2m11.2927.2

Université de Montréal

Étude exploratoire sur l'utilisation du *Rating Scale Model* de la famille de Rasch dans le processus de validation d'un instrument de cueillette d'information

par

Patrik Maheux

Département d'études en éducation et d'administration de l'éducation

Faculté des sciences de l'éducation

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès arts (M.A.)
en mesure et évaluation en éducation

Novembre 2001

©Patrik Maheux, 2001

Faculté des Études Supérieures
Grade octroyé à compter du
07 FEV 2002
Université de Montréal

LB
5
N57
2002
v.008

Page d'identification du jury

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :
Étude exploratoire sur l'utilisation du *Rating Scale Model* de la famille de Rasch dans le processus de validation d'un instrument de cueillette d'information

présenté par

Patrik Maheux

a été évalué par un jury composé des personnes suivantes :

Jean-Guy Blais

Michel Laurier

Serge Racine

Mémoire accepté le : _____

SOMMAIRE

La modélisation de scores à l'aide de la théorie de la réponse (TRI) à l'item offre une avenue intéressante pour la mesure. Ainsi, cette modélisation offre la possibilité de construire un instrument pour lequel les scores recueillis seraient linéaires et invariants d'un échantillon à l'autre pour une même population. Toutefois, la présence de fonctionnement différentiel de l'item (FDI) entre les échantillons pourrait potentiellement avoir un impact sur cette modélisation. Les objectifs de cette recherche consiste à faire l'exploration de l'utilisation de cette forme de modélisation et à constater l'impact que peut avoir le FDI sur celle-ci.

Pour réaliser ces objectifs, le *Rating Scale Model* est appliqué aux réponses fournies par des étudiants[i] pour vingt énoncés provenant d'un questionnaire. L'échelle d'appréciation utilisée est de type Likert et la population est formée d'étudiants de deux programmes (enseignement secondaire et éducation préscolaire et primaire) de la formation des maîtres pour deux années différentes (2000 et 2001). La prémisse de la recherche est que, dans la mesure où les données recueillies s'ajustent au modèle, ces données devraient répondre aux caractéristiques offertes de la TRI. La méthodologie consiste premièrement à vérifier l'ajustement des données aux modèles et, ensuite, à éliminer les items pour lesquels les données ne s'ajustent pas au modèle pour chaque sous-groupe de la population de référence. Ainsi, le reste des items devrait répondre aux caractéristiques de la TRI. Il faut aussi étudier l'ajustement des données aux catégories de l'échelle. Ensuite, les sous-groupes de la population sont analysés dans le but de détecter les items pour lesquels il y a FDI. Le questionnaire a été modifié intentionnellement dans le but d'induire un FDI. Le croisement de ces résultats donne l'impact du FDI sur la modélisation. Finalement, pour les données estimées par le modèle, une vérification de l'invariance de ces données est effectuée.

L'analyse de l'ajustement des données au modèle suggère qu'il faudrait éliminer des items pour obtenir un ajustement entre les données et le *Rating Scale Model* peu importe le sous-groupe analysé. En moyenne, 30 % des items doivent être éliminés. Pour les catégories de l'échelle, les données ne s'ajustent pratiquement jamais, ce qui suggère que le nombre de catégories à l'échelle serait à réviser. Pour la détection du FDI, il n'a pas été possible de savoir si les modifications apportées au questionnaire ont réussi à produire un FDI. Tout de même, des items

i Simplement dans le but d'alléger la lecture du texte, le masculin est utilisé.

ont démontré du FDI de certains sous-groupes à d'autres. Ceux-ci ne semblent toutefois pas avoir eu d'impact sur la modélisation à l'aide de *Rating Scale Model*, c'est-à-dire sur l'ajustement des données. Dans le même ordre d'idée, la modélisation à partir du *Rating Scale Model* ne semble pas avoir permis d'obtenir des valeurs estimées invariantes d'un sous-groupe à l'autre.

Ces résultats ne signifient toutefois pas que l'utilisation de la TRI dans un processus de validation ne comporte pas d'avantages. La TRI permet tout de même d'identifier les items à examiner pour améliorer le questionnaire. Elle permet aussi d'étudier le nombre de catégorie de l'échelle d'appréciation et de s'interroger sur l'homogénéité de la population et sur la possibilité de distribuer des questionnaires différents aux sous-groupes qui forment la population. En plus, avec un processus de validation à long terme, il est possible que les données de l'instrument finissent par répondre aux caractéristiques de la TRI.

Avant de conclure aux avantages de l'utilisation de la TRI dans un processus de validation, d'autres recherches seront nécessaires. Reprendre cette même recherche avec des outils différents, une population plus grande ou un autre modèle constituerait une piste intéressante. Une autre piste pourrait être de continuer le processus de validation en lumière des résultats de cette recherche.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

REMERCIEMENTS

Premièrement, j'aimerais profiter de ces quelques lignes pour remercier mon directeur de recherche, M.Jean-Guy Blais, pour sa grande disponibilité, pour la justesse de ses commentaires, pour son sens critique et sa rigueur. Mais, surtout, il a su à quel moment une direction était nécessaire et à quel moment une réflexion autonome devait faire son chemin.

Ensuite, j'aimerais remercier le Centre de Formation Initiale des Maîtres de la Faculté des sciences de l'éducation de l'Université de Montréal pour m'avoir donné la permission d'utiliser les données provenant de leur questionnaire.

J'aimerais aussi remercier les personnes de mon entourage qui m'ont offert un support inconditionnel tout au long de ce travail grâce à leur patience, leur encouragement et leur amour.

INTRODUCTION

Depuis maintenant deux ans, le Centre de Formation Initiale des Maîtres (CFIM) de l'Université de Montréal distribue un questionnaire à ses étudiants pour s'enquérir de leur opinion ou de leur attitude face à leur programme universitaire de formation des maîtres. Ce questionnaire (annexe A), divisé en sept sections, touche différentes parties de la formation reçue avec une dernière section sur les caractéristiques des étudiants qui répondent au questionnaire. Parmi ces sections, la deuxième interroge les étudiants au sujet de « la préparation à l'enseignement » reçue jusqu'ici dans leur formation. Cette section, et plus particulièrement sa validation, constitue l'objet cette recherche.

Selon Messick (1989), le processus de validation d'un instrument consiste principalement à recueillir des indices de la validité de l'instrument à partir de six sources potentielles : 1) la pertinence du contenu des items de l'instrument, 2) la constance dans la réponse des candidats, 3) la structure interne ou la relation entre la substance et la pertinence, 4) la structure externe ou la comparaison avec d'autres instruments externes, 5) la possibilité de généraliser les résultats à d'autres situations et, finalement, 6) les conséquences sociales de l'utilisation de cet instrument. Chacune de ces sources donne une indication de la validité de l'instrument et celles-ci sont présentées au chapitre III. La recherche qui suit s'intéresse à deux d'entre elles : la constance dans la réponse des candidats et la structure interne.

Ces deux sources de validité peuvent s'étudier par l'entremise d'une modélisation des réponses cueillies par l'instrument. Parmi les modélisations possibles, une première modélisation « traditionnelle » s'inspire de la théorie classique des tests, mais certaines lacunes lui sont associées tel qu'illustré au chapitre I. Dans un effort pour combler ces lacunes, il est possible de s'inspirer d'une modélisation alternative qui porte le nom de « théorie de la réponse à l'item ». Théoriquement, cette modélisation offre une solution intéressante face aux lacunes et limites associées à la théorie classique des tests. La théorie de la réponse à l'item est présentée au chapitre II. Parmi la théorie de la réponse à l'item, une famille s'est développée autour du modèle de Rasch. Dans cette famille, on trouve un modèle qui porte le nom de *Rating Scale Model* (Andrich, 1978; 1978b). À l'aide de ce modèle, nous allons étudier la structure interne du questionnaire et l'impact de la présence d'un biais de mesure sur celle-ci.

Un biais de mesure est une caractéristique externe qui influence la façon de répondre des candidats et qui peut avoir des conséquences sur l'interprétation des résultats ou sur la modélisation des données. Pour étudier ce phénomène, nous avons volontairement essayé d'insérer un biais de mesure dans la composition du questionnaire pour ensuite en évaluer l'impact sur la modélisation des réponses.

L'étude exploratoire de ces deux volets du processus de validation à l'aide du *Rating Scale Model* de la famille de Rasch constitue l'objet principal de cette recherche qui a comme point de départ un questionnaire d'enquête destiné à des étudiants des programmes de la formation des maîtres.

Le présent document se divise en six chapitres. Le premier chapitre situe la recherche par rapport à son contexte et définit le problème et les questions auxquels la recherche propose des éléments de réponse. Le deuxième chapitre présente les concepts importants de la théorie de la réponse à l'item et le troisième chapitre décrit certaines caractéristiques du processus de validation d'un instrument et de ses sources de validité. Le quatrième chapitre présente la méthode utilisée pour procéder à la recherche et répondre aux questions qu'elle pose. Les deux derniers chapitres analysent les résultats obtenus et discutent de ceux-ci en fonction des questions de la recherche. Finalement, la conclusion résume les résultats de la recherche et propose quelques pistes pour de futures recherches sur le sujet.

CHAPITRE I

CONTEXTE DE LA RECHERCHE

L'utilisation d'instruments de cueillette d'information, tels des questionnaires de sondage ou des tests de connaissance, fait partie du quotidien des chercheurs en sciences sociales, maisons de sondage, institutions scolaires et autres. Par exemple, une institution scolaire peut souhaiter obtenir l'opinion des étudiants au sujet de son programme de formation dans le but d'y apporter des modifications. Dans ces circonstances, les concepteurs de l'instrument peuvent demander aux répondants d'encercler le choix qui correspond le mieux à leur opinion par rapport à un énoncé. Par exemple, à l'énoncé « Je consacre vingt heures par semaine à mes études », l'étudiant donne son opinion à partir de choix offerts comme : 1=Très favorable; 2=Plutôt favorable; 3=Plutôt défavorable; et 4=Très défavorable. Cette forme de présentation des choix constitue une échelle de référence et elle est connue sous le nom d'échelle de Likert. À partir des réponses données, il arrive que les concepteurs et les utilisateurs résument l'information obtenue en additionnant les codes numériques qui correspondent aux différentes étiquettes de l'échelle, pour créer un score total. Les concepteurs utilisent très souvent ces scores comme une mesure à partir de laquelle ils peuvent faire des inférences au sujet des répondants en regard de l'information que l'instrument cherchait à recueillir. Dans plusieurs cas, ces inférences sont à la source de différentes prises de décision.

La théorie classique des tests (TCT) propose permet de faire une interprétation de ce score total. Fondamentalement, la TCT affirme que la seule chose qui différencie le score réel (SR)[ii] du score observé à l'instrument (SO) est une erreur de mesure (E). Cette erreur serait aléatoire,

ii Dans certains textes, on peut aussi retrouver « score vrai » en rapport avec l'expression anglaise True Score. Pour ce texte, « score réel » sera maintenu.

indépendante de toute autre variable et non systématique. C'est-à-dire qu'elle n'entretient aucune relation avec le score réel, l'erreur de mesure à d'autres tests ou les caractéristiques du candidat. Avec la TCT, lorsqu'un candidat répond plusieurs fois au même instrument, l'erreur de mesure devrait s'estomper et le score observé s'approcherait idéalement du score réel. Ceci reste toutefois difficile à vérifier dans la pratique. Plusieurs auteurs (Embretson, 1999; Hambleton,1981; Laurencelle, 1998; Laveault et Grégoire, 1997) utilisent l'équation suivante pour décrire la relation entre le score réel, le score observé et l'erreur :

$$SR = SO + E. \quad (1)$$

Spector (1992) reprend l'équation (1) pour y inclure un deuxième type d'erreur, c'est-à-dire le biais de mesure (B) :

$$SO = SR + E + B. \quad (2)$$

Cette erreur est plus systématique et généralement associée à un sous-groupe de la population. Par exemple, il pourrait y avoir un biais si les filles réussissaient systématiquement mieux que les garçons à certaines questions d'un examen de classement qui utilisent accidentellement des termes plus familiers aux filles. Nous reviendrons plus en détail sur le biais de mesure à la section 3.3.

1.1 PROBLÉMATIQUE DE LA RECHERCHE

Hambleton, Swaminathan et Rogers (1991) et Embretson (1999) soulèvent la présence de certains problèmes reliés aux deux équations ci-haut et à l'utilisation du score total comme source première des inférences. Premièrement, la considération du seul score total ne permet pas de sortir du contexte de l'instrument et des répondants. Par exemple, dans le cas d'un test, le score (SO) des sujets dépend de la difficulté des questions, et l'évaluation de la difficulté des questions dépend du thêta [iii] des sujets. Tout dépend des sujets et des questions, l'information obtenue se limite au contexte spécifique de cet instrument de cueillette d'information. Deuxièmement, le score total à un instrument n'apporte aucune information sur la performance

iii Pour le reste de ce texte, thêta se définit comme la valeur (ou la quantité) attribuée à un répondant peu importe ce que l'instrument cherche à recueillir comme information. Pour un test, thêta équivaut à l'habileté; pour un sondage, thêta équivaut à son degré d'accord avec les énoncés, etc.

du sujet à une question (item) [iv] précise. Troisièmement, l'erreur de mesure est la même pour chaque sujet, c'est-à-dire qu'elle ne dépend pas de la position relative de ce dernier par rapport aux autres répondants. Quatrièmement, la TCT ne serait pas toujours très efficace pour identifier les biais de mesure. Finalement, la fidélité d'un instrument se calcule à partir d'administrations répétées de l'instrument. Pour plus de détails sur ces problèmes, il est suggéré de se référer aux auteurs cités au début de ce paragraphe.

En plus de ces problèmes, la TCT n'apporte pas d'information qui assure que les unités du score total sont équivalentes d'un répondant à l'autre, d'un instrument à l'autre. Prenons l'exemple spécifique d'un test de connaissance pour illustrer cette variation dans l'unité. Deux étudiants forts et deux étudiants faibles passent deux tests, un facile et un difficile, et ils obtiennent les résultats suivants : au test facile, les étudiants faibles obtiennent 7 et 8 (sur 10) et les étudiants forts 10 et 10; ensuite au test difficile, les étudiants faibles obtiennent 0 et 0 et les étudiants forts 8 et 9. La différence de 1 pour les étudiants faibles au test facile n'a pas la même valeur que la différence de 1 pour les étudiants forts au test difficile. Pour franchir l'échelon 1, la marche est plus grande pour les étudiants forts. C'est pourquoi, selon Wright (1979; p. xi), un score ne peut avoir une interprétation valable hors de son contexte parce que notre unité de mesure change d'un instrument à l'autre. Il critique ce manque d'objectivité dans son introduction : « *Comment pouvons-nous faire une mesure mentale objective et construire une science du développement mental en travaillant avec une échelle élastique ?* ».

En réponse à ces problèmes et à ces questions, Andrich (1988) suggère d'utiliser une modélisation différente de celle proposée par la TCT. Comme Hambleton, Swaminathan et Rogers (1991 : p. 5) l'indiquent, il serait préférable d'avoir un modèle mathématique qui réponde aux besoins suivants : « *a) les caractéristiques des items ne dépendent pas des groupes, b) les scores qui décrivent la performance ne dépendent pas du test, c) le modèle s'exprime au niveau des items et non pas au niveau du test, d) le modèle ne nécessite pas de tests parallèles pour calculer la fidélité, et e) le modèle donne une mesure de précision pour chaque score d'habileté.* » Il faudrait aussi chercher un modèle qui stabilise l'unité de mesure de l'instrument. Il faudrait en plus que ces modèles prennent en considération le biais de mesure (B), pour ne pas perdre de vue l'influence qu'il peut avoir sur la validité de l'interprétation du score.

iv Selon Wainer et Thissen (2001), il faut utiliser « item » lorsqu'une échelle ou un système de codage accompagne l'énoncé et question quand la réponse est ouverte.

Au cours des années 1950-1960, des chercheurs ont développé des modèles qui devraient théoriquement permettre de répondre à l'ensemble de ces demandes. Actuellement, ces modèles se regroupent sous l'appellation de théorie de la réponse à l'item (TRI). Parmi les modèles de la TRI, il y a les modèles de la famille de Rasch qui sont parmi les plus simples et les plus faciles à utiliser (Gustaffson, 1991). Les modèles de Rasch ne sont pas les seuls modèles qui existent mais ils offrent plusieurs avantages importants par rapport aux autres modèles de la TRI comme la possibilité de stabiliser l'unité sur l'échelle (Embretson, 1999). Ces particularités du modèle de Rasch en font un outil théorique intéressant. Ces modèles de la famille de Rasch et de la TRI pourraient potentiellement servir à améliorer le processus de validation des instruments de mesure. Parmi les modèles de la famille de Rasch, le *Rating Scale Model* (RSM : Andrich, 1978a et 1978b) est celui choisi pour cette recherche parce qu'il permet de modéliser des données recueillies à l'aide d'une échelle de Likert.

1.2 OBJECTIFS ET QUESTIONS DE LA RECHERCHE

L'objectif principal de la recherche consiste à explorer l'apport que peut avoir l'utilisation de la TRI dans un processus de validation d'un instrument de cueillette d'information. Dans ce cas précis, l'instrument en question est un questionnaire distribué aux étudiants des programmes de la formation des maîtres au cours des années 1999/2000 et 2000/2001. Face à cet objectif principal, il est maintenant possible de définir une question de recherche précise :

- Quels sont les avantages et les désavantages d'utiliser les modèles de la TRI dans un processus de validation d'un instrument de cueillette d'information ?

Cette question de recherche se décompose en deux questions plus spécifiques :

- Dans quelle mesure le *Rating Scale Model* de Rasch permet-il de construire une échelle de mesure à partir d'un instrument de cueillette d'information ?

- Dans quelle mesure la présence d'un biais de mesure influence-t-elle l'utilisation du *Rating Scale Model* dans cette construction ?

