente lorsque la compétence est estimée avec les 15 items des tâches discrètes (tableau 5.15). En revanche, il n'y a pas de différences significatives entre les moyennes des groupes linguistiques lorsque la compétence est calculée avec les 15 items des tâches intégrées (tableau 5.15). Enfin, on ne trouve aucune différence significative pour la moyenne de la différence des scores (calculée à partir des 30 items du test) pour les groupes linguistiques composés de candidats de « niveau 2 » (tableau 5.15).

Pour les différents groupes linguistiques de « niveau 2 », la différence de classement s'explique donc par le type de tâche utilisé. La moyenne de la différence des scores est la même pour les groupes linguistiques de « niveau 2 ». Ces deux résultats semblent nous indiquer que la différence est bien due au type de tâche. La différence de difficulté entre les tâches discrètes et les tâches intégrées est constante pour les trois groupes linguistiques. On peut ajouter à cela, comme le signalent Jang et Roussos (2007 : 4), que Wainer et Wang (2001) ont trouvé que l'utilisation de minitest ne modifie pas l'estimation de la difficulté. En revanche, cela provoque une sous-estimation des paramètres de la discrimination, de la pseudo-chance et une sous-estimation de l'erreur de mesure de l'estimation de la compétence.

En résumé, la différence de moyenne des rangs pour les groupes linguistiques de « niveau 2 » trouvée avec les 15 items des tâches discrètes pourrait donc bien être liée à la différence du type de tâche (discrètes et intégrées) et non pas à l'utilisation de minitest.

Dans le contexte du testing adaptatif ce résultat est important. En effet, il montre qu'à niveau de compétence équivalent, les candidats n'obtiennent pas toujours les mêmes

résultats selon qu'on utilise des tâches discrètes ou intégrées. Ce résultat semble justifier le besoin d'utiliser une plus grande variété de tâches dans les tests en langue seconde.

6.3 La difficulté perçue par les candidats

La dernière question de recherche portait sur la perception subjective de la difficulté. La question était posée de savoir si la perception de la difficulté des tâches discrètes et intégrées était différente pour l'ensemble des candidats et comment elle était interprétée par chacun des groupes linguistiques.

Pour l'ensemble des candidats, les tâches discrètes sont perçues comme étant plus faciles que les tâches intégrées. Les textes sont perçus comme étant plus faciles que les questions. La perception de la difficulté semble traduire la perception des candidats quant à l'organisation du test. Pour les candidats, il est clair que le test propose deux séries de tâches (trois tâches discrètes et trois tâches intégrées). Pour eux, chacune des deux séries de trois textes est ordonnée par ordre croissant de difficulté.

Le classement des groupes linguistiques, selon que l'on prend en considération la difficulté objective ou subjective, n'est pas identique. Le groupe des langues latines qui est le groupe le plus compétent est celui qui a perçu le test comme étant le plus facile. Pour les autres groupes la correspondance n'est pas aussi évidente. Alors que le groupe des langues slaves est le deuxième groupe le plus compétent, ils ont une perception de la difficulté qui est du même niveau que celui des sinophones et qui, lui-même, est très proche de celui du groupe des multilingues (moyennes obtenues sans prendre en considération les cas extrêmes) (tableau 5.23 et tableau 5.24). Lorsque l'on analyse la moyenne du groupe des langues latines avec celle des autres groupes, on découvre que cette moyenne est significativement différente. Ils perçoivent le test comme ayant été beaucoup plus facile que les autres groupes. Il faut cependant faire attention à ce résultat, car cela ne signifie pas qu'ils pensent que le test était facile. La seule certitude est que pour eux, le degré de difficulté perçu était moindre que pour les autres groupes et qu'ils réussissent.

L'analyse de la difficulté perçue par chacun des groupes, pris isolément, indique que la perception n'est pas exactement la même. Cependant, il faut modérer ces interprétations

car les groupes linguistiques ne sont pas composés de beaucoup de candidats et les erreurs de mesure restent encore relativement importantes.

Il semble, cependant, qu'il soit possible de soutenir l'hypothèse selon laquelle l'appartenance à un groupe linguistique conditionne une perception de la difficulté. Ainsi pour le groupe des langues latines (qui regroupe environ la moitié des candidats), le lien entre la difficulté perçue des textes et la nature des tâches (discrètes et intégrées) n'est pas le même que pour l'ensemble des candidats. Pour le groupe des langues latines, le texte C est bien plus difficile que les textes E et D, même si ce dernier n'est pas issu des tâches intégrées. En revanche, pour ce groupe, les questions des tâches intégrées sont systématiquement plus difficiles que les questions des tâches discrètes. Pour eux, il y a donc un écart entre la difficulté perçue des textes et des questions des tâches discrètes et intégrées.

Pour ce qui est des candidats de « niveau 2 » (tableau 5.30), leurs résultats sont très proches de ceux de l'ensemble des candidats (tableau 5.25). Pour les candidats de « niveau 2 » comme pour l'ensemble des candidats, les textes et les questions des tâches discrètes sont plus faciles que les textes et les questions des tâches intégrées. Ce qui différencie les deux groupes, c'est le positionnement de la tâche E. Pour les candidats de « niveau 2 », le texte et les items de la tâche E sont perçus comme étant plus difficiles. Pourtant, il est très difficile d'en déduire que la difficulté perçue par les candidats de « niveau 2 » diffère de celle perçue par l'ensemble des candidats.

6.4 Synthèse et perspectives pour les tests adaptatifs sur ordinateur

Pour la conception des tests pour l'évaluation de la compétence de lecture en langue seconde, et plus spécifiquement pour les tests adaptatifs par ordinateur visant le positionnement de candidats dans des cours de langue, les résultats de l'étude de la dimensionnalité semblent indiquer qu'il est sans doute utile de prendre en considération l'existence de tâches discrètes et intégrées lors de la conception de test évaluant la compétence linguistique. Ces résultats nous indiquent que les tâches discrètes et intégrées ont bien un facteur général corrélé à des facteurs secondaires. Les tâches proposées dans

le test, évaluent la compétence générale mais aussi certains aspects particuliers de la lecture.

Les résultats obtenus à l'issue de la recherche montrent encore que, comme le signalent Bachman (2002), Carr (2006) mais aussi Nunan et Keobke (1995), la difficulté est le fruit de l'interaction entre les différentes caractéristiques de la tâche et du candidat. Sans doute que dans des recherches futures, il serait intéressant de vérifier le lien entre la difficulté et les deux types de tâches. Si, dans cette recherche, les difficultés objectives (mesurées) des tâches discrètes et intégrées sont très semblables lorsqu'elles sont calibrées ensemble, au jugé des résultats d'autres recherches (Jang et Roussous, 2007 ; Trites et Groarty, 2005 ; Wainer et Wang, 2001), il est possible de penser que les tâches intégrées ne sont pas, « par nature », plus difficiles que les tâches discrètes. La difficulté variant en fonction de plusieurs paramètres, il est légitime de s'attendre à ce que les tâches varient en fonction de la complexité linguistique, du but du lecteur, de l'activité à accomplir, du type de document à lire, des candidats et de la nature des liens entre les documents (textes liés ou indépendants).

D'un point de vue opérationnel, plus qu'une attention portée exclusivement à la difficulté, le concepteur de test devra se poser la question de l'utilité et du rôle des tâches discrètes ou intégrées pour la validité de contenu et la qualité de la mesure. On devra opérer des choix qui correspondent au construit du test.

Enfin, dans la mesure où le lien entre la perception de la difficulté des tâches, des questions et des items n'est pas représentatif de la réalité de la difficulté objective (pour les candidats la difficulté est liée à la nature des tâches), il est sans doute utile d'informer les candidats de l'organisation du test concernant la difficulté des items et des tâches. Ce conseil est d'autant plus important que certains groupes linguistiques ont une perception de la difficulté différente de celle de l'ensemble des candidats. Par ailleurs, il est probable que l'appartenance culturelle définisse leur perception de la difficulté. Le problème est, comme le disent Nunan et Keobke (1995), qu'on ne sait pas quelles sont les interactions entre les variables de la difficulté objective et de la difficulté subjective et, surtout, quel est l'impact de la difficulté subjective sur la difficulté objective.

Plus spécifiquement, pour les tests adaptatifs, dans le cadre d'une évaluation qui vise le positionnement de candidats dans des cours de langue, on peut penser qu'il serait utile d'évaluer le candidat à la fois avec des tâches discrètes et avec des tâches présentant un niveau d'intégration plus important. Cependant, il convient de définir le niveau d'intégration visée pour les tâches (du moins intégré au plus intégré : textes courts, textes longs, deux ou plusieurs textes courts liés les uns aux autres par le contenu, deux ou plusieurs textes longs liés les uns aux autres par le contenu, voire dans un contexte universitaire, des tâches demandant la compréhension de supports écrits et oraux pour vérifier un niveau de compréhension général en langue). Il faut encore décider d'un dispositif qui respecte un minimum de faisabilité. Autrement dit, il faut choisir un dispositif qui permette, à la fois, de proposer un contenu diversifié et une bonne qualité de mesure, sans pour autant augmenter le temps de conception du test ou encore de remise des résultats. Pour ce qui est de la mesure, il convient de définir un construit unidimensionnel (ce qui a des conséquences sur le type de tâche proposée, le type d'item, de textes), puis de vérifier la dimensionnalité à l'issue des résultats. Pour la compétence de lecture cela est d'autant plus important que la présence de multidimensionnalité faible à forte est mentionnée dans les résultats de la recherche. Et puis, pour un test de positionnement, il ne s'agit pas de faire un diagnostic précis des compétences des candidats mais de proposer des items avec un contenu beaucoup plus diversifié. Il faudra donc veiller à ne pas trop s'écarter de l'unidimensionnalité pour ne pas à avoir à utiliser des scores composites, et surtout, des scores dont l'interprétation est complexe. On peut penser qu'il serait sans doute utile, dans un test adaptatif, de proposer à la fois des items de lecture issus de tâches discrètes et des items issus de tâches intégrées pour répondre à aux contraintes de la qualité du contenu mais aussi à celles de la mesure. La variété des tâches est nécessaire pour faire des inférences valides sur la compétence en langue à partir des résultats au test et ce quelle que soit la qualité de la mesure. Cependant, cette même variété ne doit pas nuire à la faisabilité du test ou encore à sa mesure. Le défi consistera à pouvoir proposer un ensemble d'items au plus proche du niveau du candidat en sélectionnant à la fois des items discrets et des items intégrés. Les travaux menés sur l'assemblage automatisé (et surtout optimisé) des items dans le cadre des tests adaptatifs pourraient apporter des solutions à ce problème de faisabilité (Belov & Armstrong,

2008). Dans un test, on peut imaginer qu'à partir d'une estimation préalable du niveau de lecture (avec des items discrets ou encore à partir d'une autoévaluation du niveau de compétence), les candidats reçoivent des items intégrés (liés à des textes), puis, que leur niveau de compétence soit « affiné » avec des items issus de tâches discrètes. Les tâches discrètes parce qu'elles sont indépendantes sont plus faciles à utiliser quand on travaille avec une banque d'items ou des minitests. L'utilisation d'items issus de tâches discrètes et intégrées doit permettre de mieux répondre aux exigences de contenu sans pour autant négliger les aspects liés à la précision de la mesure ou encore à la faisabilité.

Evidemment, dans la mesure où les candidats pensent que les tâches intégrées sont plus difficiles que les tâches discrètes (tableaux 5.23 et 5.24), il faudra veiller à les informer des avantages et des inconvénients de l'utilisation des deux types de tâches. Par exemple, on peut leur expliquer que, dans le contexte d'un test adaptatif sur ordinateur, répondre à des questions sur des textes liés (même si cela n'est pas différent d'une tâche où l'on demande à un individu de répondre à plusieurs questions sur un texte moyen ou long) est sans doute à leur avantage. Dans une administration sur ordinateur, la possibilité de répondre à un ensemble de questions sur un même texte en ayant la possibilité de répondre aux questions dans un ordre qui convienne à chacun des candidats, peut être perçu comme un atout, notamment vis-à-vis de l'administration d'items discrets constitués d'un texte court et d'un seul item. En effet, souvent, lorsqu'on administre des items discrets sur ordinateur, les candidats se plaignent de ne pas pouvoir revenir en arrière pour modifier leurs réponses. Cela provoque chez eux un certain stress et peut engendrer de mauvaises réponses (Papanastasiou et Reckase, 2008).

Pour les tâches intégrées avec des textes dont le contenu est lié, on devra surtout présenter les avantages en termes d'authenticité. Il convient d'expliquer l'intérêt (utilisation de connaissances « antérieures » acquises au cours de la lecture pour comprendre de nouvelles informations) d'évaluer à la fois des compétences de lecture sur les textes courts, moyens ou longs reliés entre eux. Pour bien expliquer l'intérêt des tâches intégrées, en termes d'authenticité, il serait bon de rappeler aux candidats que si la découverte d'une langue (voire d'une culture) se fait par la découverte de documents différents, sa pratique quotidienne ou encore sa connaissance plus approfondie se fait par la capacité à faire des liens entre ces différentes informations. Cet aspect des choses est

très certainement à leur avantage. Il faut également leur expliquer que sans tâches intégrées, les résultats manqueraient très certainement de validité de contenu. On risquerait de mesurer des compétences en dehors de tout contexte d'utilisation de la langue. S'il n'est pas question de prendre en considération des variables contextuelles trop précises, il s'agit surtout de centrer la conception du test autour de deux aspects jugés comme essentiels pour le construit des tests de langue, soit l'intégration des composantes de la langue et le « contexte » d'utilisation (ici, la notion de « contexte » d'utilisation renvoie à un usage de langue regroupant une très large famille de situations).
Du côté des spécialistes en évaluation, proposer des minitests aux candidats, selon leur niveau de compétence, avec la possibilité de réviser uniquement les questions du seul minitest « en cours de passation » permettra de laisser l'opportunité aux candidats de réviser leurs réponses sans trop de stress. Par ailleurs, cela permettra d'éviter trop de problèmes de tricheries dus à un temps d'exposition des items trop important au cours de la passation. Cela peut être le cas lorsque, dans un test (adaptatif, mais pas uniquement), les candidats ont la possibilité de réviser toutes leurs réponses, et ce durant toute la durée de la passation (Wainer, 1993). Evaluer des items à partir de textes composés d'items évaluant le candidat autour de son niveau de compétence, évitera encore que les candidats mettent en place des stratégies destinées uniquement à utiliser les faiblesses de l'administration informatisée des items. En effet, parfois, les candidats, parce qu'ils se voient proposer des questions plus faciles que celles auxquelles ils ont déjà répondu, reviennent en arrière parce qu'ils comprennent alors qu'ils ont mal répondu (Kingsbury, 1996). Ici, en utilisant des textes accompagnés de plusieurs questions, les candidats ne connaîtront pas leurs résultats immédiatement. Les minitests étant composés de questions de niveau différent, on peut supposer que l'attention du candidat sera moins focalisée sur la difficulté des questions que sur sa capacité à répondre à l'ensemble des questions. Contrairement à un test linéaire, dans lequel le candidat identifie les questions les plus difficiles de part leur simple positionnement (le candidat est informé que les questions sont placées dans un ordre de difficulté croissant), on peut supposer que dans une telle configuration le candidat focalise moins sur la difficulté des questions (et le besoin de « gagner des points ») que sur le contenu des textes. En effet, l'information présentée dans les textes précédents permettra de mieux aborder les autres textes, même si

l'information demandée pour répondre aux questions portant sur un texte se situe toujours dans ce texte.

Pour ce qui de la conception des tâches intégrées, afin de bâtir une batterie de textes reliés les uns aux autres, on peut penser comme nous l'avons proposé dans cette recherche à utiliser des réactions de lecteurs à des articles, mais aussi à une information qui revient plusieurs fois dans la presse au cours de la semaine, ou bien à des documents liés entre eux par une thématique que l'on retrouve fréquemment dans la presse québécoise.

Pour respecter un minimum de faisabilité, il faudra être très vigilant sur le choix des documents et sur la nécessité de respecter le principe d'indépendance locale. Pour le choix des documents, à défaut de présenter des thèmes différents, on devra veiller à ce que les textes soient de nature différente (informatif, argumentatif, narratif) afin de respecter un minimum de variété pour la couverture de contenu. Pour la question de l'indépendance locale, il faudra veiller à ce que les questions ne provoquent pas d'effets d'apprentissage trop importants et que l'on puisse toujours répondre à une question sans avoir répondu aux autres.

6.5 Limites de la recherche

Tout d'abord, il convient de rappeler que le test qui a été créé n'est qu'un test en rodage. Pour obtenir de meilleurs résultats, il conviendrait de remplacer les questions qui fonctionnent le moins bien. On pourrait alors reproduire la recherche et vérifier si les résultats que l'on obtient sont cohérents avec ceux que l'on a obtenus ici.

Le nombre de candidats, s'il était suffisant pour un test en rodage, n'est pas suffisant pour obtenir des résultats dont on soit absolument certain. Il conviendrait, notamment, que les groupes linguistiques aient beaucoup plus de candidats et notamment plus de candidats de niveau intermédiaire. Avec plus de candidats dits de « niveau 2 », il serait alors plus facile de parvenir à des résultats possédant une erreur de mesure moins importante. Il serait également possible de réduire l'étendue du « niveau 2 » pour garder les candidats au plus proche du niveau des items. Pour ce qui est des items, le test étant composé de 30 items, il semble difficile de parvenir à obtenir autant, voire plus d'items, au plus près du niveau des candidats. En effet, l'usage d'items liés à un passage permet difficilement de

calibrer la difficulté des items pour avoir un maximum d'items dans un minimum d'étendue.

Outre ces deux premières limites, une autre limite importante tient au fait que cette étude est isolée. Pour que les résultats soient généralisables, il faudrait que l'on produise plusieurs jeux de tests pour vérifier si les résultats obtenus ne sont pas dus aux textes et aux items utilisés dans cette recherche en particulier. Certes, le cadre conceptuel et les avancées de la recherche en matière de tâches discrètes et intégrées nous donnent des indications (Jang & Roussos, 2007) mais pas vraiment de certitudes.

Pour ce qui est des items, il ne faut pas oublier que la position qui a été adoptée pour la recherche était de travailler à partir d'items dont les textes sont concaténés, mais dont les items restent relativement indépendants. Probablement qu'une recherche menée avec des textes reliés les uns aux autres et des items demandant exclusivement et de manière directe aux candidats de prendre de l'information dans plusieurs textes donnerait des résultats quelque peu différents.

Enfin, comme l'étude des variables sociodémographiques a pu le montrer, la population suivant les cours du MICC a un profil assez particulier. Dans la mesure où les politiques migratoires évoluent et les arrivées des immigrants peuvent changer le profil de la population, il est possible de se demander si les résultats trouvés ici seraient identiques avec une population d'immigrants quelque peu différente.

Pour ce qui est de la perception de la difficulté, il aurait été intéressant d'interroger les candidats au sujet de leur perception de la difficulté pour chacun des items. Sans cette information, il est difficile de savoir comment les candidats du même niveau que le niveau des items interprètent la difficulté et quelles sont les conséquences sur les résultats. Ceci dit, une telle entreprise n'était pas possible avec des candidats volontaires qui avaient déjà accordé une demi-journée de leur formation linguistique à la participation à la recherche.

Pour finir, le lien entre les trois questions de recherche peut interroger le chercheur. En effet, on ne sait pas encore jusqu'à quel point la proximité linguistique a un lien avec la dimensionnalité. Par ailleurs, le concept de « proximité linguistique » reste difficile à manier notamment dans les milieux multilingues. Par exemple, dans ces milieux, chez un même locuteur, la distance entre la langue maternelle et la langue cible peut être grande

alors que la distance entre une langue apprise (et maîtrisée) et la langue cible est plus étroite. Le lien entre la perception de la difficulté et la dimensionnalité est lui aussi difficile à faire et, ce, même si on situe les deux concepts sur des plans totalement opposés. On est en droit de se demander quels liens unissent la difficulté objective et subjective et jusqu'à quel point il y a bien des interactions entre les deux.

Conclusion

La recherche se donnait comme objectif d'analyser les dimensionnalités psychométrique et psychologique des tâches de compréhension écrite en français selon leur degré d'intégration. Pour l'analyse de la dimensionnalité psychométrique, les tâches ont été analysées, non seulement, du point de vue de leur dimensionnalité psychométrique, mais également, du point de vue des variations de leur difficulté. Pour étudier ces variations, ont été considérés l'appartenance à un groupe linguistique et le fonctionnement des items lorsque le niveau des candidats et des items sont proches. Enfin, la dimensionnalité psychologique des items (difficulté subjective telle que perçue par les candidats) a été étudiée. *In fine*, il s'agissait de tirer des conclusions sur la possibilité d'utiliser des tâches de lecture discrètes et intégrées dans les tests adaptatifs visant le positionnement dans des cours de langue, pour une population culturellement et linguistiquement hétérogène, à partir d'informations issues de l'étude des dimensionnalités psychométrique et psychologique.

S'il est prudent de rappeler que les résultats obtenus sont ceux d'un test en « rodage », ces résultats nous indiquent, toutefois, des pistes qui pourraient être étudiées dans le futur.

Tout d'abord, même si la consistance interne des tâches discrètes et intégrées est bonne, il semble bien que les concepteurs de test soient en droit de s'interroger sur le lien entre le type de tâche et la compétence linguistique. Si les variances des tâches discrètes et des tâches intégrées s'expliquent par un facteur général, la variance est cependant mieux expliquée, pour les tâches discrètes, lorsqu'on regroupe les items en fonction du texte sur lequel ils portent, et pour les tâches intégrées sur un facteur général plus un facteur spécifique regroupant l'ensemble des items. Que ce soit pour les tâches discrètes ou pour les tâches intégrées, ces solutions permettent non seulement d'expliquer plus de variance mais aussi d'obtenir une meilleure saturation sur le facteur général. La dimensionnalité des tâches discrètes utilisant des textes indépendants et des tâches intégrées utilisant des textes dépendants n'ayant pas la même structure, il conviendra, pour la validité de

contenu, de s'interroger sur la pertinence à proposer à la fois des tâches discrètes et intégrées dans un test ou uniquement des tâches discrètes ou intégrées. De ce choix découleront des décisions quant à l'inférence de la compétence des candidats, quant à la nécessité d'utiliser ou pas des scores composites.

Pour ce qui est de la difficulté objective pour des groupes linguistiques distincts et des items de niveaux différents, elle varie selon le type de tâche utilisé. Le classement des candidats de même niveau est différent selon que l'on calcule la compétence avec des tâches discrètes ou intégrées. Pour ce qui est des interactions entre le type de tâches et l'appartenance à un groupe linguistique avec des items et des candidats de niveau similaire, la recherche a montré des différences de fonctionnement pour les locuteurs des langues latines et les sinophones. Il semble légitime de penser qu'il reste utile de continuer les recherches dans ce sens.

En ce qui concerne la difficulté subjective, il est intéressant de constater que les candidats se livrent bien à une « interprétation » de la difficulté du test. Cette interprétation diffère selon l'appartenance à un groupe linguistique. S'il est clair que la difficulté subjective n'a pas de lien direct avec la difficulté objective, pour les candidats, l'ensemble des groupes linguistiques, les tâches discrètes sont plus faciles que les tâches intégrées. Parce que dans le test, les difficultés objectives des tâches discrètes et des tâches intégrées étaient comparables, on peut supposer que les tâches intégrées risquent d'être moins bien perçues par les candidats que les tâches discrètes. Concevoir des tests avec des tâches intégrées demandera donc de dissocier ce qui est de l'ordre de la complexité perçue et ce qui est de l'ordre de la difficulté objective.

Pour ce qui est des tests adaptatifs dans une perspective de positionnement, l'utilisation de tâches discrètes et intégrées permettra d'améliorer la validité de contenu. Pour ce qui est de la mesure, pour les deux types de tâche, si la saturation sur le facteur général n'est pas très importante, les facteurs secondaires ne sont pas assez importants pour considérer que les tâches ne sont pas unidimensionnelles. Il semble donc qu'on puisse les utiliser dans le cadre du positionnement d'apprenants dans des cours de francisation. Même si les

résultats demanderaient à être confirmés, il est intéressant de constater que les items de tâches intégrées, lorsqu'ils sont calibrés seuls, présentent une meilleure saturation sur le facteur général que les items des tâches discrètes.

Pour la définition du construit d'un test de positionnement adaptatif de compréhension écrite, il conviendra d'utiliser à la fois des tâches discrètes et intégrées, et ce, pour plusieurs raisons. Tout d'abord, lorsque les tâches intégrées et discrètes sont utilisées conjointement pour le calcul d'un seul score, la fidélité du test et sa dimensionnalité sont suffisantes pour assurer la qualité de la mesure. La dimensionnalité des tâches discrètes et intégrées n'étant pas strictement équivalente, l'utilisation de tâches discrètes et intégrées permettra une meilleure couverture de contenu. Enfin, l'utilisation de tâches de nature différente permettra d'équilibrer les interactions entre le type de tâches et le type de candidat. Ces aspects doivent impérativement être pris en considération dès la rédaction du construit du test. En effet, l'utilisation de différents types de tâches permet non seulement une meilleure couverture de contenu mais encore d'éviter de défavoriser systématiquement les candidats d'un même groupe linguistique et donc, *in fine*, un meilleur respect de la diversité culturelle des candidats. Trop souvent, pour des raisons liées à la faisabilité des tests, mais aussi à d'autres facteurs, les concepteurs de test font le choix d'utiliser un seul type d'item (souvent des items discrets). Ils vérifient alors le fonctionnement différenciel de ces items selon l'appartenance à des groupes linguistiques. N'utilisant que des items discrets dans leur test, il est alors impossible de vérifier les interactions dues au type d'item. Cela ne signifie pas pour autant que ces interactions n'existent pas et qu'il ne faille pas s'en préoccuper. La diversification du type de tâche sera d'autant plus nécessaire que la difficulté perçue du type de tâche par les groupes de candidats ne reflète pas nécessairement leurs niveaux de compétence. Ainsi, dans la présente recherche, les sinophones de niveau intermédiaire ne sont pas conscients du gain qu'ils obtiennent lorsqu'ils répondent aux items des tâches intégrées.

La recherche, parce qu'elle a été faite à partir de données réelles (d'un test qu'il aurait été difficile, voire impossible, de trouver parmi les outils déjà existants et d'un échantillon de candidats correspondant à la population ciblée) permet d'obtenir des connaissances sur un type de tâche qui est très peu utilisé dans les tests adaptatifs sur ordinateur ayant des

objectifs de placement. Les résultats sont encourageants et laissent à penser que de nouvelles solutions pourraient être trouvées pour répondre aux contraintes, parfois en contradiction, de qualité de la mesure, de validité de contenu et de faisabilité.

Annexes

Annexe 1

Test de Compréhension écrite
Durée 1H15 Maximum

Livret du candidat

2006

Recherche menée par Vincent Folny
dans le cadre de la Maîtrise en mesure et évaluation en éducation de l'Université de Montréal

Fonctionnement de tâches discrètes et intégrées pour l'évaluation de la lecture en français langue seconde des nouveaux arrivants au Québec.

A remplir avant de commencer l'examen

Feuille de renseignements sociodémographiques

Cochez la ou les casses (☑) correspondant à votre profil ou écrivez quand c'est nécessaire.

RENSEIGNEMENTS PERSONNELS

Prénom :_____

Âge : _____**ans.**

Sexe : ☐ Homme ☐ Femme

NIVEAU D'ÉTUDE

☐ Secondaire (10 ou 11 années d'école)

☐ Collégial (12 ou 13 années de scolarité, niveau CEGEP au Québec)

☐ Premier cycle universitaire (15 ou 16 années de scolarité, niveau baccalauréat au Québec)

☐ Deuxième cycle universitaire (17 ou 18 années de scolarité, niveau maîtrise, DESS,… au Québec)

☐ Troisième cycle universitaire (20 ou 21 années de scolarité, niveau doctorat au Québec)

LANGUE(S)

Langue maternelle : _____

Langue(s) parlée(s) dans la rue à Montréal : ☐ Français ☐ Anglais ☐ Langue maternelle ☐ Autre

Langue(s) parlée(s) au travail : ☐ Français ☐ Anglais ☐ Langue maternelle ☐ Autre

Langue(s) parlée(s) à la maison : ☐ Français ☐ Anglais ☐ Langue maternelle ☐ Autre

Ancienneté au Québec
Date d'arrivée au Québec : _____

Courriel (facultatif) :_____**@**_____

Feuille de Réponses

Espace Réservé au Chercheur :
Numéro : ……..

Document A

	A	B	C	D
Exemple	☑	☐	☐	☐
Question 1	☐	☐	☐	☐
Question 2	☐	☐	☐	☐
Question 3	☐	☐	☐	☐
Question 4	☐	☐	☐	☐
Question 5	☐	☐	☐	☐

Document B

	A	B	C	D
Question 1	☐	☐	☐	☐
Question 2	☐	☐	☐	☐
Question 3	☐	☐	☐	☐
Question 4	☐	☐	☐	☐
Question 5	☐	☐	☐	☐

Document C

	A	B	C	D
Question 1	☐	☐	☐	☐
Question 2	☐	☐	☐	☐
Question 3	☐	☐	☐	☐
Question 4	☐	☐	☐	☐
Question 5	☐	☐	☐	☐

Document D

	A	B	C	D
Question 1	☐	☐	☐	☐
Question 2	☐	☐	☐	☐
Question 3	☐	☐	☐	☐
Question 4	☐	☐	☐	☐
Question 5	☐	☐	☐	☐

Document E

	A	B	C	D
Question 1	☐	☐	☐	☐
Question 2	☐	☐	☐	☐
Question 3	☐	☐	☐	☐
Question 4	☐	☐	☐	☐
Question 5	☐	☐	☐	☐

Document F

	A	B	C	D
Question 1	☐	☐	☐	☐
Question 2	☐	☐	☐	☐
Question 3	☐	☐	☐	☐
Question 4	☐	☐	☐	☐
Question 5	☐	☐	☐	☐

Réussir ses plats mijotés du temps des Fêtes

Jacinthe Côté
Choisir son mode de cuisson

Quand vient le temps de faire son choix de mode de cuisson, différents impératifs s'imposent.

Il est important de s'assurer que la température de la cuisson sera suffisamment élevée pour obtenir une belle couleur dorée mais aussi des odeurs et des saveurs complexes. Pour cela, la température de cuisson doit atteindre au moins 154 degrés C (310 F). Cette température permet aussi d'attendrir la chair. La température utilisée pour la cuisson des viandes et de la volaille doit également être suffisamment élevée pour détruire les microorganismes dangereux.

Les bienfaits de la cuisson lente

De manière générale, il était préférable de cuire les viandes et la volaille à des températures basses, maintenues constantes, pendant plus longtemps. On améliore ainsi la texture et les saveurs, et on réduit la perte d'eau et de nutriments. S'il y avait une température universelle de cuisson pour faire rôtir les viandes et la volaille, ce devrait être 165 degrés C (325 F).

Des accompagnements santé et gourmands

En cours de la cuisson des rôtis de viande ou de volaille, il est de coutume d'ajouter des fruits frais ou secs tels que des figues, des canneberges, des pommes, des pruneaux, etc. On peut également servir les rôtis avec des fruits préparés en sauce ou en compote.

Ces fruits, en plus d'apporter d'agréables saveurs acidulées et sucrées, contribuent à bonifier la qualité nutritive du plat final.

Bonne cuisine du temps des Fêtes!

Source :
http://www.cyberpresse.ca/article/20051230/CPSPECIAL02/512300426/5344/CPSPECIAL02

Article modifié et adapté à partir de l'original.

Document A

Selon les informations contenues dans le texte, choisissez la bonne réponse (*A, B, C, D*) et reportez-la sur votre feuille de réponse (*<u>une seule</u> *bonne réponse par question).

Question 1
☞***Quel type de conseils donne-t-on dans cet article ?***
A- Des conseils hygiéniques.
B- Des conseils médicaux.
C- Des conseils nutritionnels.
D- Des conseils préventifs.

Question 2
☞***Selon l'auteure, comment doit-on cuire les viandes et les volailles ?***
A- En baissant continuellement la température.
B- En faisant varier la température.
C- En utilisant des températures différentes.
D- En utilisant toujours la même température.

Question 3
☞***Quel résultat obtient-on si on cuit la viande et les volailles à 165 degrés C ?***
A- Une chair aussi nutritive que la chair non cuite.
B- Une chair pauvre en saveurs complexes.
C- Une chair qui a gardé toute son eau.
D- Une chair saine, sans microbes.

Question 4
☞***Selon l'auteure, qu'améliore-t-on quand on accompagne les rôtis de figues, canneberge, pommes, pruneaux, etc. ?***
A- La quantité à manger.
B- La texture de la chair.
C- Le goût de la préparation.
D- Le temps de cuisson.