1.2.1 Pertinence des questions de recherche

Linn (1990) s'est déjà posé la première question pour finalement conclure qu'il faudrait plus de recherches sur le sujet. Ces recherches tardent à venir pour l'instant. D'ailleurs, dans la pratique quotidienne de la recherche ou de l'utilisation d'instruments de cueillette d'information, les conditions d'utilisation des modèles de la TRI restent à clarifier malgré le bon nombre d'études présentées dans la littérature. De plus, la littérature n'indique pas si la présence d'un biais de mesure affecte les conditions d'application de la TRI. Cette recherche est donc pertinente puisque l'utilisation de la TRI peut encore être approfondie dans le contexte du processus de validation, parce qu'il manque de recherche concrète sur l'impact de la présence d'un biais dans ces conditions et parce que les conséquences peuvent être importantes pour affirmer la validité des scores de l'instrument.

1.2.2 Choix méthodologiques pour répondre aux questions

Pour répondre aux questions de la recherche, nous avons dû faire certains choix d'ordre méthodologique. Premièrement, nous avons choisi d'utiliser le *Rating Scale Model* parmi les différents modèles de la TRI. Deuxièmement, nous avons choisi un logiciel informatique (CONQUEST) parmi d'autres pour faire la modélisation des données à partir du RSM. Troisièmement, pour étudier le processus de validation et, plus précisément, le biais de mesure, nous avons choisi de modifier l'ordre de présentation des questions ce qui représente une façon de faire parmi beaucoup d'autres. D'autre part, nous n'avons pas privilégié une présentation ou un traitement technique et mathématique des concepts. L'objet de cette recherche n'est pas d'étudier la question sous cet angle.

CHAPITRE II

LA THÉORIE DE LA RÉPONSE À L'ITEM

Ce chapitre est le premier de deux chapitres servant à décrire les concepts clefs de cette recherche. Comme la théorie de la réponse à l'item est relativement récente dans le domaine de la mesure en sciences sociales, il semble pertinent de se familiariser avec ses fondements avant de voir comment elle peut être utile au processus de validation d'instruments de cueillette d'information. Les modèles les plus utilisés de la TRI (tels le RSM et les modèles de la famille de Rasch) ont comme fondements deux hypothèses et une propriété de base. Les hypothèses de l'unidimensionalité et de l'indépendance locale se présentent comme deux conditions à remplir pour utiliser les modèles de la TRI. La propriété de l'invariance est plutôt une caractéristique de ces modèles. En plus de ces hypothèses et de cette propriété, le chapitre introduit les modèles et leurs paramètres, les méthodes d'estimation des paramètres, l'ajustement des données au modèle et, finalement, les programmes informatiques qui facilitent l'application des modèles de la TRI. Cette introduction s'inspire principalement de deux textes (Embretson et Reise, 2000; Hambleton, Swaminathan et Rogers, 1994) et de certains articles sur le sujet.

2.1 HYPOTHÈSES ET PROPRIÉTÉ DE LA TRI

2.1.1 Unidimensionalité

Au moment de construire un instrument de cueillette d'information, le concepteur possède généralement une bonne idée de la variable qu'il souhaite mesurer. Dans certains cas, une théorie de l'objet existe déjà et assure une définition assez claire de la variable. Toutefois, il arrive que des variables ne s'observent pas directement. Dans ce cas, la variable que l'instrument

cherche à mesurer prend le nom de trait latent. L'attitude d'une personne, son intelligence ou son habileté à résoudre un problème mathématique sont des exemples de traits latents.

De préférence, un instrument de mesure cherchera à obtenir de l'information sur un seul trait latent à la fois. Par exemple, un test de géographie s'attarde uniquement à vérifier des connaissances sur la géographie. Malgré cela, il faut tout de même que le candidat possède un minimum de connaissances en langue écrite et en lecture pour répondre aux questions. En plus, d'autres variables comme la durée de l'épreuve, l'environnement ou la nervosité peuvent influencer la réponse du candidat au test. Conséquemment, il est rare qu'un seul trait influence la réponse des candidats. La clé du succès réside alors dans l'isolation du trait. Par exemple, en donnant tout le temps voulu à des répondants qui maîtrisent bien la langue, le concepteur d'un test serait en état de croire que seulement les connaissances de la personne en géographie contribuent aux réponses.

L'hypothèse d'unidimensionalité peut aussi se résumer à ces deux conditions : l'instrument de cueillette vise un seul trait et aucun autre facteur n'influence la réponse des sujets aux items. Pratiquement, une formulation plus souple est aussi acceptée. D'ailleurs Blais et Laurier (1997, p. 66-7) suggèrent : « *Dans cette perspective, il semble qu'il soit beaucoup plus réaliste et pratique de considérer que l'hypothèse d'unidimensionalité est vérifiée lorsqu'on peut montrer qu'une dimension dominante explique ou est responsable de la performance et des réponses des candidats.* » Humphreys (1984) formule une suggestion qui va dans le même sens. La vérification de l'hypothèse d'unidimensionalité est une condition nécessaire à une application adéquate des modèles de la TRI où il n'y a qu'une seule dimension dans la modélisation.

Dans la pratique, quelques recherches ont étudié les effets de la violation de l'hypothèse d'unidimensionalité. Selon Blais (1987), l'apparition d'un deuxième trait dans les réponses influence l'estimation des paramètres pour les trois catégories de modèles de la TRI. Blais et Laurier (1997) arrivent aux mêmes conclusions à partir de quelques études supplémentaires. Henning (1988) arrive aux mêmes résultats pour un test de langue avec le modèle de Rasch. Dans ces conditions, il serait préférable de vérifier cette hypothèse avant d'utiliser les modèles qui ne prennent en considération qu'une seule dimension. Hattie (1985) fait une recension de ces méthodes statistiques, Blais (1987) fait une présentation des méthodes utilisées jusqu'en 1987 et Hambleton, Swaminathan et Rogers (1991) reprennent essentiellement le travail de Blais (1987) et Hattie (1985). Embretson et Reise (2000) et Blais et Laurier (1997) complètent le tour

d'horizon depuis la recension de Hattie. Ces derniers font même une classification des méthodes. À partir de ces textes, il est possible de construire une brève liste d'auteurs qui ont développé des méthodes statistiques de détection de la dimensionalité d'un instrument de mesure :

- Reckase (1979);
- Bejar (1980);
- Drasgow et Parsons (1983);
- Doody-Bogan et Yen (1983);
- Ansley et Forsyth (1985);
- McDonald et Mok (1995);
- Stout (1987;1990).

Au sujet de ces méthodes, Blais et Laurier (1997) et Embretson et Reise (2000) soulignent qu'aucune méthode n'est absolue, c'est-à-dire qu'il n'y a pas de guide qui permette de choisir une méthode appropriée en fonction de la situation ou qui prévienne contre certaines méthodes inaptes dans certains contextes. En plus, certaines de ces méthodes auraient encore des preuves à faire.

2.1.2 Indépendance locale

L'hypothèse de l'unidimensionalité nous amène logiquement à présenter l'hypothèse de l'indépendance locale. Techniquement, l'indépendance locale est respectée lorsque la covariance entre différents items d'un instrument tend vers zéro. C'est-à-dire lorsque la probabilité de répondre d'un candidat à un item n'a aucun impact sur sa probabilité de répondre à un autre item. Le concept s'explique peut-être mieux à l'inverse. Il y aura dépendance locale entre deux items lorsque, par exemple, la formulation d'un item donne un indice de réponse pour l'item suivant. Théoriquement, l'indépendance locale devrait être obtenue lorsqu'il y a unidimensionalité puisqu'il n'y a qu'une seule dimension responsable de la réponse d'un sujet à chaque item. Toutefois, l'inverse ne va pas de soi. Il pourrait y avoir présence d'indépendance locale même si l'instrument de cueillette tente de mesurer plus d'un trait pourvu que chacun des traits soit identifié (Hambleton et Swaminathan, 1985). Statistiquement, l'indépendance locale peut se définir comme suit, soit $\mathbf{U}= (u_1, u_2, ..., u_N)$ les réponses d'un candidat à N items dichotomiques (0 ou 1); alors pour chaque trait latent θ possible :

$$P(\mathbf{U} \mid \theta) = P(u_1 \mid \theta)P(u_2 \mid \theta)....P(u_N \mid \theta) \qquad (3)$$

McDonald (1981) qualifie de stricte cette définition mathématique. Une interprétation plus souple consisterait à comparer les items deux à la fois.

Yen (1993) identifie les principales situations qui peuvent causer de la dépendance entre les items comme, par exemple, le temps alloué pour le test, l'assistance externe d'un professeur, la fatigue, la pratique, etc. Certaines de ces situations peuvent être contrôlées, ce qui est d'ailleurs suggéré. Yen propose aussi quelques trucs pratiques pour prévenir les problèmes de dépendance comme construire des items indépendants, construire des échelles séparées, etc. Pour détecter l'indépendance locale, Thompson et Pommerich (1996) présentent cinq auteurs qui ont travaillé sur la détection de dépendance entre les items et comment ils s'y prennent :

- Stout, Nandakumar, Junker, Chang, et Steidinger (1991);
- Yen (1984);
- Chen et Thissen (1997)[v];
- Wilson, Wood, et Gibbons (1987);
- Fraser (1988).

Quelques études (Sireci, Thissen, et Wainer, 1991; Thissen, Steinberg, et Mooney, 1989; Thompson et Pommerich, 1996; Yan, 1997; et Yen, 1984;1993) illustrent que, sous certaines conditions, l'absence d'indépendance locale entre les items peut avoir un effet sur l'estimation des paramètres ou sur la fidélité de l'instrument. À ce jour, il existe peu d'études sur les effets de la violation de cette hypothèse sur l'utilisation des modèles de Rasch. D'ailleurs, il n'existe pas suffisamment d'études pour permettre de retenir des suggestions communes qui pourraient régir la vérification de cette hypothèse et ses effets sur l'utilisation des modèles de la TRI. Les méthodes pour vérifier l'indépendance ne sont pas totalement éprouvées et doivent être utilisées avec prudence.

Jusqu'ici, il semble que ces deux hypothèses de base doivent être respectées avant de pouvoir utiliser sans problème certains des modèles de la TRI comme ceux de Rasch. Ces deux hypothèses sont difficiles à remplir, à vérifier, et l'effet de la violation de celles-ci reste à clarifier. Ces hypothèses sont exigeantes d'un point de vue théorique et pratique, mais dans la

v Au moment de la publication du texte de Thompson et Pommerich en 1996, le texte de Chen et Thissen était en impression. Nous savons maintenant qu'il a été publié en 1997.

mesure où elles sont vérifiées, elles devraient contribuer à la démonstration de la propriété de l'invariance.

2.1.3 Invariance

Dans la mesure où un modèle de la TRI s'ajuste bien aux données (voir section 2.4), il est possible de penser qu'il y a présence de la propriété d'invariance. Essentiellement, l'invariance se définit par un calibrage des items indépendants de l'échantillon et par une estimation des paramètres des individus indépendante des items. Dans le premier volet, l'influence du groupe dans l'estimation des paramètres des items est contrôlée pour obtenir une estimation qui ne variera pas d'un groupe de personnes à l'autre ou d'un sous-groupe de personnes à l'autre. En d'autres mots, peu importe les personnes choisies pour estimer l'item, l'estimation de la valeur des paramètres demeurerait toujours la même. Donc, les caractéristiques du groupe (race, sexe) ne devraient pas, en théorie, influencer le calibrage de chacun des items de l'instrument pour une population déterminée. Dans le deuxième volet, l'effet des items est contrôlé. Conséquemment, thêta ne dépend pas des items ou des sous-groupes d'items répondus par le candidat. La propriété d'invariance permet alors d'obtenir une estimation de la valeur des paramètres qui demeurent stables d'un groupe à l'autre pour l'ensemble des items. Cette propriété donne à la mesure en sciences sociales une certaine forme d'objectivité (Wright et Stone, 1979). Une objectivité spécifique au contexte d'utilisation, c'est-à-dire pour la population visée et pour l'instrument distribué (Andrich, 1988).

Dans la pratique, certains auteurs se sont attardés à vérifier si le modèle de Rasch permettait vraiment d'obtenir des estimations de paramètre invariantes. Pour le modèle de Rasch, Dong et al. (1983) concluent à des estimations de paramètres invariantes. Ozçelzik et Berberoglu (1995) arrivent à la même conclusion. Fan (1998) confirme les résultats de Ozçelzik et Berberoglu pour le modèle de Rasch et il généralise ses conclusions aux modèles à deux et trois paramètres lorsque les données s'ajustent aux modèles. D'un autre côté, Whitely et Dawis (1974) émettent certaines réserves. Selon eux, l'invariance peut être affectée par l'hétérogénéité du groupe utilisé pour estimer les items, par la qualité des estimateurs et par l'ajustement du modèle avec les données. D'ailleurs, Ozçelzik et Berberoglu (1995) et Dong et al. (1983) donnent peu d'information sur la qualité des estimateurs ou sur l'ajustement des données au modèle. En plus, Blais et Ajar (1992) soulèvent le problème du manque d'études empiriques démontrant des valeurs estimées invariantes. Ils situent aussi le concept de l'invariance dans une perspective

plus globale. Selon eux, il faut alors tenir compte des candidats, du contexte, des items, de la méthode d'estimation utilisée, etc. Ils vont dans le même sens que les recommandations de Whitely et Dawis (1974).

Breithaupt et Zumbo (2000) soulèvent le problème des méthodes utilisées pour vérifier cette hypothèse. Selon eux, la méthode de Fan (1988) ne serait pas appropriée. Le problème principal dans la vérification de l'hypothèse d'invariance est un problème d'arrimage de l'échelle. Il faut arbitrairement fixer un 0 soit pour les items, soit pour le thêta des sujets. Conséquemment, d'un groupe de données à l'autre, le 0 fixé peut varier et il devient difficile de comparer l'estimation des paramètres parce que le référentiel n'est pas le même. En plus, il se peut que la valeur de l'estimation des paramètres soit affectée par cette difficulté dans l'arrimage. Embretson et Reise (2000) discutent plus longuement de ce problème dans leur texte.

Bien que la condition d'invariance soit théoriquement respectée dans les études mentionnées plus haut, si, dans la pratique les données ne s'ajustent pas bien avec le modèle, l'hypothèse d'invariance ne devrait pas être présente pour cette estimation (Swaminathan, Hambleton et Rogers, 1991). En somme, il faut demeurer vigilant et explorer attentivement les données avant d'affirmer la présence d'invariance pour des valeurs estimées à l'aide de la théorie de la réponse à l'item.

2.2 LES MODÈLES DE LA THÉORIE DE LA RÉPONSE À L'ITEM

Un modèle (ex : les modèles de la TRI) peut se définir ou s'interpréter comme une équation mathématique qui permet de décrire des données. Par exemple, la formule de la vitesse permet de décrire des observations sur le temps et la distance. Avec la TCT, la plupart des équations (ou modèles) se construisent à partir des codes numériques qui correspondent aux étiquettes de l'échelle de référence utilisée avec l'instrument. La somme de ces codes numériques constitue le score total attribué au candidat. À l'aide de ce score, il est ensuite possible de comparer les candidats entre eux. Avec la TRI, plutôt que de faire la somme des codes numériques, l'accent est plutôt mis sur la capacité d'un candidat à répondre à chacun des items de l'instrument. Les modèles de la TRI se définissent alors comme la probabilité de répondre correctement à un item en fonction des caractéristiques de l'item et du thêta des sujets. La fonction mathématique qui traite les codes numériques et explique la relation entre ces deux paramètres est une forme de régression logistique. Pour plus de détails sur les liens à faire entre la régression et la TRI, voir

Blais (1987); Hambleton et Swaminathan (1985); ou Lord (1980). La représentation graphique qui illustre cette fonction pour chaque item s'appelle la *Courbe Caractéristique de l'Item* (CCI).

2.2.1 Courbe caractéristique de l'item

La figure 2.1 est un exemple de trois CCI pour trois items différents pour le modèle de Rasch. Ces courbes en forme de *S* donnent la probabilité (en ordonnée) de répondre correctement à un item en fonction d'un prédicteur (en abscisse). Pour chaque CCI, une augmentation sur l'abscisse conduit à une augmentation de la probabilité de répondre correctement à l'item. Au milieu des courbes, une petite différence sur le continuum augmente davantage la probabilité de répondre correctement tandis qu'aux deux extrémités un grand changement sur le continuum est nécessaire pour que la probabilité de réponse augmente légèrement. L'endroit où se situe la courbe sur le continuum (c'est-à-dire plus à gauche ou plus à droite) donne une indication de la difficulté de l'item. La difficulté de l'item s'interprète comme le point d'inflection sur la courbe où un sujet a une probabilité de répondre correctement de 0,5. Par exemple, plus la courbe est à gauche, plus l'item est considéré comme facile parce que le sujet a besoin d'un moins grand thêta pour avoir 50 % de chance de répondre correctement à l'item. Sur la figure 2.1, la difficulté de l'item 1 est de –1,0 et celle de l'item 2 est de 0. La difficulté de l'item est le premier paramètre qui peut affecter l'emplacement et la forme de la CCI. Les deux autres paramètres sont la discrimination de l'item et le facteur chance [vi].

vi Dans les écrits anglophones, le facteur chance est aussi appelé « pseudo-chance ».