Question 5
☞***Qu'est-ce que pourrait dire Jacinthe Côté de la cuisine à Noël ?***
A- « A Noël, il faut innover et proposer des plats originaux ».
B- « L'important c'est le résultat, pas le temps passé à cuisiner ».
C- « Un plat qui sent bon, c'est un plat idéal ».
D- « Peu importe l'esthétique, l'essentiel c'est la quantité ».

Document B

Vendredi 13 janvier 2006, p. A18

LA PRESSE

Violent incendie à Valleyfield

Meunier, Hugo

Un violent incendie a occupé les pompiers pendant plus de six heures, hier, dans un immeuble à logement de Valleyfield. Dans des circonstances encore nébuleuses, le sinistre a pris naissance en fin de nuit, autour de 5 h, hier, rue Centenaire. Un résidant de l'immeuble a appelé le 911, après avoir vu de la fumée épaisse. Une fois sur les lieux, les pompiers ont déployé une grande échelle et commencèrent à éteindre le feu. À ce moment, une femme en détresse au troisième étage a basculé dans le vide. " Elle semblait désorientée et, malheureusement, a brisé la fenêtre. Elle a perdu connaissance et est alors tombée sur le sol", a rapporté l'agent Jayson Gauthier, de la Sûreté du Québec. La femme a subi des brûlures sur 20 % de son corps. On craint pour sa vie.

Alors que les activités des pompiers se poursuivaient à l'avant du bâtiment, l'agente de police Nicole Champagne et sa consœur Sandra Morin ont pris l'initiative d'entrer dans l'immeuble par un autre accès. Deux personnes âgées, un homme et une femme, se trouvaient alors toujours dans leur logement respectif !

Rapidement, l'agente Champagne aide la vieille dame encore en jaquette à s'habiller, pour ensuite la transporter vers l'extérieur et fuir les flammes. Dans la cage d'escalier, elle retrouve l'homme âgé, étendu inerte sur le sol.

Il "était évanoui à cause d'une crise cardiaque", a indiqué Mme Champagne. La policière décide de poser la femme par terre, pour tirer l'homme à l'extérieur. Elle demande alors l'aide d'un pompier pour sortir la femme de l'immeuble. L'homme âgé était toujours inconscient au moment de mettre sous presse, alors que sa voisine a été incommodée par la fumée.

Article modifié et adapté à partir de l'original.

☞Selon les informations contenues dans le texte, choisissez la bonne réponse (A, B, C, D) et reportez-la sur votre feuille de réponse (une seule bonne réponse par question).

Question 1

☞De quoi l'auteur de l'article cherche-t-il à informer le lecteur ?

A- De la vie en province au Québec.
B- Des caractéristiques du métier de pompier.
C- D'un événement aux conséquences graves.
D- Du problème des incendies dans les immeubles.

Question 2

☞Quel est le bilan de l'incendie ?

A- Des blessés légers.
B- Deux personnes dans un état grave.
C- Un mort et un blessé grave.
D- Plusieurs brûlés.

Question 3

☞Pourquoi la femme au troisième étage a brisé la fenêtre ?

A- Elle voulait trouver de l'aide.
B- Elle pensait se suicider.
C- Elle tentait de s'échapper.
D- Elle ignorait ce qu'elle faisait.

Question 4

☞Que faisaient les pompiers quand Nicole Champagne et Sandra Morin sont entrées dans l'immeuble ?

A- Ils arrivaient devant le bâtiment.
B- Ils avaient déjà éteint le feu.
C- Ils continuaient à éteindre le feu.
D- Ils se préparaient à éteindre le feu.

Question 5

☞Que peut-on penser du travail des pompiers dans l'incendie à Valleyfield ?

A- Ils ont commis des fautes graves, presque criminelles.
B- Ils ont été incapables de maîtriser tous les événements.
C- Ils ont fait un travail d'amateurs, peu professionnel.
D- Ils ont tardé à commencer à éteindre le feu.

Document C

Texte modifié à partir d'un article de « Le Soleil »
Actualités, dimanche 18 décembre 2005, p. A9

La frénésie s'empare des centres commerciaux

Mais les commerçants s'attendent à ce que les plus grosses journées soient jeudi et vendredi.
Normandin, Pierre-André

Alors que Noël approche, la frénésie du magasinage de dernière minute s'empare des centres commerciaux. Et cette année encore, on peut constater que plusieurs attendront jusqu'au tout dernier moment pour terminer leurs achats du temps des Fêtes. Une visite du SOLEIL dans un centre commercial a permis de constater que la frénésie de Noël est bien commencée, mais qu'elle n'a pas encore atteint son sommet. "Généralement, les plus grosses journées de l'année sont plutôt le jeudi et le vendredi avant Noël", indique la copropriétaire de la boutique Kettö, Catherine Fafard.
[…].
"Les gens ne dépensent plus autant", dit Serge Fortin. Selon lui, l'achalandage n'est pas à la baisse, mais les gens planifient des sommes moins importantes pour leurs achats de Noël. "Côté volume, c'est la même chose. Mais on fait moins d'argent à la fin de la journée pour le même travail."
Le commerçant prend pour exemple les masques africains qu'il vend. "On demande toujours aux gens quel est leur budget. Avant, ils allaient jusqu'à 50 $, alors que maintenant c'est 30 ou 40 $", poursuit M. Fortin.
Selon un sondage du Conseil québécois du commerce de détail réalisé en octobre, deux fois plus de ménages comptaient réduire leurs dépenses que ceux qui prévoyaient les augmenter. Reste que la majorité des sondés souhaitaient le même budget.
Yvan Potvin et Raymonde Munger font partie de ces ménages qui ont décidé de dépenser moins en cadeaux pour ce Noël.
"J'ai dit à mes enfants que c'était fini les gros cadeaux", dit M. Potvin, qui invoque des raisons économiques. L'âge de ses enfants explique surtout qu'il ne se sente plus obligé d'en faire autant.
Reste que le couple, a pris trois jours pour terminer ses achats de Noël. LE SOLEIL les a rencontrés alors qu'ils prenaient une pause après une journée d'emplettes bien remplie. "Elle ne veut pas que je porte mon cadeau même s'il est plus lourd. Elle a peur que je devine c'est quoi", dit M. Potvin, avec un regard de complicité à sa conjointe.
PANormandin@lesoleil.com

Article modifié et adapté à partir de l'original.

☞***Selon les informations contenues dans le texte, choisissez la bonne réponse (*A, B, C, D*) et reportez-la sur votre feuille de réponse (*<u>une seule</u> *bonne réponse par question).***

Question 1
☞ ***D'après cet article, qu'est-ce que Noël représente au Québec ?***
A- Une fête destinée aux petits et grands enfants.
B- Une fête maintenue vivante par les commerçants.
C- Une tradition familiale toujours appréciée.
D- Un moment de stress et de désespoir financier.

Question 2
☞***Cette année, à quel(s) moment(s) les Québécois achèteront-ils le plus de cadeaux ?***
A- Tous les jeudis et vendredis de décembre.
B- La semaine avant Noël.
C- Le jour de Noël.
D- Quelques minutes avant Noël.

Question 3
☞***D'après cet article, quelle est la situation des commerçants au Québec ?***
A- Ils font face à des problèmes entre les générations.
B- Ils doivent répondre à une nouvelle demande ethnique.
C- Ils ont de graves problèmes économiques.
D- Ils vivent un changement des habitudes de consommation.

Question 4
☞ ***D'après cet article, quelle description correspond le mieux au profil des consommateurs québécois à Noël ?***
A- Ils achètent les cadeaux qu'ils veulent.
B- Ils détestent dépenser leur argent.
C- Ils font attention à leur argent.
D- Ils négocient l'achat de leurs cadeaux.

Question 5
☞ ***D'après cet article, qu'est-ce que les Québécois sont habitués de faire à Noël ?***
A- Dépenser plus pour les petits que pour les grands enfants.
B- Faire leurs achats de Noël en secret, sans les autres.
C- Partir en voyage pendant la période des fêtes.
D- Supprimer les cadeaux pour les enfants plus âgés.

17 novembre 2005 **Jeunes musiciens du monde**

Refaire le monde

Frédéric Denoncourt

Cette année encore, les spectateurs assisteront à un spectacle avec des artistes québécois de renom avec les Ariane Moffatt, Daniel Bélanger, Polémil Bazar, Colectivo et les artistes indiens **Jagjit Singh** et **Gurpreet Chana**. Des participants du Cirque du monde (programme social du Cirque du Soleil) offriront aussi une prestation. Les soirées seront animées par **Monique Giroux** de Radio-Canada et la journaliste **Karina Marceau**. De 20 h 30 à 3 h, tout comme l'an dernier, les artistes se relaieront sur scène. Ils s'uniront pour célébrer le pouvoir rassembleur de la musique par-delà les différences ethniques ou culturelles.

UN PROJET ANCRÉ DANS LE MILIEU ARTISTIQUE

Jeunes musiciens du monde (JMM) est un organisme de bienfaisance. Il œuvre à la fondation d'écoles de musiques traditionnelles gratuites pour les jeunes des quartiers populaires du Québec et d'ailleurs. Il mise sur la magie de la musique pour aider les jeunes à grandir. Après avoir fondé une école en Inde, pays où ils résident pendant la plus grande partie de l'année, les frères Mathieu et Blaise Fortier, chez eux, à Québec, ouvrent l'École des musiques traditionnelles de Saint-Sauveur. Motivés par les succès obtenus et l'enthousiasme suscité, les frères Fortier décident de poursuivre l'aventure familiale avec l'École de musique traditionnelle La Bolduc, qui vient à peine d'ouvrir ses portes dans le quartier Hochelaga-Maisonneuve à Montréal.

Soirée-bénéfice : le 25 novembre Au Métropolis. *Billets en prévente au coût de 20 $ (étudiants) et 25 $ (général). À la porte: 25 $ (étudiants) et 30 $ (général). La soirée débute à 20 h par un cocktail dînatoire incluant canapés et vins de qualité.*

Article modifié et adapté à partir de l'original.

Source :
http://www.voir.ca/actualite/actualite.aspx?iIDArticle=39108

réactions des membres | **Réagissez à ce texte !** Lisez les réactions des membres [17]

Document D

☞***Selon les informations contenues dans le texte, choisissez la bonne réponse (*A, B, C, D*) et reportez-la sur votre feuille de réponse* <u>une seule</u> *bonne réponse par question).***

Question 1

☞***Quel est l'objectif principal de Jeunes musiciens du monde ?***

A- Aider les jeunes en difficulté.
B- Augmenter le taux de scolarisation.
C- Lutter contre le racisme.
D- Organiser des spectacles.

Question 2

☞***Dans quel ordre aura lieu la soirée proposée par JMM ?***

A- D'abord des concerts puis une dégustation de vin.
B- D'abord un repas puis une dégustation de vin.
C- D'abord un dîner puis un spectacle.
D- D'abord un spectacle puis des concerts.

Question 3

☞***Où et quand les frères Fortier résident-ils ?***

A- Dans deux pays, en alternance.
B- Dans deux pays jusqu'à la fin de l'année.
C- Dans un seul pays, depuis peu de temps.
D- Dans plusieurs pays, depuis toujours.

Question 4

☞***Quel type de réaction suscite le projet JMM au Québec ?***

A- L'adhésion de ceux qui ont découvert JMM.
B- L'énervement, on attend l'ouverture d'autres écoles.
C- Une augmentation du nombre de soirées bénéfices.
D- Une nouvelle mode de la musique traditionnelle.

Question 5

☞***Quel est le public majoritairement attendu pour le spectacle de JMM ?***

A- Le public appréciant les artistes québécois peu connus.
B- Le public aimant les activités artistiques et culturelles.
C- Le public des professionnels de la culture.
D- Le public des enfants des quartiers populaires.

Document E

réactions des membres

Être là c'est bien, écouter c'est mieux

Ma compagne et moi étions présents au Métropolis. L'ambiance était chaleureuse et bon enfant. Alors que les gens mangeaient déjà, on pouvait entendre en musique de fond de la musique québécoise. Un bref discours de Monique Giroux accueillit le premier groupe sur scène. Il y avait environ deux cents personnes. L'enthousiasme pour une belle cause semblait partagé par les artistes et les spectateurs. Partagé ?... Hum, peut-être pas. Une rumeur presque constante montait de l'immense salle. Les gens parlaient entre eux. Pas tout bas. Assez fort pour déranger leurs voisins et même les artistes sur scène. J'aurais souhaité que Monique Giroux leur explique que la musique naît du silence, mais comment lui reprocher de ne pas jouer au préfet de discipline? Passée la prestation des artistes indous, près de nous, une dame racontait à sa voisine des détails de son quotidien... Pourquoi ne pas quitter la salle s'il est si urgent de parler ? Un peu plus tard, au cours de la prestation de Monica Freire, on a commencé à diffuser simultanément des images du documentaire sur JMM tourné en Inde, ce qui contribuait à la distraction généralisée. A la vérité de belles images bien que le bruit des uns et des autres nous empêchait de les apprécier à leur juste valeur. En quittant la salle, deux couples très bavards racontaient qu'ils retourneraient très certainement en Inde. On est contents d'avoir participé à l'événement. Mais on est déçus par le manque de respect de nombreux spectateurs. Est-ce que d'autres ont connu ce genre d'expérience ? Manifestez-vous, s'il vous plaît!

Article modifié et adapté à partir de l'original.

Source : http://www.voir.ca/actualite/actualite.aspx?iIDArticle=39108

Yan Steven	{25 votes}	**18 novembre 2005**

☞***Selon les informations que <u>vous avez déjà lues dans ce test et celles contenues dans ce document</u>, choisissez la bonne réponse (*A, B, C, D*) et reportez-la sur votre feuille de réponse (<u>une seule</u> bonne réponse par question).***

Question 1

☞***Qu'est-ce que cette réaction dénonce ?***

A- L'attitude du public pendant le spectacle.
B- L'organisation désastreuse du spectacle.
C- Une manifestation contre Monique Giroux.
D- Un scandale dont est responsable JMM.

Question 2

☞***Dans le texte, qu'est-ce qu'une « rumeur » ?***

A- Le bruit fait par les spectateurs.
B- Le mécontentement des spectateurs.
C- L'opinion de l'ensemble des spectateurs.
D- Une fausse nouvelle.

Question 3

☞***A la connaissance de Yan Steven, combien de personnes sont déçues par le spectacle ?***

A- Une personne.
B- Deux personnes.
C- Plus de deux personnes.
D- Plusieurs groupes de personnes.

Question 4

☞***Qu'est-ce que Yan Steven déteste dans les événements culturels ?***

A- La présentation des artistes par des animateurs.
B- La projection d'images sur de la musique.
C- Les demandes de silence adressées aux spectateurs.
D- Les manifestations de manque d'intérêt du public.

Question 5

☞***Qu'est-ce que Yan Steven a pensé du spectacle ?***

A- Il coûtait trop cher**.**
B- Il était difficile d'en évaluer la qualité.
C- Il pouvait être amélioré.
D- Il représentait une bonne distraction.

réactions des membres

Quel projet merveilleux! remède à la délinquance juvénile ?

Je ne suis pas allé au show de Montréal il y a 1 semaine. Je serai au show vendredi prochain. Peut-être que l'attitude du public sera différente de celle qu'il a eue lors du dernier spectacle, on verra bien. Ceci dit, parmi ceux qui ont écrit dans ce forum de discussion, pas grand monde semble penser comme Yan Steven. Visiblement, le jugement des uns n'est pas le jugement des autres !

Le projet est magnifique, surtout pour les plus démunis qui, souvent, n'ont pas d'espoir d'aller à l'école longtemps, qui ont l'impression de ne pas pouvoir entrer dans le moule. Remarquez une chose, les enfants et adolescents qui ont une passion, un intérêt marqué pour quelque chose de concret et créateur, seront moins délinquants que ceux qui n'ont rien à quoi se raccrocher. Par ailleurs, c'est tout un défi pour ces jeunes de faire de la musique dite traditionnelle quand la mode veut du moderne.

Donc des projets comme celui-ci qui permet de donner une chance à tout le monde de créer, de se réaliser, c'est fantastique. Les enfants ont besoin d'un toit sur la tête, de nourriture et de soins médicaux, cela nous en sommes tous conscients. Mais ils ont besoin aussi d'un peu plus. La culture, c'est important. Les jeunes ont des talents qui ne demandent qu'à être développés car sinon ces talents, loin d'être perdus, serviront plutôt à faire de « mauvaises choses ».

Je comprends que les fondateurs de cette belle initiative œuvrent comme ils le font ailleurs, c'est louable, mais il faut penser aussi à notre jeunesse qui, elle aussi, veut pouvoir s'exprimer par la musique. De plus, notre jeunesse, qui vient souvent de quartiers moins favorisés, peut profiter d'un camp d'été où en plus du plein air elle peut mettre en pratique ce qu'elle apprend et surtout vivre de belles expériences.

Ah ! Si d'autres frères, d'autres passionnés, pouvaient suivre leur idée et ouvrir ici et là-bas d'autres écoles. Espérons que l'attitude du public n'empêchera pas à l'avenir la tenue de spectacle-bénéfice. Ici, l'argent a toujours été plus difficile à trouver pour les cours de musique ou de sport, ou de dessin, que pour les maths et le français.

Article modifié et adapté à partir de l'original.

Source :
http://www.voir.ca/actualite/actualite.aspx?iIDArticle=39108

Michel Noël	{16 votes}	**25 novembre 2005**

Document F

☞***Selon les informations* que vous avez déjà lues dans ce test et celles contenues dans ce document*, choisissez la bonne réponse (*A, B, C, D*) et reportez-la sur votre feuille de réponse (*une seule* bonne réponse par question).***

Question 1

☞***Qu'est-ce que Michel Noël pense de l'attitude du public lors du dernier spectacle des JJM ?***

A- Elle était normale, rien d'extraordinaire.
B- Il est difficile de se faire une opinion.
C- Le public a été très désagréable.
D- Seule une partie du public a posé problème.

Question 2

☞***Selon Michel Noël, que propose JMM aux élèves de ses écoles ?***

A- Améliorer leurs conditions de vie quotidienne.
B- Leur fournir une éducation traditionaliste.
C- Leur permettre de développer leur potentiel.
D- Travailler leurs habiletés relationnelles.

Question 3

☞***Pour Michel Noël, à quoi peut servir la culture ?***

A- À améliorer les performances scolaires.
B- À éviter que les enfants soient sans talents.
C- À lutter contre les inégalités économiques.
D- À valoriser les talents des jeunes démunis.

Question 4

☞***Selon Michel Noël, au Québec, pour l'éducation des enfants, dans quoi investit-on le plus ?***

A- L'expression artistique.
B- L'éducation morale.
C- Les activités sportives.
D- Les matières scolaires.

Question 5

☞***Que pense Michel Noël ?***

A- Les religieux devraient ouvrir des écoles.
B- Il est nécessaire de changer le monde.
C- On peut progresser dans la vie.
D- Un Québécois doit travailler au Québec.

A remplir après avoir fini l'examen

Annexe 1

Évaluation du Niveau de difficulté du test

☞ ***Après avoir terminé le test**, dites, selon vous, quel est le **niveau de difficulté** du texte et des questions associés à chaque document, puis de l'ensemble des documents* **A, B, C** *et des documents* **D, E, F.** *Le chiffre* 1 *correspond à « très facile » et* 5 *à « très difficile ».*

1 **2** **3** **4** **5**

1= très facile
2= facile
3= difficulté moyenne
4= difficile
5= très difficile

Document A

	1	2	3	4	5
Exemple	☐	☑	☐	☐	☐
Texte	☐	☐	☐	☐	☐
Questions	☐	☐	☐	☐	☐

Document B

	1	2	3	4	5
Texte	☐	☐	☐	☐	☐
Questions	☐	☐	☐	☐	☐

Document C

	1	2	3	4	5
Texte	☐	☐	☐	☐	☐
Questions	☐	☐	☐	☐	☐

Document D

	1	2	3	4	5
Exemple		☑	☐	☐	☐
Texte	☐	☐	☐	☐	☐
Questions	☐	☐	☐	☐	☐

Document E

	1	2	3	4	5
Texte	☐	☐	☐	☐	☐
Questions	☐	☐	☐	☐	☐

Document F

	1	2	3	4	5
Texte	☐	☐	☐	☐	☐
Questions	☐	☐	☐	☐	☐

	1	2	3	4	5
Difficulté moyenne des documents A, B, C	☐	☐	☐	☐	☐
Difficulté moyenne des documents D, E, F	☐	☐	☐	☐	☐

Annexe 2

Test de compréhension écrite- Document professeur

Description du test

Matériel

- 6 textes pour un total de 30 questions.
- 1 feuille de renseignements personnels -à remplir avant la passation- accompagnée d'une feuille de report des réponses et d'évaluation de la difficulté du test.
- 1 jeu de 6 textes avec des questions à choix multiples.

Durée totale : 1H15 (15 minutes certificat éthique + 1H00 test)

- 15 minutes pour la signature du certificat d'éthique.
- 1H00 Questionnaire de renseignements personnels + passation du test + évaluation de la difficulté des tâches.

Déroulement de la passation

1) Explication aux candidats du principe de la passation.

- Expliquer aux candidats qu'ils vont passer un test de compréhension écrite d'une durée d'une heure dans le cadre de la recherche d'un étudiant de maîtrise en mesure et évaluation en éducation de l'Université de Montréal.

Les candidats devront :

- Remplir un formulaire de consentement éthique (avant la passation, obligatoire).
- Remplir un questionnaire de renseignements personnels.
- Répondre à des questions de compréhension écrite.
- Remplir un questionnaire sur leur perception de la difficulté du test.

- But de la recherche : « cette recherche a pour objectif de comprendre comment les résultats de l'évaluation de la lecture en français langue seconde varient en fonction du type d'activité utilisé pour la population des immigrants au Québec ».

Pendant la passation, aucun autre matériel que le stylo ne sera autorisé et le candidat ne pourra pas communiquer avec d'autres personnes que le professeur. Ce professeur ne répondra qu'à des questions pourtant sur le déroulement de la passation.

2) Certificat d'éthique.

-Distribuer un certificat d'éthique à chacun des candidats.

-Expliquer, à l'oral, les points principaux du certificat d'éthique :

- La participation des candidats est volontaire. S'ils en éprouvent le besoin, ils peuvent arrêter le test quand ils le veulent (droit de retrait).
- Les renseignements personnels qu'ils fournissent au chercheur seront tenus confidentiels. Les informations personnelles utilisées pour la recherche resteront anonymes.

- Inconvénient à participer à la recherche :

« Répondre à des questions trop faciles ou trop difficiles. »

- Avantages à participer à la recherche :

« Participer à « l'amélioration » des services de francisation offerts aux immigrants, notamment à l'évaluation de leur niveau en français. »

« Faire évaluer son niveau de compréhension écrite en français. Si vous le souhaitez, vous pourrez donner votre adresse courriel (cf. certificat d'éthique) pour que le résultat au test vous soit communiqué. Si vous n'êtes pas intéressé par cette option, n'écrivez pas votre adresse courriel sur le formulaire qui vous sera remis. »

- Demander aux candidats s'ils ont des questions avant qu'ils ne signent le certificat.

3) Avant la distribution du livret d'examen.

- Expliquer aux candidats (à l'aide d'un livret d'examen) qu'ils devront tout d'abord remplir la feuille avec les renseignements personnels. Les réponses devront être écrites sur la feuille de réponse et non à côté du texte.
- Après avoir répondu aux questions –et uniquement après- ils devront remplir la feuille concernant la difficulté du test (en fin de livret).

4) Distribution des livrets d'examen : textes et questions, feuille de renseignements personnels, feuille de report des réponses et évaluation de la difficulté des questions

- S’assurer que le candidat a signé le certificat avant de lui laisser le livret d’examen.

- Poser le test sur le coin de la table des candidats et leur demander d’attendre avant de retourner la feuille.
- Demander aux candidats de retourner la feuille et leur annoncer qu’ils ont 1 heure pour compléter le test.
- Écrire sur le tableau l’heure du début et de la fin de la passation.

5) Pendant la passation.

- Vérifier que les candidats répondent bien sur la feuille de réponse prévue à cet effet. En cas de besoin, merci de demander au candidat de bien vouloir répondre sur la feuille de réponses prévue à cet effet.
- En cas de besoin, se renseigner auprès de Vincent Folny qui sera présent lors de la passation.

6) Avant la fin de la passation.

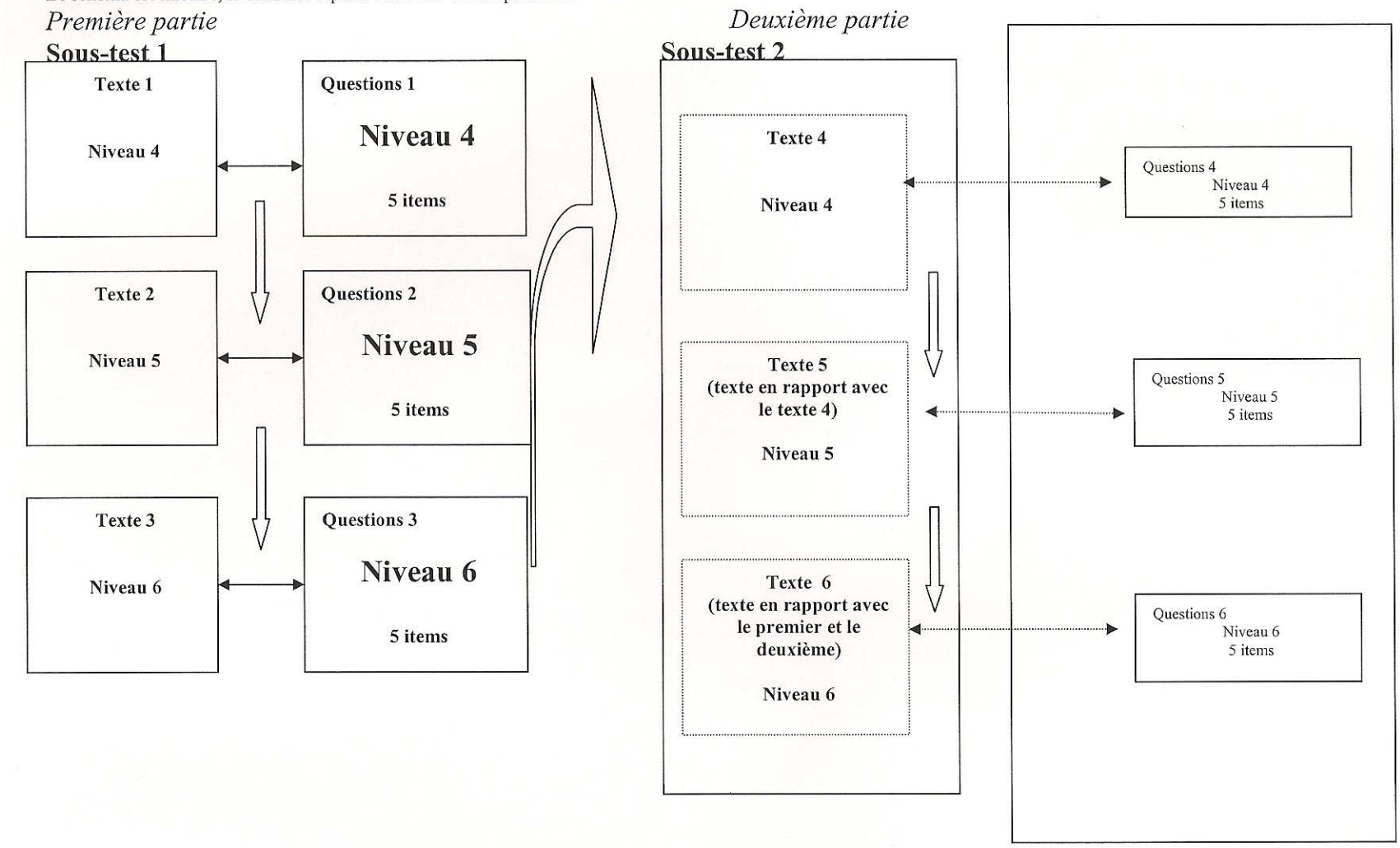
- Annoncer « il reste 15 minutes ».
- Annoncer « il reste 5 minutes ».

7) Fin de la passation.

- Annoncer « fin de l’examen ».

ANNEXE 3 Description du test.

Le schéma est linéaire, le candidat répond donc aux trente questions.



ANNEXE 4

Comparaison des textes proposés dans le test.

	Type de texte /Thème	**Public visé par l'auteur du texte**	**Longueur (nombre de mots et de paragraphes du corps de texte[1])**	**Co-texte**	**Contexte**
Texte 1	Informatif / programmatif Conseils pour la cuisson de la viande et des volailles	Grand public (ce texte ne demande pas de fortes connaissances préalables pour sa compréhension)	234 mots 8 paragraphes	Absence On ne dispose pas du journal où l'article a paru.	Textes sortis de leur contexte « historique » mais pas socioculturel puisque ce document a été écrit pour être lu au Québec
Texte 2	Informatif /chronologique Fait divers : Incendie en province au Québec	Grand public (fait divers qui intéresse autant les habitants de la province du Québec que les amateurs de faits divers)	280 mots 4 paragraphes	Absence On ne dispose pas du journal où l'article a paru.	
Texte 3	Informatif / analytique La consommation à Noël	Grand public (texte destiné aux consommateurs avant noël)	334 mots 8 paragraphes	Absence On ne dispose pas du journal où l'article a paru.	

[1] Statistiques fournies par le logiciel Word.

Texte 4	Informatif / programmatif Information culturelle	Grand public (texte destiné aux amateurs d'événements artistiques et récréatifs)	218 mots (+ 46 mots d'informations complémentaires en fin de texte) 4 paragraphes	Absence presque totale, on n'a pas accès au site internet où l'article a paru mais on a la présence de deux réactions.	Textes sortis de leur contexte « historique » mais pas socioculturel puisque ces document a été écrit pour être lu au Québec
Texte 5	« réactif » /chronologique Réaction à un événement culturel	-Lecteurs du texte 4 - spectateurs du spectacle	259 mots 1 paragraphe	Présence du texte déclencheur qui permet de connaître le contexte de l'événement culturel dont on parle	
Texte 6	« réactif »/ analytique Réaction à un événement culturel	-Lecteurs du texte 4 et 5 - spectateurs du spectacle - auteur du texte 4	365 mots 5 paragraphes	Présence du texte déclencheur qui permet de connaître le contexte de l'événement culturel dont on parle et de la réaction qui permet de comprendre le contexte d'écriture de cette réaction	

Annexe 5

Formulaire de consentement

Titre de la recherche : ***Fonctionnement de tâches discrètes et intégrées pour l'évaluation de la lecture en français langue seconde des nouveaux arrivants au Québec.***

Chercheur *:* Vincent Folny

Co-chercheur *: sans objet*

Directeur de recherche : Michel Laurier

A) RENSEIGNEMENTS AUX PARTICIPANTS

1. Objectifs de la recherche.

Ce projet de recherche a pour objectif de comprendre comment les résultats de l'évaluation de la lecture en français langue seconde varient en fonction du type d'activité utilisé pour la population des immigrants au Québec.

2. Participation à la recherche.

Votre participation à cette recherche consiste
À répondre à des questions à choix multiples portant sur des textes en français.
À donner votre avis sur la difficulté des activités d'évaluation.

3. Confidentialité.

Les informations que vous nous donnerez seront confidentielles. Chaque participant à la recherche se verra attribuer un numéro et seul le chercheur principal aura la liste des participants et du numéro qui leur aura été accordé. De plus, les informations seront gardées dans un classeur sous clé situé dans un bureau fermé. Aucune information permettant de vous identifier d'une façon ou d'une autre ne sera publiée. Les informations personnelles seront détruites au plus tard le 30 juin 2006. Seules les données ne permettant pas de vous identifier pourront être conservées après cette date.

4. Avantages et inconvénients.

En participant à cette recherche, vous pourrez contribuer au développement des connaissances et à l'amélioration des services de francisation offerts aux nouveaux arrivants au Québec. Votre participation à la recherche pourra également vous permettre de faire évaluer votre niveau de compréhension écrite en français. Si vous le souhaitez, vous pourrez donner votre adresse courriel pour que le résultat au test vous soit communiqué. Si vous n'êtes pas intéressé par cette option, n'écrivez pas votre adresse courriel sur les formulaires qui vous seront remis.

Par contre, Il est possible que le test que vous passerez vous demande de répondre à des questions d'un niveau supérieur au votre ou encore d'un niveau inférieur. La note que vous

recevrez pour ce test, ne sera pas prise en compte pour le calcul des notes de vos cours. Le test que vous passerez n'a aucun caractère officiel.

5. Droit de retrait

Votre participation est entièrement volontaire. Vous êtes libre d'arrêter le test en tout temps par avis verbal, sans préjudice et sans devoir justifier votre décision. Si vous décidez de vous retirer de la recherche, vous pouvez communiquer avec le chercheur, au numéro de téléphone indiqué à la dernière page de ce document. Si vous vous retirez de la recherche, les renseignements personnels vous concernant et qui auront été recueillis au moment de votre retrait seront détruits.

B) CONSENTEMENT

Je déclare avoir pris connaissance des informations ci-dessus, avoir obtenu les réponses à mes questions sur ma participation à la recherche et comprendre le but, la nature, les avantages, les risques et les inconvénients de cette recherche.

Après réflexion et un délai raisonnable, je consens librement à prendre part à cette recherche. Je sais que je peux me retirer en tout temps sans préjudice et sans devoir justifier ma décision.

Signature : _____ Date : _____

Nom : _____ Prénom : _____

Je déclare avoir expliqué le but, la nature, les avantages, les risques et les inconvénients de l'étude et avoir répondu au meilleur de ma connaissance aux questions posées.

Signature du chercheur_____ Date : _____
(ou de son représentant)

Nom : _____ Prénom : _____

Pour toute question relative à la recherche, ou pour vous retirer de la recherche, vous pouvez communiquer avec XXX, (chercheur principal), au numéro de téléphone suivant : XXX ou à l'adresse courriel suivante : XXX

Toute plainte relative à votre participation à cette recherche peut être adressée à l'ombudsman de l'Université de Montréal.

Un exemplaire du formulaire d'information et de consentement signé doit être remis au participant.

ANNEXE 6

Analyse des items présentant des indices de corrélation « point-bisérielle » insatisfaisants.

Item 20

Tableau A.6.1: fonctionnement des leurres de l'item 20

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	item	
20	A c	0	1	1	-.49	.4	-.09		D5 Intégrées3	c
	a	0	9	8	.09	.23	.9	-.09		a
	d	0	48	43	.22	.12	1.2	-.15		d
	b	1	54	48	.60	.15	1.3	.21		b
	MISSING	***	1	1*	-.52			-.09		

Quel est le public majoritairement attendu pour le spectacle de JMM ?

A- Le public appréciant les artistes québécois peu connus.
B- Le public aimant les activités artistiques et culturelles.
C- Le public des professionnels de la culture.
D- Le public des enfants des quartiers populaires.

Quoique l'item 20 présente des carrés moyens *infit* et *outfit* d'une valeur inférieure à 1,3, les valeurs standardisées sont supérieures à 2 (respectivement, 2,4 et 2,1). Si ces valeurs sont élevées, elles n'invitent pas à rejeter l'item directement mais plutôt à étudier son fonctionnement. L'examen attentif de l'item 20 (tableau A6.1) indique que l'item semble fonctionner correctement mais qu'un leurre (le d) attire presque autant de candidats que la bonne réponse. Pour améliorer l'item, il faudrait réécrire le leurre "le public des enfants des quartiers populaires". Pour l'analyse, cet item quoique peu satisfaisant peut être conservé.

Item 23

Tableau A.6.2 : fonctionnement des leurres de l'item 23

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	item	
23	B d	0	32	29	.05	.17	1.0	-.24	E3 Intégrées3	d
	c	0	56	51	.33	.10	.9	-.08		c
	a	0	5	5	1.57	.35	3.1	.27		a
	b	1	16	15	1.00*	.28	1.3	.25		b
	MISSING	***	4	4*	-.26	.30		-.13		

A la connaissance de Yan Steven, combien de personnes sont déçues par le spectacle ?

A- Une personne.
B- Deux personnes.
C- Plus de deux personnes.
D- Plusieurs groupes de personnes.

L'item 23 (tableau A.6.2) se singularise par sa difficulté. En effet, il s'agit de l'item le plus difficile parmi les 30 items du test (2,45 *logits*). Si ses statistiques de l'ajustement des données par rapport au modèle ont des valeurs acceptables (tableau 5.6), on peut détecter cependant plusieurs dysfonctionnements. Tout d'abord, cinq personnes ont choisi le leurre "a" alors que leur niveau de compétence est supérieur à celui des personnes qui ont choisi la bonne option (on imagine que les candidats les plus forts ont considéré que la consigne « à la connaissance de » signifiait « à l'exception de »). Le leurre « une personne » attire les candidats les plus forts. Dans le but d'améliorer le test, sans doute conviendrait-il de remplacer ce choix de réponse ou de reformuler la question dans le but d'éviter les mauvaises interprétations. Pour la présente analyse, le fonctionnement de cet item est toutefois satisfaisant.

Item 30

Tableau A.6.3 : fonctionnement des leurres de l'item 30

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	item	
30	a	0	21	20	-.11	.16	.9	-.24	F5 Intégrées	a
	d	0	4	4	.02	.88	2.1	-.07		d
	b	0	13	13	.46	.22	1.6	.03		b
	c	1	66	63	.58	.12	1.1	.24		c
	MISSING	***	9	8*	.10	.29		-.08		

Que pense Michel Noël ?
A- Les religieux devraient ouvrir des écoles.
B- Il est nécessaire de changer le monde.
C- On peut progresser dans la vie.
D- Un Québécois doit travailler au Québec.

Si l'item 30 a une corrélation point-bisérielle d'une valeur de 0,22, les leurres de l'item fonctionnent convenablement. Il est intéressant de constater que 20 personnes ont choisi le leurre « Les religieux devraient ouvrir des écoles ». Cette option avait été rédigée dans le but de vérifier si les candidats faisaient le lien entre le premier texte et le troisième texte pour écarter ce choix de réponse (le mot « frère » dans le premier texte étant clairement explicité). On peut constater que 20 % des candidats avec un niveau moyen n'ont pas fait ce lien. Une proportion importante de candidats, neuf candidats soit 8% de l'échantillon, n'a pas répondu à la question. Sur neuf réponses manquantes, cinq ont eu la

version 2 du test (questions 16-30 puis 1-15) et quatre la version 1 (questions 1-30). Il semble donc que le choix de ne pas répondre à cette question ne s'explique pas uniquement pas un éventuel manque de temps.

Item 12

Tableau A.6.4 : fonctionnement des leurres de l'item 12

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	item	
12 D	a	0	18	17	-.26	.18	.8	-.31	C2 Discrètes2	a
	d	0	10	9	.06	.33	1.3	-.12		d
	c	0	4	4	1.31	.28	3.2	.18		c
	b	1	75	70	.57*	.11	1.0	.25		b
	MISSING	***	6	5*	-.21	.21		-.15		

☞Cette année, à quel(s) moment(s) les Québécois achèteront-ils le plus de cadeaux ?

A- Tous les jeudis et vendredis de décembre.
B- La semaine avant Noël.
C- Le jour de Noël.
D- Quelques minutes avant Noël.

Cet item, de difficulté moyenne, fonctionne correctement. La seule véritable ombre au tableau vient du choix de réponse effectué par quatre candidats parmi les meilleurs, qui, pour une raison peu compréhensible, ont fait le choix de réponse « c ». Sans doute conviendrait-il de les interroger pour comprendre pourquoi ils ont fait ce choix et pouvoir juger de la valeur réelle du leurre. En l'absence d'une telle information, il est difficile d'aller plus avant dans l'analyse.

Item 5 et 25

Tableau A.6.5 : fonctionnement des leurres des items 5 et 25

Item 5

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	item	
5 E	d	0	9	8	-.21	.34	1.1	-.19	A5 Discrètes2	d
	a	0	15	14	.05	.22	1.0	-.14		a
	c	0	28	26	.31	.16	1.3	-.06		c
	b	1	56	52	.64	.13	1.1	.25		b
	MISSING	***	5	4*	-.09	.20		-.11		

☞Qu'est-ce que pourrait dire Jacinthe Côté de la cuisine à Noël ?

A- « A Noël, il faut innover et proposer des plats originaux ».
B- « L'important c'est le résultat, pas le temps passé à cuisiner ».
C- « Un plat qui sent bon, c'est un plat idéal ».
D- « Peu importe l'esthétique, l'essentiel c'est la quantité ».

Item 25

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	item	
25 F	a	0	2	2	-1.35	.50	.2	-.25	E5 Intégrées3	a
	d	0	23	21	.16	.19	1.1	-.13		d
	c	0	36	33	.30	.15	1.3	-.08		c
	b	1	47	44	.69	.14	1.1	.26		b
	MISSING	***	5	4*	-.24	.22		-.14		

Qu'est-ce que Yan Steven a pensé du spectacle ?

A- Il coûtait trop cher**.**
B- Il était difficile d'en évaluer la qualité.
C- Il pouvait être amélioré.
D- Il représentait une bonne distraction.

Alors que les deux items ont des valeurs de carrés *infit* et *oufit* acceptables (tableau A.6.5), la valeur standardisée de l'*infit* dans le cas de l'item 25 est égale à 2 et celle de l'item 5 à 1,9. Il convient donc de réviser ces deux items pour vérifier s'ils sont écrits correctement et pour tenter de comprendre pourquoi ils présentent de telles valeurs. L'examen du fonctionnement des leurres (tableau A6.5) nous apprend que ceux de ces questions fonctionnent correctement. Cependant, ces deux questions de difficulté moyenne ne permettent pas de distinguer fortement les bons des mauvais lecteurs. Ne faisant pas face à un problème d'écriture majeure, il semble que les deux items ne soient pas dénoués d'intérêt et qu'il soit possible de les conserver dans l'échantillon. Enfin, ces deux items évaluent la même compétence (savoir inférer l'opinion de l'auteur), mais à des niveaux différents (0,29 et 0,70 *logit*) et avec des résultats différents pour les candidats.

Item 1

Tableau A.6.6 : fonctionnement des leurres de l'item 1

ENTRY NUMBER	DATA CODE	SCORE VALUE	DATA COUNT	%	AVERAGE MEASURE	S.E. MEAN	OUTF MNSQ	PTMEA CORR.	item	
1 G	b	0	2	2	-.47	.43	.6	-.12	A1 Discrètes	b
	a	0	8	7	-.44	.22	.7	-.24		a
	d	0	11	10	.26	.20	1.4	-.04		d
	c	1	90	81	.49	.10	1.1	.23		c
	MISSING	***	2	2*	.22	.44		-.02		

Question 1

Quel type de conseils donne-t-on dans cet article ?

A- Des conseils hygiéniques.
B- Des conseils médicaux.
C- Des conseils nutritionnels.
D- Des conseils préventifs.

Cet item est un item facile (-1,32 *logits*). Il est donc « normal » que la corrélation point-bisérielle ait une valeur basse. Malgré une corrélation point-bisérielle faible, l'item fonctionne de manière satisfaisante. Tous les leurres sont choisis (Tableau A.6.6), même si on peut constater que le leurre « Des conseils médicaux » n'est que très peu choisi (2%). Dans une version améliorée du test, on pourrait éventuellement penser à changer ce leurre.

ANNEXE 7

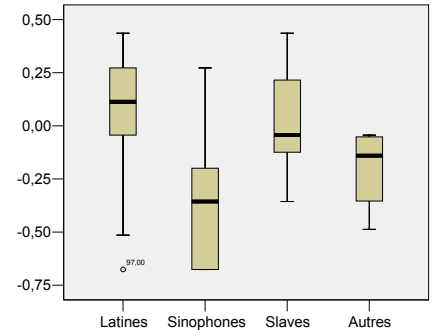
Répartition des candidats et des items en groupes de niveaux comparables

Création du groupe de niveau intermédiaire, groupe de « niveau 2 »

Pour constituer le premier groupe (« niveau 2 » dans la figure A.7.1), la moyenne des personnes a servi de référence. On a décidé de garder toutes les personnes comprises dans un intervalle d'environ 0,5 *logit* autour de la moyenne. L'intervalle [0,44,-0,68], niveau intermédiaire ou « niveau 2 », a été choisi pour trouver le meilleur compromis entre un nombre suffisant de candidats et d'items pour ce niveau, un nombre suffisant de représentants de chacun des groupes linguistiques et la possibilité de constituer un autre groupe ayant des propriétés plus ou moins similaires. Pour choisir ce niveau, on a également veillé à ce que le découpage soit cohérent en termes de compétence de lecture.

Le « niveau 2 » a une médiane autour de 1 *logit*. Il comprend 49 candidats dont on peut voir la dispersion en fonction du groupe de langue dans la figure A.7.1 et le tableau A.7.1.

Figure A.7.1 : boîtes à moustaches des groupes linguistiques pour le « niveau 2 »



Afin de vérifier si les moyennes des personnes du « niveau 2 » appartenant à des groupes linguistiques différents variaient significativement, on a procédé à une ANOVA (tableau 4.47). Bien que l'interprétation de ces résultats requière une certaine prudence (il y a moins de 30 observations par catégories, les groupes ne sont pas de tailles identiques, en revanche les variances sont égales), on constate que l'ANOVA est significative (tableau, F= 6,195, ddl=48, sig. 0,001). La moyenne qui diffère des autres est celle des sinophones. L'appartenance des personnes aux groupes linguistiques expliquerait encore 29 % de la variabilité (cf. êta-carré, tableau A.7.2). Le groupe

Tableau A.7.1 : répartition des personnes par groupe linguistique pour le niveau intermédiaire, « niveau 2 »

groupe linguistique

		Fréquence	Pour cent
Valide	Latines	17	34,7
	Chinois	17	34,7
	Slaves	11	22,4
	Autres	4	8,2
	Total	49	100,0

intermédiaire, ne se définit donc pas par l'égalité des moyennes des groupes linguistiques mais bien par leur comparabilité.

Tableau A.7.2 : résultats de l'ANOVA pour vérifier l'égalité des moyennes des groupes linguistiques du « niveau 2 »

Test d'homogénéité des variances

Estimated personne Measure: UMEAN=.00 USCALE=1.00

Statistique de Levene	ddl1=	ddl2	Signification
,800	3	45	,500

ANOVA

Estimated personne Measure: UMEAN=.00 USCALE=1.00

	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Inter-groupes	1,779	3	,593	6,195	,001
Intra-groupes	4,308	45	,096		
Total	6,087	48			

Comparaisons multiples

Variable dépendante: Estimated personne Measure: UMEAN=.00 USCALE=1.00

Bonferroni

(I) groupe linguistique	(J) groupe linguistique	Différence de moyennes (I-J)	Erreur standard	Signification
Chinois	Latines	-,41218*	,10612	,002
	Slaves	-,39768*	,11972	,011

*. La différence de moyennes est significative au niveau .05.

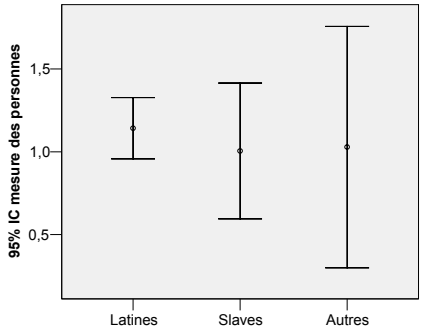
Mesures des associations

	Eta	Eta carré
Estimated personne Measure: UMEAN=.00 USCALE=1.00 * groupe linguistique	,541	,292

Analyse des deux autres groupes

Après avoir crée le groupe intermédiaire, nous avions à définir le groupe des candidats les plus forts et celui des plus faibles. Le niveau le plus faible, « niveau 1 », contient cinq items et 12 candidats (en majorité composé de sinophones et personnes d'autres langues). Il a une étendue de 1 *logit* et une médiane autour de – 1 *logit* (figure A.7.3). Quoique ce groupe comporte peu d'items et de personnes, il a été décidé de ne pas tenter d'en élargir la composition. La stratégie retenue consiste en la création d'un groupe de « niveau 2 » comprenant plus de personnes. Cela permettra de mener à bien des analyses sur le niveau intermédiaire et avancé. Le « niveau 1 », niveau faible, ne servira dans l'analyse qu'à définir le seuil du groupe intermédiaire.

Figure A.7.2 : intervalles de confiance des moyennes des groupes linguistiques du « niveau 3 »



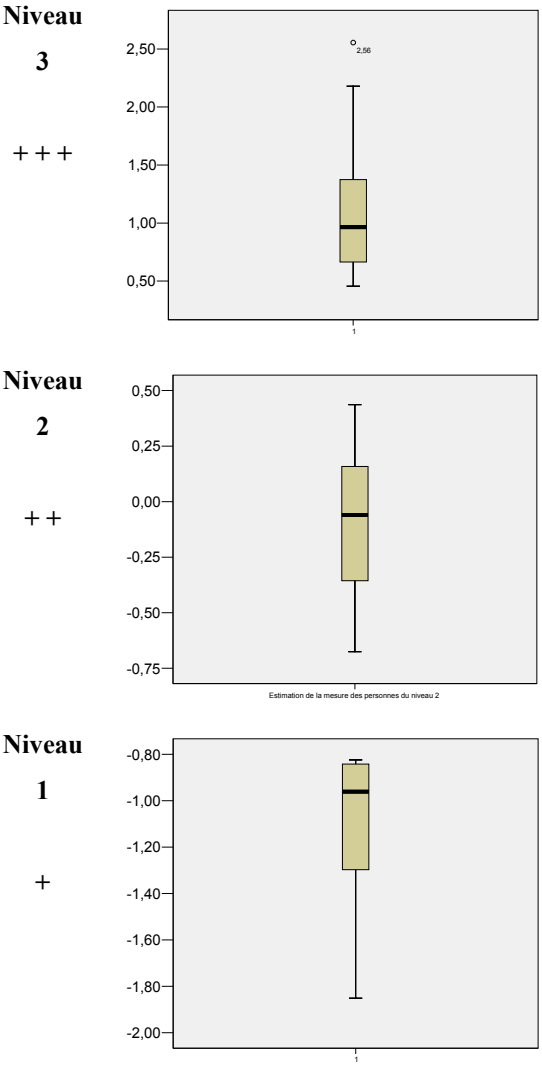
Le « niveau 3 » regroupe 50 personnes et huit items. Il a une étendue de 2,10 *logits*. Comme on peut le voir sur la carte des personnes et des items (figure 5.13), deux personnes ont été exclues de ce groupe (personnes 36 et 101). Le critère, qui a été retenu pour faire ce choix, a été d'exclure les personnes dont le niveau était supérieur au niveau du dernier item, soit l'item 23. Parce que le « niveau 3 » ne comprend qu'une seule personne sinophone et quatre personnes ayant une « autre langue » (tableau A.7.3), seules les personnes des groupes des langues latines et des langues slaves pourront être comparées au « niveau 3 ».

Tableau A.7.3 : fréquence des groupes linguistiques pour le « niveau 3 »

groupes linguistiques

		Fréquence
Valide	Latines	36
	Slaves	9
	Chinois	1
	Autres	4
	Total	50

Figure A.7.3 : boîtes à moustaches des candidats pour les trois niveaux empiriques

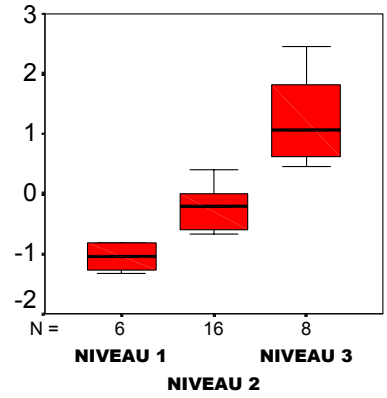


N'ayant pas des conditions suffisantes pour pratiquer une ANOVA afin de vérifier l'égalité des moyennes, il a été décidé de les étudier à partir des intervalles de confiance à 95 % autour de la moyenne des groupes linguistiques (figure A.7.2). Les moyennes des personnes du groupe des langues latines et du groupe des langues slaves étant quasiment identiques (1, 14 *logits* pour le groupe des langues latines et 1 *logit* pour le groupe des langues slaves), l'intervalle de confiance du groupe des langues latines étant superposé à celui des langues slaves, on décide que les conditions sont suffisantes pour comparer les résultats des personnes de ces deux groupes linguistiques.

Vue d'ensemble des trois niveaux de compétence des candidats.

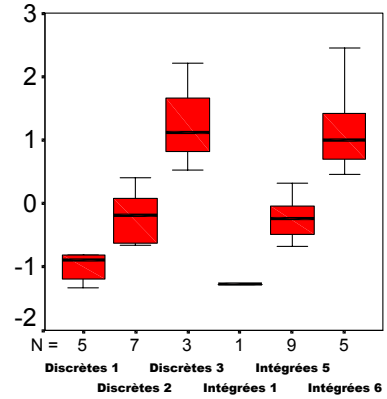
Afin d'avoir une vue d'ensemble du découpage en trois niveaux, il a été décidé de juxtaposer les boîtes à moustaches des échantillons de personnes de chacun des niveaux (figure A7.4).

Figure A.7.4 : boîte à moustaches pour les questions de niveau 1, 2 et 3



Il appert que le « niveau 2 », le niveau intermédiaire, est distribué à peu près normalement, même s'il est affecté par une légère asymétrie. Des trois groupes, il s'agit certainement de celui qui a les meilleures qualités et celui à partir duquel on pourra effectuer les analyses. En effet, c'est celui qui nous permet de comparer, les trois principaux groupes linguistiques, qui a l'étendue la plus faible, qui bénéficie du plus d'items (16 items couvrent ce niveau).

Figure A.7.5 : boîte à moustaches pour les questions de niveau 1, 2 et 3 pour les tâches discrètes et intégrées



Les deux autres groupes, ne sont pas distribués normalement, le « niveau 3 » est affecté par une valeur extrême. Le « niveau 1 » a ses deux premiers quartiles proches du niveau seuil avec le « niveau 2 » et le « niveau 3 » ses deux derniers quartiles proches du niveau seuil avec le « niveau 2 ».

L'examen de ces boites à moustaches montre que le découpage en trois niveaux tel qu'il est proposé autorise la constitution d'un groupe de très bonne qualité au centre et l'établissement des seuils regroupant des candidats autour des extrémités supérieures et inférieures du « niveau 2 ». Cela permettra d'étudier l'impact du type de tâche sur le classement de candidat à la limite entre deux niveaux (surtout que, comme nous le verrons plus loin, ce découpage correspond à un découpage en contenu).

Vue d'ensemble des trois niveaux de difficulté des questions

Le découpage en trois niveaux aura permis de regrouper les questions autour de trois niveaux, soit –1, 0 et 1 *logit* (figure A.7.4). Si la dispersion est réduite pour les « niveaux 1 » et « niveau 2 », en revanche, pour le « niveau 3 », elle est plus grande. Cela ne constitue pas en soit un problème, puisque l'objectif qui avait été fixé lors de la conception de l'examen était de rédiger des questions qui soient autour du niveau du candidat, soit les questions du « niveau 2 ».

Si on examine les distributions des questions des trois niveaux pour les questions discrètes et intégrées, on constate que globalement, les questions ont des distributions similaires et des moyennes comparables (tableau A.7.5).

Pourtant, pour les questions de « niveau 1 », il ne sera pas possible de comparer le fonctionnement pour les tâches discrètes et pour les tâches intégrées, puisqu'il n'y a qu'une seule question pour les tâches intégrées. Mais, encore, une fois, l'objectif étant de comparer le niveau moyen des candidats, le découpage qui a été privilégié permet d'avoir sept questions de « niveau 2 » pour les tâches discrètes et neuf pour les tâches intégrées. Les distributions de ces questions sont relativement semblables.

Tableau A.7.4 : indices de tendances centrales des questions de niveau 1, 2 et 3 pour les tâches discrètes et intégrées

Subtotal specification is: ISUBTOTAL=$S4E13
ALL SCORES ARE NON-EXTREME

item COUNT	MEAN MEASURE	S.E. MEAN	OBSERVED S.D.	MEDIAN SEPARATION	REAL	CODE
30	.00	.17	.93	-.16	3.87	***
5	-1.01	.10	.21	-.89	.00	Discrètes1
7	-.22	.17	.41	-.18	1.55	Discrètes2
3	1.28	.49	.70	1.11	2.73	Discrètes3
1	-1.27	-	.00	-1.27	.00	Intégrées1
9	-.23	.11	.30	-.25	.96	Intégrées2
5	1.20	.35	.70	1.01	2.71	Intégrées3

Correspondance entre le découpage en trois niveaux et les compétences des candidats

Après avoir étudié la distribution des personnes et des items pour chacun des niveaux, cette partie de l'analyse a pour objectif de vérifier si le découpage en trois niveaux correspond bien à trois niveaux de compétence différents. Comme cela a déjà expliqué, la répartition en trois niveaux ne s'est pas faite uniquement à partir de l'analyse de la distribution des personnes ou des items ou encore à partir de l'analyse de contenu des items, mais en fonction des deux et, ce, de manière simultanée. Si la partie précédant

l'analyse a permis d'expliquer quels arguments ont été retenus pour le découpage de l'échantillon des personnes en trois niveaux, il reste à expliquer pourquoi, on considère que ce découpage permet de distinguer des niveaux de compétence de lecture réellement différents.

Tout d'abord, il convient de rappeler comment les items ont été rédigés. Le choix a été fait de poser des questions à partir de textes sélectionnés selon les descriptions des niveaux 4, 5 et 6 des *Niveaux de compétence en français langue seconde pour les immigrants adultes* (1998).(tableau A.7.5).

Tableau A.7.5 : description des niveaux 4, 5 et 6 des *Niveaux de compétence en français langue seconde pour les immigrants adultes* (1998) en compréhension écrite

Niveau 6	**Compréhension partielle de Textes de complexité moyenne d'environ une page dont le contenu et le contexte sont familiers.**
DESCRIPTION GÉNÉRALE Comprend la plupart des informations et des idées exprimées de façon explicite dans divers types de textes d'environ une page. Parvient à saisir le sens du texte même si des mots ou des expressions lui sont inconnus. Comprend des mots, des expressions et des formules peu ou non utilisés à l'oral mais qui le sont fréquemment à l'écrit. Sait établir les liens appropriés en reconnaissant les synonymes et en s'appuyant sur certains mots qui assurent la cohésion du texte tel le pronom	
Niveau 5	**Compréhension limitée de textes de complexité moyenne de quelques paragraphes dont le contenu et le contexte sont familiers.**
DESCRIPTION GÉNÉRALE Comprend l'essentiel d'un texte de quelques paragraphes portant sur un sujet concret et familier malgré la présence de phrases complexes et de mots peu courants. Suit le déroulement des événements en s'appuyant sur différents indices de temps tels que des expressions courantes ou certains temps de verbe. Comprend un mode d'emploi détaillé ou une directive comportant plusieurs étapes.	
Niveau 4	**Compréhension entière de textes simples de quelques paragraphes dont le contenu et le contexte sont familiers.**
DESCRIPTION GÉNÉRALE Saisit les idées tant générales que spécifiques d'un court texte de quelques paragraphes rédigé dans un langage simple et courant. Saisit l'enchaînement des étapes d'une directive ou celui d'événements relatés en s'appuyant sur des mots qui marquent la chronologie. Compare plusieurs informations factuelles de même nature dans le but de faire un choix. Parcourt rapidement un court texte ou un document simple pour y trouver l'information spécifique recherchée. Lit les textes simples lisiblement écrits à la main.	

Ensuite, dans un souci d'authenticité, il a été décidé, autant que peut ce faire, de poser des questions que les candidats seraient susceptibles de se poser en lisant les textes. Il est donc important de vérifier, si un décalage n'est pas apparu entre le niveau visé par les textes et les niveaux réellement évalués par les questions.

Pour les besoins de l'analyse, on a placé les items par ordre de difficulté. On constate que la répartition prévue des items en trois niveaux selon des tâches de niveaux différents n'a

pas fonctionnée (tableau A.7.6). Très peu de questions (7 sur 30) correspondent au niveau pour lequel elles avaient été conçues initialement (chiffres de la diagonale en gras). Si ce résultat semble peu encourageant, il convient toutefois de le nuancer. En effet, si une seule des dix questions qui avaient été écrites à partir des textes les plus faciles est une question classée facile, on constate que six de ces questions sont des questions de niveau moyen. Si seulement deux des dix questions écrites pour avoir un niveau fort ont effectivement un niveau fort, six ont un niveau intermédiaire. Vu sous cet angle, l'utilisation de textes dont la difficulté se trouve autour de la moyenne de l'ensemble des personnes, a permis d'écrire 16 questions sur 30 se situant proche de la moyenne de l'ensemble des candidats ce qui est plus que les dix questions prévues initialement.

Tableau A.7.6 : comparaison des prévisions faites pour la difficulté des items et la difficulté réelle

	Niveaux prévus selon les *Niveaux de compétence en français langue seconde pour les immigrants adultes*.			
Niveaux empiriques	**Niveau 4** +	**Niveau 5** ++	**Niveau 6** +++	**Total**
Niveau 3 +++	3/10**(1)**	3/10	**2/10**	8
Niveau 2 ++	6/10	**4/10**	6/10	16
Niveau 1 +	**1/10**	3/10	2/10	6
Total	10	10	10	

(1) lire : fréquence attendue 10, observée 3

À présent, si on procède à une méta-analyse du contenu des items (établie par induction après écriture et administration des items) répartis dans chacun des niveaux (tableau A.7.7), on constate que les niveaux de compétence 4, 5 et 6 du M.I.C.C. ne correspondent pas tout à fait aux trois niveaux définis empiriquement.

En réalité, si les niveaux de compétences visés au moment de l'écriture des questions étaient 4, 5 et 6, les niveaux constitués à partir de la difficulté réelle des items correspondent aux niveaux 5, 6 et 7 (tableau A.7.8)

Tableau A.7.7: répartition des questions par niveau

Question	Difficulté en *logits* (ordre décroissant de difficulté)	But de la question	Activité cognitive (mise en œuvre analyser l'option qui est la bonne réponse)
Niveau visé 6 **Synthèse de la description empirique des items (« niveau 3 »)** **Le candidat comprend des textes de divers types de complexité moyenne, d'une longueur d'une page portant sur la compréhension d'informations non-explicite, socioculturelle demandant au candidat la mise en œuvre d'inférences simples (un paragraphe) ou complexes (plusieurs paragraphes ou entre plusieurs textes).**			
Question 23 E3	2,45	Vérifier la compréhension d'une information non explicite	Inférence Locale multiple (I.L.M.)
Question 11 C1	2,21	Vérification de l'acquisition d'une connaissance socioculturelle non explicite	Inférence Globale (I.G.)
Question 19 D4	1,41	Vérification de l'acquisition d'une information non-explicite	Inférence locale (I.L.)
Question 3 A3	1,11	Vérification acquisition d'un savoir	Inférence multi-paragraphe (I.M.P), mise en relation d'information présente dans plusieurs paragraphes.
Question 26 F1	1.01	Reconnaître l'opinion, la sensibilité de l'auteur d'un article	Inférence intertextuelle (information du texte 5 couplée à l'information du texte 6)
Question 25 E5	0.7	Vérifier la compréhension d'une information non explicite	Inférence locale (I.L.)
Question 8 B3	0,52	Vérifier la compréhension d'une information non explicite	Inférence locale (I.L.)
Question 20 D5	0,45	Vérifier la capacité à reconnaître le type de lecteur ciblé par l'auteur	Inférence locale (I.L.)
Niveau visé : 5 **Synthèse de la description empirique des items (« niveau 2 »)** **Le candidat comprend des textes de divers types, de complexité moyenne, d'une longueur d'une page portant sur la compréhension d'informations le plus souvent explicites ou demandant parfois au candidat de faire des inférences soit localisée sur un passage dans un texte ou portant sur tout un texte afin de comprendre une information ou encore sur l'opinion de l'auteur.**			
Question 15 C5	0,41	Vérification de l'acquisition d'une connaissance socioculturelle explicite	Inférence locale (I.L.)
Question 29 F4	0,32	Reconnaître l'opinion, la sensibilité de l'auteur d'un article	Inférence locale (I.L.)
Question 5 A5	0,29	Reconnaître l'opinion, la sensibilité de l'auteur d'un article	Inférence Globale (I.G.)
Question 17 D2	0,04	Vérification de l'acquisition d'une information explicite	Paraphrase / reformulation (P-R) texte 4 ou début du texte 5
Question 28 F3	-0,03	Reconnaître l'opinion, la sensibilité de l'auteur d'un article	Inférence locale (I.L.)
Question 16 D1	-0,10	Vérification de l'acquisition d'une information	Paraphrase / reformulation (P-R)

		explicite	
Question 10 B5	-0,14	Vérifier la compréhension d'une information non explicite	Inférence Globale (I.G.)
Question 2 A2	-0,18	Vérification acquisition d'un savoir	Paraphrase / reformulation (P-R)
Question 18 D3	-0,25	Vérification de l'acquisition d'une information explicite	Paraphrase / reformulation (P-R) texte 4
Question 30 F5	-0,27	Reconnaître l'opinion, la sensibilité de l'auteur d'un article	Inférence Globale (I.G.)
Question 27 F2	-0,48	Vérifier la compréhension d'une information non explicite	Inférence locale (I.L.)
Question 22 E2	-0,59	Vérification de la compréhension d'un élément de vocabulaire en contexte	Inférence locale (I.L.)
Question 12 C2	-0,61	Vérification de l'acquisition d'une information non explicite	Inférence multi-paragraphe (I.M.P)
Question 9 B4	-0,63	Vérifier la compréhension d'une information explicite	Paraphrase / reformulation (P-R)
Question 4 A4	-0,66	Vérification de l'acquisition d'un savoir	Inférence locale (I.L.), à partir de plusieurs informations contenues dans un même paragraphe.
Question 24 E4	-0,68	Vérifier la compréhension d'une information non explicite	Inférence locale (I.G.)
Niveau visé : 4 **Synthèse de la description empirique des items (« niveau 1 »)** **Le candidat comprend des textes ou des passages de textes de divers types, de complexité moyenne, d'une longueur d'une page. Peut répondre à des questions portant sur la reconnaissance de l'intention de communication de l'auteur d'un texte et à l'occasion à des questions faciles portant sur la compréhension d'informations le plus souvent explicites ou demandant parfois au candidat de faire des inférences le plus souvent globales.**			
Question 6 B1	-0,82	Reconnaître l'intention de communication de l'auteur	Inférence Globale (I.G.)
Question 14 C4	-0,82	Vérification de l'acquisition d'une connaissance socioculturelle non explicite	Paraphrase / reformulation (P-R-M) de plusieurs passages
Question 7 B2	-0,89	Vérifier la compréhension d'une information non explicite	Inférence multi-paragraphe (I.M.P)
Question 13 C3	-1,19	Vérification de l'acquisition d'une connaissance socioculturelle non explicite	Inférence Globale (I.G.)
Question 21 E1	-1,27	Vérification de l'intention de communication de l'auteur	Inférence Globale (I.G.) pour le texte 5 Inférence locale (I.L.) pour le texte 6, premier paragraphe
Question 1 A1	-1,32	Reconnaître intention de communication de l'auteur	Inférence Globale (I.G.), après la lecture complète du texte.