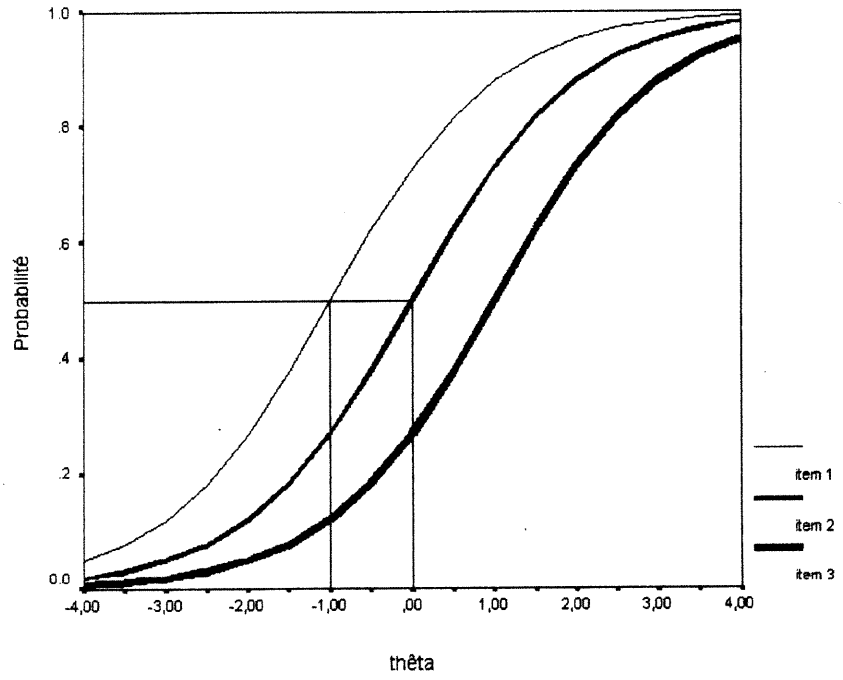

Figure 2.1 Courbes caractéristiques (CCI) de trois items avec des indices de difficulté différents.

La discrimination de l'item se définit par sa capacité à différencier deux candidats sur le continuum. Mathématiquement, la discrimination d'un item correspond à la pente de sa CCI. Visuellement, la CCI de l'item 1 de la figure 2.2, à la page suivante, monte très progressivement (pente faible) le long du continuum thêta. Même si thêta augmente beaucoup, la probabilité de répondre correctement augmente très peu. Deux sujets avec des thêta très différents (-2,0 et 3,0 par exemple) ont une probabilité similaire de répondre à l'item. À l'opposé, la CCI de l'item 2 fait une montée très rapide (pente élevée) le long du continuum et deux sujets avec des thêta peu différents (−1,0 et 0,0) ont une probabilité de répondre qui diffère plus que dans le cas de la courbe de l'item 1. Il faut alors comprendre que l'item 2 discrimine davantage que l'item 1. C'est-à-dire que l'item 2 permet plus clairement d'identifier lequel des deux sujets a une plus grande probabilité de répondre correctement à un item. Il est intéressant de noter que, dans le cas où seulement le paramètre de la difficulté influence la probabilité de répondre correctement, les courbes pour les différents items ne se recoupent jamais.

La figure 2.2 montre aussi la différence entre les CCI si on considère un troisième paramètre. Sur la figure 2.2, la CCI de l'item 3 se distingue des deux autres par la queue extrême gauche de sa courbe. Pour les deux autres courbes, l'extrémité gauche tend vers zéro à $-\infty$. Pour la courbe

de l'item 3, l'extrémité gauche tend plutôt vers 0,3 à $-\infty$. Cet ajustement est attribuable au facteur chance. C'est-à-dire que, pour certains items tels les items à choix multiples, le facteur chance a pour effet d'augmenter la probabilité de répondre correctement du sujet le plus faible sur le continuum. Il serait naïf de croire que la probabilité de répondre d'un sujet très faible est de zéro puisqu'il est quand même possible que le sujet obtienne la bonne réponse à l'item par hasard. Il faut alors estimer la valeur de ce facteur chance. En résumé, la CCI est la représentation graphique du modèle et elle s'ajuste en fonction du modèle mathématique choisi, c'est-à-dire en fonction du nombre de paramètres inclus dans celui-ci.

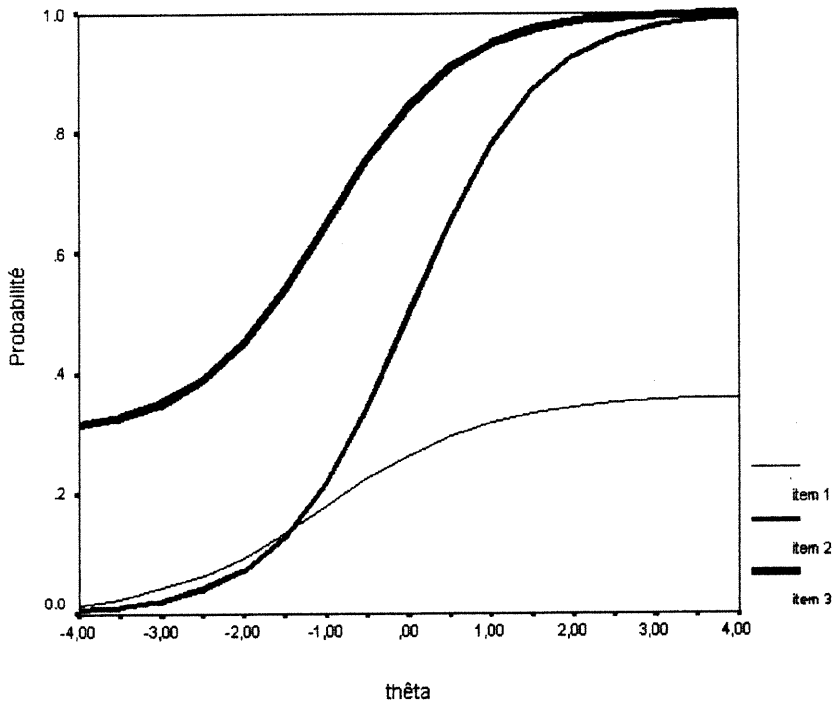

Figure 2.2 Courbes caractéristiques (CCI) de trois items avec des indices différents de discrimination et l'indice pseudo-chance.

2.2.2 L'échelle thêta

La quantité sur l'abscisse (sur le continuum de thêta) est appelée un *logit* lorsque le modèle se définit comme une régression logistique (ce qui est le cas de plusieurs modèles de la TRI). Généralement, elle prend une valeur entre –3 et 3 mais, la plupart du temps, le continuum inclut les valeurs possibles entre –4 et 4 pour souligner la qualité asymptotique des deux extrémités de la courbe comme il est possible de le voir sur les figures 2.1 et 2.2. Dans une utilisation quotidienne, des quantités comme -3,24 ou 3,2 ne sont pas très conviviales pour des personnes

habituées à travailler avec des pourcentages ou des scores bruts. Heureusement, il est possible d'effectuer des transformations linéaires ou non linéaires sur les *logit*[vii] afin d'obtenir des résultats semblables à ceux avec lesquels les praticiens travaillent dans le quotidien (Wright et Stone, 1979; Hambleton, Swaminathan et Rogers, 1991; Ludlow et Haley, 1995 et Embretson et Reise, 2000). Il faut également mentionner que les modèles de la famille de Rasch devraient permettre d'obtenir une unité à intervalle sur laquelle des additions sont possibles. Fischer (1995) et Roskam & Jansen (1984) présentent la preuve de cette propriété particulière du continuum de thêta pour les modèles de la famille de Rasch. Bartholomew (1996) fait la même démonstration pour la famille plus large des modèles exponentiels.

2.2.3 Présentation des modèles de la TRI

Nous avons vu à l'alinéa 2.2.1 que la première caractéristique d'un modèle de la TRI est son nombre de paramètres. Mais, il existe d'autres caractéristiques qui influencent la forme du modèle. Parmi ceux-ci, il y a le nombre de dimensions (traits latents), le type de modélisation et le type de réponses obtenues à l'instrument. Voici une liste plus précise des caractéristiques de base qui permettent de différencier les modèles entre eux :

- nombre de paramètres (difficulté, discrimination, facteur chance);
- nombre de dimensions (unidimensionnel, multidimensionnel);
- types de modélisation (ogive normale, logistique, non paramétrique);
- types de réponses (nominal, ordinal [dichotomique, polytomique], choix multiples, etc.).

Le but de cette recherche n'est pas de faire une présentation exhaustive de tous les modèles de la TRI et de leurs particularités. Un lecteur intéressé peut se référer à van der Linden (1996) pour une description complète de la plupart des modèles recensés jusqu'ici. Dans le contexte de ce travail, quelques modèles de base seront présentés, suivis d'une liste des modèles les plus utilisés et les plus souvent mentionnés dans la littérature.

La présentation des quelques modèles de base qui suivent a pour but de donner un bref aperçu de la logique qui sous-tend la construction des différents modèles de la TRI. La notation suivante aide à décrire les caractéristiques et les modèles et elle sera maintenue tout au long de ce texte :

$P_i(\theta)$ Probabilité qu'un sujet θ réponde correctement à un item i;

N Nombre d'items à un test;

vii Le Logit est une abbréviation pour « logistic probability unit » et il représente le logarithme naturel d'une probabilité

n	Nombre de sujets;
θ	Thêta du sujet;
b_i	Difficulté de l'item i;
a_i	Discrimination de l'item i;
c_i	Facteur pseudo-chance de l'item i;
u_i	Réponse d'un sujet à l'item i;
x	Score total d'un sujet à un instrument;
π	3,1416;
e	2,718;
D	1,7 (Facteur d'équivalence entre les courbes ogive normale et logistique).

Historiquement, un des premiers modèles de la TRI à faire son apparition est le modèle à ogive normale à deux paramètres de Lord (1952). Il se définit mathématiquement comme suit :

$$P(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \partial z \qquad (4)$$

où z réfère aux scores z d'une distribution normale réduite.

Pour des raisons pratiques et techniques les modèles à ogive normale sont moins utilisés que les modèles logistiques. Ce sont ces derniers qui sont généralement présentés dans la littérature puisque la plupart des modèles plus complexes sont une extension de ceux-ci. Il y a trois modèles logistiques de base définis en fonction du nombre de paramètres. Le modèle à un paramètre est plutôt connu sous le nom de modèle de Rasch en mention de son concepteur, le Danois Georg Rasch (Rasch, 1960). Il se définit :

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}. \qquad (5)$$

Les modèles logistiques à deux paramètres et trois paramètres sont des propositions de Birnbaum (1968). Le modèle à deux paramètres peut prendre la forme suivante :

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad (6)$$

Le modèle à trois paramètres est :

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}. \quad (7)$$

Les équations 5,6 et 7 sont des modèles conçus pour des données dichotomiques à une seule dimension. Pour les modèles plus complexes, il suffit de faire une extension à partir de ces modèles. Comme il est mentionné plus haut, l'objectif ici n'est pas de décrire tous les modèles existants. Tout de même, il paraît intéressant de faire une liste synthèse des modèles les plus utilisés et les plus pertinents. Dans la liste qui suit, les lettres indiquent la nature du modèle (L pour logistique, ON pour ogive normale et P pour le nombre de paramètres), le nombre de dimensions expliquées par le modèle (UD pour unidimensionnel ou MD pour multidimensionnel) et le type de réponses (D pour dichotomique, L pour Likert , N pour nominal ou É pour réponse partielle en étape et CM pour choix multiple) :

- 1PON, 2PON et 3PON (Lord, 1952; UD,D);
- 1PL (Rasch, 1960, UD,D);
- 2PL et 3PL (Birnbaum, 1968; UD,D);
- Multidimensionnel de Rasch (McKinley et Reckase, 1982; MD,D);
- Multidimensionnel 1PON, 2PON et 3PON (Bock, Gibbons et Muraki, 1988, MD,D);
- Graded Response, GRM (Samejima, 1969; 1996; UD,L);
- Modified Graded Response, M-GRM (Muraki, 1990; UD,L);
- Partial Credit , PCM (Masters, 1982; UD, É);
- Generalized Partial Credit, GPCM (Muraki, 1992; 1993; UD,É);
- Rating Scale (Andrich, 1978a; 1978b; UD, L);
- Nominal Reponse (Bock, 1972; UD,N);
- Multiple Choice (Thissen et Steinberg, 1984; UD,CM).

Bien qu'ils soient représentatifs d'une bonne partie du type de données rencontrées en mesure, les modèles présentés dans cette liste ne sont pas les seuls qui existent. Il faut aussi mentionner que ces modèles sont tous des modèles paramétriques, c'est-à-dire que des paramètres définis

expliquent la forme que prend la distribution des estimations. Pour une description plus détaillée des différents modèles, le lecteur peut se référer, en plus de van der Linden (1996), à DeAyala (1993); Thissen (1986); Embretson et Reise (2000); ou encore Sijtsma et Hemker (2000).

Parmi les modèles énumérés ci haut, le *Rating Scale Model* est un cas particulier des modèles de la famille de Rasch que l'on peut appliquer à une échelle de type Likert. C'est ce modèle précis qui nous intéresse dans le cadre de cette recherche parce qu'il répond aux caractéristiques des données recueillies par notre instrument. Prenons une échelle de Likert sur laquelle il y a k catégories et m étapes à franchir entre la première et dernière catégorie. Ainsi, m + 1 = k.

1----------2----------3----------4 (k=4 catégories)
étape 1 étape 2 étape 3 (m=3 étapes)

Dans le cas où ces catégories demeurent les mêmes d'un item à l'autre, il y a k courbes et le paramètre b_i (présenté en 2.2.1) est remplacé par le paramètre δ_{ix}. Comme le paramètre b_i, le paramètre δ_{ix} situe la courbe sur le continuum (plus à gauche ou plus à droite). Mais parce que ce modèle inclut k catégories (et k courbes), le paramètre δ_{ix} indique aussi le point sur le continuum où il devient plus probable que le candidat ait franchi une des m étapes. Pour ce faire, δ_{ix} se décompose en deux paramètres : δ_i et τ_x. Le paramètre δ_i remplace b_i et situe l'ensemble des quatre courbes sur le continuum. Le paramètre d'intersection des catégories τ_x situe les k catégories les unes par rapport aux autres et correspond au point d'intersection entre les courbes des différentes catégories. δ_{ix} se lit alors comme la probabilité de franchir x des m étapes pour l'item i. Formellement, le modèle ressemble à ceci :

$$P_{ix} = \frac{e^{\left|\sum_{s=1}^{x}(\theta-\delta_i-\tau_s)\right|}}{\sum_{q=0}^{m} e^{\left|\sum_{s=1}^{x}(\theta-\delta_i-\tau_s)\right|}}. \qquad (8)$$

Dans la formule, il faut lire x (le score du sujet) comme le nombre des m étapes qui ont été franchies par le candidat (x = 1,..., m) et m comme le nombre total de catégories à l'échelle – 1 (k - 1 comme expliqué plus haut). C'est à partir de cette formule que les valeurs des paramètres de ce modèle de Rasch sont estimées.

2.3 ESTIMATION

L'estimation des paramètres de l'item et du sujet est un aspect crucial de la TRI, car, sans une estimation qui donne une bonne approximation de la valeur réelle des paramètres, il faudrait remettre en question l'utilisation du modèle. Dans certains cas, l'estimation donne de bons résultats, mais, dans d'autres circonstances, il peut être difficile d'obtenir une bonne estimation. C'est pourquoi il faut s'assurer que la méthode choisie pour faire l'estimation est efficace en fonction des caractéristiques de la situation (comme le nombre de sujets et le nombre d'items) et les caractéristiques du modèle utilisé (le nombre de paramètres). Les paragraphes qui suivent présentent les méthodes d'estimation des paramètres et les critères qui permettent de juger de leur efficacité.

Les méthodes d'estimation les plus utilisées peuvent se classer en deux catégories principales. Il y a les méthodes d'estimation par maximum de vraisemblance *(Maximum Likelihood Estimation - MLE)* et les méthodes d'estimation Bayesienne. Les méthodes d'estimation par maximum de vraisemblance se divisent en trois : la méthode d'estimation par maximum de vraisemblance conjointe *(Joint Maximum Likelihood Estimation - JMLE)*, la méthode d'estimation par maximum de vraisemblance marginale *(Marginal Maximum Likelihood Estimation - MMLE)* et la méthode d'estimation par maximum de vraisemblance conditionnelle *(Conditional Maximum Likelihood Estimation - CMLE)*. Les méthodes d'estimation Bayesienne se divise en deux : les méthodes de « maximum a posteriori » *(Maximum A Posteriori - MAP)* et les méthodes « a posteriori attendu » *(Expected A Posteriori - EAP)*. Dans certains cas, les méthodes Bayesiennes sont utilisées conjointement avec les méthodes de maximum de vraisemblance. Ces méthodes sont brièvement présentées dans les pages suivantes. Pour le reste du texte, les acronymes anglais présentés entre parenthèses plus haut seront utilisés pour identifier les méthodes. À ces deux catégories s'ajoutent des méthodes qui prennent moins de place dans la littérature comme les méthodes d'estimation heuristique (Wright et Stone, 1979; Embretson et Reise, 2000), d'estimation par paire (Andrich, 1988; Zwinderman, 1995) et les méthodes non paramétriques (Levine et Williams, 1991; Samejima, 1994). Ces méthodes ne seront pas présentées en détail ici. Le lecteur peut se référer à ces auteurs pour en obtenir un aperçu.

2.3.1 Qualité des estimations

Au moment de faire l'estimation des paramètres d'un modèle, certaines caractéristiques de base permettent de nuancer la qualité des estimations obtenues. Les caractéristiques recherchées ici sont la consistance, l'efficacité et l'applicabilité. La consistance indique que la valeur de l'estimation converge vers sa valeur réelle à mesure que le nombre de répondants et de questions augmente. Une estimation efficace est une méthode d'estimation dont la valeur obtenue a une faible variance. L'applicabilité indique, par exemple, la capacité à estimer les scores parfaits et nuls ou la capacité à estimer plusieurs paramètres d'un modèle. Pour les méthodes d'estimation présentées ici, les auteurs s'entendent assez bien sur les qualités statistiques que possède chacune d'elles (voir Hambleton et Swaminathan, 1985; Baker, 1987; Hambleton, Swaminathan et Rogers, 1991; Molenaar, 1995; Embretson et Reise, 2000).