Tableau A.7.8 : correspondance entre les descriptions empiriques et les *Niveaux de compétence en français langue seconde pour les immigrants adultes* (M.R.C.I., 1998) pour la compréhension écrite	
Synthèse de la description empirique	***Niveaux de compétence en français langue seconde pour les immigrants adultes* (M.R.C.I., 1998) pour la compréhension écrite**
« Niveau 3 » Le candidat comprend des textes de divers types de complexité moyenne, d'une longueur d'une page ou de plusieurs pages portant sur la compréhension d'informations non-explicites, socioculturelles, demandant au candidat la mise en œuvre d'inférences simples (un paragraphe) ou complexes (plusieurs paragraphes) ou intertextuelles (entre plusieurs textes).	 **Compréhension satisfaisante de textes de complexité moyenne allant jusqu'à environ deux pages dont le contexte facilite la lecture.** **DESCRIPTION GÉNÉRALE** Comprend les informations détaillées, les idées formulées explicitement et quelques idées implicites dans une ou deux pages de texte dont la présentation et l'organisation favorisent la compréhension. Comprend une anecdote ou un récit comportant des éléments humoristiques. Distingue les opinions des faits. Déduit ou précise le sens d'un mot en utilisant ses connaissances en français, son expérience de lecteur ou le contexte.
« Niveau 2 » Le candidat comprend des textes de divers types, de complexité moyenne, d'une longueur d'une page portant sur la compréhension d'informations le plus souvent explicites ou demandant parfois au candidat de faire des inférences soit localisées sur un passage dans un texte ou portant sur tout un texte afin de comprendre une information ou encore sur l'opinion de l'auteur.	 **Compréhension partielle de Textes de complexité moyenne d'environ une page dont le contenu et le contexte sont familiers.** **DESCRIPTION GÉNÉRALE** Comprend la plupart des informations et des idées exprimées de façon explicite dans divers types de textes d'environ une page. Parvient à saisir le sens du texte même si des mots ou des expressions lui sont inconnus. Comprend des mots, des expressions et des formules peu ou non utilisés à l'oral mais qui le sont fréquemment à l'écrit. Sait établir les liens appropriés en reconnaissant les synonymes et en s'appuyant sur certains mots qui assurent la cohésion du texte tel le pronom.
Niveau 1 Le candidat comprend des passages de textes de divers types, de complexité moyenne, d'une longueur d'une page. Peut répondre à des questions portant sur la reconnaissance de l'intention de communication de l'auteur d'un texte et à l'occasion à des questions faciles portant sur la compréhension d'informations le plus souvent explicites ou demandant parfois au candidat de faire des inférences le plus souvent globales.	 **Compréhension limitée de textes de complexité moyenne de quelques paragraphes dont le contenu et le contexte sont familiers.** **DESCRIPTION GÉNÉRALE** Comprend l'essentiel d'un texte de quelques paragraphes portant sur un sujet concret et familier malgré la présence de phrases complexes et de mots peu courants. Suit le déroulement des événements en s'appuyant sur différents indices de temps tels que des expressions courantes ou certains temps de verbe. Comprend un mode d'emploi détaillé ou une directive comportant plusieurs étapes.

Annexe 8

Programme pour l'unidimensionnalité des tâches discrètes et intégrées calibrées ensemble, 15 items discrets et 15 items intégrés.

```
>Title
 Fonctionnement différentiel
 entre des tâches intégrées et discrètes;
>PROBLEM NITEM=30,RESPONSE=5,SUBTEST=2;
>NAMES A1,A2,A3,A4,A5,B1,B2,B3,B4,B5,C1,C2,
C3,C4,C5,D1,D2,D3,D4,D5,E1,E2,E3,E4,E5,F1,F2,F3,F4,F5 ;
>RESPONSES ' ','a','b','c','d';
>KEY cddcbcbdcbcbdcaacaabaabdbbcddc;
>RELIABILITY ALPHA;
>PLOT BISERIAL,NOCRITERION,FACILITY;
>BIFACTOR NIGROUPS=2,
IGROUPS=(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
2,2,2,2,2,2,2,2,2,2,2,2,2,2,2),LIST=2,CPARMS=(0.25(0)30),OMIT=MISS;
>SCORE CHANCE;
>SUBTEST BOUNDARY=(15,30), NAME=(discrètes,Intégrées);
>INPUT NIDW=3,SCORE,FILE='donnees.DAT';
(3A1,1X,30A1)
>CONTINUE;
>STOP
```

Programme pour l'unidimensionnalité des tâches discrètes et intégrées, six textes

```
>Title
 Fonctionnement différentiel
 entre des tâches intégrées et discrètes
(6 textes);
>PROBLEM NITEM=30,RESPONSE=5,SUBTEST=6;
>NAMES A1,A2,A3,A4,A5,B1,B2,B3,B4,B5,C1,C2,
C3,C4,C5,D1,D2,D3,D4,D5,E1,E2,E3,E4,E5,F1,F2,F3,F4,F5 ;
>RESPONSES ' ','a','b','c','d';
>KEY cddcbcbdcbcbdcaacaabaabdbbcddc;
>RELIABILITY ALPHA;
>PLOT BISERIAL,NOCRITERION,FACILITY;
>BIFACTOR NIGROUPS=6,
IGROUPS=(1,1,1,1,1,2,2,2,2,3,3,3,3,3,
4,4,4,4,4,5,5,5,5,6,6,6,6,6),LIST=2,CPARMS=(0.25(0)30),OMIT=MISS;
>SCORE CHANCE;
>SUBTEST BOUNDARY=(5,10,15,20,25,30), NAME=(T1,T2,T3,T4,T4,T5,T6);
>INPUT NIDW=3,SCORE,FILE='donnees.DAT';
(3A1,1X,30A1)
>CONTINUE;
>STOP
```

Programme pour l'unidimensionnalité des tâches discrètes et intégrées, trois textes discrets et un ensemble de trois textes intégrés

```
>Title
 Fonctionnement différentiel
 entre des tâches intégrées et discrètes
(3 textes discrets et 1 ensemble textes intégrés);
>PROBLEM NITEM=30,RESPONSE=5,SUBTEST=4;
>NAMES A1,A2,A3,A4,A5,B1,B2,B3,B4,B5,C1,C2,
C3,C4,C5,D1,D2,D3,D4,D5,E1,E2,E3,E4,E5,F1,F2,F3,F4,F5 ;
>RESPONSES ' ','a','b','c','d';
>KEY cddcbcbdcbcbdcaacaabaabdbbcddc;
>RELIABILITY ALPHA;
>PLOT BISERIAL,NOCRITERION,FACILITY;
>BIFACTOR NIGROUPS=4,
IGROUPS=(1,1,1,1,1,2,2,2,2,3,3,3,3,3,
4,4,4,4,4,4,4,4,4,4,4,4,4,4,4),LIST=2,CPARMS=(0.25(0)30),OMIT=MISS;
>SCORE CHANCE;
```

```
>SUBTEST BOUNDARY=(5,10,15,30), NAME=(T1,T2,T3,TI);
>INPUT NIDW=3,SCORE,FILE='donnees.DAT';
(3A1,1X,30A1)
>CONTINUE;
>STOP
```

Programme pour l'unidimensionnalité des tâches discrètes et intégrées avec un facteur général

```
>Title
 Fonctionnement différentiel
 entre des tâches intégrées et discrètes;
>PROBLEM NITEM=30,RESPONSE=5,SUBTEST=2;
>NAMES A1,A2,A3,A4,A5,B1,B2,B3,B4,B5,C1,C2,
C3,C4,C5,D1,D2,D3,D4,D5,E1,E2,E3,E4,E5,F1,F2,F3,F4,F5 ;
>RESPONSES ' ','a','b','c','d';
>KEY cddcbcbdcbcbdcaacaabaabdbbcddc;
>RELIABILITY ALPHA;
>PLOT BISERIAL,NOCRITERION,FACILITY;
>SELECT 1(1)30
>BIFACTOR NIGROUPS=1,
IGROUPS=(0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),LIST=2,CPARMS=(0.25(0)30),OMIT=MISS;
>SCORE CHANCE;
>SUBTEST BOUNDARY=(15,30), NAME=(discrètes,Intégrées);
>INPUT NIDW=3,SCORE,FILE='donnees.DAT';
(3A1,1X,30A1)
>CONTINUE;
>STOP
```

Programme pour l'unidimensionnalité des tâches discrètes, trois textes discrets

```
>Title
 Fonctionnement
 des tâches discrètes (3 textes);
>PROBLEM NITEM=15,RESPONSE=5,SUBTEST=3;
>NAMES A1,A2,A3,A4,A5,B1,B2,B3,B4,B5,C1,C2,
C3,C4,C5,;
>RESPONSES ' ','a','b','c','d';
>KEY cddcbcbdcbcbdca;
>RELIABILITY ALPHA;
>PLOT BISERIAL,NOCRITERION,FACILITY;
>BIFACTOR NIGROUPS=3,
IGROUPS=(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3),LIST=3,CPARMS=(0.25(0)15),OMIT=MISS;
>SCORE CHANCE;
>SUBTEST BOUNDARY=(5,10,15), NAME=(Texte1, Texte2,Texte3);
>INPUT NIDW=3,SCORE,FILE='donnees.DAT';
(3A1,1X,15A1)
>CONTINUE;
>STOP
```

Programme pour l'unidimensionnalité des tâches discrètes, un facteur général et un ensemble de trois textes

```
>Title
 Fonctionnement
 des tâches discrètes (1 ensemble de textes);
>PROBLEM NITEM=15,RESPONSE=5,SUBTEST=1;
>NAMES A1,A2,A3,A4,A5,B1,B2,B3,B4,B5,C1,C2,
C3,C4,C5,;
>RESPONSES ' ','a','b','c','d';
>KEY cddcbcbdcbcbdca;
>RELIABILITY ALPHA;
>PLOT BISERIAL,NOCRITERION,FACILITY;
>BIFACTOR NIGROUPS=1,
IGROUPS=(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),LIST=3,CPARMS=(0.25(0)15),OMIT=MISS;
>SCORE CHANCE;
>SUBTEST BOUNDARY=(15), NAME=(3 textes);
```

```
>INPUT NIDW=3,SCORE,FILE='donnees.DAT';
(3A1,1X,15A1)
>CONTINUE;
>STOP
```

Programme pour l'unidimensionnalité des tâches discrètes, un facteur général

```
>Title
 Fonctionnement
 des tâches discrètes (1 ensemble de textes);
>PROBLEM NITEM=15,RESPONSE=5,SUBTEST=1;
>NAMES A1,A2,A3,A4,A5,B1,B2,B3,B4,B5,C1,C2,
C3,C4,C5,;
>RESPONSES ' ','a','b','c','d';
>KEY cddcbcbdcbcbdca;
>RELIABILITY ALPHA;
>PLOT BISERIAL,NOCRITERION,FACILITY;
>BIFACTOR NIGROUPS=1,
IGROUPS=(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),LIST=3,CPARMS=(0.25(0)15),OMIT=MISS;
>SCORE CHANCE;
>SUBTEST BOUNDARY=(15), NAME=(3 textes);
>INPUT NIDW=3,SCORE,FILE='donnees.DAT';
(3A1,1X,15A1)
>CONTINUE;
>STOP
```

Programme pour l'unidimensionnalité des tâches intégrées, trois textes intégrés.

```
>Title
 Fonctionnement
 des tâches intégrées (3 textes);
>PROBLEM NITEM=15,RESPONSE=5,SUBTEST=3;
>NAMES D1,D2,D3,D4,D5,E1,E2,E3,E4,E5,F1,
F2,F3,F4,F5 ;
>RESPONSES ' ','a','b','c','d';
>KEY acaabaabdbbcddc;
>RELIABILITY ALPHA;
>PLOT BISERIAL,NOCRITERION,FACILITY;
>BIFACTOR NIGROUPS=3,
IGROUPS=(1,1,1,1,1,2,2,2,2,2,3,3,3,3,3),
LIST=3,CPARMS=(0.25(0)15), OMIT=MISS;
>SCORE CHANCE;
>SUBTEST BOUNDARY=(5,10,15), NAME=(Texte1, Texte2, Texte3);
>INPUT NIDW=3,SCORE,FILE='integrees.DAT';
(3A1,1X,15A1)
>CONTINUE;
>STOP
```

Programme pour l'unidimensionnalité des tâches intégrées avec un facteur général et un ensemble de trois textes.

```
>Title
 Fonctionnement
 des tâches intégrées (1 ensemble de 3 textes);
>PROBLEM NITEM=15,RESPONSE=5,SUBTEST=0;
>NAMES D1,D2,D3,D4,D5,E1,E2,E3,E4,E5,F1,
F2,F3,F4,F5 ;
>RESPONSES ' ','a','b','c','d';
>KEY acaabaabdbbcddc;
>RELIABILITY ALPHA;
>PLOT BISERIAL,NOCRITERION,FACILITY;
>BIFACTOR NIGROUPS=1,
IGROUPS=(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),
LIST=3,CPARMS=(0.25(0)15), OMIT=MISS;
>SCORE CHANCE;
>SUBTEST BOUNDARY=(15), NAME=(3Textes);
>INPUT NIDW=3,SCORE,FILE='integrees.DAT';
```

```
(3A1,16X,15A1)
>CONTINUE;
>STOP
```

Programme pour l'unidimensionnalité des tâches intégrées avec un facteur général

```
>Title
 Fonctionnement
 des tâches intégrées (1 ensemble de 3 textes);
>PROBLEM NITEM=15,RESPONSE=5,SUBTEST=0;
>NAMES D1,D2,D3,D4,D5,E1,E2,E3,E4,E5,F1,
F2,F3,F4,F5 ;
>RESPONSES ' ','a','b','c','d';
>KEY acaabaabdbbcddc;
>RELIABILITY ALPHA;
>PLOT BISERIAL,NOCRITERION,FACILITY;
>BIFACTOR NIGROUPS=1,
IGROUPS=(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
LIST=3,CPARMS=(0.25(0)15), OMIT=MISS;
>SCORE CHANCE;
>SUBTEST BOUNDARY=(15), NAME=(intégrées);
>INPUT NIDW=3,SCORE,FILE='donnees.DAT';
(3A1,16X,15A1)
>CONTINUE;
>STOP
```

ANNEXE 9

Résultats de la calibration des items et des personnes du questionnaire portant sur la perception subjective de la difficulté

TABLE 3.1 Test de compréhension écrite ZOU032WS.TXT Aug 22 20:36 2006
INPUT: 118 personnes, 14 questions MEASURED: 99 personnes, 14 questions, 5 CATS

SUMMARY OF 98 MEASURED (NON-EXTREME) personnes

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	41.2	13.3	.32	.45	.98	-.2	.99	-.2
S.D.	9.9	1.8	1.69	.09	.60	1.6	.60	1.6
MAX.	65.0	14.0	4.70	1.28	2.80	3.4	2.78	3.4
MIN.	9.0	2.0	-3.32	.42	.21	-3.2	.20	-3.2
REAL RMSE	.51	ADJ.SD	1.61	SEPARATION	3.16	person RELIABILITY		.91
MODEL RMSE	.46	ADJ.SD	1.62	SEPARATION	3.51	person RELIABILITY		.92
S.E. OF personne MEAN = .17								

MAXIMUM EXTREME SCORE: 1 personnes
LACKING RESPONSES: 14 personnes
DELETED: 5 personnes
VALID RESPONSES: 95.1%

SUMMARY OF 99 MEASURED (EXTREME AND NON-EXTREME) personnes

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	41.4	13.3	.39	.47				
S.D.	10.0	1.8	1.83	.17				
MAX.	65.0	14.0	7.73	1.85				
MIN.	9.0	2.0	-3.32	.42				
REAL RMSE	.54	ADJ.SD	1.75	SEPARATION	3.25	person RELIABILITY		.91
MODEL RMSE	.50	ADJ.SD	1.77	SEPARATION	3.56	person RELIABILITY		.93
S.E. OF personne MEAN = .19								

personne RAW SCORE-TO-MEASURE CORRELATION = .70 (approximate due to missing data)
CRONBACH ALPHA (KR-20) personne RAW SCORE RELIABILITY = .95 (approximate due to missing data)

SUMMARY OF 14 MEASURED (NON-EXTREME) questions

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	288.7	93.2	.00	.17	.99	-.1	1.00	.0
S.D.	28.0	4.2	.77	.00	.15	1.0	.13	.9
MAX.	335.0	97.0	1.46	.18	1.27	1.8	1.25	1.6
MIN.	244.0	83.0	-1.27	.16	.75	-1.8	.76	-1.8
REAL RMSE	.17	ADJ.SD	.75	SEPARATION	4.38	questi RELIABILITY		.95
MODEL RMSE	.17	ADJ.SD	.75	SEPARATION	4.50	questi RELIABILITY		.95
S.E. OF question MEAN = .21								

UMEAN=.000 USCALE=1.000
question RAW SCORE-TO-MEASURE CORRELATION = -.84 (approximate due to missing data)
1305 DATA POINTS. APPROXIMATE LOG-LIKELIHOOD CHI-SQUARE: 2433.47

Annexe 10
Analyses préliminaires

Dans un but de clarté, les analyses liées à la calibration des items, à l'analyse des variables sociodémographiques ou encore à l'effet de fatigue ont été placées dans un chapitre différent de celui des analyses reliées directement aux questions de recherche. Ce travail d'analyses préliminaires est essentiel avant même de pouvoir répondre aux questions de recherche.

Calibration des questions et des items pour l'ensemble des questions et des items

Résultats après la première calibration

Tableau A.10.1 : mesures de tendance centrale, de dispersion et de fiabilité pour les personnes et les questions après une première calibration

SUMMARY OF 118 MEASURED personnes

	RAW			MODEL		INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	16.4	28.8	.34	.45	1.00	.1	.99	.0
S.D.	5.6	3.7	.95	.09	.15	.8	.30	.9
MAX.	29.0	30.0	3.78	1.04	1.33	2.1	2.50	3.2
MIN.	5.0	10.0	-1.83	.39	.63	-1.6	.21	-1.6
REAL RMSE	.47	ADJ.SD	.83	SEPARATION	1.78	person	RELIABILITY	.76
MODEL RMSE	.45	ADJ.SD	.84	SEPARATION	1.85	person	RELIABILITY	.77
S.E. OF personne MEAN = .09								

VALID RESPONSES: 96.1%

personne RAW SCORE-TO-MEASURE CORRELATION = .93 (approximate due to missing data)

CRONBACH ALPHA (KR-20) personne RAW SCORE RELIABILITY = .82 (approximate due to missing data)

SUMMARY OF 30 MEASURED items

	RAW			MODEL		INFIT		OUTFIT	
	SCORE	COUNT	MEASURE	ERROR	MNSQ	ZSTD	MNSQ	ZSTD
MEAN	64.4	113.4	.00	.22	1.00	.0	.99	.0
S.D.	18.9	2.2	.87	.02	.10	1.2	.17	1.1
MAX.	94.0	118.0	2.39	.28	1.15	2.0	1.42	1.8
MIN.	17.0	109.0	-1.36	.20	.80	-2.8	.71	-2.3
REAL RMSE	.22	ADJ.SD	.84	SEPARATION	3.75	item	RELIABILITY	.93
MODEL RMSE	.22	ADJ.SD	.84	SEPARATION	3.83	item	RELIABILITY	.94
S.E. OF item MEAN = .16								

UMEAN=.000 USCALE=1.000

Une première calibration des données a été faite à partir des données brutes, soit 30 items et 118 candidats[1]. Cette calibration est le fruit de cinq itérations menées avec le logiciel WINSTEPS. Comme on peut le voir dans le tableau A.10.1, pour les personnes, la mesure maximale est de 3,78 et la mesure minimale de -1,83 *logits*. Pour ce qui est des items, le maximum de 2,39 et le minimum de -1,36. L'indice de fidélité KR-20[2] des personnes (« *reliability* ») signalé par WINSTEPS est de 0,82. Pour les items, le logiciel TESTFACT signale une valeur de l'indice alpha de 0,78 (l'indice de fidélité calculé par WINSTEPS signale une valeur de 0,93).

Ajustement des items après la première calibration

Après avoir consulté le résumé des statistiques, l'ajustement des données par rapport au modèle pour les items (tableau A.10.2) sera étudié.

Tableau A.10.2 : ajustement des données par rapport au modèle pour les items après la première calibration

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	item
23	17	114	2.39	.28	1.10	.6	1.42	1.4	A .22	84.2	85.9	E3
11	23	113	1.95	.25	1.09	.6	1.29	1.3	B .26	82.3	81.2	C1
20	**54**	**117**	**.51**	**.20**	**1.15**	**2.0**	**1.19**	**1.8**	**C .24**	**59.8**	**66.7**	**D5**
5	**58**	**112**	**.26**	**.21**	**1.15**	**2.0**	**1.18**	**1.6**	**D .25**	**62.5**	**66.7**	**A5**
30	69	109	-.30	.22	1.11	1.3	1.17	1.2	E .26	65.1	69.3	F5
3	41	112	.96	.21	1.09	1.1	1.15	1.2	F .30	64.3	70.1	A3
25	**48**	**113**	**.70**	**.21**	**1.15**	**1.9**	**1.11**	**1.0**	**G .26**	**58.4**	**67.4**	**E5**
12	76	112	-.53	.22	1.08	.9	1.15	.9	H .27	70.5	71.5	C2
26	44	113	.87	.21	1.08	.9	1.11	.9	I .32	66.4	69.0	F1
1	94	116	-1.36	.25	1.11	.8	1.08	.4	J .20	77.6	81.2	A1
8	53	114	.50	.20	1.08	1.0	1.09	.9	K .32	58.8	66.8	B3
16	69	116	-.14	.21	1.06	.9	1.04	.4	L .32	62.1	67.9	D1
28	64	112	.01	.21	1.04	.5	.99	.0	M .36	66.1	67.5	F3
15	54	110	.38	.21	1.01	.1	1.02	.3	N .39	60.9	66.7	C5
4	78	114	-.58	.22	.93	-.7	1.02	.2	O .40	71.9	71.7	A4
22	79	116	-.56	.21	1.00	.1	.95	-.3	o .36	73.3	71.5	E2
19	32	113	1.39	.23	.99	.0	.96	-.2	n .39	72.6	75.0	D4
17	67	118	-.01	.20	.99	-.1	.95	-.4	m .40	62.7	67.1	D2
9	76	113	-.51	.22	.98	-.1	.92	-.5	l .39	70.8	71.0	B4
7	82	112	-.84	.23	.96	-.3	.88	-.5	k .38	75.9	74.8	B2
14	80	112	-.73	.22	.95	-.5	.89	-.6	j .40	74.1	73.7	C4
13	86	111	-1.11	.24	.94	-.4	.76	-1.0	i .40	76.6	78.2	C3
18	74	117	-.32	.21	.94	-.7	.88	-.9	h .43	70.1	69.0	D3
27	75	113	-.46	.21	.94	-.7	.87	-.8	g .43	74.3	70.6	F2
29	55	110	.35	.21	.93	-.9	.88	-1.2	f .47	67.3	66.7	F4
10	66	112	-.09	.21	.89	-1.5	.82	-1.5	e .50	70.5	67.9	B5
6	80	114	-.68	.22	.88	-1.3	.76	-1.4	d .49	75.4	72.7	B1
21	91	114	-1.25	.25	.87	-.9	.72	-1.2	c .45	83.3	80.2	E1
2	70	116	-.18	.21	.80	-2.8	.73	-2.3	b .57	75.9	68.1	A2
24	78	113	-.60	.22	.80	-2.3	.71	-1.8	a .56	75.2	72.1	E4
MEAN	64.4	113.4	.00	.22	1.00	.0	.99	.0		70.3	71.6	
S.D.	18.9	2.2	.87	.02	.10	1.2	.17	1.1		7.1	5.1	

[1] Chaque changement apporté aboutit à une calibration différente des candidats et des questions.
[2] Cet indice de fidélité est conçu pour des données dichotomiques.

Tout d'abord, il convient de faire le constat qu'aucun des items n'a de corrélations point-bisérielles négatives avec le score total (noté « *PTMEACORR.* » dans le logiciel, Linacre (2005 : 308)). En revanche, six items ont des corrélations ayant une valeur de moins de 0,3, soit les items 23, 11, 20, 5, 30 et 12. Aucun des items ne présente de valeur du carré moyen de l'*infit* en dehors de l'intervalle [0,75 ;1,3]. La valeur la plus élevée est de 1,15 pour les items 5, 20 et 25. On constate que les items 20, 5, 25, 2 et 24 présentent des valeurs standardisées signalant plus de variabilité qu'attendu pour les items 20, 5 et 25 et moins de variabilité qu'attendu pour les items 2 et 24.

Compte tenu des objectifs de la recherche, des contraintes imposées et des valeurs de l'indice d'*infit*, il ne nous apparaît pas pertinent d'envisager l'élimination des items moins adéquats par rapport au modèle.

Ajustement des personnes après la première calibration et calibrations suivantes

La corrélation point-bisérielle, indique que les scores des personnes 18 et 107 ne sont pas reliés à l'ensemble du test. Les indices de l'*outfit* du candidat 18 et de l'*infit* et de l'*outfit* pour le candidat 107 nous signalent plus de variabilité qu'attendu. Par conséquent, ils sont éliminés de l'échantillon. Le sujet 81, ayant un carré moyen *infit* d'une valeur de 1,31, est lui aussi ôté de l'échantillon.

Tableau A.10.3 : ajustement des personnes après la première calibration

INPUT118 personnes, 30 items MEASURED 118 personnes, 30 items, 12 CATS 3.60.1

personne: REAL SEP..78 REL.: .76 ... item: REAL SEP.: 3.75 REL.: .93

personne STATISTICS: MISFIT ORDER

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%
18	**8**	**30**	**-1.18**	**.43**	**1.33**	**1.6**	**2.50**	**3.2**	**A -.23**	**73.3**	**73.7**
107	**7**	**30**	**-1.37**	**.45**	**1.15**	**.7**	**1.97**	**2.1**	**B -.02**	**76.7**	**76.6**
27	7	22	-.82	.48	1.08	.5	1.89	2.3	C .11	68.2	70.3
113	10	30	-.82	.41	1.18	1.2	1.82	2.5	D .04	60.0	68.7
93	12	30	-.50	.40	1.27	2.0	1.53	2.1	E .00	50.0	65.2
1	10	30	-.82	.41	1.12	.8	1.43	1.5	F .14	73.3	68.7
13	7	18	-.52	.52	1.24	1.4	1.38	1.4	G .04	61.1	66.2
81	**17**	**30**	**.28**	**.40**	**1.31**	**2.0**	**1.38**	**1.9**	**H .04**	**56.7**	**67.1**
85	14	29	-.13	.40	1.28	2.1	1.37	1.8	I .05	44.8	65.0
MEAN	16.4	28.8	.34	.45	1.00	.1	.99	.0		70.0	71.5
S.D.	5.6	3.7	.95	.09	.15	.8	.30	.9		10.7	6.9

Après le retrait de l'échantillon de ces trois personnes, les données sont à nouveau calibrées. Les candidats 93 et 85 montrent de sérieux problèmes d'ajustements (tableau A.10.4). Par ailleurs, le candidat 93 a une corrélation point-bisérielle négative. Il a donc été décidé d'ôter ces deux candidats de l'échantillon.

Tableau A.10.4 : ajustement des données par rapport au modèle pour les personnes après la deuxième calibration

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%
27	7	22	-.83	.49	1.08	.5	1.98	2.4	A .11	68.2	70.4
113	10	30	-.83	.41	1.19	1.2	1.92	2.6	B .03	60.0	68.8
93	**12**	**30**	**-.51**	**.40**	**1.30**	**2.1**	**1.61**	**2.3**	**C-.02**	**50.0**	**65.7**
1	10	30	-.83	.41	1.12	.8	1.50	1.6	D .14	73.3	68.8
85	**14**	**29**	**-.13**	**.40**	**1.31**	**2.2**	**1.43**	**2.0**	**E .05**	**48.3**	**65.3**
13	7	18	-.52	.52	1.27	1.5	1.43	1.4	F .03	55.6	66.6
MEAN	16.5	28.8	.37	.45	1.00	.1	.98	.0		69.9	71.8
S.D.	5.6	3.7	.95	.09	.15	.8	.26	.9		10.9	6.9

A l'issue de la troisième calibration, un nouvel examen des statistiques de l'ajustement des données par rapport au modèle pour les personnes montre que toutes les personnes ont des corrélations point-bisérielles positives (tableau A.10.5).

Tableau A.10.5 : ajustement des données par rapport au modèle des personnes après la troisième calibration

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%
27	7	22	-.82	.49	1.08	.5	1.97	2.4	A .12	68.2	70.3
113	10	30	-.84	.41	1.18	1.2	1.92	2.6	B .05	60.0	68.8
1	10	30	-.84	.41	1.14	1.0	1.61	1.8	C .12	73.3	68.8
13	**7**	**18**	**-.52**	**.52**	**1.30**	**1.6**	**1.45**	**1.4**	**D .02**	**55.6**	**67.1**
66	15	30	-.04	.40	1.13	1.0	1.38	1.8	E .21	66.7	65.9

Toutefois, le sujet 13 a atteint la limite de 1,30 pour le carré moyen de l'infit. Le candidat ayant un niveau de compétence moyen (-0,52), il semble que l'on puisse se fier à la valeur de l'*infit* standardisé. Cette valeur (1,6) n'étant pas supérieure à 2, la décision est prise de laisser le sujet dans l'échantillon. Par ailleurs, l'élimination de l'échantillon du candidat, n'améliore ni l'ajustement des personnes et ni celui des questions.

Les personnes 27, 113, 1, 13 et 66 présentent de sérieux problèmes d'ajustement pour la valeur du carré moyen de la statistique *outfit*. Les personnes 27 et 113 ont même des valeurs standardisées supérieures à 2. Toutefois, parce que ces candidats présentent des valeurs *infit* satisfaisantes, ils n'ont pas été retirés de l'échantillon.

Afin de vérifier l'éventuel impact que pourrait avoir le retrait des cinq candidats (18, 81, 85, 93, 107) sur la qualité des données, il a été décidé de vérifier l'appartenance à des groupes linguistiques des candidats. Comme on peut le constater dans le tableau A.10.6, le retrait des cinq candidats ne semble pas lié à leurs profils linguistiques.

Tableau A.10.6 : appartenance à des groupes linguistiques des candidats retirés de l'échantillon final

Groupe Linguistique	Numéro du candidat
Langue latine	81
Langues slaves	85 – 107
Sinophones	18
Autre langue (bilingue roumain - ukrainien)	93

Description de l'échantillon et des items après la calibration finale

A l'issue de la dernière calibration, les données des personnes (113 individus au total) s'ajustent de manière satisfaisante au modèle. Pour ce qui est des items, tous ont des valeurs pour les carrés moyens *infit* satisfaisantes, comprises dans l'intervalle [0,75 ; 1,3] (tableau A.10.7). Le résumé des statistiques pour les personnes et les données fournies pour les items par le logiciel TESTFACT révèle que les indices de fidélité KR-20 des personnes et l'indice alpha pour les items présentent des valeurs similaires à celles obtenues après la première calibration, soit 0,82 pour les personnes et 0,78 pour les items. Concernant la calibration des items (tableau A.10.7), l'ensemble des items présente des indices d'ajustement acceptables. Toutefois, cinq items présentent des corrélations point-bisérielles d'une valeur de moins de 0,3. Il a été décidé de procéder à un examen attentif de ces items à partir des statistiques de l'ajustement des données par rapport au modèle et

de l'examen du fonctionnement des leurres (Annexe 6). Après un examen attentif de ces cinq items, il a été décidé qu'ils avaient des qualités suffisantes pour être conservés.

Tableau A.10.7 : m*isfit* des items après la calibration finale

INPUT 118 personnes, 30 items MEASURED 113 personnes, 30 items, 12 CATS 3.60.1

personne: REAL SEP..77 REL.: .76 ... item: REAL SEP.: 3.87 REL.: .94

ENTRY NUMBER	RAW SCORE	COUNT	MODEL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	EXACT MATCH OBS%	EXACT MATCH EXP%	item	G
20	54	112	.45	.21	1.18	**2.4**	1.23	**2.1**	**A .21**	57.1	66.5	D5 Intégrées	4
23	16	109	2.45	.29	1.09	.5	1.22	.8	**B .25**	84.4	86.2	E3 Intégrées	5
30	66	104	-.27	.22	1.13	1.5	1.19	1.3	**C .24**	64.4	69.5	F5 Intégrées	6
12	75	107	-.61	.23	1.08	.8	1.18	1.0	**D .25**	72.9	72.8	C2 Discrètes	3
5	56	108	.29	.21	1.15	**1.9**	1.16	1.4	**E .25**	60.2	66.7	A5 Discrètes	1
25	47	108	.70	.21	1.16	**2.0**	1.13	1.1	**F .26**	57.4	67.2	E5 Intégrées	5
1	90	111	-1.32	.26	1.08	.6	1.04	.2	**G .23**	78.4	81.4	A1 Discrètes	1
16	66	111	-.10	.21	1.07	1.0	1.05	.4	H .31	61.3	67.9	D1 Intégrées	4
8	51	109	.52	.21	1.06	.9	1.07	.7	I .34	58.7	66.8	B3 Discrètes	2
4	77	109	-.66	.23	.95	-.5	1.07	.4	J .38	72.5	72.9	A4 Discrètes	1
28	63	107	-.03	.21	1.06	.8	1.01	.1	K .34	66.4	67.9	F3 Intégrées	6
26	40	108	1.01	.22	1.03	.4	1.05	.5	L .37	70.4	70.0	F1 Intégrées	6
15	52	105	.41	.21	1.02	.3	1.05	.4	M .38	61.0	66.7	C5 Discrètes	3
3	37	107	1.11	.22	1.03	.4	1.01	.2	N .37	66.4	71.3	A3 Discrètes	1
22	77	111	-.59	.22	1.02	.3	.97	-.1	O .34	73.0	72.2	E2 Intégrées	5
11	19	108	2.21	.27	1.01	.1	1.00	.1	o .35	81.5	83.5	C1 Discrètes	3
9	76	108	-.63	.23	1.01	.2	.95	-.2	n .35	71.3	72.9	B4 Discrètes	2
19	31	108	1.41	.23	1.00	.1	.98	-.1	m .39	72.2	74.8	D4 Intégrées	4
7	80	107	-.89	.24	.99	.0	.90	-.4	l .36	75.7	76.0	B2 Discrètes	2
14	79	107	-.82	.24	.98	-.1	.93	-.3	k .36	74.8	75.3	C4 Discrètes	3
17	64	113	.04	.21	.98	-.3	.94	-.5	j .41	62.8	67.2	D2 Intégrées	4
27	73	108	-.48	.22	.97	-.3	.89	-.6	i .40	73.1	71.2	F2 Intégrées	6
13	84	106	-1.19	.25	.94	-.4	.75	-1.0	h .39	78.3	79.7	C3 Discrètes	3
29	54	105	.32	.21	.93	-.9	.88	-1.1	g .47	65.7	66.8	F4 Intégrées	6
18	70	112	-.25	.21	.92	-1.1	.85	-1.1	f .46	72.3	68.8	D3 Intégrées	4
6	80	109	-.82	.23	.90	-.9	.76	-1.2	e .46	77.1	74.9	B1 Discrètes	2
10	65	107	-.14	.21	.90	-1.3	.83	-1.3	d .48	71.0	68.5	B5 Discrètes	2
21	88	109	-1.27	.26	.90	-.7	.73	-1.0	c .42	83.5	81.1	E1 Intégrées	5
24	77	108	-.68	.23	.82	-1.9	.72	-1.6	b .53	74.1	73.4	E4 Intégrées	5
2	68	111	-.18	.21	.76	-3.4	.68	-2.7	a .62	78.4	68.5	A2 Discrètes	1
MEAN	62.5	108.4	.00	.23	1.00	.1	.97	-.1		70.5	72.3		
S.D.	19.0	2.1	.93	.02	.10	1.2	.15	1.0		7.6	5.4		

Calibrations séparées des items des tâches discrètes et intégrées

L'objectif de cette partie de l'analyse est de savoir si les calibrations, « tout item » et « items des tâches discrètes et des tâches intégrées » produisent des résultats différents ou similaires ce qui donnera une orientation différente à leur interprétation. Pour les calibrations séparées des items des tâches discrètes et intégrées, il a été décidé d'utiliser l'échantillon des 113 personnes retenues à l'issue de la calibration de l'ensemble des items. Quand bien même des items ou des personnes présenteraient des problèmes d'ajustement des données par rapport au modèle, on ne procédera à aucune suppression

de données supplémentaire. Ces problèmes d'ajustement seront interprétés comme des indications de fonctionnements différents pour des personnes, des groupes de personnes selon le type d'items administrés, soit des items issus des tâches discrètes ou des items issus des tâches intégrées.

Calibration séparée des items des tâches discrètes

Lorsque les items des tâches discrètes sont calibrés séparément, ils ont tous des statistiques de l'ajustement des données par rapport au modèle satisfaisantes (tableau A.10.8). Seul l'item 2 a une valeur proche du seuil de 0,75, soit 0,77 et une valeur standardisée supérieure à –2, soit –2,7. L'item 2 présente donc moins de variabilité qu'attendu par le modèle. L'indice de fidélité alpha des items, signalé par le logiciel TESTFACT, a une valeur de 0,66. Pour ces 15 items, les candidats ont une compétence moyenne de 0,58 *logit*.

Tableau A.10.8 : statistiques générales des items calibrés séparément

SUMMARY OF 111 MEASURED (EXTREME AND NON-EXTREME) personnes

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	8.9	14.6	.58	.68				
S.D.	3.1	1.7	1.17	.20				
MAX.	15.0	15.0	4.46	1.88				
MIN.	1.0	4.0	-1.66	.56				

REAL RMSE	.74	ADJ.SD	.91	SEPARATION	1.24	person RELIABILITY	.61
MODEL RMSE	.71	ADJ.SD	.94	SEPARATION	1.33	person RELIABILITY	.64

S.E. OF personne MEAN = .11

personne RAW SCORE-TO-MEASURE CORRELATION = .95 (approximate due to missing data)

CRONBACH ALPHA (KR-20) personne RAW SCORE RELIABILITY = .72 (approximate due to missing data)

SUMMARY OF 15 MEASURED (NON-EXTREME) items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	63.9	105.9	.00	.24	1.00	.0	.99	-.1
S.D.	18.9	1.6	.97	.02	.09	1.0	.15	1.0
MAX.	88.0	109.0	2.53	.28	1.14	1.7	1.26	1.1
MIN.	17.0	103.0	-1.24	.22	.77	-2.7	.71	-2.5

REAL RMSE	.24	ADJ.SD	.94	SEPARATION	3.89	item RELIABILITY	.94
MODEL RMSE	.24	ADJ.SD	.94	SEPARATION	3.98	item RELIABILITY	.94

S.E. OF item MEAN = .26

Les personnes ont un indice de fidélité KR-20 signalé par WINSTEPS d'une valeur de 0,72. Pour les personnes, les statistiques d'adéquation nous signalent que sept individus ont des mesures qui contredisent l'ensemble des mesures du test et ont des corrélations point-bisérielles négatives.

Bien que ce groupe de sept personnes soit important (il représente 5% de l'échantillon) aucun schéma ne semble se dégager. Deux personnes sont des sinophones, trois appartiennent au groupe des langues latines. Les deux personnes restantes appartiennent au groupe des langues slaves. L'examen de leurs scalogrammes (figure A.10.1) nous apprend que les candidats 41, 89, 78 semblent avoir commis des fautes d'étourderie. Le candidat 96 semble avoir focalisé son attention sur un texte et une tâche en particulier puisqu'il a été essentiellement capable de répondre aux items 11 à 14. Si ces items correspondent aux items du texte le plus difficile, ce sont aussi les premiers items auxquels le candidat devait répondre. Le candidat 14, lui, n'a pas réussi à répondre aux items les plus faciles. Peut-être a-t-il décidé de consacrer plus de temps aux questions difficiles. Concernant les autres candidats, il est difficile de se prêter à une interprétation.

Figure A.10.1 : scalogramme des personnes signalées misfit pour les tâches discrètes calibrées séparément

```
GUTTMAN SCALOGRAM OF
RESPONSES:
personne |item
         | 1  1  1 1 1  1
         |137644922055831
         |---------------
       1 +101000010100001
      10 +000000111011000
      41 +111011101111011
      89 +110111011101111
      78 +011101011100011
      14 +000001111111100
      96 +010010110110001
         |---------------
         | 1  1  1 1 1  1
         |137644922055831
```

Tableau A.10.9 : moyenne aux tâches discrètes par groupe de langue

person COUNT	MEAN MEASURE	S.E. MEAN	OBSERVED S.D.	MEDIAN	REAL SEPARATION	CODE
111	.58	.11	1.17	.43	1.24	*
11	-.08	.35	1.12	.10	1.10	A
25	-.54	.13	.65	-.53	.29	C
56	1.14	.14	1.01	1.17	.83	L
19	.76	.22	.93	.78	.77	S

UMEAN=0 USCALE=1

légende

A=multilingues ; C=sinophones ; L= Langues latines ; S= langues slaves

Concernant le niveau des différents groupes linguistiques (la répartition en groupes linguistiques est expliquée au point 5.3.4), le groupe des langues latines (tableau 5.9) est celui a qui a le meilleur niveau pour les tâches discrètes calibrées séparément. Suivent, dans l'ordre, le groupe des langues slaves, le groupe multilingue et enfin celui des sinophones.

Calibration séparée des tâches intégrées

Lorsque les items des tâches intégrées sont calibrés séparément, ils présentent des statistiques d'adéquation, *infit* et *outfit* satisfaisantes (tableau A.10.10). Seuls les items 8 et 15 présentent des statistiques *outfit* avec des valeurs élevées (respectivement, 1,41 et 1,35) mais ces valeurs ne sont pas associées à des statistiques de l'*infit* significativement élevées. L'indice de fidélité alpha signalé par le logiciel TESTFACT pour les items a une valeur de 0,62. Pour les 15 items, la moyenne des candidats est de 0,21 *logit*.
L'indice de fidélité alpha signalé par WINSTEPS (tableau A.10.10) pour les personnes a une valeur de 0,66.

Tableau A.10.10 : statistiques générales pour les items des tâches intégrées calibrées séparément

```
INPUT18 personnes, 15 items  MEASURED 113 personnes, 15 items, 2 CATS   3.60.1
-------------------------------------------------------------------------------
     SUMMARY OF 15 MEASURED items
+-------------------------------------------------------------------------------+
|           RAW                           MODEL         INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-------------------------------------------------------------------------------|
| MEAN      59.1     108.9         .00     .23      1.00     .1   1.00     .0  |
| S.D.      18.5       2.5         .93     .02       .09    1.0    .18     .8  |
| MAX.      88.0     113.0        2.34     .30      1.14    1.7   1.41    1.9  |
| MIN.      16.0     104.0       -1.51     .21       .84   -1.6    .75   -1.1  |
|-------------------------------------------------------------------------------|
| REAL RMSE    .23  ADJ.SD     .90  SEPARATION  3.85  item   RELIABILITY  .94  |
|MODEL RMSE    .23  ADJ.SD     .90  SEPARATION  3.94  item   RELIABILITY  .94  |
| S.E. OF item MEAN = .25                                                       |
+-------------------------------------------------------------------------------+
UMEAN=.000 USCALE=1.000
item RAW SCORE-TO-MEASURE CORRELATION = -.99 (approximate due to missing data)
1633 DATA POINTS. APPROXIMATE LOG-LIKELIHOOD CHI-SQUARE765.35

     SUMMARY OF 113 MEASURED personnes               DELETED:      5 personnes
+-------------------------------------------------------------------------------+
|           RAW                           MODEL         INFIT        OUTFIT     |
|          SCORE     COUNT     MEASURE   ERROR      MNSQ   ZSTD   MNSQ   ZSTD  |
|-------------------------------------------------------------------------------|
| MEAN       7.8      14.5         .21     .64      1.00     .1   1.00     .0  |
| S.D.       2.9       1.7        1.07     .13       .20     .8    .45     .8  |
| MAX.      14.0      15.0        3.05    1.16      1.53    2.1   4.04    2.9  |
| MIN.       1.0       4.0       -2.94     .56       .57   -2.0    .18   -1.6  |
|-------------------------------------------------------------------------------|
| REAL RMSE    .68  ADJ.SD     .83  SEPARATION  1.22  person RELIABILITY  .60  |
|MODEL RMSE    .65  ADJ.SD     .85  SEPARATION  1.30  person RELIABILITY  .63  |
| S.E. OF personne MEAN = .10                                                   |
+-------------------------------------------------------------------------------+
        VALID RESPONSES:  96.3%
personne RAW SCORE-TO-MEASURE CORRELATION = .95 (approximate due to missing data)
CRONBACH ALPHA (KR-20) personne RAW SCORE RELIABILITY = .66 (approximate due to missing
data)
```

En ce qui concerne les personnes, huit ont une corrélation point-bisérielle négative ou égale à zéro. Aucune ne fait partie de celles qui présentaient des corrélations point-

bisérielles négatives pour les tâches discrètes. Comme pour les tâches discrètes calibrées séparément, aucun groupe linguistique ne semble présenter de problèmes d'ajustement des données par rapport au modèle systématiques.

L'examen du scalogramme (figure A.10.2) des candidats présentant des corrélations point-bisérielles négatives nous indique que, si les candidats 3 et 34 sont signalés comme présentant des problèmes d'ajustement par WINSTEPS, c'est parce qu'ils n'ont pas su répondre à des questions faciles alors qu'ils ont des scores élevés. Il semble que ces candidats aient fait preuve d'étourderie. Le candidat 13 a pu répondre à deux des questions les plus difficiles. Cependant, il n'est pas aisé d'interpréter ce résultat. Le candidat a-t-il choisi de répondre au hasard aux questions auxquelles il a choisi de répondre ? Pour le candidat 68, l'essentiel de ses bonnes réponses correspondent au premier et dernier document. Ses mauvaises réponses se concentrent sur le deuxième document. Le candidat 84, lui, a été capable de répondre aux questions des deux premiers documents. Dans son cas, on peut penser qu'il a pu manquer de temps. Le candidat 79 voit ses bonnes réponses concentrées dans le deuxième document et le 42 sur le premier document. Le candidat 113 pourrait avoir répondu au hasard.

Figure A.10.2 : scalogramme des candidats présentant des corrélations point-bisérielles négatives ou égales à zéro pour les tâches intégrées calibrées isolément

```
GUTTMAN SCALOGRAM OF RESPONSES:
personne |item
         |   11  1 1 11
         |697253132450148
         |---------------
       3 +111101111111111  12514ES1 52EL3
      34 +101111011111111  32214RU1452SS3
      13 +   0 11010  11   12503CH5235CC2
      68 +001110110110110  51513ES4431EL2
      84 +110001110011011  52304RO3 31RL2
      79 +111100010001011  52603ES5 32EL3
      42 +001001001000110  31513RO1132RL2
     113 +001001000000001  61 14CH4 52CC1
         |---------------
         |   11  1 1 11
         |697253132450148
```

Pour ce qui est de la compétence moyenne des groupes linguistiques aux tâches intégrées (tableau A.10.11), le groupe des langues latines obtient le meilleur résultat, suivi des slaves, du groupe multilingue et des sinophones. Vu sous cet angle, le classement est identique à celui des tâches discrètes.

Tableau A.10.11 : moyenne aux tâches intégrées par groupe de langue

```
Subtotal specification is: PSUBTOTAL=$S13W1
ALL SCORES ARE NON-EXTREME UMEAN=0 USCALE=1
```

person COUNT	MEAN MEASURE	S.E. MEAN	OBSERVED S.D.	MEDIAN SEPARATION	REAL	CODE
113	.21	.10	1.07	.12	1.22	*
12	-.22	.29	.97	-.30	1.10	A
25	-.54	.16	.81	-.51	.70	C
56	.64	.14	1.03	.61	1.13	L
20	.21	.20	.88	.12	.78	S

légende
A=multilingue ; C=sinophones ; L= Langues latines ; S= langues slaves

Comparaison de la calibration pour tous les items et des deux calibrations séparées

Pour comparer les différences, on a calculé la difficulté des items avec tous les items et ensuite uniquement avec ceux des tâches discrètes et intégrées. On a attribué des rangs aux items de chaque type de tâche selon leur difficulté. Il ressort de cette analyse que les items, quel que soit le type de calibration, occupent toujours strictement le même rang. Que les items soient calibrés tous ensemble où séparément ne change pas le classement des items des taches discrètes d'une part et des items des tâches intégrées d'autre part. De ce point de vue, le résultat des trois calibrations est homogène.

Corrélations entre les différents calibrations du test

L'objectif, ici, est de vérifier, à la fois, la consistance interne du test, mais aussi, le lien entre les mesures associées à des tâches de nature différente. Ces mesures sont établies à partir de calibrations « tout item » ou bien séparées pour les items des tâches discrètes et intégrées ou encore « items pairs » et « items impairs ». Les scores calibrés en *logits* des candidats ont servi à calculer les corrélations.

Tout d'abord, on peut faire le constat que toutes les corrélations sont positives et significatives (tableau A.10.12). Les corrélations entre la compétence des candidats estimée à partir des 30 items du test et les 15 items pairs d'un côté et les 15 items impairs de l'autre sont élevées (les deux corrélations ont une valeur de 0,92). Ces deux corrélations sont légèrement plus élevées que celles trouvées entre l'estimation de la compétence des candidats éffectuée à partir des 30 du test et les estimations faites à partir

des calibrations séparés des tâches discrètes et intégrées (les deux corrélations ont une valeur de 0,89).

Tableau A.10.12 : corrélations de Pearson entre la compétence des candidats estimée à partir de l'ensemble des items (30 items), des 15 items des tâches discrètes ou des 15 items des tâches intégrées et des 15 items pairs et des 15 items impairs

Correlations

		mesure de l'habileté avec l'ensemble des items	mesure de l'habileté avec les items pairs	mesure de l'habileté avec les items impairs	mesure de l'habileté avec les tâches discrètes	mesure de l'habileté avec les tâches intégrées
mesure de l'habileté avec l'ensemble des items	Pearson Correlation	1	,925**	,920**	,897**	,892**
	Sig. (2-tailed)	,	,000	,000	,000	,000
	N	113	113	113	111	113
mesure de l'habileté avec les items pairs	Pearson Correlation	,925**	1	,714**	,822**	,841**
	Sig. (2-tailed)	,000	,	,000	,000	,000
	N	113	113	113	111	113
mesure de l'habileté avec les items impairs	Pearson Correlation	,920**	,714**	1	,837**	,802**
	Sig. (2-tailed)	,000	,000	,	,000	,000
	N	113	113	113	111	113
mesure de l'habileté avec les tâches discrètes	Pearson Correlation	,897**	,822**	,837**	1	,614**
	Sig. (2-tailed)	,000	,000	,000	,	,000
	N	111	111	111	111	111
mesure de l'habileté avec les tâches intégrées	Pearson Correlation	,892**	,841**	,802**	,614**	1
	Sig. (2-tailed)	,000	,000	,000	,000	,
	N	113	113	113	111	113

**. Correlation is significant at the 0.01 level (2-tailed).

La corrélation tâches discrètes / tâches intégrées est la corrélation la plus faible (0,614). On remarquera que le calcul des corrélations avec les items pairs et impairs permet de tenir compte de la longueur du test et donc de fournir des indices comparables à ceux obtenus pour chaque type de tâche. Il appert donc que la corrélation la plus basse est la corrélation entre les items des tâches discrètes et intégrés.

Comme il est possible de le constater dans le diagramme de dispersions (figure A.10.3), le lien entre la mesure de la compétence langagière avec des tâches discrètes et intégrées est plus fort pour le groupe des langues latines et des langues slaves. Pour le groupe des sinophones et des multilingues, la mesure de la compétence pour les tâches discrètes n'offre pas de conditions suffisantes pour faire une prédiction de la mesure des tâches intégrées. Pour ce qui est de la prédiction de la compétence des candidats au test complet par groupe linguistique (figures A.10.4 et A.10.5), on constate que les tâches intégrées sont un peu meilleures que les tâches discrètes (« R carré » plus élevée pour tous les groupes), même si la différence est très faible. Les prédictions sont meilleures pour les

groupes des langues latines et des Slaves que pour les groupes des sinophones et des multilingues. Selon le groupe linguistique, les variables « test complet », « tâches discrètes » et « tâches intégrées » ne permettent donc pas de prédire avec la même précision la compétence des candidats pour l'ensemble du test. Ces trois figures (A.10.3, A.10.4 et A.10.5) montrent encore que le test complet permet d'accroître la fidélité des prédictions du simple fait qu'il est plus long.

Figure A.10.3 : diagramme de dispersion de la mesure de la compétence langagière pour les tâches discrètes et pour les tâches intégrées pour les différents groupes linguistiques

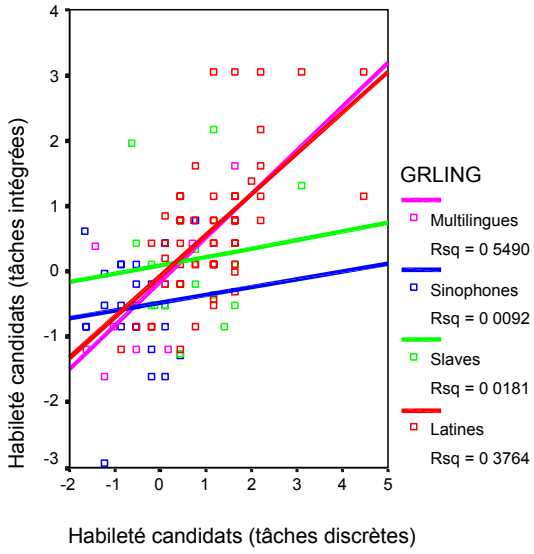


Figure A.10.4 : diagramme de dispersion de la mesure de la compétence pour le test complet et pour les tâches intégrées pour les groupes linguistiques

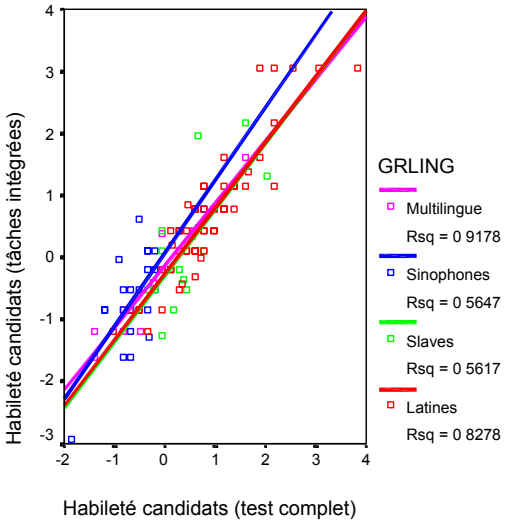
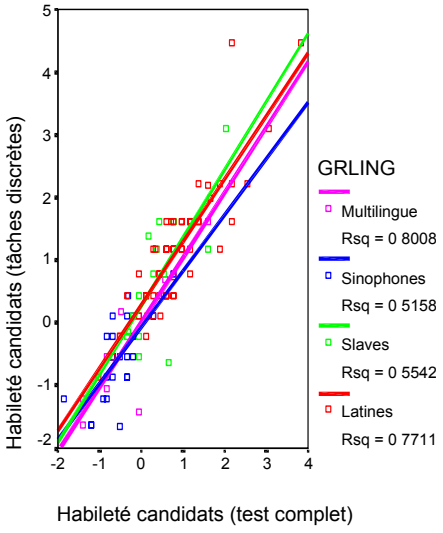


Figure A.10.5 : diagramme de dispersion de la mesure de la compétence pour le test complet et pour les tâches intégrées pour les groupes linguistiques



Normalité des variables mesure des personnes et des items en *logits*

L'objectif de cette partie des analyses préliminaires est de statuer sur la normalité de la distribution des scores, exprimée en *logits*, des personnes et des items calibrés pour l'ensemble des items du test. Pour la distribution des scores en *logits*, il s'agit de savoir si la condition de normalité est suffisante pour pouvoir procéder à des tests statistiques (tests *t*, ANOVA). Pour les items, il s'agit essentiellement de procéder à une description de la distribution pour savoir si on a bien pour les tâches discrètes et intégrées des items de niveau similaire. Il s'agit également de comprendre quels items, quelles tâches sont comparables, ou encore, si les deux parties du test (tâches intégrées et tâches discrètes) ont un niveau de difficulté comparable.

Analyse de la variable score des personnes exprimée en *logits*

Figure A.10.6 : histogramme de la variable score des personnes en *logits*

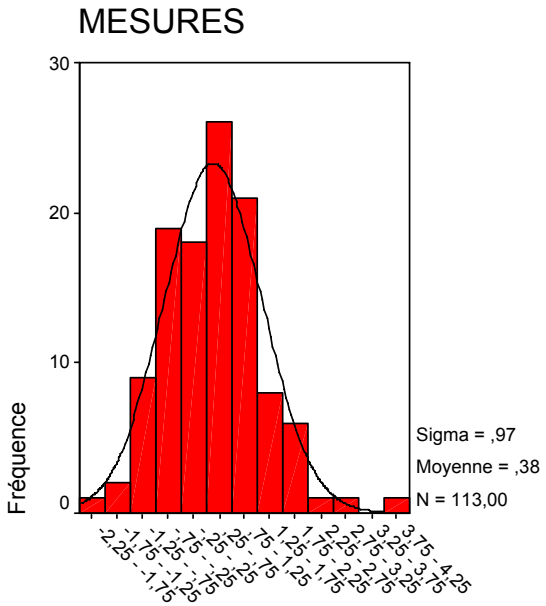


Tableau A.10.13 : variable score des personnes en *logits*

Statistiques

MESURES

N	Valide	113,00
	Manquante	,00
Moyenne		,38
Médiane		,38
Mode		,78
Ecart-type		,97
Variance		,93
Asymétrie		,57
Aplatissement		,94
Minimum		-1,85
Maximum		3,84

Comme on peut le constater dans le tableau A.10.13, la moyenne et la médiane de la distribution des scores des personnes ont la même valeur. En revanche, le mode, a une valeur légèrement différente ce qui semble indiquer un pic dans la distribution (visible sur la figure A.10.6). L'asymétrie et l'aplatissement de la distribution sont ceux d'une distribution normale. La distribution des scores est affectée par des valeurs extrêmes (candidats 36, 101, 111).

Le test de normalité de Kolgomorov-Smimov (tableau 5.14) indique que la distribution des scores n'est pas normale (sig.=0,02). Cependant, le test de Shapiro-Wilk nous indique que la variable est normalement distribuée. Au vu de ces informations, étant donné que l'échantillon n'est que de 113 personnes, on considère que la distribution des scores est « suffisamment » normale pour procéder à l'analyse des données.

Tableau A.10.14 : test de normalité de la variable score des personnes en *logits*

Tests de normalité

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistique	ddl	Signification	Statistique	ddl	Signification
MESURES	,091	113	,022	,978	113	,065

a. Correction de signification de Lilliefors

Les scores étant regroupés autour de la médiane, il est possible d'en conclure que les sujets retenus pour la recherche ont bien un niveau comparable. Toutefois, la distribution des scores en *logits* pour l'ensemble des candidats est suffisamment dispersée pour que l'on considère que les scores sont distribués normalement.

Distribution du niveau des items exprimé en *logits*

Les mesures de tendance centrale (tableau A.10.15) indiquent que la moyenne et la médiane ont des valeurs presque identiques. Si la distribution des items est multimodale, elle est symétrique (valeur proche de 1) et l'aplatissement (kurtosis) est normal.

Tableau A.10.15 : statistiques générales de la variable niveau des items exprimé en *logits*

Statistics

Estimated item Calibration: UMEAN=.00 USCALE=1.00

N	Valid	30
	Missing	0
Mean		,00
Median		-,16
Mode		-1,32[a]
Std. Deviation		,94
Variance		,89
Skewness		,97
Std. Error of Skewness		,43
Kurtosis		,79
Std. Error of Kurtosis		,83
Minimum		-1,32
Maximum		2,45

a. Multiple modes exist. The smallest value is shown

Les tests de normalité (tableau A.10.16) ne concordent pas puisque le test de Kolmogorov-Smimov signale que la variable des scores des items est distribuée normalement, alors que le test de Shapiro-Wilk signale le contraire (le test est toutefois proche du niveau de signification, soit 0,5). Là encore, pour l'analyse, il est possible de dire que les items ont une distribution présentant des conditions de normalité suffisantes.

Tableau A.10.16 : tests de normalité, variable mesure des items

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Estimated item Calibration: UMEAN=.00 USCALE=1.00	,116	30	,200*	,930	30	,048

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Comparaison entre la distribution des items et celle des personnes

Un des objectifs fixés de la recherche était de pouvoir comparer des candidats et des items de niveaux similaires. Il est donc indispensable en préalable à une analyse plus poussée des données de vérifier si les candidats et les items sont bien comparables. L'examen des boîtes à moustaches des variables mesure en *logits* des « items » et « personnes » (figures A.10.7 et A.10.8) nous apprend que toutes deux ont leurs deuxième et troisième quartiles situés entre 1 et -1 *logit*. On constate que la dispersion est moindre pour les personnes que pour les items.