2.3.2 Principes de base de l'estimateur MLE

Le principe de base de la méthode MLE consiste à maximiser une fonction mathématique qui donne la probabilité de répondre d'un candidat en fonction de sa séquence de réponse à des items et en fonction de thêta. La valeur maximale de la fonction mathématique s'obtient en fixant à 0 la dérivé première de la fonction suivante où $P_i(\theta)$ correspond à un modèle de la TRI et $Q_i(\theta)=1-P_i(\theta)$:

$$\mathbf{L}(u_1, u_2, \ldots, u_N \mid \theta) = \prod_{i=1}^{N} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} . \qquad (9)$$

Baker (1987) et Hambleton, Swaminathan et Rogers (1991) démontrent qu'il peut être plus approprié d'utiliser la version logarithmique de l'équation 9. Cette dernière prend alors la forme suivante :

$$\ln \mathbf{L}(u_1, u_2, \ldots, u_N \mid \theta) = \sum_{i=1}^{N} u_i \ln P_i(\theta) + (1-u_i) \ln Q_i(\theta) . \qquad (10)$$

Les valeurs obtenues avec la méthode MLE sont reconnues pour être consistantes et efficientes. L'estimation faite à l'aide de la méthode MLE présente toutefois certaines limites. Elle ne permet pas d'obtenir de valeur pour les candidats qui ont obtenu toutes les bonnes réponses ou toutes les mauvaises réponses. De plus, l'estimation est moins efficace pour des instruments de

moins de 50 items. Pour répondre à ces limites, les méthodes d'estimation Bayesiennes peuvent être utilisées.

La procédure utilisée avec la méthode MLE et les méthodes Bayesiennes est la même, c'est-à-dire que la fonction mathématique est maximisée à partir de la dérivée première. Parce que la méthode d'estimation Bayesienne suppose que la distribution des répondants est normale, il est alors possible de combiner les deux méthodes (MLE et Bayesienne) pour obtenir des valeurs estimées pour les répondants ayant répondu correctement à tous les items. Cette combinaison des deux méthodes ne permet toutefois pas d'obtenir une estimation consistante pour de petits échantillons. En plus, pour utiliser l'estimateur MLE, il faut connaître préalablement un des deux paramètres. La méthode JMLE est une solution à ce dernier problème.

2.3.3 Méthode d'estimation par maximum de vraisemblance conjointe (JMLE)

Lord (1974; 1980) a développé une procédure itérative qui permet d'estimer les paramètres de l'item et le paramètre thêta conjointement. La procédure est semblable à celle décrite plus haut. Mais, comme aucun des deux paramètres n'est connu, il faut fixer arbitrairement un point de départ pour un des deux paramètres et procéder à l'estimation de l'autre paramètre en fonction de ce point de départ. Cette étape permet d'obtenir une première estimation d'un des deux paramètres (celui de l'item, par exemple). Dans une deuxième étape, la même procédure est reprise en utilisant les résultats obtenus en première étape pour faire l'estimation du deuxième paramètre. La procédure est répétée jusqu'à ce que les valeurs convergent convenablement.

Selon Molenaar (1993) et Baker (1987), lorsque le nombre d'items est fixe, les valeurs estimées pour les paramètres ne sont pas consistantes, même lorsque le nombre de sujets augmente. Tout comme pour la méthode MLE, il n'est pas possible d'obtenir une valeur précise pour les sujets ayant répondu correctement à tous les items (ou l'inverse). Malgré tout, cette méthode s'applique assez bien à tous les modèles et, comme mentionné plus haut, il est toujours possible d'utiliser une des deux méthodes bayesiennes pour contrer le problème des personnes ayant répondu correctement à tous les items. Il faut noter que, pour le modèle à un paramètre, la méthode JMLE peut aussi prendre le nom de méthode de vraisemblance inconditionnelle (UCON). Cette deuxième appellation est de moins en moins utilisée et les auteurs semblent maintenant s'entendre pour utiliser JMLE pour le modèle à un paramètre.

2.3.4 Méthode d'estimation par maximum de vraisemblance marginale (MMLE)

Parce que les deux paramètres sujet et item sont estimés en même temps, la méthode conjointe offre des résultats moins intéressants (pas de consistance). Pour solutionner ce problème, il faut trouver un moyen de faire disparaître un des deux paramètres de la fonction. Bock et Aitkin (1981) suggèrent de spécifier préalablement la distribution d'un des deux paramètres. À partir d'une procédure itérative assez complexe, ils ont réussi à intégrer le paramètre thêta hors de la fonction. Les paramètres ne sont alors plus estimés en même temps. Grâce à cette procédure, les estimations obtenues sont consistantes et efficientes peu importe la longueur du test ou le score obtenu par la personne (c'est-à-dire les scores parfaits ou nuls). Il faut toutefois affecter une distribution à l'un des deux paramètres.

2.3.5 Méthode de maximum de vraisemblance conditionnelle (CMLE)

Les travaux de Andersen (1972;1973) et de Rasch (1960) constituent une alternative intéressante à la méthode de maximum de vraisemblance marginale. Au lieu d'exiger que les paramètres répondent à une distribution, ils utilisent plutôt la statistique exhaustive. Dans le cas du modèle à un paramètre, le score total à un instrument est une statistique exhaustive parce qu'elle contient toute l'information sur le sujet et peut être utilisée comme une estimation du paramètre thêta. C'est-à-dire qu'en le remplaçant par le score d'un sujet à l'instrument, le paramètre thêta ne fait plus partie de la fonction. Les paramètres peuvent donc être estimés un à la fois. Lord (1980, p. 57) décrit la statistique exhaustive ainsi :

> *La propriété clé de la statistique exhaustive S provient de l'indépendance de thêta par rapport à la distribution conditionnelle des observations étant donné S. Ceci signifie qu'à partir du moment où S est décrit, les données ne contiennent plus aucune information de thêta. Ceci justifie l'affirmation habituelle que la statistique S contient, dans les données, toute l'information concernant thêta.*

L'utilisation de cette méthode est conditionnelle à la présence de la statistique exhaustive. Cette statistique est seulement valable pour les modèles de la famille de Rasch. Les principes de base de la procédure sont équivalents à ce qui a été présenté jusqu'ici. Les estimations faites à partir de la méthode CMLE sont consistantes et efficientes mais il n'est pas possible d'obtenir une estimation pour les personnes avec des scores parfaits ou nuls. Il y aurait aussi certains problèmes avec de longs instruments (Embretson et Reise, 2000).

Parmi les méthodes moins exploitées, la méthode d'estimation par paire est régulièrement mentionnée. Cette méthode compare les paires d'items entre elles. Cette procédure s'utilise très bien et donne de très bonnes estimations selon Andrich (1988). Zwinderman (1995) confirme en effet que, sur de petits échantillons, cette méthode d'estimation est un bon outil. Toutefois, elle ne peut être utilisée que pour le modèle dichotomique de Rasch et, partiellement, pour son extension aux données polytomiques.

Théoriquement, les méthodes d'estimation présentées ci-dessus possèdent des caractéristiques différentes mais, dans la pratique, plusieurs travaux seront encore nécessaires avant de pouvoir savoir quelle méthode est préférable selon le contexte d'utilisation. Pour les modèles de la famille de Rasch, la méthode CMLE serait préférable et plus facile à utiliser grâce à l'existence de la statistique exhaustive. Mais, selon Baker (1987, p. 135) : *« Malgré tout, les résultats empiriques pour le modèle de Rasch suggèrent que l'estimation du paramètre de l'item obtenue à partir de la méthode de vraisemblance conjointe* **JMLE** *ne diffère pas de celle obtenue à l'aide de la statistique exhaustive* **CMLE.** »[viii] Encore plus, dans sa recension des méthodes pour le modèle de Rasch, Linacre (1999, p. 402) conclut : *« Néanmoins, en prenant en considération la précision et l'exactitude d'une estimation, toutes les méthodes produisent des estimations statistiquement équivalentes. »* Pour le moment, il semble bien que, pour couvrir l'ensemble des possibilités, la méthode MMLE possède les meilleures qualités (Wainer et Thissen : chap. 2, 2001).

2.4 AJUSTEMENT ENTRE LES DONNÉES ET LE MODÈLE

À la section 2.1, il est mentionné qu'un bon ajustement entre les données et le modèle utilisé serait une condition nécessaire à la présence de certaines propriétés des modèles de la TRI (invariance des paramètres, échelle à intervalle pour Rasch, etc.). Dans ces circonstances, il est pertinent de se demander quels sont les effets d'un manque d'ajustement sur la valeur de l'estimation des paramètres des modèles de la TRI. Linacre (1995) et Adams et Wright (1994) prétendent que le manque d'ajustement peut avoir des effets sur la valeur de l'estimation des paramètres avec les modèles de Rasch. Fan (1999) prétend qu'un mauvais ajustement n'aurait pas d'effets précis sur la propriété d'invariance pour les paramètres thêta, difficulté de l'item ou discrimination de l'item. Toutefois, dans la pratique, il manque d'études qui étudient les effets et

viii Les sigles en gras sont ajoutés pour faciliter la compréhension de la citation parce que la traduction peut porter à confusion entre la terminologie anglaise et française.

la nature de ceux-ci sur la valeur de l'estimation et sur les hypothèses des modèles de la TRI. Malgré cela, le présupposé théorique de la TRI est que, dans la mesure où les données s'ajustent bien au modèle, les hypothèses de base devraient être respectées et la propriété d'invariance présente dans les données (Hambleton, Swaminathan et Rogers, 1991).

Hambleton et Swaminathan (1985) divisent en trois étapes la procédure de vérification de l'ajustement des données au modèle. Ils suggèrent de : premièrement, vérifier les hypothèses de base de l'unidimensionalité et de l'indépendance locale; deuxièmement, vérifier la propriété d'invariance dans les valeurs estimées pour les paramètres; et finalement, d'étudier le rapport entre les données empiriques et les données prédites par le modèle à l'aide de statistiques d'ajustement. La section 2.1 présente les deux premières étapes de la procédure. Cette section présente la troisième étape de cette procédure.

Selon la recension faite par Embretson et Reise (2000), il existerait trois méthodes différentes pour évaluer l'ajustement des données empiriques aux données estimées par le modèle. Parmi ces trois méthodes, une première se conduit d'abord sur les personnes et les deux autres sur les items *(person-fit analysis, item-fit analysis)*.

2.4.1 Méthodes d'ajustement sur les candidats

Pour l'ajustement du paramètre thêta associé aux candidats, l'ajustement s'évalue à partir d'une comparaison entre la séquence de réponse du candidat et la séquence normale définie par le modèle. Par exemple, un sujet qui obtient un score de 2 sur 10 devrait raisonnablement répondre correctement à deux items parmi les plus faciles. Le principe de base de la procédure consiste alors à évaluer si certaines personnes ont des séquences de réponses aberrantes. Pour plus de détails sur ces procédures, le lecteur peut se référer à la revue *Applied Psychological Measurement* (1996) qui consacre un numéro entier aux méthodes d'analyse de l'ajustement entre le candidat et le modèle. Meijer (1994) présente aussi une méthode adaptée au modèle de Rasch.

2.4.2 Méthodes d'ajustement pour le paramètre difficulté de l'item

Pour sa part, l'ajustement des items s'évalue principalement de deux façons. La première consiste à faire une analyse à partir de la représentation graphique du modèle et la deuxième à

partir d'une statistique khi-carré adaptée aux besoins. Dans la première façon de faire, le diagramme de dispersion (pour chaque item) est tracé et comparé aux courbes (CCI) prédites par le modèle. Dans la deuxième façon, il y a comptabilisation d'une statistique du khi-carré adaptée pour les besoins et utilisée comme point de référence pour évaluer l'ajustement. Pour de grands échantillons, ce calcul est très sensible et ne permet pas toujours de prendre une décision éclairée (Hambleton et Swaminathan, 1985; Embretson et Reise, 2000). Parmi les statistiques khi-carré développées spécialement pour les modèles de la TRI, il y a celles de Wright et Panchapekasan (1969); Bock (1972); Hambleton et Traub (1973); Muraki (1996) et pour les modèles de Rasch, Masters et Wright (1996). Pour une recension complète des méthodes, voir Meijer et Sijstma (1999).

Les deux prochains paragraphes présentent le rationnel derrière la statistique khi-carré de Wright et Masters (1982) utilisée pour évaluer l'ajustement des modèles de la famille de Rasch. Le point de départ de la statistique consiste à construire une matrice x par y avec les x candidats et les y items. Chaque élément de la matrice est composé de la différence entre les données observées pour chaque xy, et les données prédites par le modèle pour ces mêmes xy. Cette différence forme un résiduel. Plus le résidu est petit, plus les données observées sont semblables aux données prédites. Pour calculer la somme des carrées résiduels, le résidu est mis au carré et additionné pour les x candidats ou pour les y items. Dans certains cas, c'est plutôt la moyenne des carrés résiduels qui est calculée pour un item ou pour un candidat.

La moyenne des carrés peut être une moyenne *infit* ou une moyenne *outfit*. La moyenne *infit* prend en considération la variance de l'estimation faite pour chacune des cases xy de la matrice. L'interprétation de la valeur de la moyenne des carrés se fait à l'aide de l'équation suivante : 100 (valeur – 1). Donc, une moyenne des carrés de 2,3 signifie qu'il y a 130 % plus de variation que prévu par le modèle. Pour une moyenne des carrés de 0,7, il y a 30 % moins de variation que prévu par le modèle. Dans certains cas, la moyenne des carrés résiduels se présente sous la forme d'une distribution normale *t* avec une moyenne de 0 et une variance de 1.

Plusieurs évaluations et interprétations de cette valeur calculée à l'aide du khi-carré de Wright et Masters (1982) sont possibles dans la pratique. Smith (2000), Bond et Fox (2001) et Smith et Schumaker (1998) donnent tous des lignes directrices qui permettent d'évaluer l'ajustement des données au modèle selon le modèle utilisé et les circonstances. Au moment de procéder à l'analyse des données, nous reviendrons sur ces lignes directrices.

Finalement, Karabatsos (2000) critique l'utilisation d'une statistique khi-carré comme celle expliquée plus haut pour évaluer l'ajustement des données. Le problème est dans le calcul du résiduel en soi parce que la valeur des données observées peut seulement être discrète tandis que la valeur prédite par le modèle est continue. La valeur prédite par le modèle n'est pas discrète, donc elle ne peut être réellement obtenue à partir de l'instrument, elle est un artifice du modèle et, conséquemment, la valeur résiduelle l'est aussi. Pour ces raisons, Karabatsos rejette l'utilisation de cette statistique qui demeure toutefois la statistique utilisée par les plupart des programmes informatiques disponibles sur le marché à l'heure actuelle.

2.5 LOGICIELS INFORMATIQUES

Que ce soit pour l'estimation des paramètres ou pour les méthodes qui permettent de vérifier les hypothèses de base des modèles, les processus de calcul sont très complexes et nécessitent l'utilisation de logiciels informatiques. Il existe présentement plusieurs logiciels informatiques sur le marché qui permettent d'utiliser la TRI. Parmi les plus connus, il y a BILOG (Mislevy et Bock, 1990), MULTILOG (Thissen , 1991), PARSCALE (Muraki et Bock, 1993), WINSTEPS (Wright et Linacre, 1991), LOGIST (Wingersky et al., 1982) et CONQUEST (Wu et al., 1998). La compagnie *Assessment System Corporation* constitue une bonne source de logiciels et elle commercialise la plupart des logiciels mentionnés ci-haut.

Les logiciels informatiques se caractérisent principalement par les méthodes statistiques qu'ils utilisent 1) pour faire l'estimation des paramètres, 2) pour vérifier l'ajustement entre les données et le modèle ou 3) pour vérifier la présence de biais de mesure. Il a été suggéré à la section 1.3 de ce chapitre que la méthode MMLE donne de bons résultats dans l'ensemble. Les logiciels utilisent principalement cette méthode d'estimation. Les logiciels informatiques CONQUEST, BILOG, MULTILOG et PARSCALE utilisent tous la méthode MMLE tandis que les logiciels WINSTEPS et LOGIST utilisent la méthode JMLE. Dans le cas plus spécifique des modèles de Rasch, des logiciels moins connus, parce que plus spécifiques, existent aussi. Par exemple, le logiciel LPCM-WIN (Fischer, 1988) utilise la méthode CMLE et le logiciel RUMM (Andrich et al., 1997) utilise la méthode d'estimation par paire. La plupart de ces logiciels sont assez diversifiés dans leur façon d'estimer les paramètres ou de calculer l'ajustement et ils offrent la possibilité de travailler avec la plupart des modèles présentés à la section 1.2.3.

Le choix du programme informatique dépend évidemment du modèle mathématique utilisé et du type d'analyse et d'estimation que le chercheur désire obtenir. Il est plutôt difficile de discriminer entre les résultats ou les valeurs estimées à partir de différents programmes informatiques. Toutefois, Linacre (1999) montre que les valeurs des estimations de paramètre obtenues à l'aide des logiciels QUEST (Adams et Toon, 1994), CONQUEST, RUMM, WINSTEPS et LPCM-WIN sont très similaires même si les logiciels n'utilisent pas les mêmes méthodes d'estimation. Dans l'ensemble, il n'existe pas encore suffisamment d'articles scientifiques spécifiques à différentes situations pour aider les chercheurs à faire leur choix parmi les logiciels informatiques offerts. C'est pourquoi il est préférable, pour l'instant, de garder un œil attentif sur la littérature scientifique qui analyse les différents logiciels et les compare entre eux dans des contextes qui s'apparentent à celui de la situation étudiée.

En conclusion, il y a plusieurs caractéristiques importantes à connaître avant d'utiliser un modèle de la TRI. Il faut connaître les principes d'utilisation associés aux hypothèses de base, aux paramètres des modèles, aux différents types de modèles, aux méthodes d'estimation des paramètres et aux méthodes d'ajustement des données aux modèles. En plus, il est utile de connaître les programmes informatiques disponibles puisqu'ils sont nécessaires aux calculs. Dans le cadre de cette recherche, l'utilisation de la TRI dans le contexte d'un processus de validation d'un instrument de cueillette d'information nous intéresse particulièrement. Le chapitre suivant s'attarde à décrire les concepts clefs de la validation d'un instrument et à voir comment la TRI peut s'insérer dans ce processus.