Figure A.10.7 : boîtes à moustaches, variable mesure en *logits* des items et des personnes

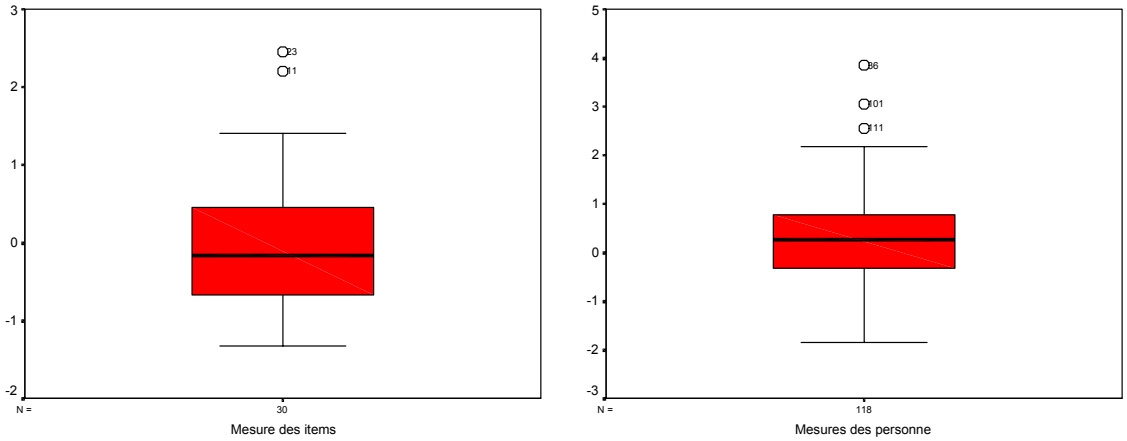
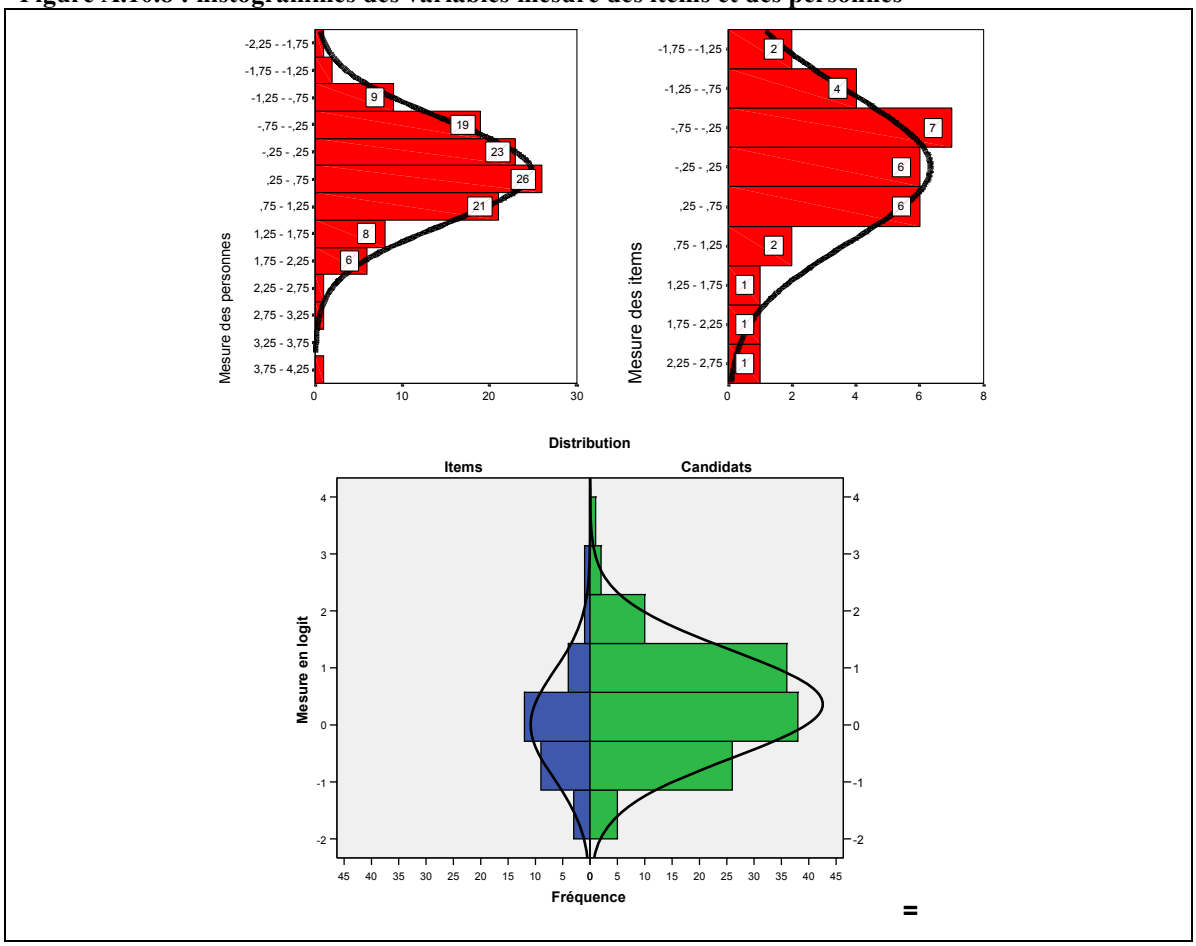


Figure A.10.8 : histogrammes des variables mesure des items et des personnes



L'observation des histogrammes des deux distributions (figure A.10.8) nous permet de constater que l'intervalle [-1,25, 1,25] *logits* est celui qui regroupe le plus de personnes

(98) et celui qui regroupe le plus d'items (25). Les items correspondent donc au niveau des personnes.

Description des caractéristiques sociodémographiques

Après avoir établi la normalité de la distribution des scores, des tests statistiques (test *t*, ANOVA et tests non-paramétriques) ont été utilisés pour procéder à la description des caractéristiques sociodémographiques des personnes de l'échantillon. Les variables « langue parlée dans la rue », « langue parlée au travail », « langue parlée à la maison » et « arrivée au Québec » n'ont pas été retenues dans l'analyse. En effet, elles n'éclairaient en rien les questions de recherche.

Variable sexe

L'échantillon est constitué de 41 hommes et de 72 femmes soit, 36,3 % d'hommes pour 63,7 % de femmes. La moyenne des mesures (en *logit*) des femmes et des hommes est sensiblement la même (hommes=0,43 ; femmes=0,35). Le test *t* pour échantillons indépendants ne permet pas de trouver de différence entre les deux moyennes (tableau A.10.17). La variable sexe ne produit donc pas de différence significative.

Tableau A.10.17 : test *t* score total des personnes-sexe

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes		
		F	Sig.	t	ddl	Sig. (bilatérale)
MESURES	Hypothèse de variances égales	1,14	,29	,41	111,00	,68
	Hypothèse de variances inégales			,43	91,76	,67

Variable âge

Les candidats, en moyenne, ont 33 ans. La moyenne des candidats de chacun des groupes linguistiques est elle aussi de 33 ans. Les moyennes des différents groupes sont statistiquement identiques (tableau A.10.18).

Tableau A.10.18 : fréquence de l'âge

Âge des candidats

	N	Mean
Langues latines	55	33,89
Langues slaves	17	33,35
Sinophones	22	33,64
Multilingues	11	32,45
Total	105	33,60

ANOVA

Âge des candidats

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	20,154	3	6,718	,274	,844
Within Groups	2473,046	101	24,486		
Total	2493,200	104			

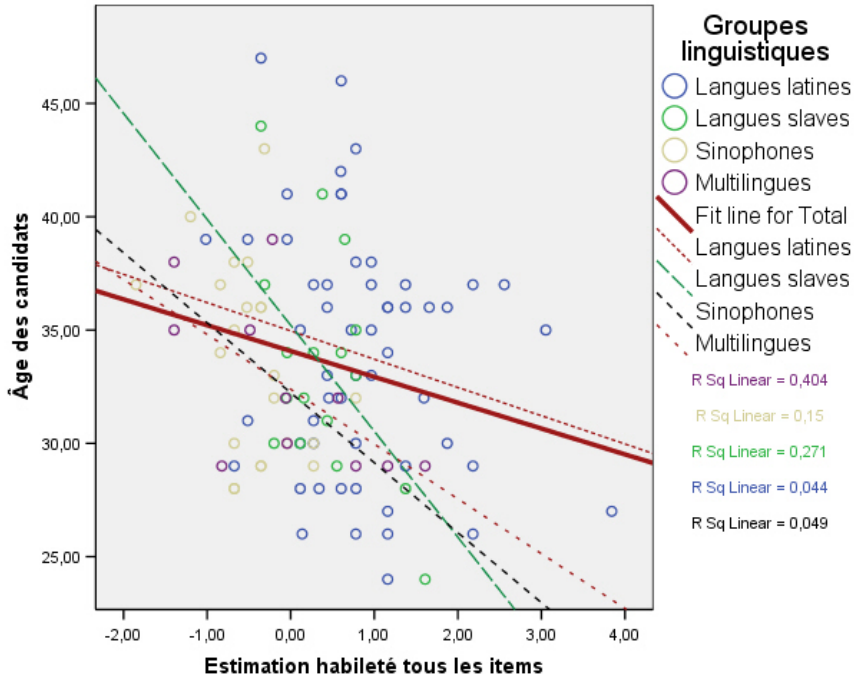
Comme nous le montre le diagramme de dispersion (figure A.10.9), la variable « âge » pour l'ensemble des candidats n'est pas un très bon prédicteur de la variable estimation de la compétence (avec les 30 items) (Rcarré = 0,05). Lorsqu'on examine le lien entre la variable âge et le niveau de compétence de chacun des groupes linguistiques (figure A.10.9) on constate que la variable âge n'est pas un bon prédicteur pour le groupe des langues latines et des sinophones (respectivement, Rcarré=0,05 et Rcarré=0,15). En revanche, pour le groupe des slaves et des multilingues, il est un assez bon prédicteur de performance (respectivement, Rcarré=0,27 et Rcarré=0,40).

Tableau A.10.19 : corrélation entre la variable âge et estimation du niveau de compétence et moyenne des groupes linguistiques

		Âge des candidats	Estimation habileté tous les items
Âge des candidats	Pearson Correlation	1	-,221*
	Sig. (2-tailed)		,023
	N	105	105
Estimation habileté tous les items	Pearson Correlation	-,221*	1
	Sig. (2-tailed)	,023	
	N	105	113

*. Correlation is significant at the 0.05 level (2-tailed).

Figure A.10.9 : diagramme de dispersion de l'âge des candidats



La corrélation, entre la variable « âge » et celle de l'« estimation du niveau de compétence » pour l'ensemble des candidats et moyenne des groupes linguistiques, est négative et significative (tableau A.10.19). Elle nous apprend donc que dans la population des candidats, plus on est âgé, plus le niveau de français est bas. Toutefois, la corrélation est basse et c'est pourquoi l'âge n'explique que très peu l'estimation de la compétence.

On peut donc conclure que la décision de ne pas considérer la variable de l'âge dans la question de recherche était fondée.

Variable langue maternelle

Au total, 12 langues sont parlées par les personnes de l'échantillon (tableau A.10.20). Au vu de la variété des langues parlées par les immigrants au Québec, on peut considérer que la variété linguistique de l'échantillon n'est pas très importante. Elle le devient encore moins si on regroupe ces langues par famille afin de constituer des groupes comparables tant quantitativement que qualitativement.

Tableau A.10.20 : fréquence des langues maternelles

LANGMAT

		Fréquence	Pour cent	Pourcentage valide	Pourcentage cumulé
Valide		2	1,8	1,8	1,8
	Arabe	2	1,8	1,8	3,5
	Bengalî	2	1,8	1,8	5,3
	Bulgare	7	6,2	6,2	11,5
	Chinois	25	22,1	22,1	33,6
	Coréen	3	2,7	2,7	36,3
	Espagnol	35	31,0	31,0	67,3
	Farsi	1	,9	,9	68,1
	PO	1	,9	,9	69,0
	Roumain	20	17,7	17,7	86,7
	Russe-roumain	1	,9	,9	87,6
	Russe	13	11,5	11,5	99,1
	TU	1	,9	,9	100,0
	Total	113	100,0	100,0	

Légende : PO=portugais, TU= turc

Les groupes linguistiques présents dans l'échantillon sont les suivants :

Groupe 1 : langues sémitiques : arabe.

Groupe 2 : langues indo-iraniennes : bengali, farsi.

Groupe 3 : langues slaves : bulgare, russe, ukrainien.

Groupe 4 : langues sino-tibétaines, chinois.

Groupe 5 : langues altaïques : coréen.

Groupe 6 : langues latines : espagnol, roumain et portugais.

Groupe 7 : langue turques : turc.

Groupe 8 : sujets bilingues (dans l'échantillon les bilingues parlent des langues de familles différentes, ce qui nous interdit de les placer dans une famille de langue ou dans une autre).

Pour les besoins de la recherche, les langues maternelles de l'échantillon ont été regroupées de la manière suivante :

Groupe 1 : langues latines : espagnol, roumain et portugais.

Groupe 2 : langues slaves : bulgare, russe.

Groupe 3 : langues sino-tibétaines, chinoises.

Groupe 4 : groupe multilingue langues sémitiques : arabe, bilinguisme non rattaché à une même famille de langue, bengalî, farsi, coréen, turc, valeurs manquantes.

Cette répartition permet d'obtenir trois groupes de langues de familles différentes et un groupe multilingue (tableau A.10.21).

Tableau A.10.21 : moyenne des scores totaux par groupe de langues

```
INPUT 118 personnes, 30 items  MEASURED 113 personnes, 30
items, 12 CATS  3.60.1
-----------------------------------------------------------
Subtotal specification is: PSUBTOTAL=$S13W1
 ALL SCORES ARE NON-EXTREME
+-----------------------------------------------------------+
| person  MEAN     S.E.  OBSERVED  MEDIAN   REAL            |
| COUNT  MEASURE   MEAN    S.D.          SEPARATION  CODE  |
|-----------------------------------------------------------|
|    113     .38     .09     .96     .38    1.77     *     |
|     12    -.10     .28     .93    -.14    1.60   multi.  |
|     25    -.52     .11     .52    -.52     .64 Sinophone|
|     56     .85     .12     .89     .78    1.54  Latines  |
|     20     .47     .14     .62     .41     .82  Slaves   |
+-----------------------------------------------------------+
UMEAN=0 USCALE=1
```

Le groupe des langues latines est le plus nombreux avec 56 personnes, suivi par le groupe des sinophones avec 25 personnes, le groupe des slaves avec 20 personnes et enfin celui des multilingues avec 12 personnes. Le groupe ayant la meilleure moyenne pour l'ensemble du test est celui des langues latines (0,85 *logit*), suivi du groupe des slaves (0,47), du groupe multilingue (-0,10) et des sinophones (-0,52).

Après avoir procédé à la répartition des groupes linguistiques, il convient de vérifier si cette répartition permet de mener à bien les questions de recherche et si les candidats ont bien des niveaux comparables dans chacun des groupes.

Tableau A.10.22 : résultats test ANOVA pour les groupes linguistiques

Test d'homogénéité des variances

Mesures

Statistique de Levene	ddl1=	ddl2	Signification
2,738	3	109	,047

Tests d'égalité des moyennes

MESURES

	Statistique[a]	ddl1=	ddl2	Sig.
Welch	26,20	3	36,80	,00
Brown-Forsythe	20,05	3	42,91	,00

a. Distribution F asymptotique.

ANOVA

MESURES

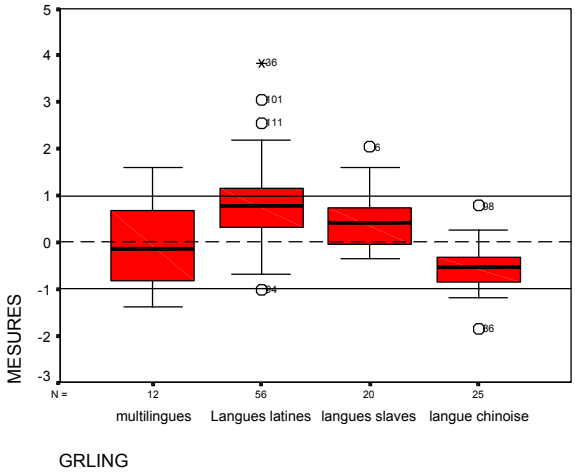
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	35,888	3	11,963	19,021	,000
Within Groups	68,553	109	,629		
Total	104,441	112			

Mesures des associations

	Eta	Eta carré
Mesures * grling	,579	,336

Comme on peut le constater (figure A.10.10), si la dispersion et le niveau de chacun des groupes ne sont pas identiques, des personnes de tous les groupes sont toutefois présentes dans l'intervalle [-0,5 ; 0,5] *logit*.

Figure A.10.10 : boîtes à moustaches de la mesure de la compétence en *logits* à partir de 30 items et des 113 personnes réparties par groupe de langue



Un test ANOVA (tableau A.10.22) indique que les moyennes sont significativement différentes Si tous les groupes n'ont pas 30 sujets, en revanche, le test de Levene pour l'homogénéité des variances indique une variance égale pour tous les groupes linguistiques Certes, le niveau de signification est à la limite mais cela n'est pas étonnant compte-tenu de la taille de l'échantillon.

L'ANOVA pour $F_{(3,109)}$ est égale à 19,021 et la signification est de 0,00. Parce que le test de l'homogénéité des variances est limite, il a été procédé à des tests des moyennes de Welch et de Brown-Forsythe. Ces tests sont préférables à l'utilisation de la statistique *F* lorsqu'on doute de l'homogénéité des variances. L'examen de ces deux tests d'égalité des moyennes confirme le résultat de l'ANOVA. En effet, les deux tests des moyennes sont significatifs à 0,00. Les moyennes sont donc significativement différentes. L'appartenance à un groupe linguistique explique 34 % de la différence des moyennes parmi les personnes (Eta carré = 0,336).

Ayant trouvé que les moyennes étaient significativement différentes, il a été procédé à une analyse *post-hoc* afin de savoir quelle(s) moyenne(s) diffère(nt) des autres (tableau A.10.23). Les différences de moyenne significatives concernent les groupes suivants :

- Celle du groupe des locuteurs d'une langue latine et celle du groupe multilingue.
- Celle du groupe des locuteurs d'une langue latine et celle des sinophones.
- Celle des sinophones et celle des slaves.

Tableau A.10.23 : comparaisons multiples pour l'ANOVA mesure de la compétence en *logits* des candidats regroupés en groupes linguistiques

Comparaisons multiples

Variable dépendante: MESURES

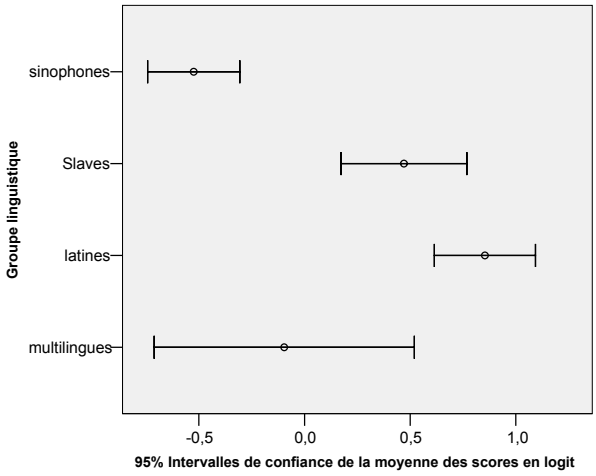
Bonferroni

(I) GRLING	(J) GRLING	Différence de moyennes (I-J)	Erreur standard	Signification
Multilingues	ILtines	-,9496*	,25227	,002
Latines	Multilingues	,9496*	,25227	,002
	Sinophones	1,3779*	,19076	,000
Slaves	Sinophones	,9947*	,23792	,000
Sinophones	Latines	-1,3779*	,19076	,000
	Slaves	-,9947*	,23792	,000

*. La différence de moyennes est significative au niveau .05.

Afin de faciliter l'interprétation des résultats et les significations du tableau, un diagramme des « barres d'erreurs » (figure 5.11) de l'intervalle de confiance à 95% des moyennes des groupes linguistiques est proposé.

Figure A.10.11 : barre des intervalles de confiance à 95% autour de la moyenne de la compétence exprimée en *logits* des personnes des groupes linguistiques



Il appert que l'intervalle de confiance de la moyenne du groupe des sinophones ne coïncide qu'avec le groupe multilingue et que l'intervalle de confiance du groupe des langues latines ne coïncide qu'avec celui du groupe des langues slaves. Il semble que l'intervalle de confiance du groupe multilingue ne soit pas facilement interprétable dans la mesure où son étendue est beaucoup plus importante.

Quoique les élèves aient été choisis en fonction de leur niveau linguistique pour intégrer les cours du M.I.C.C., on constate que, pour ce qui est de la compréhension écrite, les moyennes sont significativement différentes entre les groupes linguistiques.

Pour l'analyse, et pour avoir des candidats de niveaux comparables, il faudra donc très probablement conserver les personnes du groupe des langues latines qui ont le plus bas niveau et les meilleurs parmi les sinophones.

Effets expérimentaux, effet de fatigue

Afin de vérifier l'effet de fatigue, le test a été administré en deux versions, la version 1 proposant, dans l'ordre, les tâches A, B, C, D, E et F et la version 2 les tâches D, E, F, A, B et C. La version 1 du test (items 1-30) a été proposée à 55 personnes, soit 48,7% des candidats. La version 2 (items 16-30 puis 1-15) a été proposée à 58 personnes, soit 51,3%

des candidats[3]. L'administration de versions différentes n'a pas eu d'effet sur le score total des personnes au test, puisque les moyennes des deux groupes sont comparables (groupe version 1 = 0,35 ; groupe version 2 = 0,41). Le test *t* à échantillons indépendants ne permet pas de rejeter l'égalité de moyennes (tableau A.10.24). Il n'y a donc pas eu d'effets observables liés à l'une ou à l'autre des administrations pour l'ensemble des résultats au test.

Tableau A.10.24 : test *t* score total des personnes au test et version du test

Test d'échantillons indépendants

		Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes					
								Intervalle de confiance 95% de la différence	
		F	Sig.	t	ddl	Sig. (bilatérale)	Différence moyenne	Inférieure	Supérieure
MESURES	Hypothèse de variances égales	2,04	,16	-,24	110	,81	-,04	-,41	,32
	Hypothèse de variances inégales			-,24	104,34	,81	-,04	-,40	,32

Si le score global des candidats n'est pas affecté par l'ordre de passation des deux types de tâches, on est en droit de penser que l'interaction entre l'ordre de passage des tâches et le type de tâche a éventuellement eu un impact. Pour vérifier l'hypothèse nulle selon laquelle la moyenne obtenue pour les tâches discrètes et intégrées ne dépend pas de l'ordre de passage de ces tâches, on a utilisé la fonction « *interaction* » entre le fonctionnement différentiels dans WINSTEPS entre le type de personne (celles ayant reçu la version 1 et celles ayant reçu la version 2) et le type d'item (appartenant aux tâches discrètes ou intégrées). Les « interactions » signalées dans WINSTEPS sont en fait une étude de fonctionnement différentiel d'items (FID) et des personnes (FDP), autrement dit des biais ou encore des interactions entre des types d'items et des types de personnes (Linacre, 2005). Dans ces analyses, les personnes et les items ayant un score extrême sont exclus car ils ne devraient pas présenter de différences de compétence quel que soit le niveau des items (pour les items) ou des personnes (pour les items). Pour qu'une différence de fonctionnement soit considérée comme « notable », il faut au moins qu'elle soit de 0,5 *logit*. Linacre (2005) nous apprend que s'il n'y a pas de taille minimale pour l'échantillon des personnes et des items, il indique néanmoins qu'en dessous de 30

[3] Pour mémoire, les versions du test ont été distribuées de manière aléatoire. La première copie d'examen distribuée était une version 1, la deuxième une version 2, puis une version 1, version 2,…

personnes ou items, les résultats sont peu généralisables et dépendent beaucoup de l'échantillon.

Les résultats de l'analyse (tableau A.10.25) indiquent qu'il n'y a pas eu d'interaction (p>0,05). Il est donc possible de dire que pour l'ensemble des candidats, il n'y a pas eu d'effet de fatigue significatif pour les tâches discrètes et intégrées.

Tableau A.10.25 : interaction entre l'ordre de passage des tâches et le type de tâche

INPUT 118 personnes, 30 items MEASURED 113 personnes, 30 items, 2 CATS 3.60.1
CLASS-LEVEL BIAS/INTERACTIONS FOR DIF=$S2W1 AND DPF=$S3W9

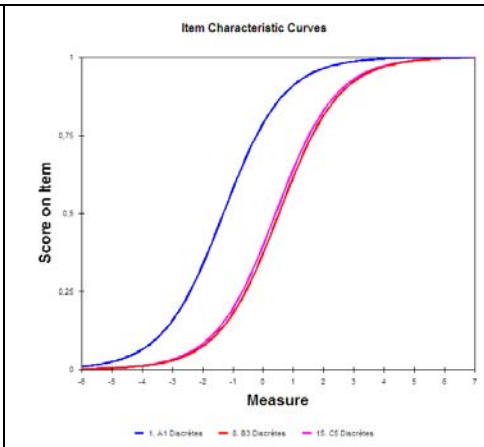
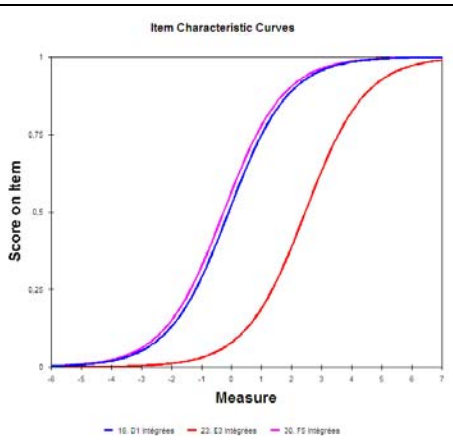
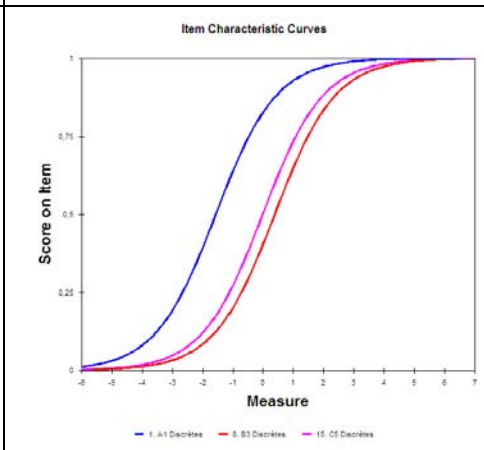
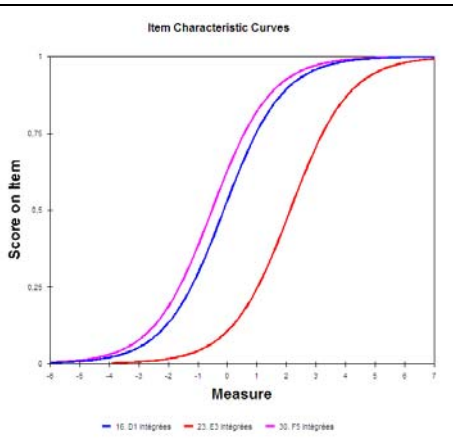
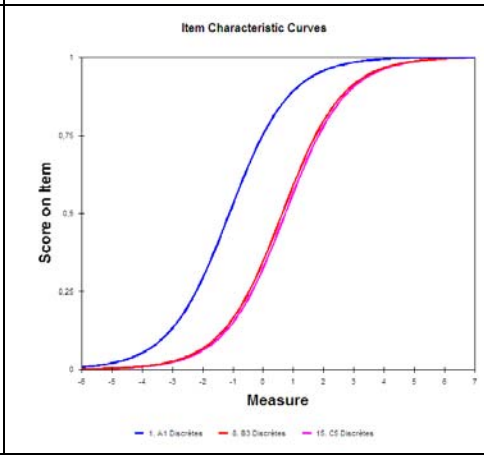
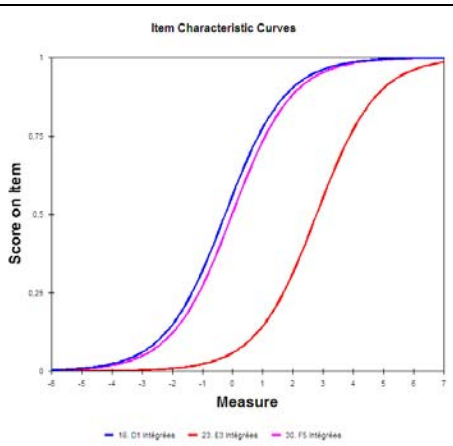
personne CLASS	DIF SIZE	DIF S.E.	personne CLASS	DIF SIZE	DIF S.E.	DIF CONTRAST	JOINT S.E.	t	d.f.	Prob.	item CLASS
1	-.06	.08	2	.06	.08	-.12	.12	.99	INF	.3212	Discrète
2	.06	.08	1	-.06	.08	.12	.12	-.99	INF	.3212	Discrète
1	.06	.08	2	-.05	.08	.11	.11	-.97	INF	.3333	Intégrée
2	-.05	.08	1	.06	.08	-.11	.11	.97	INF	.3333	Intégrée

Légende : 1 = groupe ayant reçu la version 15 items discrets puis 15 items intégrés ; 2 = groupe ayant reçu la version 15 items intégrés puis 15 items discrets

En conclusion sur ce point, l'examen des courbes caractéristiques des items confirme l'absence d'effet de fatigue (figure A.10.12). Comme on peut le constater, le niveau de difficulté des items des tâches discrètes (items 1, 8 et 15) et des tâches intégrées (items 16, 23, 30) ne fluctue que très peu selon que les candidats aient tout d'abord répondu aux items des tâches discrètes ou bien des tâches intégrées.

Figure A.10.12 : courbes caractéristiques des items des tâches discrètes et intégrées selon le type de calibration

	Items 1, 8 et 15 des tâches discrètes	Items 16, 23, 30 des tâches intégrées

Calibration avec tous les candidats		
Calibration avec les candidats ayant reçu 15 items discrets puis 15 items intégrés		
Calibration avec les candidats ayant reçu 15 items intégrés puis 15 items discrets		

Répartition des items et des personnes en groupes de niveau

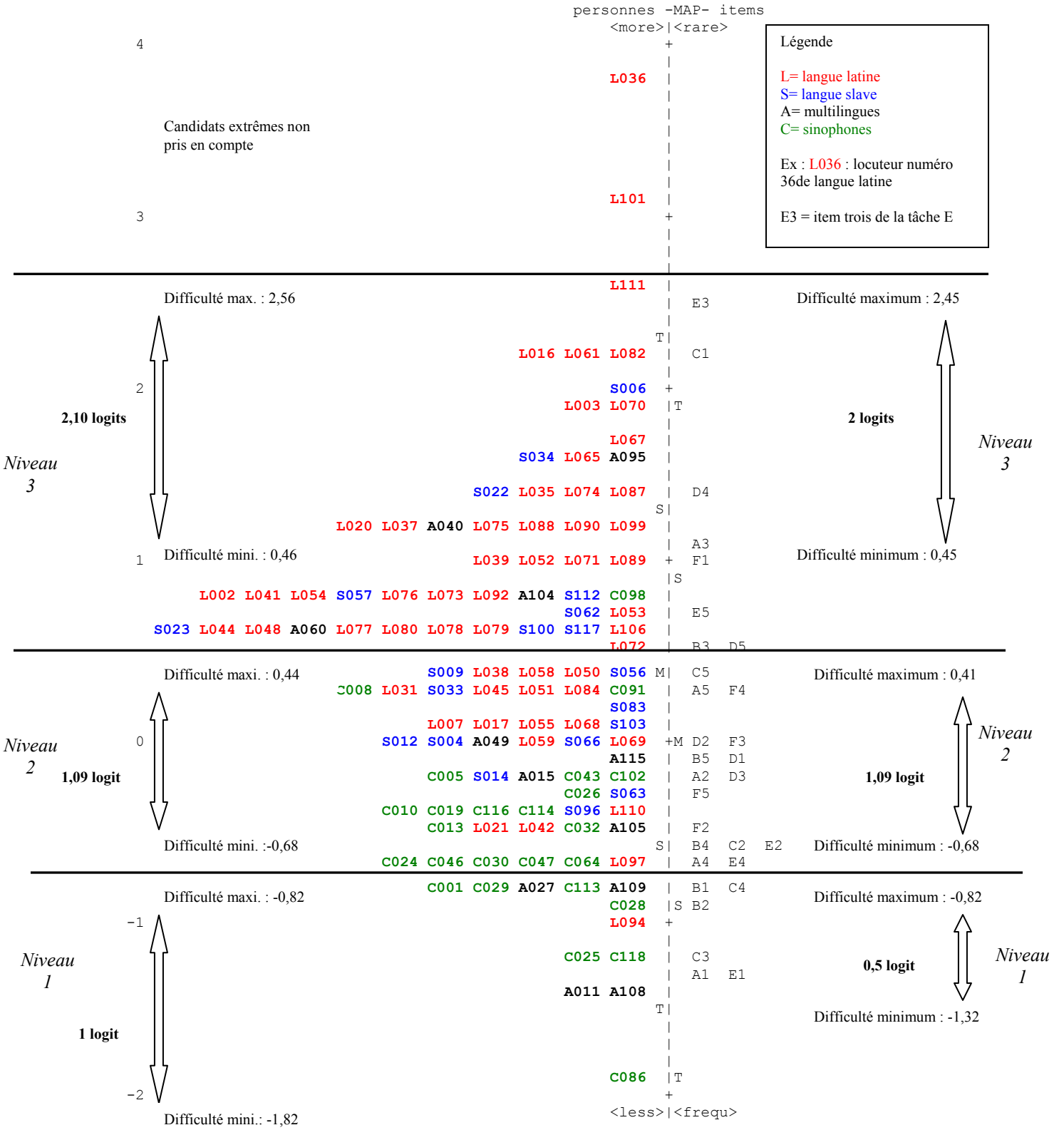
L'étude de la variable « groupe linguistique » a montré que les moyennes des groupes linguistiques diffèrent significativement. Comme il est possible de le constater sur la carte des personnes et des items (figure A.10.13), s'il y a beaucoup de sinophones au bas de l'échelle, ils sont totalement absents du haut de l'échelle. Quant aux personnes appartenant au groupe des langues latines, elles sont localisées sur le haut de l'échelle. Afin de sélectionner des candidats de niveau comparable appartenant à chacun des groupes linguistiques, nous avons décidé de répartir l'ensemble des candidats en trois niveaux. Si ce découpage est forcément arbitraire, autant que faire se peut, une tentative de rationalisation a été faite. L'objectif du découpage n'était pas d'obtenir des niveaux parfaitement bien définis, sinon de trouver un moyen, sur un plan opérationnel, de mener la recherche à son terme et de comparer des items et des personnes de niveau comparable. Il a donc été décidé de répartir les items et les personnes en trois groupes selon une logique de contenu (analyse et synthèse du contenu des items, puis comparaison avec les *Niveaux de compétence en français langue seconde pour les immigrants adultes* (M.R.C.I., 1998) et d'une répartition permettant d'avoir tous les .groupes linguistiques dans un même niveau.

Le résultat de la répartition est visible dans la figure A.10.13. L'explication complète de la répartition en niveau est présentée dans l'annexe 7.

Fonctionnement des tâches discrètes et intégrées pour l'ensemble des candidats

Cette partie a pour objectif de préparer les réponses aux questions de recherche 2 et 3 portant sur le fonctionnement différentiel des tâches discrètes et intégrées pour des candidats ayant un niveau comparable. Avant de procéder à de telles analyses, il convient de savoir si ces tâches, pour l'ensemble des candidats sont comparables. Les premières conclusions élaborées à partir de candidats de niveaux différents pourront servir pour l'analyse et l'interprétation des résultats avec des candidats de niveau comparable.

Figure A.10.13 : carte de l'estimation de la compétence en *logit*s des personnes et des items



Difficulté des tâches discrètes et intégrées calibrées ensemble

L'ensemble des 15 items des tâches discrètes (μ=-0,18 ; tableau A.10.26) est légèrement plus facile que l'ensemble des 15 items des tâches intégrées (μ= 0,18) ; tableau A.10.27). La différence entre les deux moyennes n'est que de 0,36 *logit*. La différence entre les deux médianes des tâches discrètes et intégrées est de 0,58, soit un demi-*logit*. Les difficultés maximales et minimales des items, elles aussi, montrent peu de différence, tout comme les valeurs des carrés moyens *infit* et *outfit*.

Tableau A.10.26 : statistiques générales des tâches discrètes pour l'ensemble des personnes

"Discrètes" SUBTOTAL FOR 15 NON-EXTREME items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	65.9	107.9	-.18	.23	.99	-.1	.96	-.2
S.D.	18.9	1.6	.92	.02	.09	1.2	.15	1.0
MAX.	90.0	111.0	2.21	.27	1.15	1.9	1.18	1.4
MIN.	19.0	105.0	-1.32	.21	.76	-3.4	.68	-2.7

REAL RMSE	.23	ADJ.SD	.89	SEPARATION	3.81	item	RELIABILITY	.94
MODEL RMSE	.23	ADJ.SD	.89	SEPARATION	3.87	item	RELIABILITY	.94

S.E. OF item MEAN = .25

MEDIAN = -.61

Tableau A.10.27 : statistiques générales pour les items des tâches intégrées pour tous les candidats

"Intégrées" SUBTOTAL FOR 15 NON-EXTREME items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	59.1	108.9	.18	.22	1.02	.3	.99	.0
S.D.	18.5	2.5	.90	.02	.10	1.1	.16	1.0
MAX.	88.0	113.0	2.45	.29	1.18	2.4	1.23	2.1
MIN.	16.0	104.0	-1.27	.21	.82	-1.9	.72	-1.6

REAL RMSE	.23	ADJ.SD	.87	SEPARATION	3.77	item	RELIABILITY	.93
MODEL RMSE	.23	ADJ.SD	.87	SEPARATION	3.87	item	RELIABILITY	.94

S.E. OF item MEAN = .24

MEDIAN = -.03

Il est possible de dire que, du point de vue de la mesure, les deux types de tâche sont comparables quant à leurs moyennes, leurs difficultés maximale et minimale. Afin de raffiner l'analyse, la difficulté mesurée pour chacune des six tâches a été étudiée (tableau A.10.28).

Tableau A.10.28 : difficulté de chacun des types de tâche

ALL SCORES ARE NON-EXTREME

item COUNT	MEAN MEASURE	S.E. MEAN	OBSERVED S.D.	MEDIAN	REAL SEPARATION	CODE
30	.00	.17	.93	-.16	3.87	*
5	-.15	.41	.83	-.18	3.44	A
5	-.39	.26	.53	-.63	2.11	B
5	.00	.61	1.23	-.61	4.92	C
5	.31	.30	.60	.04	2.55	D
5	.12	.67	1.33	-.59	5.23	E
5	.11	.26	.52	-.03	2.13	F

UMEAN=0 USCALE=1

Les moyennes des items associés à chacune des tâches discrètes (A, B et C) sont moins élevées que celles associée aux tâches intégrées (D, E et F). En moyenne, la tâche B est la plus facile des tâches discrètes et la C la plus difficile. Concernant la moyenne des items associés aux textes des tâches intégrées, la tâche D est la plus difficile, la F la plus facile. Les items des tâches E et F, malgré la difficulté accrue des textes auxquels ils sont associés, ont une moyenne plus basse que celle des items de la tâche D.

La comparaison de la fréquence des questions par texte, type de tâche et niveau des items (tableau A.10.29) nous apprend qu'il y a plus de questions classées dans le « niveau 1 » et moins de questions classées dans le « niveau 2 » pour les tâches discrètes que pour les tâches intégrées.

Tableau A.10.29 : fréquence des questions par texte, type de tâches et niveau des items

	Niveau 1	**Niveau 2**	**Niveau 3**
Texte A	1 20%	3 60%	1 20%
Texte B	2 40%	2 40%	1 20%
Texte C	2 40%	2 40%	1 20%
Total tâches discrètes	5 34%	7 46%	3 20%
Texte D	0 0%	3 60%	2 40%
Texte E	1 20%	2 40%	2 40%
Texte F	0 0%	4 80%	1 20%
Total tâches intégrées	1 6%	9 60%	5 34%

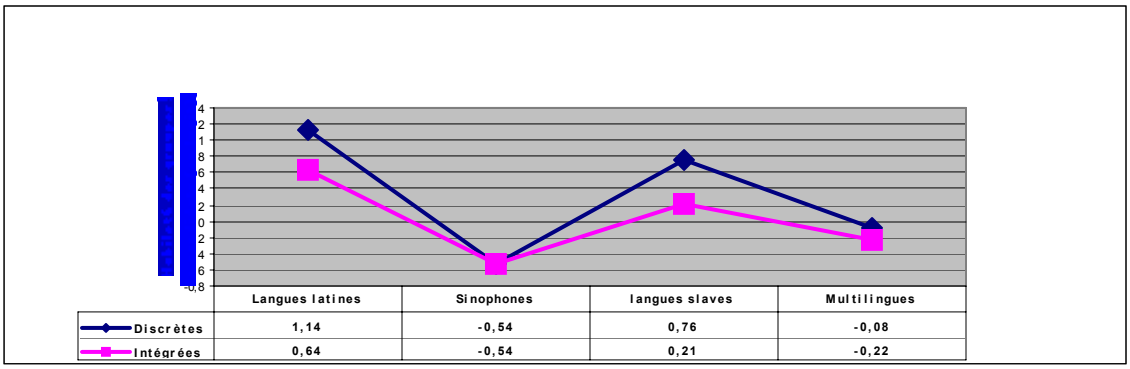
Les tâches intégrées ont plus de questions de « niveau 2 » et « niveau 3 » que les tâches discrètes. Si le premier texte (D) a été choisi parce qu'il était le plus simple, les questions,

qui lui sont associées, sont les plus difficiles. Si les deuxième et troisième textes (E et F) sont plus difficiles, plus complexes et plus longs que le texte D, les questions qui leur sont associées sont plus faciles.

Difficulté des tâches discrètes et intégrées pour les groupes linguistiques

L'examen de la compétence des groupes linguistiques avec les items des tâches discrètes et intégrées calibrées séparément (figure A.10.14) dévoile qu'à l'exception des sinophones, tous les autres groupes de candidats ont une estimation de leur niveau compétence plus élevée avec les tâches discrètes qu'avec les tâches intégrées.

Figure A.10.14 : différence de compétence des groupes linguistiques selon le type de tâches (calibrées séparément)



	Langues latines	Sinophones	langues slaves	Multilingues
Discrètes	1,14	-0,54	0,76	-0,08
Intégrées	0,64	-0,54	0,21	-0,22

La différence pour le groupe des sinophones est celle qui est la plus basse, suivie de celle du groupe des multilingues, du groupe des langues latines et enfin du groupe des langues slaves. Bien que le sous-test composé des items intégrés soit plus difficile que celui composé des items des tâches discrètes, les sinophones ont un niveau similaire.

Afin de compléter l'analyse, la différence des scores en *logits* entre les tâches discrètes et intégrées a été calculée pour les différents groupes de candidats. Un test ANOVA a servi à vérifier si la différence des scores entre les groupes était significative. L'ANOVA n'est pas significative (tableau A.10.30, F= 1,994, ddl=110, sig.= 0,119).

Tableau A.10.30 : test ANOVA de la différence de score entre les tâches discrètes et les tâches intégrées selon les groupes linguistiques

Test of Homogeneity of Variances

DIFF

Levene Statistic	df1	df2	Sig.
1,063	3	107	,368

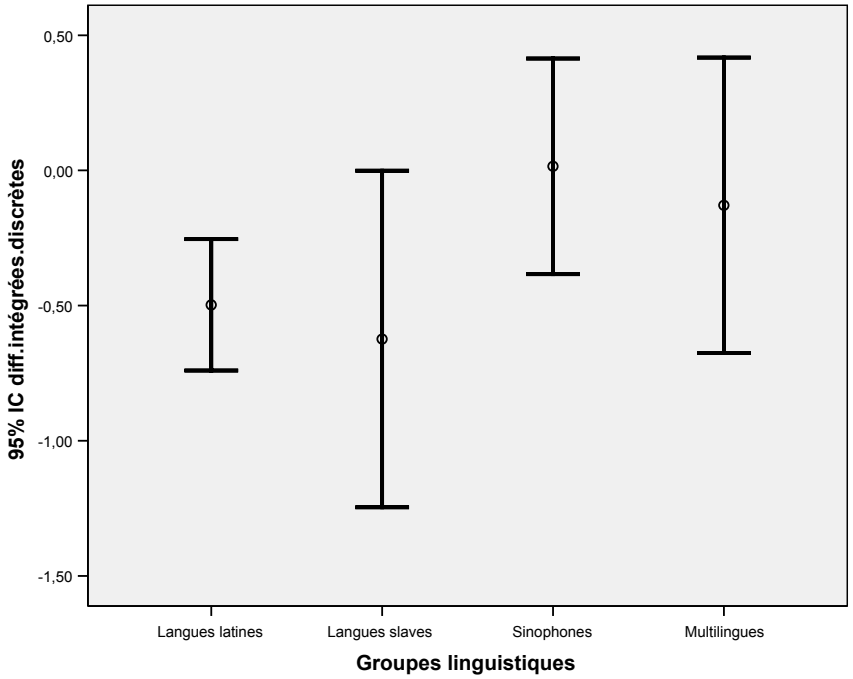
ANOVA

DIFF

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5,797	3	1,932	1,994	,119
Within Groups	103,699	107	,969		
Total	109,496	110			

Toutefois, l'examen de la figure A.10.15, nous indique que la moyenne de la différence des scores pour les deux types de tâches entre le groupe des langues latines et celui des sinophones est presque significativement différente.

Figure A.10.15 : intervalle de confiance à 95% de la différence entre la moyenne atteinte pour les tâches intégrées et celle atteinte pour les tâches discrètes pour chacun des groupes linguistiques



Différence de classement des candidats avec les items discrets et intégrés calibrés conjointement ou séparément

Cette partie de l'analyse a pour objectif de vérifier si le classement de l'ensemble des candidats et des groupes de candidats varie selon que la compétence des candidats est estimée avec l'ensemble des tâches et avec les tâches discrètes ou intégrées calibrées

séparément. Pour ce faire, la corrélation de Spearman est utilisée ainsi que des tests non-paramétriques (tableau A.10.31).

Les corrélations de Spearman, entre les différentes « versions » du test, indiquent que les scores atteints par les candidats aux deux sous-tests sont fortement corrélés à leurs scores pour l'ensemble du test (0,884 pour les tâches intégrées et 0,896 pour les tâches discrètes), Néanmoins, la corrélation de rang entre les scores des candidats aux tâches intégrées et aux tâches discrètes, même si elle est significative, est modérée (0,606 dans les deux cas). Elle est encore plus faible que la corrélation entre les scores atteints par les candidats aux items pairs et impairs (0,748).

Tableau A.10.31 : corrélations de Spearman entre la compétence des candidats calculée avec l'ensemble des tâches, les tâches intégrées seules et les tâches discrètes seules, les items pairs et impairs

Correlations

			Estimation habileté tâches intégrées	Estimation habileté tâches discrètes	Estimation habileté tous les items	Estimation habileté items pairs	Estimation habileté items impairs
Spearman's rho	Estimation habileté tâches intégrées	Correlation Coefficient	1,000	,606**	,884**	,866**	,788**
		Sig. (2-tailed)	.	,000	,000	,000	,000
		N	113	111	113	113	113
	Estimation habileté tâches discrètes	Correlation Coefficient	,606**	1,000	,896**	,825**	,835**
		Sig. (2-tailed)	,000	.	,000	,000	,000
		N	111	111	111	111	111
	Estimation habileté tous les items	Correlation Coefficient	,884**	,896**	1,000	,941**	,918**
		Sig. (2-tailed)	,000	,000	.	,000	,000
		N	113	111	113	113	113
	Estimation habileté items pairs	Correlation Coefficient	,866**	,825**	,941**	1,000	,748**
		Sig. (2-tailed)	,000	,000	,000	.	,000
		N	113	111	113	113	113
	Estimation habileté items impairs	Correlation Coefficient	,788**	,835**	,918**	,748**	1,000
		Sig. (2-tailed)	,000	,000	,000	,000	.
		N	113	111	113	113	113

**. Correlation is significant at the 0.01 level (2-tailed).

Afin de vérifier la différence de classement entre les groupes linguistiques selon le type de tâche utilisé, la variable de la différence de score aux tâches discrètes et intégrées calibrées séparément a été utilisée. Le test de Kruskall-Wallis pour tester les moyennes des rangs pour la différence des scores aux tâches discrètes et intégrées des groupes linguistiques n'est pas significatif (tableau A.10.32, chi-carré=6,824, dll.=3, sig.=0,078).

Tableau A.10.32 : test de kruskal-Wallis pour tester la moyenne des rangs des différents groupes linguistiques (composé de l'ensemble des candidats) pour la différence des scores entre les tâches discrètes et intégrées

Ranks

	GRLING	N	Mean Rank
DIFF	Latines	56	60,38
	Slaves	19	64,32
	Sinophones	25	43,36
	Multilingue	11	48,09
	Total	111	

Test Statistics[a,b]

	DIFF
Chi-Square	6,824
df	3
Asymp. Sig.	,078

a. Kruskal Wallis Test

b. Grouping Variable: GRLING

Toutefois, le test non paramétrique pour échantillons indépendants de Mann-Whitney indique que la moyenne des rangs de la différence entre les scores obtenus pour les items des tâches discrètes et les items des tâches intégrées du groupe des sinophones et du groupe des langues latines varie significativement (tableau A.10.33, *Z*=-2,178, sig.=0,029).

Tableau A.10.33 : test de Mann-Whitney pour tester la moyenne des rangs du groupe des langues latines et de celui des sinophones portant sur la différence des scores entre les tâches discrètes et intégrées

Ranks

	GRLING	N	Mean Rank	Sum of Ranks
DIFF	Latines	56	44,80	2509,00
	Sinophones	25	32,48	812,00
	Total	81		

Test Statistics[a]

	DIFF
Mann-Whitney U	487,000
Wilcoxon W	812,000
Z	-2,178
Asymp. Sig. (2-tailed)	,029

a. Grouping Variable: GRLING

Références

Abbott, M. (2004). *The identification and interpretation of group differences on the Canadian Language Benchmarks Assessment Reading Items.* C.R.A.M.E.. Université d'Alberta. Article présenté lors du congrès annuel N.C.M.E.. Consulté à http://www.education.ualberta.ca/educ/psych/crame/files/CLBA_DIF.pdf, dernière consultation le 15/12/2005.

Abbott, M. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, 24 (7), 7-36.

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language Testing in the 1990s: The Communicative Legacy* (71-86). London: Macmillan.

Alderson, J. C. (1999). Reading construct and reading assessment. In M. Chalhoub-Deville (dir.) *Issues in Computer-Adaptive Testing of Reading Proficiency.* (49-70). Cambridge: University of Cambridge Local Examinations Syndicate.

Alderson, J. C. (2000). *Assessing Reading*. Cambridge : Cambridge University Press.

Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency*. New York : Continuum.

Alderson, J. C., Clapham, C., et Wall, D. (1995). *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., et Tardieu, C. (2006). Analysing Tests of reading and listening in relation to the Common European Framework of Reference: The experience of The dutch CEFR construct project. *Language Assessment Quarterly*, *3* (1), 3–30.

Alderson, J. C., et Lukmani, Y. (1989). Cognition and reading: cognitive levels as embodied in test questions. *Reading in a Foreign Language, 5* (2), 253-270.

Anderson, A., et Lynch, T. (1988). *Listening.* New York : Oxford University Press.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika*, 43, 561-73.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, 38, 665-680.

Andrich, D. (2002). Understanding resistance to the relationship in Rasch's paradigm: A reflexion for the next generation. *Journal of Applied Measurement*, 3(3), 325-359.

Andrich, D. (2004). Controversy and Rasch model. A characteristic of incompatible paradigms ? *Medical Care*, 42 (1), 7-16.

Ariel, A, Van der Linden, W. J., et Veldkamp, B. P. (2006). A Strategy for optimizing item-pool management. *Journal of Educational Measurement.*, 43 (2), 85–96.

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

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, 19 (4), 453-476.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge : Cambridge University Press.

Bachman, L. F., Lynch, B. K., et Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing,* 12 (2), 239-257.

Bachman, L. F., et Palmer, A. S. (1996). *Language Testing in Practice*, Oxford : Oxford University Press.

Belov, D. I., et Armstrong, R. D. (2008). A monte carlo approach to the design, assembly, and evaluation of multistage adaptive tests. *Applied Psychological Measurement*, 32, 119-137

Berg, **C.** (2003). The role of grounded theory and collaborative research. *Reading Research Quarterly,* 38 (1), 105-11.

Bernhardt, **E.** (2003). Challenges to reading research from a multilingual world. *Reading Research Quarterly,* 38 (1), 112-17.

Bernhardt, E. B., et Kamil, M. (1995). Interpreting relationships between L1 and L2 reading: Consolidating the linguistic threshold and the linguistic interdependence hypotheses. *Applied Linguistics*, 16, 15-34.

Bertrand, R., et Blais J. G. (2004). *Modèles de Mesure, L'apport de la Théorie des Réponses aux Items*. Sainte-Foy : Presses de l'Université du Québec.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord et M. Novick (dir.), *Statistical Theories of Mental Test Scores.* (397–479). Reading : MA: Addison Wesley.

Blais, J. G. (1987). *Effets de la Violation du Postulat d'Unidimensionnalité dans la Théorie des Réponses aux Items*. Université de Montréal. Thèse de doctorat non publiée.

Blais, J. G., et Laurier, M. D. (1995a). The dimensionality of a placement test from several analytical perspectives. *Language Testing*, 12 (1), 72-98.

Blais, J. G., et Laurier, M. D. (1995b). La détermination de l'unidimensionnalité de l'ensemble des scores à un test. *Mesure et Evaluation en Education*, 20 (1), 65-90.

Blais, J. G., et Raîche, G. (2005). La détection des patrons de réponse inappropriés dans le contexte des tests adaptatifs par ordinateur. Actes du congrès ADMEE.

Bock, R.D., Gibbons, R., et Muraki, E. J. (1988). Full information factor analysis. *Applied Psychological Measurement*, 12, 261-280.

Bond, T. G., et Fox, C. M. (2001). *Applying the Rasch Model, Fundamental Measurement in the Human Sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bottani, N., et Vrignaud, P. (2005). *La France et les Evaluations Internationales*. Rapport numéro 16. Haut conseil de l'évaluation. Consulté le 1/12 2005 à http://cisad.adc.education.fr/hcee/documents/rapport_Bottani_Vrignaud.pdf.

Brindley, G. (1987). Factors affecting task difficulty. In D. Nunan (dir.) *Guidelines for the Development of Curriculum Resources for the Adult Migrant Education Program.* Adelaide: National Curriculum Resource centre.

Brindley, G., et Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing*, 19 (4), 369-394.

Brown, J. D., Hudson, T. D., Norris, J. M., et Bonk, W. (2002). *Investigating Task-Based Second Language Performance Assessment.* Honolulu, HW: University of Hawai'i Press.

Brown, A., et Iwashita, N. (1996). Language background and item difficulty: The development of a computer-adaptive test of Japanese. *System* 24 (2), 199-206.

Brown, G., et Yule, G. (1983). *Teaching The Spoken Language*. Cambridge : Cambridge University Press.

Bunderson V. C., Inouye, D. K., et Olsen, J. B. (1989). The Four Generations of Computerized Educational Measurement. In R. L. Linn (dir.), *Educational Measurement* (3e éd., 367-407). New York : Macmillan /American Council of Education**.**

Canale, M., et Swain, M. (1980). Theoretical bases of communicative approach to second language teaching and testing. *Applied Linguistics*, 1 (1), 1-47.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards, et R. W. Schmidt (dir.). *Language and Communication*. (2-27). London: Longman.

Candlin, C. N. (1987). Towards task-based learning. In C.N. Candlin and D.F. Murphy (eds.). *Language Learning Tasks* (5-23). Englewood Cliffs, NJ: Prentice Hall.

Carr, N. T. (2006). The factor structure of test task characteristics and examinee performance. *Language Testing*, 23 (3), 269–289.

Carroll, J. B. (1961, 1972). Fundamental considerations in testing english language proficiency of foreign students. In H. B. Allen & R. N. Campbell (Eds.). *Teaching English as a Second Language* (2e éd., 313–321). New York: McGraw-Hill.

Carver, R. P. (1990). *Reading Rate: A Review of Research and Theory*. New York : Academic Press.

Carver, R. P. (1997). Reading for one second, one minute, or one year from the perspective of rauding theory. *Scientific Studies of Reading, 1 (1)*, 3-43.

Carver, R. P. (2000). *The Cause of High and Low Reading Achievement.* Mahwah, N.J.:Erlbaum.

Celce-Murcia, M, Dornyei, Z., et Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specification. *Applied Linguistics*, 6 (2), 5-35.

C.C.L.B. (2000). *Canadian Language Benchmarks: English as a Second Language for Adults.* Consulté le 1/12/2005 à http://www.language.ca/pdfs/clb_adults.pdf.

C.C.L.B. (2002). *Les Standards Linguistiques Canadiens. Français Langue Seconde Adulte*. Consulté le 1/12/2005 à http://www.language.ca/cclb_files/doc_viewer_dex.asp?doc_id=120&page_id=383

Chapelle, C. (1999). From reading theory to testing practise. In M. Chalhoub-Deville, (dir.) *Issues in computer-Adaptive Testing of Reading Proficiency* (150-166). Cambrisge : University of Cambridge Local Examinations Syndicate.

Chen, Z., et Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, 2, 155-163.

Choi, I. C., et Bachman, L. F. (1992). An investigation into the adequacy of three IRT models. *Language Testing*, 51-78

Crehan, K. D., et Haladyna, T. M. (1991). The validity of two item-writing rules. *Journal of Experimental Education, 59*, 183–192.

Conseil de l'Europe (2001). *Cadre Européen Commun de Référence pour les Langues – Apprendre, Enseigner, Evaluer.* Paris. Didier. (3e édition revue et corrigée). Consulté le 1/12/05 à http://culture2.coe.int/portfolio//documents/cadrecommun.pdf.

Cummins, J. (1991). Conversational and academic language proficiency in bilingual contexts. In J. Hulstijn et A. Matter (dir.), *AILA Review*, 8, (75-89).

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., et James, M. (2006). *Analysis of Discourse Features and Verification of Scoring Levels for Independent and Integrated Prototype Written Tasks for the New T.O.E.F.L.*. MS-30 avril 2006.

Daoust, F. (1996). *SATO (Système d'analyse de texte par ordinateur), Version 4.0, Manuel de référence*. Service d'analyse de texte par ordinateur (ATO), Université du Québec à Montréal. Consulté le 15/12/2006 à http://www.ling.uqam.ca/ato/.

Dassa, C., et Laurier, M. (2003). Le diagnostic pédagogique comment moyen d'informer. In M. Laurier (dir.), *Évaluation et Communication, de l'Evaluation Formative à l'Evaluation Informative* (103-130). Outremont : Les éditions Québécor.

Davey, B., et Lasasso, C. (1984). The interaction of reader and task factors in the assessment of reading comprehension. *Experimental Education, 52*, 199-206.

Davis, F. B. (1968). Research in comprehension reading. *Reading Research Quarterly,* 3, 499-545.

Déclaration commune des ministres européens de l'éducation - 19 juin 1999 – Bologne, (1999). Consulté le 9/11/2005 à http://www.education.gouv.fr/realisations/education/superieur/bologne.htm.

Déclaration de l'Association Internationale des Universités : Internationalisation de l'enseignement supérieur, (1998). Consulté le 9/11/2005 à http://www.unesco.org/iau/internationalization/fre/inter_ddurban.html.

Déclaration de l'A.U.C.C. sur l'internationalisation et les universités canadiennes (1995). Au Canada, l'Association des Universités et Collèges du Canada a adopté une déclaration sur l'internationalisation dès 1995. Consulté le 9/11/2005 à http://www.aucc.ca/publications/statements/1995/intl_04_f.html.

Dooey, P. (2008). Language testing and technology : problems of transition to a new era. *ReCALL*, 20 (1), 21-34.

Dorans, N. J. (2000). Scaling and equating. In H. Wainer, D. Eignor, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg et D. Thissen (dir*.), Computerized Adaptive Testing : a Primer* (135-158). Hillsdale, NJ : Lawrence Erlbaum Associates.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge : Cambridge Language Assessment series, Cambridge University press.

Du, Y., Wright, B. D., et Brown, W.L. (1996). *Differential Facet Functioning Detection in Direct Writing Assessment*. A.E.R.A, ED 400 293.

Dunkel, P. (1999). Research and development of a computer-adaptive test of listening comprehension in the less-commonly taught language Hausa. In M. Chalhoub-Deville, (dir.) *Issues in computer-adaptive testing of reading proficiency* (91-121). Cambridge : University of Cambridge Local Examinations Syndicate.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, *2*, 197–221.

Eignor, D. (1999). Selected technical issues in the creation of computer-adaptive tests of a second language reading proficiency. In M. Chalhoub-Deville (dir.) *Issues in Computer-Adaptive Testing of Reading Proficiency* (167-181). Cambridge : University of Cambridge Local Examinations Syndicate.

Eisley, M. E. (1990). *The Effect of Sentence Form and Problem Scope in Multiple-Choice Item Stems on Indices of Test and Item Quality*. Unpublished doctoral dissertation, Brigham Young University, Provo, UT.

Elder, C. (1996). The effect of language Background on “foreign” language test performance: the case of chinese, italian and modern greek. *Language Learning*, 46 (2), 233-282.

Elder, C., Iwashita, N., et McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing*, 19 (4), 347-368.

Elder, C., McNamara, T., et Congdon, P. (2003). Rasch techniques for detecting bias in performance assessments: an example comparing the performance of native and non-native speakers on a test of academic english. *Journal of Applied measurement*, 4 (2), 181-197.

Éloy, J. M. (2004a). Langues proches : que signifie de les enseigner ?, *É.L.A.,* (136), 393-402.

Éloy, J. M. (2004b). *Politique des locuteurs descripteurs- concernant le continuum*. Communication au colloque Contacts de langues et minorisations, organisé par l'Université de Neuchâtel et l'IUKB à Sion (Suisse) les 3-4 septembre 2003.

Embretson, S. E., et Reise, S. P. (2000). *Item Response Theory for Psychologists*. Hillside, NJ: Lawrence Erlbaum Associates.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., et Schedl, M. (2000). *T.O.E.F.L. 2000 Reading Framework: a Working Paper*. T.O.E.F.L. Monograph MS–17. Princeton, NJ: Educational Testing Service.

Ercikan, K., Gierl, M. J., McCreith, T., Puhan, G., et Koh, K. (2004). Comparability of bilingual version of assessments: sources of incomparability of english and french versions of Canada's national achievement tests. *Applied Measurement in Education*, 17 (3), 301-321.

Fender, M. (2003). English word recognition and word integration skills of native Arabic- and Japanese-speaking learners of English as a second language. *Applied Psycholinguistics,* 24, 289–315.

Figueras, N., North, B., Takala, S., Verhelst, N., et Van Avermaet, P. (2005). Relating examinations to the CEFR. *Language Testing*, 22 (3), 261–279.

Fitzgerald, J. (2003). Multilingual Reading Theory**.** *Reading Research Quarterly.* 38 (1), 118-22.

Freedle, R. (1997). The relevance of multiple-choice reading test data in studing expository passage comprehension: the saga of a 15 year effort towards an experimental / Correlational Merger. *Discourses Processes*, 23, 399-440.

Freedle, R., et Kostin, I. (1993). The prediction of T.O.E.F.L. reading item difficulty: implications for construct validity. *Language Testing,* 10, 133–70.

Freedle, R., et Kostin, I. (1996).The Prediction of T.O.E.F.L. Listening Comprehension Item Difficulty for Minitalk Passages: Implications for Construct Validity. *T.O.E.F.L. research Report No. 56*. Also available as *RR-96-29*. Princeton, NJ: Educational Testing Service.

Freedle, R., et Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of T.O.E.F.L.'s minitalks. *Language Testing,* 16, 2–32.

Garcia, G.E. (1991). Factors influencing the english reading test performance of spanish-speaking hispanic children. *Reading Research Quarterly*, 26, 371-393.

Goldman, S. R. (1997). Learning from text: reflections on the past and suggestions for the future. *Discourse Processes,* 23, 357-398.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: the feasibility of verbal item generation. *Journal of Educational Measurement*, 42 (4), 351-373.

Goulier, F. (2004). L'évaluation des langues vivantes est-elle une évaluation comme les autres?, *Administrer l'Enseignement des Langues Vivantes*, *Administration et Education*, 101, 29-39.

Grabe, W. (1999). Developments in reading research and their applications for computer-adaptive reading assessment. In M. Chalhoub-Deville (dir.) *Issues in Computer-Adaptive Testing of Reading Proficiency* (11-48). Cambridge: University of Cambridge Local Examinations Syndicate.

Green, T. et Maycock, L. (2004). Computer-based IELTS and paper-based versions of IELTS. *University of Cambridge ESOL Research Notes*, 18, 3-6.

Gulliksen, H. (1950). *Theory of Mental Tests*. New York: Wiley.

Haladyna, T. M. (2004). *Developing and Validating Multiple-Choice Test Items.* Mahwah, N.J. : Lawrence Erlbaum Associates.

Haladyna, T. M., Downing, S. M., et Rodriguez, M. C. (2002). A review of multiple-choice item-wrinting guidelines for classeroom assessment. *Applied Measurement in Education*, 15 (3), 309-334.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (dir.), *Educational Measurement* (3^{e} éd., 147-200). New York: Macmillan /American Council of Education**.**

Hambleton, R. K., Merenda, P. F., et Spielberger, C. D. (2005). *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., et Xing, D. (2006). Optimal and nonoptimal computer-based test designs for making pass-fail decisions. *Applied Measurement in Education*, 19 (3), 221-239.

Hambleton, R. K., Swaminathan, H., et Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Publications.

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing*, 9, 1-11.

Hibbison, E. P. (1991). The ideal multiple choice question: A protocol analysis. *Forum for Reading*, 22(2), 36-41.

Hudson, T. (1996). *Assessing Second Language Academic Reading from a Communicative Competence Perspective: Relevance for T.O.E.F.L. 2000*. T.O.E.F.L. Monograph Series: *MS-4*. Princeton, NJ: Educational Testing Service.

Hymes, D. H. (1971). On communicative competence. In J. Pride and J. Holmes (Eds.), *Sociolinguistics.* Penguin, 1972. (extrait de l'article publié en 1971, Philadelphia, University of Pennsylvania Press.).

Ingram, D. E. (1985). Assesssing proficiency : an overview of some aspects of testing. In Hyltenstam et Pienmann (eds.) *Modelling and Assessing Second Language Acquisition*. Clevedon: Multilingual Matters. (215-276).

Iwashita, N., McNamara, T., et Elder, C. (2001). Investigating predictors of task difficulty in the measurement of speaking proficiency. *Language Learning,* 21, 401-436.

Jang, E. E., et Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, 44 (1), 1-21.

Jones, N. (2005). *Seminar to calibrate examples of spoken performance CIEP Sèvres*, 02-04.12.2004, report on analysis of rating data. Council of Europe. Consulté le 1/12/2005 à http://www.coe.int/T/E/Cultural_Co-operation/education/Languages/Language_Policy/Common_Framework_of_Reference/SevresreportNJ.pdf?L=E.

Katz, S., et Lautenschlager, G. J. (1999). The contribution of passage no-passage item performance on the SAT1 reading task. *Educational Assessment*, 7(2), 165-176.

Kingsbury,G.G. (1996). Item review and adaptive testing. Article présenté à la rencontre annuelle du National Council of Measurement in Education, NY.