CHAPITRE III

LE PROCESSUS DE VALIDATION D'INSTRUMENTS DE CUEILLETTE D'INFORMATION

Les principaux fondements de la théorie de la réponse à l'item ont été présentés au chapitre deux. Dans ce troisième chapitre, il n'est pas seulement question de définir les composantes du processus de validation, mais aussi de voir comment ce dernier s'insère dans le cadre de la recherche. Concrètement, une première définition du concept de validité est présentée par l'entremise d'une comparaison entre les types traditionnels de validité et la vision unifiée de Messick. Ensuite, un premier lien est fait entre la validité et la qualité de l'échelle de référence d'un instrument de mesure dans le contexte de la TRI. Finalement, le biais de mesure associé à l'ordre de présentation des questions est introduit dans son rapport avec la validité.

3.1 PROCESSUS DE VALIDATION

Dans la pratique, le processus de validation d'un instrument permet au concepteur d'un instrument d'améliorer la qualité de ce dernier en évaluant sa validité. Cet important processus demande vigilance et une analyse des caractéristiques de l'instrument. L'*American Educational Research Association* (1985; p. 9), en collaboration avec l'*American Psychological Association* et le *National Council on Measurement in Education* définit la validité comme : « *Le concept réfère à la convenance, la signification et l'utilité des inférences tirées des résultats du test. La validation du test est un processus d'accumulation d'indices qui viennent en support à ces inférences.* » De son côté, Messick (1993), en référence à son texte de 1989, définit la validité comme : « *La validité est un jugement intégré qui évalue à quel point les indices empiriques et le rationnel théorique viennent en support à l'adéquation et à la convenance des interprétations et*

actions qui découlent des résultats au test ou à toute autre forme de mesure. » Ces définitions soulignent que ce n'est pas l'instrument de mesure qui est validé, mais bien les inférences faites à partir des données recueillies.

Dans ce contexte, le processus de validation d'un instrument se caractérise par une démarche qui consiste à accumuler le plus d'indices empiriques possibles permettant d'affirmer que les interprétations ou inférences faites à partir des données sont adéquates. Ces indices peuvent être qualitatifs ou quantitatifs. Ce sont ces indices empiriques qui sont présentés avec les instruments comme preuves de leur validité. Dans la plupart des recherches, ces indices sont quantitatifs, tel le résultat d'une analyse factorielle ou le coefficient alpha de Cronbach. Il existe aussi des indices qualitatifs de validation d'instruments (Downing et Haladyna, 1997; 1999).

Il est possible de regrouper ces indices quantitatifs et qualitatifs sous des appellations communes en fonction de deux conceptions du concept de validité. Dans la conception traditionnelle de la validité, les regroupements se font selon les types de validité et, dans sa conception moderne, ils se font en regard des sources de la validité. La distinction entre les deux est attribuable aux travaux de Messick (1989; 1993) et à sa vision unifiée du concept de validité. En somme, les sources de validité sont une extension des types de validité.

3.1.1 Types traditionnels de validité

Traditionnellement, la validité d'un instrument se définit en trois types : la validité de contenu, la validité en référence à un critère externe et la validité du construit. Ces appellations sont celles retrouvées dans la plupart des recherches ou livres lorsque le processus de validation est abordé (par exemple : Aiken, 1997; Spector, 1992; Wilkie, 1999). Toutefois, la définition de ces types de validité a subit plusieurs modifications au cours des années. Pour un court historique, il est suggéré de consulter Angoff (1988). Dans sa version moderne, la validité du contenu se présente comme une série d'indices qui permettent d'évaluer si le contenu de l'instrument, c'est-à-dire les items, est pertinent et représentatif de la dimension théorique mesurée (du trait latent). La validité en référence à un critère externe se définit comme une corrélation entre les données d'un instrument en regard des données obtenues à l'aide d'un autre instrument utilisé comme critère externe. Ce dernier type de validité se divise en deux sous-catégories (Spector, 1992). Il y a la validité concurrente pour laquelle la corrélation se fait entre des données de l'instrument recueillies simultanément avec celles de l'instrument externe; et la validité prédictive pour

laquelle les données de l'instrument sont recueillies avant celle de l'instrument externe. Dans ces deux cas, il y aura validité convergente lorsque la corrélation entre les données est grande et il y aura validité divergente lorsque la corrélation est faible. Finalement, le troisième type traditionnel est la validité du construit. Ce type de validité cherche à savoir si la structure des items de l'instrument est explicite en regard de la conception théorique de la dimension mesurée et si elle correspond bien à ce que l'instrument donne comme résultat.

3.1.2 Sources de validation

Messick (1993) critique cette classification en trois types. Selon lui, il n'est pas possible de séparer les indices empiriques en types et de les utiliser de façon exclusive. Ce n'est pas à la définition des types que Messick s'oppose mais à l'idée de les utiliser exclusivement, d'en choisir une seule pour procéder à la validation. D'ailleurs, Spector (1992) souligne : « *La validation d'un test requiert une stratégie qui exige d'accumuler le plus grand nombre possible d'indices différents* ». Dans sa vision, Messick (1989) élargit les types de validité et il change aussi l'appellation type pour source des indices. Messick (1993) identifie six sources possibles. Ces six sources incluent indirectement tout ce qui est couvert par les trois types traditionnels. Elles sont : la pertinence, la représentativité et les qualités techniques du contenu; la substance; la structure interne; la structure externe; la généralisation de l'utilisation de l'instrument; et les conséquences pour les répondants. La pertinence fait référence au lien entre le contenu de l'instrument et la théorie sur la dimension que l'on cherche à mesurer. La substance s'occupe du rationnel théorique qui explique la constance dans les réponses des sujets. La structure interne s'intéresse aux relations entre la structure des réponses et celle du construit. La structure externe compare la structure des données avec la structure des données d'un instrument externe. La généralisation essaie de voir à quelle situation les résultats peuvent être généralisés. Finalement, les conséquences pour les répondants se résument aux implications sociales de l'utilisation des résultats obtenus à l'instrument (au niveau des valeurs par exemple). C'est surtout au niveau des conséquences que Messick apporte quelque chose de nouveau. Pour Messick, ces sources de la validité se catégorisent ensuite sous la forme de facette de la validation.

La conception traditionnelle et la conception de Messick diffèrent quelque peu, mais à partir du moment où le concepteur est conscient que plusieurs indices existent, la différence n'est plus qu'une question de taxonomie ou de classification. Le mérite de Messick est d'avoir introduit l'idée des conséquences pour les répondants et de les avoir intégrées au processus de validation,

d'avoir transcendé la division qui existait auparavant, et d'avoir fait de la validité un concept unifié. La règle d'or du processus de validation serait donc de ne négliger aucune information et d'utiliser le plus de techniques et d'indices possibles tout au long du processus. La validité d'un instrument de mesure ne se limite toutefois pas aux types traditionnels de validité et aux sources de la validation. Il y a un autre concept lié de près à ceux-ci et qui doit être mentionné ici : la fidélité.

3.1.3 La fidélité

La fidélité d'un instrument de mesure est un indice qui permet de s'assurer que l'instrument de mesure donne des résultats constants d'une administration de l'instrument à l'autre. Elle se calcule généralement à partir d'administration répétitive de l'instrument (seulement l'erreur de mesure devrait distinguer les résultats obtenus). Cet indice de la validité est lié de près aux sources de validation de la substance, de la généralisation des résultats et aux conséquences pour les répondants et est une opération qui accompagne le processus de validation. Par exemple, dans la mesure où un instrument donne des résultats stables pour un groupe donné, il peut être possible de faire des inférences sur la possibilité de généraliser ces résultats à un autre groupe avec des caractéristiques conceptuellement équivalentes. Dans le cadre de cette recherche, nous n'aborderons pas la fidélité. Pour plus de détails sur la question de la fidélité ou son rapport avec la TRI, le lecteur peut consulter Hambleton, Swaminathan et Rogers (1991) ou Laveault et Grégoire (1997).

3.2 LA VALIDITÉ DU CONSTRUIT DANS LE CADRE DE CETTE RECHERCHE

Cette recherche se concentre sur deux des six sources de validité de Messick : la structure interne ou la validité du construit dans sa compréhension traditionnelle. Dans l'introduction, les problèmes associés à l'utilisation du score total provenant de l'addition des codes numériques associés à l'échelle de référence ont brièvement été discutés. Il serait donc intéressant de trouver une solution à ces problèmes de validité des scores et de voir comment la TRI s'insère dans ce processus. Dans le cadre de cette recherche, ce sont deux aspects de la validité du construit qui nous intéressent : la construction d'une mesure fondamentale et la vérification du nombre de catégories de l'échelle d'appréciation de Likert.

3.2.1 La mesure fondamentale en sciences sociales

Depuis plusieurs années, les experts de la mesure en sciences sociales tentent de construire des échelles de mesure pour lesquelles les codes numériques auraient des propriétés (linéarité, additivité) comme celles retrouvées dans les sciences de la nature (kg, mètre, etc.) (Englehard, 1992). Pour bien comprendre comment ce type de mesure peut être obtenu en sciences humaines, il faut faire la distinction entre mesure fondamentale, mesure dérivée et mesure implicite. Van der Linden (1994) fait un bref exposé des ces trois types de mesure. L'apport de la TRI au processus de validation de l'échelle de référence passe directement par une compréhension de ces trois types de mesure.

Van der Linden (1994) s'est attardé au travail de Campbell (1928) pour faire la distinction entre mesure fondamentale et mesure dérivée. En somme, la mesure fondamentale se définit par la possibilité de mesurer directement des objets différents à l'aide d'une unité stable. Pour ce faire, il faut respecter trois principes : la relation d'ordre (linéarité), l'additivité et l'établissement d'une unité arbitraire. Dans la mesure où ces trois principes sont vérifiées empiriquement, la variable est mesurable directement, fondamentalement. Van der Linden résume : « *L'analyse que fait Campbell de la mesure se résume par le principe qu'établir une variable quantitative est un principe qui implique des lois naturelles qui doivent être vérifiées avant qu'une variable puisse être considérée comme vraiment quantitative.* » Dans l'éventualité où ces trois conditions ne sont pas respectées, il est tout de même possible d'obtenir une mesure quantitative. Il faut alors mesurer la variable indirectement à partir de la relation qu'elle entretient avec d'autres variables qui, elles, sont mesurables directement (fondamentalement). Il s'agit alors d'une mesure dérivée. Un exemple de mesure dérivée serait la vitesse qui se mesure à partir de deux mesures fondamentales que sont le temps et la distance. Selon Campbell, les mesures fondamentales et dérivées sont les deux seules mesures possibles.

Van der Linden (1994) va plus loin et décrit une procédure qui permettrait d'obtenir des mesures quantitatives en sciences sociales. Il désigne celle-ci comme la mesure implicite. Pour bien comprendre celle-ci, il faut la diviser en deux. Premièrement, la procédure consiste à construire des échelles linéaires de type quantitatif sur lesquelles des opérations de base (addition, soustraction) pourront être faites. Deuxièmement, l'objectif est de produire des instruments de mesure construits indépendamment des sujets et des items choisis pour l'estimation de thêta et de la difficulté des items. Cette deuxième étape réfère à la propriété d'invariance (voir 2.1.3). À partir du

moment où les données s'ajustent bien au modèle, une échelle invariante et de type quantitatif est obtenue. Luce et Tukey (1964) utilisent cette méthodologie pour bâtir une échelle quantitative à partir de leur modèle d'additivité conjointe. D'autres modèles, dont ceux de la TRI, répondent aux caractéristiques de la procédure décrite par van der Linden. Les modèles de Rasch, dont le *Rating Scale Model*, entrent dans cette catégorie de modèles qui permettent de produire des mesures implicites. Pour plus de détails sur la relation entre la mesure implicite et les modèles de Rasch, voir Andrich (1988) ou la section 2.2.2. C'est dans cette optique que la TRI a quelque chose à offrir dans le processus de validation d'un instrument de cueillette d'information.

Dans l'optique de cette recherche, la mesure fondamentale ou implicite se manifeste par la présence d'invariance dans l'estimation des paramètres des items. Dans la mesure où le modèle s'ajuste aux données et dans la mesure où il y a présence d'invariance dans l'estimation des paramètres, il sera sous-entendu que l'échelle possède les propriétés de la mesure implicite et, conséquemment, que la validité du construit est améliorée par l'utilisation de la théorie de la réponse à l'item.

3.2.2 Le nombre de catégorie sur l'échelle de Likert

Un autre indice de la validité du construit s'étudie par une exploration du nombre de catégories de l'échelle d'appréciation. Au début, le choix du nombre de catégories par le concepteur d'un instrument de cueillette d'information peut être assez arbitraire. Il existe certains critères qui peuvent guider le concepteur dans le choix du nombre de catégories à inclure tel que Bond et Fox (2001) l'indiquent. Mais ces deux auteurs suggèrent plutôt de vérifier la pertinence du nombre de catégories ultérieurement à partir des données empiriques recueillies avec l'instrument. Ils présentent d'ailleurs quelques suggestions qui permettent de rendre optimal le nombre de catégories à inclure dans une échelle. La stratégie consiste globalement à vérifier s'il est préférable de combiner certaines catégories en une seule. Des détails supplémentaires sont offerts au moment de la présentation des résultats à la section 5.3.

3.3 BIAIS DE MESURE, IMPACT SUR LA VALIDITÉ

En plus de faire l'exploration de la validité du construit à partir d'un modèle de la TRI, la recherche s'attarde aussi à l'impact de la présence d'un biais de mesure sur cette modélisation.

Dans la situation qui nous intéresse, le processus consiste à recueillir des indices qui permettent de vérifier si la présence d'un biais de mesure existe pour un groupe particulier. Plus précisément, il s'agit de voir si des propriétés de la TRI, telles l'invariance ou la mesure implicite, pourraient être affectées par la présence d'un biais de mesure. L'intérêt premier dans la détection du biais de mesure est d'étudier quels effets celui-ci pourrait avoir pour les répondants. Dans ce sens, c'est à la validité des conséquences (selon le sens donnée par Messick) pour les répondants qu'il faut penser. Mais dans le contexte de cette recherche, c'est l'effet de la présence d'un biais de mesure sur la modélisation à l'aide de la TRI qui nous intéresse et, plus particulièrement, son effet sur la propriété d'invariance de la TRI et sur l'ajustement des données.

3.3.1 Définition du biais de mesure

Pour définir le biais de mesure, il faut premièrement faire la différence entre le biais de mesure et le fonctionnement différentiel des items (FDI). Au chapitre I, le biais de mesure a été défini comme une erreur de mesure systématique qui est attribuable à des conditions non contrôlées par l'instrument. Il peut y avoir un biais de mesure parce qu'un groupe cible obtient systématiquement de moins bons résultats, comme il peut y avoir un biais de mesure parce que des items sont mal formulés. Pour mieux comprendre cette différence, il faut distinguer les appellations suivantes : FDI, impact sur l'item, impact défavorable sur l'item et biais de l'item. La terminologie qui suit est celle présentée dans Zumbo (1999) et inspirée de Camilli et Shepard (1994). Zumbo (1999) définit le FDI comme une différence entre deux groupes dans la probabilité de répondre à un item. Suite à la détection de FDI, il y a impact de l'item si la différence de probabilité décrite par le FDI s'avère être une différence réelle dans la performance des deux sous-groupes. Ou, il y a biais de l'item lorsqu'un groupe a une moins grande probabilité de répondre qu'un autre groupe pour des raisons autres que l'impact de l'item, pour des raisons non pertinentes à l'instrument. L'impact défavorable est un terme légal qui indique que la différence de performance à un test a des répercussions pour un groupe donné. Comme Zumbo (1999) le souligne, l'impact de l'item se différencie du biais de l'item lorsque la différence dans la probabilité de répondre des deux groupes est attribuable à des caractéristiques extérieures à l'instrument. Qu'il y ait un impact de l'item ou un biais de l'item, il faut premièrement avoir fait la preuve de FDI pour conclure à un ou l'autre. Ce court paragraphe permet de clarifier la terminologie utilisée pour la recherche. En conclusion, comme Camilli et Shepard (1994) le mentionnent, il faut se demander si la différence dans la probabilité de répondre de deux groupes

est réelle (par exemple une personne est réellement plus grande que l'autre) ou un artefact entièrement causé par un item. À noter que le biais de l'item peut se situer à un niveau autre que l'item, c'est-à-dire pour un groupe d'items ou pour l'instrument dans son ensemble. Pour plus de précisions sur l'ensemble de ces définitions, voir le premier chapitre de Camilli et Shepard (1994) ou le texte de Zumbo (1999).

Pour cette recherche, uniquement le FDI nous intéresse. Que ce FDI provienne d'une différence réelle dans la probabilité (impact de l'item) de répondre ou qu'elle soit artefact n'a pas d'importance pour la recherche. L'objectif est d'explorer l'influence que peut avoir la présence de FDI sur la modélisation, sur la validité du construit. C'est pour cette raison que, dans le reste de la recherche, pour parler de biais de l'item, l'expression « FDI » est toujours utilisée ou sous-entendue. Plus concrètement, ce qui nous intéresse est le FDI dans la réponse des étudiants de deux programmes différents du CFIM pour chacun des items de notre instrument. La définition mathématique du FDI utilisée pour cette recherche est expliquée plus longuement au prochain chapitre.

3.3.2 Ordre de présentation des items

Pour les besoins de la recherche, nous avons essayé d'induire intentionnellement un FDI dans les réponses. Schuman et Presser (1996) identifient plusieurs variables qui peuvent provoquer un FDI dans les réponses. Parmi les plus connues, il y a la désirabilité sociale lorsque la réponse est influencée par les conventions sociales, et la formulation de la question lorsque celle-ci peut influencer l'opinion du candidat. Il y a aussi l'ordre de présentation des items. Les études de Lane et al. (1987) et de Newman et al. (1988) étudient l'effet que peut avoir la présentation (en ordre de difficulté) des items pour un test à choix multiples. Ils démontrent que l'ordre des items pourrait avoir un effet (FDI) sur les réponses des groupes, mais que les études antérieures sont contradictoires et qu'il faudrait continuer la recherche. Schuman et Presser (1996) font la recension de plusieurs études effectuées antérieurement. Cette recension est intéressante parce qu'elle s'attarde à l'effet de l'ordre des items dans des sondages sur l'attitude, ce qui correspond très bien au questionnaire utilisé pour cette recherche.

Dans leur recension des études traitant de l'effet de l'ordre des items sur les réponses, Schuman et Presser (1996) ont fait une classification des différents effets possibles que peut avoir l'ordre des items. Il faudrait toutefois mentionner que l'analyse faite par Schuman et Presser utilise une

méthode de détection (khi-carré) qui, avec le recul, présente des limites et s'avère peu efficace selon Camilli et Shepard (1994). Il se pourrait donc que les résultats obtenus soient erronés. Ceci ne veut pas dire que l'étude de la question est inutile car peu d'études utilisant des méthodes alternatives existent dans la littérature. Les deux auteurs divisent les effets possibles de l'ordre des items selon les quatre catégories suivantes : la partie, le tout, la consistance et le contraste. La composition d'un item « tout » englobe le contenu de l'item « partie » qui est plus spécifique. Par exemple, un item « tout » pourrait être : « aimez-vous manger ? » tandis qu'un item « partie » serait : « aimez-vous le spaghetti ? ». La distinction entre consistance et contraste est une différence au niveau de l'effet. L'effet « consistance » est présent lorsque deux items obtiennent des réponses plus rapprochées que prévu. L'effet « contraste » est présent lorsque deux items obtiennent des réponses plus différentes que prévu. À ces catégories s'ajoutent l'effet *Salience* (effet d'une série de questions sur une autre question), l'effet rapport qui suggère de commencer avec des questions plus neutres pour se familiariser avec le répondant et ainsi obtenir plus d'honnêteté dans les réponses qui suivent, l'effet fatigue qui influencerait les dernières réponses d'un long questionnaire, et l'effet de la séquence des questions.

La recension de Schuman et Presser montre que l'intensité des effets varie selon la catégorie. Par exemple, pour le contraste partie-tout, les études recensées par Schuman et Presser suggèrent qu'il y aurait plus de support pour l'item général lorsqu'il est présenté en premier. À l'opposé, lorsque les items sont présentés dans l'ordre inverse, il y aurait une contamination de l'item spécifique sur l'item général (consistance partie-tout). La séquence dans laquelle les items sont présentés aurait un impact tant et aussi longtemps que la séquence garde un ordre logique. À l'exception de l'effet contraste partie-tout où les résultats sont plus convergents, la recension des études sur les autres effets n'est pas concluante. C'est-à-dire qu'il y a toujours une étude qui vient contredire les résultats obtenus dans une autre recherche. Il faut donc garder en tête qu'il peut y avoir des effets, mais que ce n'est pas automatique. Schuman et Presser (1996, p. 77) insistent davantage : « *Dans l'ensemble, pour nous, tous les effets de l'ordre des questions semblent constituer un des secteurs les plus importants de la recherche méthodologique. ... À ce point-ci, la recherche doit se concentrer non seulement à produire plus d'exemples, mais à comprendre ceux des études qui existent déjà.* »

Théoriquement, les fondements de la théorie de la réponse à l'item indiquent que l'ordre de présentation des questions ne devrait pas affecter l'utilisation du modèle ou l'estimation des paramètres du modèle. Pratiquement, il semble intéressant d'explorer l'effet que pourrait avoir la

présence de FDI sur les propriétés de la TRI et sur le processus de validation. Il ne reste plus qu'à voir comment détecter le FDI.

3.4 LE FONCTIONNEMENT DIFFÉRENTIEL DES ITEMS

La littérature présente plusieurs méthodes statistiques pour faire la détection du FDI. Les paragraphes qui suivent ne s'attardent pas à faire une recension complète de ces méthodes mais plutôt à dresser un bref aperçu des principales méthodes utilisées ces jours-ci. C'est-à-dire de faire un tour d'horizon des méthodes présentées dans les textes de Thissen, Steinberg et Wainer (1988); Holland et Wainer (1993); Millsap et Everson (1993); Camilli et Shepard (1994); Potenza et Dorans (1995) et Zumbo (1999). Pour une recension quasi complète des méthodes, le lecteur n'aurait qu'à ajouter le texte de Berk (1982) aux précédents. Tout au long de cette section, le groupe F (pour groupe focal) est défini comme l'échantillon complet de tous les sujets tandis que le groupe R (pour groupe de référence)est défini comme le groupe de sujets ciblé ou pour lequel il y a potentiellement un FDI.

3.4.1 Méthodes de détection inspirées de la TCT

Suite à la place grandissante qu'occupe la TRI dans la littérature, peu de méthodes récentes utilisent la TCT pour faire la détection du FDI. D'ailleurs, selon Camilli et Shepard (1994, p. 38) : « *Bien que les méthodes classiques de détection des biais soient intuitivement attirantes, le fardeau de ce chapitre a été de démontrer que les méthodes classiques – les différences valeur-p, ANOVA, et la corrélation point biséral – sont techniquement défectueuses.* » En plus, puisque c'est la TRI qui nous intéresse le plus pour cette recherche, les méthodes inspirées de la TCT (la Difficulté Transformée de l'Item [TID, ou valeur-p ou delta-plot], l'analyse de variance ou la corrélation point biséral [la corrélation item-test]) ne seront pas présentées ici. Pour plus de détails sur ces méthodes, le lecteur est redirigé vers le chapitre 2 de Camilli et Shepard (1994), le texte de Osterlind (1983) ou celui de Berk (1982). Le reste de ce chapitre présente les principales méthodes de détection de FDI associées à la TRI : la méthode Mantel-Haenszel, la méthode de standardisation, la méthode à surface et la méthode RV. Il existe aussi d'autres méthodes qui ne seront pas abordées ici. Parmi celles-ci, il y a la méthode par modèle de régression logistique (Swaminathan et Rogers, 1990) et celle de *Logistic Discriminant Function Analysis* (Miller et al., 1992; Flowers, Oshima et Raju, 1999).

3.4.2 La méthode Mantel-Haenszel

La méthode Mantel-Haenszel (Holland et Thayer, 1988) est intéressante parce qu'elle est beaucoup utilisée. Millsap et Everson (1993) mentionnent qu'elle est la plus utilisée parce qu'elle est simple et qu'elle inclut un test statistique. Hambleton et Rogers (1989) indiquent qu'un grand intérêt existe en ce moment pour cette méthode et celles inspirées de la TRI.

Camilli et Shepard (1994) décrivent cette méthode comme une approche de table de contingence. La raison est bien simple, l'analyse faite avec la méthode Mantel-Haenszel commence par la construction d'une table de contingence 2 X 2 X S. La table organise la réponse des sujets en fonction des groupes (R et F) et des réponses possibles à l'item (1 ou 0 dans le cas d'item dichotomique) pour chacun des S scores totaux possibles. Un instrument de dix items pourrait potentiellement avoir 11 tables de contingences de 2 X 2. Une table ressemble à ceci :

Tableau 3.1 Table de contingence 2 X 2 pour un score total S à un instrument de N items

	Score à l'item		
	1	0	Total
Groupe R	A	B	N (R)
Groupe F	C	D	N (F)
Total	M (1)	M (0)	T

La table de contingence est comptabilisé pour chaque score S possible (de 0 à 10 par exemple) L'intérieur des cases peut, selon le cas, représenter soit un nombre brut de sujets ayant répondu correctement ou une proportion de sujets qui ont répondu correctement à l'item dans chaque groupe. À partir de cette table, il est possible d'analyser la présence de FDI de deux façons Premièrement, il est possible de calculer l'amplitude de la différence qui existe entre les deux groupes sous la forme d'un ratio. En prenant en exemple le tableau 2.1, le ratio MH pour chaque score x est :

$$MH = \Delta_{MH} = -2{,}35 \ln \frac{\sum_{x=1}^{N} CB/T}{\sum_{x=1}^{N} AD/T}. \qquad (11)$$

Comme il s'agit d'un ratio, plus la valeur s'approche de 1, plus les deux groupes se ressemblent. Un groupe n'a pas plus de chance de réussir l'item que l'autre groupe lorsque le ratio est à 1. La deuxième possibilité consiste à procéder à un test statistique. L'hypothèse nulle stipule alors que p (R) = p (F). La statistique Mantel-Haenszel ($MH-\chi^2$)se base sur la statistique khi-carré suivante :

$$MH-\chi^2=\frac{(\left|\sum_{x=0}^{N}C-\sum_{x=o}^{N}E(C)\right|-0{,}5)^2}{\sum_{x=0}^{N}Var(C)} \qquad (12)$$

où $E(C)=\frac{N(F)M(1)}{T}$ et $Var(C)=\frac{N(F)M(1)N(R)M(0)}{T^2(T-1)}$.

La méthode de Mantel-Haenszel utilise une procédure assez simple inspirée de la statistique du khi-carré. Il y a toutefois certains reproches à faire à cette méthode. Hambleton et Rogers (1989) et Millsap et Everson (1993) indiquent que cette méthode fonctionne seulement pour détecter du FDI uniforme, où la différence demeure stable en faveur d'un groupe peu importe la valeur de thêta. En d'autres mots, les CCI ne se croisent pas. Toutefois, l'étude de Maranon, Garcia et Costas (1997) et celle de Raju, Drasgow et Slinde (1993) montrent que la méthode MH ne serait pas complètement insensible à la détection de FDI non uniforme. Millsap et Everson (1993) mentionnent aussi le problème de la représentativité des scores bruts dans l'éventualité où ceux-ci seraient obtenus à partir d'un modèle autre que le modèle de Rasch (il est ici question de la statistique exhaustive, voir l'alinéa 2.6.5), particulièrement pour des instruments de moins de 20 items. Ils mentionnent aussi que, sous certaines conditions, la méthode MH peut faussement détecter du FDI.

3.4.3 La méthode de standardisation

La méthode de standardisation a été élaborée par Dorans et Kulik (1983; 1986) et Dorans et Holland (1993). Cette méthode permet de prendre en considération la distribution du FDI en fonction du score des sujets. C'est-à-dire que, dans son calcul, elle prend en considération l'ampleur de FDI pour chaque thêta tout au long du continuum. Pour ce faire, deux courbes de régression logistique (CCI) sont produites, une pour le groupe F et une pour le groupe R. La

différence entre les deux courbes est calculée à partir d'un facteur de poids relatif. Pour un facteur de poids relatif W_K , un score thêta de x, la différence se calcule ainsi :

$$STDP - DIF = \frac{\sum_{x=0}^{N} W_k (P_{fx} - P_{rx})}{\sum_{x=0}^{N} W_k} . \quad (13)$$

Dorans et Holland (1993) suggèrent que les résultats obtenus avec la méthode de standardisation sont en lien assez étroit avec ceux de la méthode MH. Il est donc possible que les problèmes soient les mêmes. Parce que la méthode MH inclut un test statistique, il serait préférable de l'utiliser, question d'avoir le maximum d'information possible.

3.4.4 Les méthodes à surface (ou méthodes des signes)

Les méthodes à surface ont pour but de faire le calcul de la surface entre deux CCI ou deux estimation d'un paramètre. Ces méthodes sont souvent utilisées avec la méthode de Lord (1980) parce que les résultats des deux méthodes sont souvent en corrélation et parce que les méthodes à surface donnent seulement une mesure de la quantité tandis que Lord permet de procéder à un test statistique (Raju, Drasgow et Slinde, 1993). Le résultat de ce calcul peut être positif ou négatif. Par convention, un signe positif signifie qu'il y a FDI en faveur du groupe R. Le calcul de la surface est possible autant pour un FDI à signe positif ou négatif. Pour les deux prochaines formules, P représente la probabilité de répondre correctement à un item pour le groupe R ou F. La formule suivante prend en considération les signes et donne un quantité ajustée :

$$U-A=\sqrt{\int (P_r(\theta)-P_f(\theta))^2 \partial\theta} . \quad (14)$$

Dans la même optique que pour la méthode de standardisation, il peut être intéressant de pouvoir nuancer le résultat en sachant où sur le continuum la différence est la plus grande, c'est-à-dire d'identifier quel thêta ou quel groupe de sujets participe le plus dans la quantification de l'écart entre les courbes. En prenant en considération le facteur de poids relatif, la formule ressemble à ceci :

$$UPD - \theta = \frac{\sum_{X=1}^{N_f} [P_r(\theta_x) - P_f(\theta_x)]}{N_f} . \quad (15)$$

Dans la formule 14, x indique que le calcul de la différence des P pour les deux groupes est effectué pour chacun des x thêta possibles pour les n sujets groupe de référence (N_f). De cette façon, le calcul considère seulement les thêta estimés par le modèle. Il y a aussi d'autres méthodes du même type qui permettent de directement calculer la différence entre la valeur des paramètres. Mais, parmi ces méthodes à surface, Camilli et Shepard (1994) suggèrent d'utiliser l'indice UPD-θ de Shepard, Camilli et Williams (1984).

3.4.5 Méthode de ratio de vraisemblance

Avant d'entrer plus en détail dans la description de la méthode, prenons quelques lignes pour expliquer le ratio de vraisemblance. Nous avons vu à la section 2.5 comment la méthode de maximum de vraisemblance est utilisée pour faire l'estimation des paramètres. La logique est ici la même. Toutefois, dans ce cas-ci, c'est le maximum de vraisemblance du modèle en rapport avec les données qui est mesuré. Camilli et Shepard (1994, p. 74) ont défini vraisemblance comme : « *La fonction de vraisemblance est une mesure quantitative qui décrit avec quelle exactitude un modèle avec certains paramètres représente un groupe de données.* » Ils utilisent l'exemple d'une pièce de monnaie. Quel serait le maximum de vraisemblance d'utiliser p=0,5 pour représenter des données recueillies au hasard pour le tirage d'une pièce. Dans ce cas-ci, la vraisemblance serait probablement élevée. Le ratio de vraisemblance se définit alors comme le ratio entre la vraisemblance de deux modèles différents qui expliquent les mêmes données. Le premier modèle C (compact) modélise les données du groupe de référence et du groupe focal à l'aide d'une seule CCI simplifiée. Le deuxième modèle A (augmenté) modélise les données à l'aide de deux CCI, une pour le groupe R et une pour le groupe F. Ainsi, lorsque le modèle A s'ajuste mieux aux données, cela suggère que les CCI des deux groupes diffèrent suffisamment pour qu'il soit nécessaire d'utiliser ce modèle A. Donc, il y a FDI. Le ratio RV se définit mathématiquement comme :

$$RV = \frac{L(\text{ModèleC})}{L(\text{ModèleA})}. \qquad (16)$$

Ce ratio est une quantité qui compare la probabilité de la vraisemblance d'un modèle par rapport à un autre. En plus de ce ratio, il est aussi possible de calculer une statistique de khi-carré à partir du calcul du ratio et donc de faire un test de signification. L'hypothèse nulle est que RV = 1 et la statistique est :

$$\chi^2(M) = [-2 \ln L(\text{ModèleC})] - [-2 \ln L(\text{ModèleA})] \qquad (17)$$

La procédure d'ancrage constitue le problème le plus important des méthodes RV, selon Millsap et Everson (1993). C'est-à-dire que les items de l'ancrage utilisé pour l'estimation peuvent contenir du FDI et tous les calculs seraient ainsi erronés. Pour remédier à ce problème, il suffit de procéder par itération à l'élimination de tous les items de l'ancrage qui contiennent du FDI. Dans un deuxième temps, il faut aussi s'assurer d'utiliser un modèle bien ajusté dès le début. Malgré cela, Camilli et Shepard recommandent fortement l'utilisation de cette méthode.

3.4.6 Méthode SIBTEST

Similairement à la méthode RV, la procédure SIBTEST développée par Shealy et Stout (1993) commence par fixer un groupe d'items (sous-test valide) qui seraient invariables pour les deux groupes. Ensuite, les autres items sont analysés en regard de ce sous-test. Avec cette procédure, il est possible de simultanément faire l'analyse d'un item ou de plusieurs items à la fois. Dans l'analyse de plus d'un item à la fois, le score total au sous-groupe d'items est considéré comme un item en soi. Suite à cette procédure, il est aussi possible de faire un test à partir de la statistique SIBTEST. Pour plus de détails sur les aspects techniques complexes de cette méthode, le lecteur est redirigé vers Shealy et Stout (1993).

3.4.7 Extension pour les modèles polytomiques

La plupart des méthodes présentées plus haut peuvent être transposées au cas des modèles polytomiques. Il suffit de comprendre que les mêmes principes de base sont maintenus dans les versions polytomiques de ces méthodes. Dans sa version polytomique, la méthode Mantel-Haenzel se présente sous deux versions possibles : polytomique MH et MH généralisé (GMH). Les deux sont décrites dans Zwick, Donoghue et Grima (1993). Pour la méthode de standardisation (STND), quelques petites modifications sont apportées et elle porte le nom de méthode SMD (moyenne standard de différence). Il existe deux versions différentes de cette

méthode : SMD-H et SMD-M. Les deux sont décrites dans Zwick et Thayer (1996) et de façon générale dans Potenza et Dorans (1995). La méthode SIBTEST est adaptée dans Chang, Mazzeo et Roussos (1996) et, finalement, les méthodes RV sont adaptées par Wainer, Sireci et Thissen (1991) et Muraki (1993). Les méthodes à surface sont plus difficiles à adapter étant donné la présence de plusieurs courbes associées à chaque niveau possible de réponses.