Kingston, N. M., et Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptative testing. *Applied Psychological Measurement*, 8, 146-154.

Koda, K. (2005). I*nsights into Second Language Reading: A Cross-Linguistic Approach*. Cambridge : Cambridge University Press.

Kondo-Brown, K. (2002). An analysis of rater bias with FACETS in measuring Japanese L2 writing performance. *Language Testing*, 19 (1), 1-29.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model – Some critical suggestions on traditional approaches. *International Journal of Testing*, 5(4), 377-394.

Kuhn, T. S. (1961/1977). The function of measurement in modern physical science. Isis, 52, 161-190. Reproduit dans Kuhn, T. S. (1977). *The Essential Tension.* Chicago, IL: The University oh Chicago Press.

Kunnan, A. J. (2000). *Fairness and Validation in Language Assessment*. Cambridge, UK: Cambridge University Press.

Lado, R. (1961). *Language Testing*. New York: McGraw-Hill.

Laurier, M. (1993). *L'informatisation d'un Test de Classement en Langue Seconde.* Publication B-190 CIRAL.

Laurier, M. (1999). The development of an adaptive test for placement in french. In M. Chalhoub-Deville (dir.) *Issues in Computer-Adaptive Testing of Reading Proficiency* (122-135). University of Cambridge Local Examinations Syndicate.

Laurier, M. (2004). *Proposition de projet en vue de l'informatisation du test de positionnement du Ministère des Relations avec les Citoyens et de l'Immigration par l'université de Montréal*. Document de travail.

Laveault, D., et Grégoire, J. (2002). *Introduction aux Théories des Tests.* 2ème édition. Bruxelles : De Boeck.

Lee, Y- W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, 23 (2), 131-166.

Lewkowitcz, J. A. (1997). The integrated testing of a second language. In Clapham, C. et Corson, D. editors, *Encyclopedia of Language and Education : volume 7. Language Testing and Achievement* (121-130). Dordrecht: Kluwer Academic.

Li, W. (1992). *What is a Test Testing? An Investigation of the Agreement Between Student's Test-Taking Processes and the Constructor's Presumptions*. Unpublished MA thesis, Lancaster University.

Linacre, J. M. (1989). *Many-Faceted Rasch Measurement.* MESA Press, Chicago, IL.

Linacre, J. M. (1999a). A measurement approach to computer-adaptive testing of reading comprehension. In Chalhoub-Deville (1999) *Issues in Computer-Adaptive Testing of Reading Proficiency*. Studies in language testing 10. Cambridge University Press.

Linacre, J. M. (1999b). Investigating rating scale category utility. *Journal of Outcome Measurement*, 2(3), 103-122.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean ? *Rasch Measurement Transactions,* 16 (2), 878.

Linacre, J. M. (2005). *WINSTEPS Rasch measurement computer program*. Chicago: WINSTEPS.com. Consulté le 1/12/2005 à www.WINSTEPS.com.

Linacre, J. M. (2006a). *Facets Rasch Measurement Computer Program.* Chicago, IL: WINSTEPS.com.

Linacre, J. M. (2006b). Demarcating category intervals. *Rasch Measurement Transactions*, 19 (3), 341-43. Consulté le 1/12/2005 à *www.Rasch.org/rmt/rmt194f.htm.*

Long, M. (2003). Español para fines específicos: ¿textos o tareas? *Actas del secundo Congreso Internacional de Español para Fines Específicos*. Consulté le 1/12/2005 à http://www.sgci.mec.es/be/media/pdfs/icefe/Actas_II_CIEFE.pdf.

Long, M. H., et Norris, J. M. (2000). Task-based language teaching and assessment. In Byram, M. (editor). *Encyclopedia of language teaching (597-603).* London: Routledge.

Lopez, W. A. (1996). *The Resolution of Ambiguity: An Example from Reading Instruction.* Doctoral dissertation, University of Chicago. Dissertation Abstract International, 57 (07), 2986A.

Lord, F. M., et Novick, M. (1968). *Statistical Theories of Mental Tests Scores*. *Reading :* MA: Addison-Wesley.

Lumley, T. (1993). The notion of subskills in reading comprehension tests: An EAP example. *Language Testing*, 3, 211-234.

Luecht, R. (1999). The practical utility of Rasch measurement models. In M. Chalhoub-Deville (dir.) *Issues in Computer-Adaptive Testing of Reading Proficiency* (196-223). Cambridge: University of Cambridge Local Examinations Syndicate.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

McNamara, T. F. (1991). Test dimensionality: IRT analysis of ESP listening test. *Language Testing*, 8 (2), 139-159.

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

McNamara, T. F. (1999). Computer-adaptive testing: A view from outside. In M. Chalhoub-Deville (dir.). *Issues in Computer-Adaptive Testing of Reading Proficiency* (136-149). Cambridge : University of Cambridge Local Examinations Syndicate.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin*, 89, 575-88.

Messick, S. (1989). Validity. In R.L. Linn (dir.), *Educational Measurement* (3e éd., 13-103). New York : Macmillan /American Council of Education**.**

M.I.C.C. (2004). *Plan d'action 2004-2007*. Consulté le 1/12/2005 à http://www.M.I.C.C..gouv.qc.ca/publications/pdf/PlanAction20042007_integral.pdf.

M.I.C.C. (2005). *Plan d'action 2005-2008.* Consulté le 26/08/2008 à http://www.micc.gouv.qc.ca/publications/pdf/PlanStrategique20052008_Integral.pdf

M.I.C.C. (2008a). *Tableau sur l'immigration au Québec 2003-2007.* Consulté le 26/08/08 à http://www.micc.gouv.qc.ca/publications/fr/recherches-statistiques/Immigration_Qc_2003-2007.pdf

M.I.C.C. (2008b). Pour enrichir le Québec. Franciser plus intégrer mieux. Consulté le 26/08/2008 à http://www.micc.gouv.qc.ca/publications/fr/mesures/Mesures-Francisation-Brochure2008.pdf

Millman, J., et Greene, J. (1989). The specification and developpement of test of achievement and ability. In R.L. Linn (dir.), *Educational Measurement* (3e éd., 335-366). New York : Macmillan /American Council of Education**.**

Molenaar, I. W. (1983). *Item Step.* Hetmans Bulletin HB-83-63- EX, University of Groningen, Vkgroep Statistick en Meeteorie FSW, Grote Kruisstraat 2/1, Groningen : The Netherland.

Mokhtari, K., et Reichard, C. (2004). Investigating the strategic reading processes of first and second language readers in two different cultural contexts. *System*, 32 (3), 379-394.

Morissette, D. (1996). *Guide Pratique de l'Evaluation Sommative. Gestion des Epreuves et des Examens.* Bruxelles : De Boeck Université.

M.R.C.I. (1998). *Niveaux de Compétence en Français Langue Seconde pour les Immigrants Adultes*. Montréal : Gouvernement du Québec, Service de ressources matérielles du MRCI,.

Munby, J. (1978). *Communicative Syllabus Design.* Cambridge: Cambridge University Press.

Nilsen, E. (2003). *Apprendre une Langue en Ligne dans une Perspective Actionnelle. Effets de l'Interaction Sociale*, thèse de doctorat, Université Strasbourg 1, Louis

Pasteur. Consulté le 1/12/2005 à http://eprints-scd-ulp.u-strasbg.fr:8080/archive/00000339/01/Nissen_th%C3%A8se.pdf.

Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing*, 19 (4), 337-346.

Norris, J., Brown, J. D., Hudson, T., et Yoshioka, J. (1998). *Designing Second Language Performance Assessments* (Technical Report 18). Honolulu, HW: University of Hawaii.

Norris, J. M., Brown, J. D., Hudson T. D., et Bonk W. (2002). Examinee Abilities and Task Difficulty in Task-Based Second Language Performance Assessment. *Language Testing*, 19 (4), 395-418.

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency.* New York: Lang.

Nunan, D. (1989). *Designing Tasks for the Communicative Classroom*. Cambridge: Cambridge University Press.

Nunan, D. (1995). Closing the Gap Between Learning and Instruction. *TESOL Quarterly*, 29 (1); 133 58.

Nunan, D. (2004). *Task-Based Language Teaching.* Cambridge University Press.

Nunan, D., et Keobke, K. (1995). Task difficulty from the learners' perspective: perceptions and reality. *Linguistic and Language Teaching*, 18, 1-12.

Papanastasiou, E. C., et Reckase, M. D. (2008). A « rearrangement procedure » for scoring adaptive tests with review options. *International Journal of Testing*, 7(4), 387-407.

Parry, K. (1996). Culture, literacy and L2 reading. *TESOL Quarterly,* 30, 665–92.

Pawlikowska–Smith, G. (2002). *Canadian Language Benchmarks 2000: Theoretical Framework*, Centre for Canadian Language Benchmarks. Consulté le 1/12/2005 à http://www.language.ca/pdfs/final_theoreticalframework3.pdf**.**

Pearson, P. D., et Johnson, D. D. (1978). *Teaching Reading Comprehension*. New York, Holt, Rinehart et Winston.

Perfetti, C. A. (1997). Sentences, individual differences, and multiple texts: Three issues in text comprehension. *Discourse Processes*, 2, 337-355.

Phelps, R. P. (2005). *Defending Standardized Testing.* Hillsdale, NJ: Lawrence Erlbaum.

Puren, C. (2000). *Champ sémantique de « tâche ».* Polycopié distribué lors du séminaire pour doctorants « Didactique des langues et technologies éducatives ». Université Technologique de Compiègne.

Puren, C. (2001a). La problématique de l'évaluation en didactique scolaire des langues. *Les Langues Modernes,* 2, 12-29.

Puren, C. (2001b). *Évolution historique des types de cohérence en didactique scolaire des langues étrangères en France*. Document distribué dans le cadre du séminaire « Didactique des langues et technologies éducatives » 2001/2002. Université Technologique de Compiègne.

Puren, C. (2002a). Innovation et cohérence didactique en langue. *New Standpoints,* 12, 3-7.

Puren, C. (2002b). Perspectives actionnelles et perspectives culturelles en didactique des langues-cultures : vers une perspective co-actionnelle co-culturelle. *Langues Modernes*, 3, 55-71.

Puren, C. (2003). De l'entrée par les tâches à la perspective co-actionnelle co-culturelle. Conférence inaugurale au XXVe Congrès de l'APLIUT *A la recherche de situations communicatives authentiques : l'enseignement des langues par les tâches*, Auch, IUT Paul Sabatier, 5 – 7 juin 2003, à paraître.

Puren, C. (2004). De l'approche par les taches à la perspective co-actionnelle; à la recherche de situations communicatives authentiques : l'apprentissage des langues par les taches. *Les Cahiers de l'APLIUT,* 23 (1), 10-26.

Raîche, G. (2004). Le testing adaptatif. In Bertrand, R., et Blais J. G. (2004). *Modèles de Mesure, l'Apport de la Théorie des Réponses aux Items* (317-348). Sainte-Foy : Presses de l'Université du Québec.

Raschor, R. E., et Gray, G. T. (1996). *Must all stems be green? A study of two guidelines for writing multiple choice stems.* Communication présentée au congrès annuel de l'American Educational Research Association, New York.

Rasch measurement transaction. (2005). Rasch dichotomous model vs. One-parameter Logistic Model. *Rasch Measurement Transaction*, 19 (3). *Consulté le 1/06/2006 à* www.Rasch.org/rmt/rmt193h.htm.

Reckase, M. D. (1990). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests*. Article présenté à la réunion annuelle de American Educational Research Association, Boston.

Reckase, M. D. (1997). The past and the future of multidimensional item response theory. *Applied Psychological Measurement*, 21, 25-36.

Richards, J. C., et Rodgers T. S. (2001). *Approaches and methods in language Teaching, second edition*. Cambridge: Cambridge Language Teaching Library, Cambridge University Press.

Robinson, P. (2001). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics, 21* (1), 27-57.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. *Educational Measurement : Issues and practice*, 24 (2), 3-13.

Ryan, K. et, Bachman, L. (1992). Differential item functioning on two test of E.F.L. proficiency. *Language Testing*, 9, 12-29.

Savignon, S. J. (1983).*Communicative Competence: Theory and Classroom Practice: Texts and Contexts in Second Language Learning*. Reading : MA: Addison-Wesley.

Sasaki, M. (1991). A comparison of two methods for detecting differential items functioning in an E.S.L. placement test. *Language Testing*, 8 (2), 95-111.

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*, 83-117.

Shimizu, Y., et Zumbo, B. D. (2005). A logistic regression for differential item functioning primer. *Japan Language Testing Association Journal,* 7, 110-124.

Shohamy, E. (2001). *The power of test, a critical perspective on the use of language test.* Harlow : Longman.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*(1), 38-62.

Skehan, P., et Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research,* 1, 185–211.

Skehan, P., et Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning,* 49, 93–120.

Spaan, M. (2006). Test and Item Specifications Development. *Language Assessment Quarterly*, *3*(1), 71–79.

Smith, R. M., Schumacker, R. E., et Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2, 66-78.

Stanovich, K. E. (1980). Towards an interactive compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*, 16 (1), 32-71.

Stanovich, K. E. (2000). *Progress in understanding reading : scientific foundations and new frontiers.* New York: Guilford Press.

Stenner, A. J. (1996). *Measuring Reading Comprehension With the Lexile Framework*. N.A.C.A.

Stenner, A. J., et Stone, M. H. (2004). *Does the reader comprehend the text because the reader is able or because the text is easy?* Communication faite à International Reading Association Reno–Tahoe, Nevada May 4, 2004. Consulté en ligne le 12/12/2006 à : http://www.lexile.com/LexileArticles/ReaderAbilityvReadabilitly.pdf

Swan, M. (2005). Legislation by Hypothesis: The Case of Task-Based Instruction *Applied Linguistics* 26, 376-401.

Springer, C. (2002). Recherche sur l'évaluation en langue 2 : de quelques avatars de la notion de compétence. In Castellotti, V. et Py, B. *Notions En Question, La Notion de Compétence en Langue.* Lyon : ENS Edition.

Thissen, D. (2000). Reliability and measurement precision. In Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., et al. (Eds.). *Computerized adaptive testing: A primer* (2e éd.) (159-184). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., et Mislevy, R. J. (2000). Testing Algorithms. In Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., et al. (Eds.). *Computerized Adaptive Testing: A Primer* (2e éd.) (101-134). Hillsdale, NJ: Lawrence Erlbaum.

Trites, L., et McGroarty, M. (2005). Reading to learn and reading to intergrate : new tasks for reading comprehension tests ? *Language Testing,* 22 (2), 174-210.

Van der Maren, J. M. (1996). *Méthodes de recherche pour l'éducation*. Presses de l'Université de Montréal, De Boeck Université.

Van der Linden, W. (2005). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement*, 42 (3), 283-302.

Vygotsky, L. S. (1985). *Pensée et Langage*. Paris, La Dispute.

Wainer, H., et Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implement large-scale computerized testing. In Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., et al. (Eds.). *Computerized Adaptive Testing: A Primer* (2e éd., 271-301). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., Dorans, N. J., Green, B. F., Mislevy, R. J., Steinberg, L., et Thissen, D. (2000). Future challenges. In Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., et al. (Eds.). *Computerized Adaptive Testing: A Primer* (2e éd., 271-301). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., et Kiely, G. L. (1987). Item clusters and computerized adaptive testing : a case for testlets. *Journal of Educational Measurement*, 24 (3), 185-201.

Wainer, H., et Mislevy, R. J. (2000). Item response theory, item calibration and proficiency estimation. In H. Wainer, N. J. Dorans, D. Eignor, R. Flaugher, B. F. Green, R. J. Mislevy, et al. (dir.). *Computerized Adaptive Testing: A Primer* (2e éd., 61-100). Hillsdale, NJ: Lawrence Erlbaum.

Wainer, H., et Wang, X. (2001). *Using a new statistical model for testlets to score TOEFL.* (TOEFL Technical Report No. 16). Princeton, NJ: Educational Testing Service.

Wang, W. C. et Chen, C. T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models. *Educational and Psychological Measurement*, 65 (3), 376-404.

Weir, J. C. (2005a). Limitations of the common european framework for developing comparable examinations and tests. *Language Testing*, 22 (3), 281–300.

Weir, J. C. (2005b). *Language Testing and Validation : An evidence-based Approach.* Research and practice in Applied linguistics. Basingstoke. Palgrave Mc Millan.

Widdowson, H. G. (1981). *Une approche Communicative de l'Enseignement des Langues*. Paris : Hatier-Crédif, Coll. LAL.

Wilson, M. (2005). *Constructing Measures: An Item Response Modeling Approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wright, B. D., et Masters, G. N. (1982). *Rating Scale Analysis*. Chicago, IL: MESA Press,.

Wright, B. D., et Stone, M. H. (1979). *Best Test Design*. Chicago, IL: MESA Press.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement,* 30 (3), 187–213.

Young, Y., Shermis, M. D., Brutten, S. R., et Perkins, K. (1996). From conventional to computer-adaptive testing of ESL reading comprehension. *System,* 24, 23-40.

Zarate, G. (2004). Les langues vivantes : d'une vision nationale de l'identité française à une vision européenne et internationale de la France, *Administrer l'Enseignement des Langues Vivantes*. *Administration et Education*, 101, 7-19.

Zanón, J. (1999). *La Enseñanza del Español Mediante Tareas*. Madrid: Edinumen.

TRADUCTIONS

Les traductions qui sont proposées ici ne sont absolument pas l'œuvre d'un professionnel. Elles ont pour objectif modeste d'aider le lecteur qui n'atteindrait pas un niveau de compréhension suffisant avec le texte en anglais.

---

[1] « C'est une nouvelle encourageante pour les développeurs de test en langue intéressés par la création de descripteurs plus complets et valides. Eventuellement, si une nouvelle version du CECR, qui décrive mieux le contenu par niveau, est développée par les spécialistes en évaluation […], la couverture de contenu d'un examen pourra être décrite avec cet outil. Après la calibration et les procédures d'étalonnages, les résultats de l'analyse de contenu pourront être analysés au regard des valeurs psychométriques. La validité fondée sur la théorie d'une échelle de niveau, bien que de moindre importance, pourrait prendre plus de temps à être définie. La nature des processus cognitifs et métacognitifs en production écrite et orale est moins ouverte et susceptible d'être l'objet de recherche que les variables contextuelles, qui sont plus facilement assujetties à l'examen de l'expert et aux recherches empiriques. »

[2] « Le deuxième postulat de la T.R.I. est l'homogénéité de la population de candidats. Autrement dit, le seul trait qui cause des différences entre les résultats des candidats est leur compétence sur le trait mesuré par le test. »

[3] « Étant donné la preuve (pas très forte) de l'impact des variables linguistiques, comme la connaissance de la syntaxe, en lecture langue maternelle, les concepteurs de test devraient examiner attentivement le langage utilisé pour les questions, les consignes et les textes pour s'assurer qu'ils sont adéquats pour évaluer l'étendue de compétence de la population visée . Aussi une stratégie possible, si les textes sont jugées trop difficiles pour un groupe donné d'apprenants, est de simplifier les textes. Non seulement, cela rendra caduque l'authenticité du texte, mais, risque également de rendre le texte plus difficile à comprendre. En outre, une compétence à lire des textes simplifiés est peu facilement généralisable, à une compétence à lire des textes authentiques. Un moyen plus approprié pour ajuster la difficulté du texte pourrait être de développer des tâches ou des questions d'examen plus faciles. »

[4] « […] les tests informatisés et les algorithmes de sélection d'items menacent la représentativité du contenu si les décisions de sélection sont basées uniquement sur les indices statistiques d'indices de qualité (par exemple, la difficulté de l'item, la discrimination de l'item. »

[5] «Les tests de performance, dans un sens faible, semblent avoir eu une validité apparente si convaincante qu'ils ont réussi à convaincre les concepteurs de tests eux-mêmes qu'ils étaient plus que des tests de compétence linguistique. »

[6] « Dans un test, l'unidimensionnalité psychologique implique que les résultats aux tests sont destinés à être interprétés comme étant représentatifs du degré de présence de certains construits ou traits psychologiques connus. […] Ainsi, à partir de la performance aux items, des inférences peuvent être faites sur le degré de présence du construit visé. »

« L'unidimensionnalité psychométrique est similaire à l'unidimensionnalité psychologique, en ce que les deux font référence à la capacité d'un test de mesurer une dimension ou un trait majoritaire, mais l'unidimensionnalité psychométrique diffère de l'unidimensionnalité psychologique dans plusieurs aspects importants que j'ai l'intention de démontrer dans cette étude. Tout d'abord, l'unidimensionnalité psychométrique peut être présente lorsque le test mesure une variété de dimensions psychologiques sous-jacentes corrélées. En outre, l'unidimensionnalité psychométrique peut être présente même en l'absence de possibilité d'interprétation explicite de la dimension primaire prétendument mesurée, mis à part la définition opérationnelle fournie par les items eux-mêmes. »

[7] « L'unidimensionnalité n'est pas une question du type oui / non, c'est plutôt une question de degré, à interpréter en fonction de l'objectif du test. Dans quelle mesure, dans une situation donnée, un écart vis-à-vis de l'unidimensionnalité amènera à exclure l'utilisation d'un test ? En fait, nous savons qu'un test de langue n'est jamais complètement unidimensionnel. En outre, les aspects liés à l'unidimensionnalité ne devraient pas obliger les développeurs de test à restreindre la nature et l'éventail des tâches alors que la validité a besoin de diversité et de complexité. L'analyse de la dimensionnalité dans le cadre de l'étude de la validité du test est un processus à long terme, à la recherche d'un construit suffisant, mais perfectible. »

[8] « De manière générale, plus l'échelle des items est strictement unidimensionnelle, moins ambiguë est l'interprétation des résultats bruts associés à l'échelle et les corrections pour l'atténuation sont légitimes (Smith, 1996). Aussi, l'application des modèles de mesure de la T.R.I unidimensionnels est plus valide et raisonnable. Si les réponses aux items sont influencées par deux ou plusieurs facteurs communs, le chercheur devra envisager de considérer la création et la notation avec des sous-échelles ou l'application de modèles de la T.R.I. multidimensionnels (Reckase, 1997). »

[9] « **La difficulté sur un continuum.** Dans une certaine mesure, les quatre buts du lecteur forment une sorte de continuum de la difficulté. Pour en être certain, des tâches faciles pourraient être conçues pour la lecture destinée à apprendre, la lecture destinée à intégrer de l'information et des tâches difficiles demandant aux candidats de trouver des informations discrètes ou pour lire en mettant en œuvre une compréhension de base pourraient être élaborées en manipulant les variables linguistiques / syntaxiques des tâches. Pourtant, nous sommes plus susceptibles de trouver les objectifs « lire pour apprendre » et « lire pour intégrer de l'information » dans des textes associés à des tâches académiques plus difficiles et demandant des processus associés à des compétences plus sophistiquées que les compétence de lecture à trouver de l'information discrète ou la compétence de lecture utilisant la compréhension générale. Le résultat est que, selon que le but du lecteur est de lire pour trouver de l'information, lire pour une compréhension de base, lire pour apprendre ou lire pour intégrer de l'information après avoir lu plusieurs textes, plus de compétence de lecture et de stratégies efficaces sont requises. Nous pensons donc que le but du lecteur lui-même doit être une des variables qui peuvent contribuer à la difficulté de la tâche lorsqu'il est combiné avec des textes et des tâches appropriés. Il sera important d'examiner cela dans le cadre du programme de recherche. »

[10] « Considérant les résultats pour, à la fois, les locuteurs langue étrangère et des locuteurs natifs, nous concluons que les nouvelles tâches évaluent quelque chose de différent de la maîtrise de base de l'anglais universitaire. »

[11] « Nous avions espéré trouver des preuves évidentes d'une hiérarchie suggérant que la lecture pour apprendre était manifestement plus difficile que la compréhension de base de lecture et que la lecture pour intégrer des informations était manifestement plus difficile que la lecture pour apprendre, mais les résultats n'ont pas permis de dégager une hiérarchie aussi évidente. »

[12] « Cependant, les formules de lisibilité donnent uniquement la mesure brute de la difficulté du texte, et sont rarement adaptées pour les lecteurs en langue étrangère ou en langue seconde [...]. Étant donné l'éventail de variables qui influent sur la difficulté du texte (le sujet, la complexité syntaxique, la cohésion, la cohérence, le vocabulaire et la lisibilité) les concepteurs de tests en langue devraient se méfier d'une approche simpliste de la difficulté en langue lorsqu'ils concoivent leurs tests [...]. Dans de nombreux cas, la difficulté du texte ne sera pas définie en termes absolus, et au contraire les concepteurs de tests préfèreront identifier un éventail de textes authentiques susceptibles d'être lus dans des situations qui seront vécues par la personne qui passe le test. »

[13] « La plupart des chercheurs convient que la lecture est rapide, dirigée vers un but et interactive (Grabe, 1999 ; Alderson, 2000). Plus précisément, cette interactivité intervient à deux niveaux. Tout d'abord, quels que soient les composants ou des niveaux exacts qu'ils positionnent dans leur test, les modèles de processus de lecture décrivent ces composants ou niveaux comme interagissant les uns avec les autres. Ensuite, et ce qui est encore plus pertinent pour cette étude, la lecture est interactive en ce sens que les connaissances

antérieures du lecteur et d'autres caractéristiques interagissent avec le contenu des textes. Étant donné que la lecture implique l'interaction entre les lecteurs et les textes, il s'ensuit logiquement que les caractéristiques à la fois du lecteur et du texte auront une incidence sur le processus de lecture. »

[14] « Une tâche discrète est une tâche qui met l'accent sur un élément de la langue isolé, généralement les caractéristiques de surface de la phonologie, morphologie, syntaxe, ou du lexique [...] Dans leur forme la plus pure, les tâches discrètes incluent, un seul canal (par voie orale ou écrite) et une seule direction (production ou réception), c'est pourquoi, ils testent "séparément" les compétences d'écoute de document, de lecture, d'expressions orale et écrite. »

[15] « Dans des approches privilégiant les aspects discrets, l'intention est de tester une« chose » à un moment donné, dans les approches visant l'évaluation d'éléments intégrés, les concepteurs de tests visent une perspective beaucoup plus générale pour avoir une idée de la façon dont les élèves lisent. Dans ce dernier cas, les choses sont ainsi car on reconnaît que l'ensemble est plus grand que la somme des parties. »

[16] « Selon mon appréciation des différences individuelles, on peut prévoir que les individus diffèrent dans leur capacité d'adaptation pour lire des textes dans de multiples environnements, tout comme ils diffèrent dans leur capacité à comprendre des textes simples. »

[17] « Ils ont constaté que certains élèves pouvaient répondre en raison de leur expérience en dehors de l'école et de leur expérience avec les examens, ce qui jette un doute sur la capacité des QCM à mesurer la compréhension en lecture. Le problème avec la mesure de la compréhension écrite n'est pas avec le format utilisé, mais avec l'écriture d'items qui dépendent véritablement du passage à lire ou encore auxquels on répond indépendamment [des autres items, des connaissances antérieures]. »

[18] «Malheureusement, ils ont aussi l'inconvénient qu'il est possible pour un candidat de bluffer (par exemple, un étudiant qui est intelligent peut construire une réponse linguistique sophistiquée et bien organisée qui cependant ne correspond pas à la tâche à accomplir, et obtenir ainsi une partie ou la totalité des points), l'administration et la notation prennent beaucoup de temps, et la notation est à la fois difficile et quelque peu subjective. »

[19] «La sélection du modèle peut être facilitée par une révision des principaux postulats qui sous-tendent les modèles de réponses à l'items unidimensionnels les plus courants. Deux postulats communs à tous ces modèles sont que les données sont unidimensionnelles et la passation du test n'était pas trop rapide. Un autre postulat pour les modèles à deux paramètres est les réponses trouvées au hasard sont très peu nombreuses ; un autre postulat pour le modèle à un paramètre est que tous les indices de la discrimination sont égaux. »

[20] «L''hypothèse d'absence de réponse trouvées au hasard est plus plausible avec des réponses ouvertes, mais cette hypothèse peut souvent être partiellement satisfaite avec des questions à choix multiples quand un test n'est pas trop difficile pour les candidats. »

[21] «Toutefois, en choisissant des données qui collent au modèle, il est possible pour le modèle de révéler des anomalies dans les données, des anomalies qui doivent être interprétées, mais pour lesquelles le modèle ne fournit aucune réponse de fond – on obtient essentiellement des indices où chercher ».

[22] «Au lieu de simplement décrire les données, les modèles de Rasch fournissent une occasion de comprendre les données par la découverte d'anomalies, ce qui est la fonction première de la mesure dans la recherche en sciences physiques (Kuhn 1961/1977). La raison pour laquelle le modèle de Rasch peut être utilisé de cette façon est que le modèle ne décrit pas toutes les données, mais qu'il formalise les conditions d'invariance, qui conduisent à des propriétés de mesure. Ainsi, lorsque des données s'écartent du modèle de Rasch, elles s'écartent des exigences de mesure. Considérer, lorsqu'il y a un décalage entre les données et le modèle, qu'il pourrait s'agir d'un problème avec les données plutôt que le modèle, est en soi un

considérable changement de perception de la perspective traditionnelle des relations entre le modèle et les données. »

[23] « Les statisticiens peuvent trouver difficile de s'adapter à la méthodologie de Rasch. Ils ont tendance à croire que les données disent la vérité et que c'est la tâche des statisticiens de trouver des modèles, qui les expliquent, et de trouver les variables latentes, qui les sous-tendent. La méthodologie de Rasch prend une position inverse. Elle dit que la variable latente est la vérité, et lorsque cette variable latente est exprimée en termes linéaires, c'est le modèle de Rasch qui est nécessaire et suffisant pour la décrire. Par conséquent, les données, qui ne sont pas conformes au modèle de Rasch, donnent une image déformée de la variable latente. Elles semblent nous dire des choses très importantes, par exemple, «les étudiants ont été peu intéressés", "la bonne réponse est fausse" - mais cela ne peut pas être associé à la variable centrale. »

[24] «C'est donc le travail du développeur de test de déterminer empiriquement le nombre optimal de catégories de réponses à chaque fois qu'une nouvelle échelle de notation est développée ou quand une échelle de notation déjà existante est modifiée ou quand une échelle de notation est utilisée avec une nouvelle population. Ainsi, l'analyste doit découvrir empiriquement, plutôt que de le présupposer, le nombre optimal de catégories pour une échelle d'évaluation mesurant un construit donné (Lopez, 1996). »

[25] « La probabilité d'une réponse correcte est fonction de la compétence de la personne et de la difficulté de l'item, avec une correction faite pour la sévérité de l'évaluateur, et pour une déclinaison particulière de test qui a été passée (la probabilité de réponse = fonction (compétence + difficulté + évaluateurs +test). »

[26] « Aujourd'hui, […] en plus des questions concernant les biais, la technologie FID est utilisée afin d'aider à répondre à une variété de questions de recherches fondamentales et appliquées sur les aspects liés à la mesure, notamment, là où l'on veut comparer la performance à l'item entre ou parmi les groupes lorsque l'on tient compte de la distribution de la compétence. Sur ce point, les méthodes développées pour le FID ont plus en commun avec la méthodologie de recherche alignée avec l'analyse de la covariance (ANCOVA) ou de l'interaction des caractéristiques par le traitement que le biais de test en soi. »

[27] « Pour une précision de la mesure égale, les items doivent être choisis de telle manière que la courbe d'information du test soit relativement plate à travers tout l'éventail du trait mesuré. Cela signifie qu'un ensemble d'items très discriminants avec un large éventail de niveau du paramètre de la difficulté doivent être identifiés. »

[28] « Les implications de l'inaptitude des étudiants a bien percevoir la difficulté posent la question de ce que chaque élève apporte à une tâche en terme d'effort, et si cet effort est tempérée par la perception qu'ils ont de la tâche. Plus simplement dit, si la question est plus difficile, l'élève fera-t-il plus d'efforts ? »

[29] « La commande BIFACTOR est utilisée pour avoir des estimations de l'information complète de la saturation sur un facteur général en présence de facteurs portant sur des groupes d'items. »

[30] « TESTFACT comprend encore une autre méthode de l'information complète qui fournit une forme importante d'analyse factorielle confirmatoire dénommée «analyse bifactorielle ». Le patron de facteurs dans l'analyse bifactorielle est constitué d'un facteur général sur lequel toutes les questions ont une certaine saturation, ainsi qu'un certain nombre de facteurs appelés « facteurs de groupe » composés de sous-ensembles d'items sans chevauchement, sous-ensemble attribués par l'utilisateur. Les sous-ensembles représentent en général un petit nombre d'items qui se rapportent à un ensemble plus grand comme un passage de lecture ou des exercices de résolution de problèmes. »

[31] « Ces résultats constituent de solides preuves de structures de la dimensionnalité de la compréhension écrite identifiables et significatives qui va au-delà de l'effet des minitests et avec des implications importantes pour la mesure. »