3.4.8 Évaluation des méthodes

Suite à l'introduction de ces différentes méthodes, il importe de savoir quelle méthode est la plus « adéquate ». Hambleton, Swaminathan et Rogers (1991) favorisent les méthodes RV et MH. Raju, Drasgow et Slinde (1993) montrent que les méthodes à surface, de Lord et MH ont détecté sensiblement un FDI pour les mêmes items. Hambleton et Rogers (1989) supportent ces conclusions pour les méthodes à surface et la méthode MH. Mais, de façon générale, en fonction des caractéristiques des données, des ressources et du temps, il serait probablement idéal d'utiliser plus d'une méthode de détection comme Camilli et Shepard (1994, p. 131) l'indiquent : « *Premièrement, parce qu'une méthode unique ne ressort pas dans la littérature de recherche, une personne devrait utiliser des techniques complémentaires. Certaines penchent plus vers l'inférence, d'autres vers la description, certaines vers la différence de difficulté entre les groupes, et d'autres vers la discrimination.* »

Pour les méthodes polytomiques de détection de FDI, le développement de ces méthodes est assez récent. Pour ces raisons, ces méthodes n'ont pas encore été étudiées beaucoup dans la pratique quotidienne de détection de FDI (Potenza et Dorans, 1995). Potenza et Dorans (1995) proposent plutôt d'utiliser les méthodes SMD et SIBTEST parce qu'elles sont plus près des besoins pratiques. Chang, Mazzeo et Roussos (1996) montrent que la procédure SIBTEST s'avère efficace et même plus efficace que la méthode GMH sous certaines conditions. Zwick, Thayer et Mazzeo (1997) comparent eux aussi SIBTEST avec quatre autres méthodes. Dans l'ensemble, la plupart des méthodes comparées donnent les mêmes résultats bien que SIBTEST soit un peu plus efficace selon eux. Potenza et Dorans (1995) avaient d'ailleurs fait le parallèle entre les méthodes SMD et SIBTEST. Dans ces circonstances, SIBTEST nous semble être un bon outil à privilégier dans la détection de FDI pour des items à réponses polytomiques. C'est d'ailleurs dans cette direction que nous allons continuer avec cette recherche.

3.4.9 Taille de l'échantillon

La question de la taille de l'échantillon soulève des interrogations dans la détection de FDI. Quelle est la taille minimale que doit avoir notre échantillon ? Zieky (1993) prend une position philosophique quant à la grandeur de l'échantillon. Selon lui, même si de petits échantillons peuvent mener à de fausses détections, ce serait tout de même mieux de faire cette fausse détection que de ne pas essayer du tout. Il suffit d'exercer son bon jugement pour distinguer une détection insensée d'une détection réelle. Cette position n'est pas partagée par tous. Plus concrètement, à la phase de la construction de l'instrument, Zieky (1993) suggère un échantillon minimum de 100 personnes pour le groupe F et de 500 pour le groupe R. Ces chiffres montent à 200, 500 et 600, 2000 respectivement pour les groupes F et R pour la phase d'administration du test et pour la phase du rapport final. Linn (1993) émet certaines réserves suite à l'article de Zieky. Il pense que ces chiffres seraient trop bas et pourraient avoir des effets non négligeables.

3.5 SYNTHÈSE DU CADRE CONCEPTUEL

La figure 3.1 à la page suivante résume visuellement ce qui a été vu jusqu'ici. Les flèches (identifiées par un numéro) expliquent le lien entre ces concepts. Il y a deux thèmes clefs : la TRI et le processus de validation. Premièrement pour la TRI, un modèle organise les données obtenues à partir de l'instrument de cueillette d'information (1). Cette organisation se fait à partir d'une estimation des paramètres du modèle (2). Ensuite, des vérifications sont faites pour voir si les données et le modèle s'ajustent bien ensemble (3). Finalement, si le modèle s'ajuste bien, les hypothèses de base de la TRI devraient théoriquement être confirmées (4) et, pour la famille de Rasch, il serait possible d'obtenir une mesure fondamentale (5). Pour le processus de validation, il y a la cueillette d'indices qui démontrent la validité d'un instrument. Ces indices se classent sous six sources ou types de validité (6) parmi lesquels il y a la validité du construit. Incluse dans le concept de la validité du construit, il y a la construction d'une échelle de mesure à partir de la TRI (7) et la détection de FDI (8). De plus, nous allons aussi étudier l'impact du FDI sur le modèle de la TRI (9).

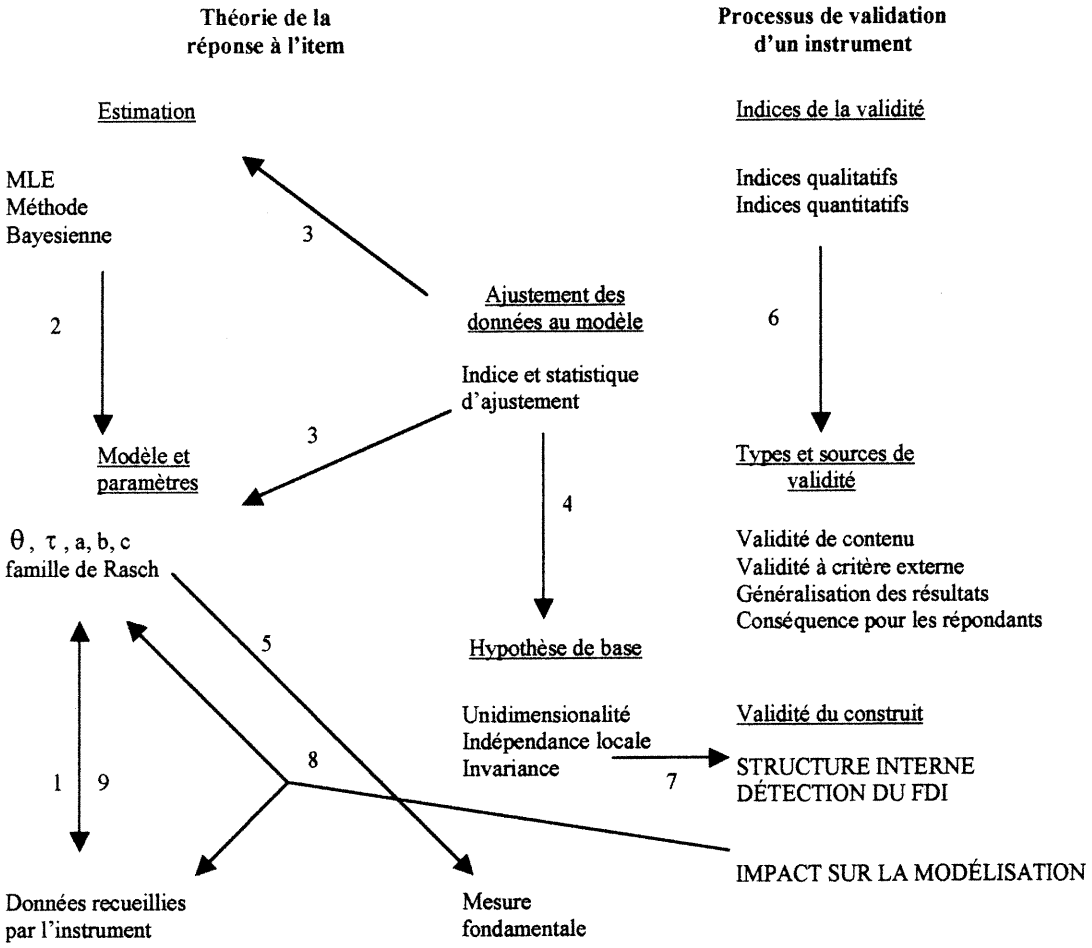

Figure 3.1 Synthèse des chapitres deux et trois.

CHAPITRE IV

CADRE MÉTHODOLOGIQUE

Les chapitres deux et trois ont servi à définir certains concepts clefs de la théorie de la réponse à l'item et du processus de validation d'instruments de cueillette d'information, c'est-à-dire à établir un cadre conceptuel pour la recherche. Ce quatrième chapitre présente la démarche de la recherche, la nature précise de la recherche, les deux versions de l'instrument de cueillette d'information, l'échantillon et la procédure de distribution de l'instrument, le choix des outils statistiques (modèles, méthodes, etc.) et, enfin, le choix des logiciels informatiques utilisés. Nous en profiterons aussi pour présenter la démarche d'analyse des données. Le contenu des deux chapitres précédents sert à justifier l'ensemble de la démarche de cette recherche.

4.1 NATURE DE LA RECHERCHE

Pour décrire la nature de cette recherche, la terminologie de Van der Maren (1999) est utilisée. Selon Van der Maren (1999), les recherches en éducation s'inscrivent principalement dans un des quatre types d'enjeux suivants : nomothétique, pragmatique, politique, ontogénique.

4.1.1 L'enjeu nomothétique

La principale différence entre l'enjeu nomothétique et les autres enjeux est dans l'implication de la recherche pour la pratique. L'enjeu nomothétique est théorique et vise le développement de la

science. Van der Maren (1999, p. 23) décrit cet enjeu :

> *Le but de ce type de recherche est le développement et le raffinement des connaissances théoriques. Il est nomothétique : proclamer des lois, des principes généraux, des théories. Sa démarche est nomologique : elle part d'énoncés généraux (lois) pour produire des énoncés ayant la forme et les caractéristiques des énoncés généraux (lois). Ce processus implique une attitude critique à l'égard des énoncés antérieurs afin de faire progresser la connaissance.*

L'enjeu de cette recherche est de faire progresser les connaissances sur le processus de validation d'un instrument de cueillette d'information à l'aide de la théorie de la réponse à l'item. Cet enjeu se distingue de recherches qui ont comme point de départ la résolution de problèmes pratiques (enjeu pragmatique), le changement de pratiques institutionnelles (enjeu politique) ou le perfectionnement professionnel par la réflexion (enjeu ontogénique). Pour reprendre une terminologie connue, la recherche à enjeu nomothétique correspond à de la recherche fondamentale tandis que les trois autres enjeux sont de l'ordre de la recherche appliquée.

4.1.2 Recherche empiriste et type de discours de la recherche

Bien que cette recherche soit fondamentale, elle s'inspire de données réelles recueillies dans le cadre de l'évaluation de programmes de formation universitaire. Dans ce cas précis, il est alors question d'une recherche empiriste hypothético-déductive au sens de Van der Maren (1999, p. 24) :

> *La recherche porte sur des collections de faits obtenus à partir de populations ou d'échantillons de ces populations et procède en deux phases. D'abord une phase inductive, ou exploratoire, qui permet de générer des hypothèses à la suite de l'observation et de l'analyse de plusieurs séries d'événements; puis une phase déductive, ou vérificative, qui tente de mettre ces hypothèses à l'épreuve dans une expérience critique, une expérience comparative (groupe témoin ou contrôle comparé à un groupe expérimental ou pilote) ou une application technologique.*

Ce type de recherche est différent de la recherche spéculative qui a pour but unique la formulation et la reformulation de théories à partir de théories. Parce que cette recherche a comme point de départ des données quantitatives sur une population, elle est donc empirique et exploratoire parce qu'elle ne vérifie pas d'hypothèses établies à l'avance mais cherche plutôt à observer les résultats obtenus.

La dernière précision à apporter sur la nature de cette recherche est dans le type de discours qui est produit. Parce que cette recherche s'inscrit dans une perspective exploratoire, sa fonction est de décrire comment il est possible d'utiliser la TRI dans le processus de validation d'un instrument et d'étudier comment la TRI se comporte lorsqu'il y a un biais de mesure dans les réponses des candidats. Le discours est donc descriptif.

En somme, notre recherche est : 1) une recherche fondamentale à enjeu nomothétique parce qu'elle vise le développement des connaissances sur l'utilisation de la TRI dans le processus de validation d'instrument de cueillette d'information; 2) une recherche empirique quantitative parce qu'elle a comme point de départ des données codées; 3) une recherche exploratoire parce qu'elle tente de faire des liens entre les résultats obtenus; et finalement 4) une recherche descriptive parce qu'elle tente de voir dans quelles conditions s'inscrivent les résultats de la recherche.

4.2 DESCRIPTION DE L'INSTRUMENT DE CUEILLETTE D'INFORMATION

Le questionnaire utilisé pour recueillir les données a été construit par le Centre de Formation Initiale des Maîtres dans le but de faire l'évaluation de ses programmes de formation des maîtres et il a été distribué aux étudiants pour une deuxième année en 2001. Ce questionnaire comporte huit sections différentes : la perception générale de la formation, la préparation à l'enseignement, les stages, les études et la gestion du temps, la poursuite des études, commentaires sur votre programme d'études, renseignements généraux et volontaires recherché(e)s. Le questionnaire en entier est présenté à l'annexe A. La section « préparation à l'enseignement » est la section retenue pour cette recherche. La consigne est la suivante : « *Cette section a pour objet de connaître ce que vous pensez de votre degré de préparation pour la réalisation de certaines tâches. Vous devez répondre en tenant compte de la phrase d'introduction suivante : Je considère que mon programme d'études m'a préparé(e) adéquatement pour...* ». La consigne est suivie de vingt items auxquels les étudiants répondent à partir d'une échelle de Likert. L'échelle de référence est en quatre points : tout à fait d'accord, plutôt d'accord, plutôt en désaccord, tout à fait en désaccord codés respectivement 1, 2, 3, 4. L'ordre des questions dans la version de l'an dernier avait été déterminé par les concepteurs du questionnaire. La version de l'an dernier, donc la version première, est identifiée A pour ce texte. Pour répondre aux objectifs de cette recherche une deuxième version a été construite. Elle est étiquetée B. Dans cette deuxième version, l'ordre

de présentation des items de la section II a été modifié dans le but d'induire un biais de mesure dans la réponse des étudiants.

4.2.1 La détermination de l'ordre des questions pour la version B

L'alinéa 3.4.2 suggère que la façon de répondre à un item peut être influencée par l'ordre de présentation de l'ensemble des items. Une façon pour essayer d'induire intentionnellement un FDI dans les données consiste à placer les items en ordre croissant de l'item ayant obtenu le plus petit degré d'accord jusqu'à l'item ayant obtenu le plus grand degré d'accord. Pour déterminer l'ordre de présentation des items, les données recueillies l'an dernier sont utilisées comme point de référence. Comme la recherche se concentre principalement sur les programmes d'enseignement secondaire et d'éducation préscolaire et enseignement primaire (ce choix est justifié à l'alinéa 4.5.3), la détermination de l'ordre croissant des questions se fait surtout à partir de ces deux groupes. Tout de même, les données recueillies pour l'ensemble des programmes seront elles aussi prises en considération. Voici un tableau synthèse des données de l'an dernier.

Tableau 4.1 Tableau des résultats de l'an dernier (en ordre croissant pour le secondaire) à la section II : préparation à l'enseignement.

	Pourcentage en accord (favorable) [ix]		
	SEC	PRI	TOUS
12.Répondre aux demandes des parents	6,6 %	12,8 %	18,4 %
30.Établir des relations avec les parents	8,6 %	16,7 %	22,0 %
11.Identifier les contenus difficiles à faire apprendre	36,2 %	44,1 %	45,8 %
29.Diriger les élèves vers les services d'aide personnelle	36,7 %	28,3 %	37,5 %
16.Corriger la langue orale des élèves	37,7 %	73,9 %	58,5 %
15.Corriger la langue écrite des élèves	38,7 %	76,1 %	62,1 %
17.Intervenir individuellement auprès des élèves à risque d'échouer	45,5 %	48,6 %	54,6 %
14.Maîtriser les contenus que j'enseignerai, en conformité avec les programmes du ministère	46,7 %	73,7 %	57,1 %
24.Établir des relations avec les membres de l'équipe-école	46,7 %	41,7 %	51,1 %
28.Sensibiliser les élèves à la discrimination envers les autres	47,2 %	55,6 %	51,0 %
26.Aider les élèves à développer leurs méthodes de travail	50,0 %	50,0 %	56,6 %
27.Sanctionner les problèmes de discipline chez les élèves	53,0 %	52,2 %	57,8 %
23.Identifier les points forts et faibles des élèves	54,3 %	72,5 %	68,7 %

ix Le pourcentage d'accord correspond au pourcentage des étudiants qui ont répondu dans les catégories tout à fait d'accord et plutôt d'accord de l'échelle de référence.

13.Construire des outils pour l'évaluation sommative	59,8 %	65,9 %	59,4 %
25.Construire des outils pour l'évaluation formative	63,8 %	60,9 %	60,7 %
21.Motiver les élèves à s'engager dans leur apprentissage	69,7 %	76,0 %	74,7 %
19.Adapter mes activités d'enseignement aux caractéristiques des élèves	73,9 %	80,0 %	79,7 %
20.Établir les règles de fonctionnement de la classe	76,4 %	79,8 %	80,6 %
22.Respecter les différences ethniques ou culturelles des élèves	80,4 %	88,3 %	83,8 %
18.Planifier le déroulement d'activités d'apprentissage	88,4 %	96,7 %	92,8 %

Ce tableau montre que les items avec le plus petit degré d'accord (DA) sont les quatre mêmes pour les programmes TOUS (tous les programmes), SEC (enseignement secondaire) et ces quatre items sont parmi les cinq plus difficiles pour le programme PRI (éducation préscolaire et enseignement primaire). Comme il n'y a presque pas de différence en pourcentage entre l'item 29 et 11 au SEC, les items avec le plus petit DA sont en ordre croissant 12-30-29 et 11. À l'autre extrémité, les cinq items avec le plus grand DA sont exactement les mêmes pour les programmes SEC et TOUS et ils se retrouvent assez similairement dans les six items avec grand DA pour le programme PRI (item 15 apparaît en plus dans la liste PRI). Parce que l'item 15 a un moins grand DA pour le SEC, il est mis de côté pour l'instant. La faible différence de 0,2 % (aux items 19 et 20 pour le PRI) permet de négliger la différence entre l'ordre des items pour le PRI par rapport à l'ordre des deux autres programmes. Conséquemment, les items avec le plus grand DA sont en ordre croissant 21-19-20-22 et 18.

Le reste des items est plus difficile à placer en ordre croissant parce que les DA sont plus variables d'un programme à l'autre. Parmi ces items, il y a une différence importante (jusqu'à 38 %) entre le DA des programmes pour les items 15 et 16. Il y a aussi une bonne différence aux items 23 et 14 entre le programme SEC et les deux autres. Ce groupe de quatre items diffère beaucoup d'un programme à l'autre. Les sept autres items parmi ceux qui restent sont un bloc d'items de difficulté moyenne qui maintiennent un ordre assez similaire pour les trois programmes à l'exception de l'item 28 qui bouge d'un programme à l'autre. Pour scinder, cet item a été placé à un endroit moyen en fonction de l'ordre qu'il occupait dans les trois programmes. Ainsi, l'ordre croissant pour ces sept items est : 24-17-28-26-27-25-13. Maintenant pour les quatre items qui diffèrent beaucoup d'un programme à l'autre, l'item 23 est plus facile que l'item 13 pour les programmes SEC et TOUS. L'item 14 est à peu près facile comme l'item 26 et les items 17/28. En insérant l'item 14 entre 28 et 26 la liste devient : 24-17-28-14-26-27-25-13-23. Pour les items 15 et 16, ils pourraient être placés soit au début (selon SEC) ou à la fin

(selon PRI) de cette dernière liste d'items. Comme l'objectif est de générer un biais de mesure suite à plusieurs items difficiles dès le début du questionnaire, ils seront insérés au début. Suite à cette analyse, l'ordre des questions modifié pour le questionnaire B est : 12-30-29-11-15-16-24-17-28-14-26-27-25-13-23-21-19-20-22 et finalement 18. Évidemment, il aurait été possible d'adopter une approche différente pour modifier l'ordre de présentation des questions et celle préférée ici est arbitraire. Une autre approche aurait pu essayer d'isoler certains items précis à la suite d'un bloc d'items en analysant qualitativement le contenu de ces items. Mais dans l'optique de cette recherche, les données de l'an dernier ont été préférées à l'analyse du contenu.

4.3 ÉCHANTILLON ET CUEILLETTE DES DONNÉES

4.3.1 L'échantillon

La population pour la recherche est formée des étudiants de cinq programmes des maîtres sur deux années universitaires, c'est-à-dire la troisième et la quatrième année des programmes à l'exception du programme de français langue seconde et d'éducation physique qui en sont seulement à leur troisième année d'existence. Le plus grand nombre d'étudiants se retrouve dans les programmes d'enseignement secondaire et d'éducation préscolaire et enseignement primaire. C'est la raison pour laquelle la recherche se concentre sur ceux-ci. L'an dernier (2000), le CFIM avait recueilli au total 387 questionnaires pour le programme d'enseignement secondaire et d'éducation préscolaire et enseignement primaire ensemble dont 199 pour le secondaire avec 107 en troisième année et 92 en quatrième année; et 180 pour le primaire avec 19 en troisième année et 161 en quatrième année. Cette année (2001), nous avons recueilli 537 questionnaires, soit 150 de plus que l'an dernier pour les mêmes programmes. Pour la version 1, nous avons obtenu : 126 répondants pour le secondaire avec 57 en troisième année et 69 en quatrième année; 117 pour le primaire avec 51 en troisième année et 66 en quatrième année. Pour la version 2, nous avons obtenu : 103 pour le secondaire avec 35 en troisième année et 68 en quatrième année; 105 pour le primaire avec 45 en troisième année et 60 en quatrième année. Pour 2001, nous avons réussi à rejoindre une bonne partie des étudiants de ces deux programmes : 537 étudiants pour environ 632 inscrits selon les données fournies par le CFIM. En somme, environ 85% des étudiants inscrits ont répondu au questionnaire.

4.3.2 La cueillette des données

L'année dernière, la stratégie utilisée pour recueillir les questionnaires n'a pas permis de rejoindre tous les étudiants. Une approche différente a donc été pensée pour cette année (2001) dans l'objectif de rejoindre le plus grand nombre possible de répondants. Dans le cadre de leur formation, il y a quelques moments précis où la grande majorité des étudiants est présente. Les activités reliées à l'inscription des stages et aux séminaires de stage sont deux activités obligatoires pour les étudiants et puisque les stages sont obligatoires dans ces programmes, la plupart des étudiants se présentent à ces activités, à quelques exceptions près (voyage, travail, etc.). Il est ainsi facile de rencontrer les étudiants de troisième année à l'occasion de l'inscription aux stages et de rencontrer les étudiants de quatrième lors de leur dernier séminaire de stage.
Dans le cas de l'inscription au stage pour les étudiants de troisième année, les coordonnatrices des stages organisent une ou deux rencontres où elles expliquent le déroulement du stage, les modalités d'inscription et l'attribution des places de stage. Suite à une entente avec les coordonnatrices de stage, nous avons pu assister à ces rencontres. À la fin de ces rencontres, les étudiants complètent des feuilles d'inscription pendant que les coordonnatrices demeurent disponibles pour des questions. Nous avons profité de ce moment pour faire la présentation du questionnaire et pour demander aux étudiants de le compléter en même temps qu'ils remplissent leur formulaire d'inscription au stage. Pour s'assurer que les deux versions du questionnaire sont distribuées le plus aléatoirement possible, nous avons fait des piles de questionnaires où une version A et une version B se suivent une après l'autre dans la pile. Ceci veut dire que le premier étudiant reçoit la première version, le deuxième l'autre version, le troisième la première version et ainsi de suite.

Dans le cas des rencontres de séminaire de stage pour les étudiants de quatrième année, la procédure diffère puisque les étudiants se divisent en plusieurs petits groupes avec plusieurs superviseurs de stage différents dans plusieurs salles. Entente fut prise avec ces superviseurs pour qu'ils demandent aux étudiants de remplir le questionnaire. Pour assurer le retour des questionnaires, nous avons inséré un nombre suffisant de questionnaires dans des enveloppes avec des consignes brèves (voir appendice B) que la coordonnatrice des stages remet aux superviseurs. Ces derniers devaient rapporter les questionnaires dans le casier de la coordonnatrice à la fin de la rencontre. Par la suite, les questionnaires nous ont été remis par la coordonnatrice. Le taux de présence à ces activités est élevé parce que ces rencontres sont obligatoires.

Toutefois notre contrôle sur les conditions d'administration est plutôt limité. Par exemple, il se pourrait que plusieurs étudiants complètent ensemble le questionnaire pendant une pause.

La liste de présence des coordonnatrices indique que très peu d'étudiants se sont absentés aux rencontres pour les troisièmes années. Lors de ces rencontres, la complétion des questionnaires a paru se faire avec sérieux, certains prenaient même beaucoup de temps à remplir le questionnaire. Dans l'ensemble, la cueillette des données s'est déroulée normalement sans événements particuliers à l'exception de quelques petites anecdotes. En une occasion, deux étudiantes assises près l'une de l'autre ont choisi de répondre au questionnaire ensemble. Elles ont remarqué que l'ordre de présentation des items n'était pas le même. Cette situation ne s'est pas reproduite par la suite. Il est difficile d'évaluer l'impact de cette situation sur les réponses aux items du questionnaire.
Pour les rencontres avec les quatrièmes années, nous n'avons pas été informé d'événements particuliers. Nous savons toutefois que certains ont complété le questionnaire le lendemain des rencontres et les ont remis par la suite directement à la coordonnatrice. Ceci laisse croire que certains n'ont peut-être pas remis le questionnaire parce qu'ils ne sont pas retournés à l'université ou parce qu'ils ont tout simplement oublié le lendemain de la rencontre.

Après avoir recueilli tous les questionnaires, ces derniers ont été envoyés à une firme qui se spécialise dans l'entrée de données. La partie quantitative est revenue codée en caractères ASCII sous un format compatible avec la plupart des logiciels informatiques disponibles sur le marché.

4.4 CHOIX MÉTHODOLOGIQUE POUR LA RECHERCHE

Le premier critère à prendre en considération dans les choix méthodologiques est la faisabilité Les calculs nécessaires pour faire l'estimation des paramètres, pour ajuster le modèle, pour détecter les biais, etc. sont très complexes et demanderaient une expertise non acquise pour être programmés et comptabilisés. C'est pourquoi l'utilisation de programmes informatiques existants est souhaitable. Il faut donc comprendre que les choix faits en la matière dépendent de la disponibilité et de l'accessibilité des programmes informatiques. De même, les résultats dépendent aussi des méthodes de calcul utilisées par ces logiciels.

4.4.1 Choix du modèle

La première étape de la méthodologie consiste à choisir un modèle. Avec tous les modèles disponibles, il faut se demander lequel convient le mieux. Le choix du modèle est d'ailleurs un enjeu d'importance puisqu'il peut influencer l'ajustement des données. D'ailleurs, comme Embreston et Reise (2000; p. 75) l'indiquent : *« En résumé, la valeur relative de chacun des critères détermine quel modèle s'applique le mieux à une situation particulière. Aucun des critères n'est suffisant en soi. »* Ils ajoutent plus loin (p. 246) que le choix du modèle est généralement assez évident en fonction des caractéristiques des données recueillies.

Le choix du modèle dépend ainsi du type d'instrument utilisé, du type de données recueillies (c'est-à-dire de l'échelle de référence) et des objectifs de la recherche. Une des visées de la recherche est l'utilisation d'une échelle qui réponde aux caractéristiques de la mesure implicite. Dès lors, seulement les modèles de la famille de Rasch possèdent cette propriété (voir section 2.2). Ces modèles imposent évidemment certaines restrictions comme l'utilisation d'un seul paramètre pour décrire les caractéristiques des items. Quoi faire alors de la discrimination et de la chance ? Comme le paramètre thêta de l'instrument est l'attitude des étudiants en regard de leur programme d'étude, il est raisonnable de croire que le facteur chance n'a pas d'impact sur la réponse des sujets. Quant à la discrimination des items, nous supposons que les items discriment tous également parce que nous avons choisi de privilégier l'angle de la mesure implicite dans l'exploration du fonctionnement de la TRI pour notre recherche. Ensuite, le nombre de catégories de l'échelle de type Likert reste le même pour les 20 items. Finalement, il est présupposé que les items au questionnaire font appel à une seule dimension. Parmi les modèles unidimensionnels de la TRI, le *Rating Scale Model* de Andrich (1978a et 1978b) répond à tous les critères décrits dans ce paragraphe. Il faut souligner que le *Partial Credit Model* de Masters (1982) pourrait aussi être utilisé si le nombre de catégories de l'échelle de Likert variait d'un item à l'autre ou si l'importance des catégories n'était pas la même d'un item à l'autre.

4.4.2 Choix des logiciels informatiques

Le choix du RSM a un impact immédiat sur le choix du logiciel informatique parce qu'il faut que le programme estime les paramètres à partir d'une méthode éprouvée théoriquement pour le RSM; qu'il offre la possibilité de vérifier l'ajustement des données et qu'il offre une méthode de calcul pour le FDI. La section 2.3 suggère que, théoriquement, la méthode CMLE serait la

méthode préférable pour estimer les paramètres des modèles de Rasch. D'un autre côté, la MMLE est aussi une méthode possédant de bonne qualité et la méthode d'estimation par paire comparée serait aussi une méthode efficace. De plus, la plupart des logiciels qui permettent d'estimer les paramètres des modèles de la famille de Rasch donneraient des valeurs semblables pour les estimations (section 2.5). Donc, n'importe laquelle de ces méthodes d'estimation est appropriée. Les trois logiciels suivants utilisent ces méthodes pour le RSM : LPCM-WIN, CONQUEST et RUMM2010. Dans un deuxième temps, tous ces logiciels offrent aussi la possibilité de vérifier l'ajustement des items et des sujets. Dans ces circonstances, n'importe lequel de ces programmes fait l'affaire.

Comme nous possédons déjà le logiciel CONQUEST, il est choisi pour cette recherche. Bien que ce logiciel permette de détecter le fonctionnement différentiel des items, il n'utilise pas une des méthodes recensées à la section 3.4. Le logiciel CONQUEST utilise sa méthode spécifique. La méthode ressemble à la méthode par ratio de vraisemblance, c'est-à-dire que le logiciel CONQUEST compare la statistique d'ajustement d'un item pour deux versions différentes du modèle (groupe R et F) à partir d'une procédure inspirée de la modélisation à facettes multiples *(multifaceted)*. Il comptabilise ensuite une statistique de déviation entre les deux modèles. Le manuel du logiciel explique avec un peu plus de détails cette procédure. À ce moment-là, il serait préférable d'avoir un logiciel qui utilise une des méthodes suggérées en 3.4 comme la méthode de comparaison par ratio de vraisemblance, la méthode STDN ou la méthode SIBTEST. Ainsi, le logiciel POLYSIBTEST qui utilise la méthode SIBTEST est choisi pour faire la détection du FDI.

4.5 DÉMARCHE DE L'ANALYSE

La démarche de l'analyse est conforme avec ce qui est décrit aux sections 3.2 et 3.3. En ordre, les étapes de la démarche de l'analyse consiste à vérifier l'ajustement des données au modèle et l'effet du FDI sur celle-ci, à étudier la pertinence du nombre de catégories de l'échelle de référence, à détecter la présence de FDI et finalement, à étudier la propriété d'invariance et l'effet du FDI sur cette dernière.

4.5.1 L'ajustement des données aux modèles

Dans la mesure où les données recueillies pour un item ne s'ajustent pas aux données, il y a deux décisions possibles ou deux philosophies qui motivent la prise de décision. Soit le modèle est rejeté parce que les données sont considérées comme plus importantes et ne peuvent être rejetées, soit les items sont modifiés ou éliminés pour qu'il y ait un meilleur ajustement du modèle. Il existe un débat important entre les deux approches et le but de ce travail n'est pas de résoudre ce débat. Pour trancher, nous allons prendre en considération le présupposé théorique qui avance que, dans la mesure où le modèle s'ajuste bien aux données, les hypothèses de base de la TRI devraient être respectées. Conséquemment, dans l'éventualité où les données ne s'accordent pas avec le modèle, la deuxième approche sera préférée. Il est aussi intéressant de s'enquérir de l'effet du FDI sur l'ajustement. Pour faire l'analyse de l'ajustement, le logiciel CONQUEST utilise la statistique d'ajustement de Wu (1997). Celle-ci est une extension de la statistique de Wright et Masters (1982). Pour guider la prise de décision et pour interpréter les sorties du logiciel, nous utiliserons les suggestions proposées par Bond et Smith (2001).

4.5.2 Le nombre de catégorie de l'échelle de référence

L'objectif de cette vérification est de déterminer le nombre optimal de catégorie à inclure dans l'échelle de référence. Avant de procéder à la cueillette des données, il est difficile de savoir quel serait le nombre optimal de catégories à utiliser. Il faut donc s'en remettre au bon jugement du concepteur et à ses intentions. Dans ce cas ci, les concepteurs désiraient forcer les candidats à faire un choix entre deux pôles (accord et désaccord) et ils ont choisi une échelle en quatre points pour le faire. La question est maintenant de savoir si une façon plus appropriée de combiner les catégories pour représenter les données existerait et, conséquemment, pour améliorer l'ajustement des catégories au modèle. Cette vérification se fait à partir des données recueillies. C'est encore une fois à partir de la discussion de Bond et Fox (2001) que l'étude du nombre de catégorie de l'échelle se fait.

4.5.3 Le Fonctionnement Différentiel des Items

La détection du FDI est une étape qui se fait assez directement à l'aide du logiciel POLYSIBTEST. Chaque item est étudié pour s'assurer de détecter tous les items pour lesquels il y a un FDI. Il n'est pas important pour cette recherche de savoir si le FDI est causé par la

modification de l'ordre des items ou non. L'ordre des items a été modifié uniquement pour améliorer nos chances de voir apparaître un FDI dans les réponses des candidats. L'important à cette étape est seulement de détecter des items qui permettent ensuite d'évaluer l'impact de la présence de FDI sur la modélisation avec la TRI.

4.5.4 La propriété d'invariance

Après l'analyse faite dans les trois premières étapes décrites plus haut, il reste encore à étudier la propriété d'invariance des paramètres des items. Les méthodes utilisées pour vérifier la propriété d'invariance des données demandent davantage de recherche. L'utilisation d'un coefficient de corrélation n'est pas toujours appropriée (Zumbo, 2000). L'alinéa 2.1.3 explique ce qui rend difficile cette analyse. C'est pour cette raison que nous nous limiterons à vérifier si l'ordre de la position des items sur le continuum change d'un groupe à l'autre. À partir de cette information, nous discuterons la propriété d'invariance et l'effet de la présence de FDI sur celle-ci.

4.5.5 Schéma d'analyse de la recherche

L'alinéa 4.3.1 décrit l'échantillon disponible pour cette recherche. Mais, avant de procéder à l'analyse, il faut comprendre que les étudiants de chaque programme se distinguent qualitativement. Ils ne suivent pas les mêmes cours, ils n'ont pas les mêmes enseignants, ils ne sont pas tous à la même année du programme, etc. Malgré cela, les items du questionnaire ont été composés en fonction de ces différences et la plupart de ces items constituent une généralisation de l'expérience vécue et réfèrent à l'attitude de l'étudiant par rapport à son programme. Il n'en reste pas moins que les étudiants ne vivent pas exactement la même chose. C'est dans cette mesure qu'il est possible qu'un biais de mesure existe entre les différents programmes. Ce biais pourrait possiblement affecter la modélisation des données. Toutefois, selon la logique du questionnaire, l'attitude des étudiants devrait transgresser les différences attribuables au programme d'appartenance. C'est-à-dire que l'instrument du CFIM a été conceptuellement construit en considération de ces différences et les items cherchent à refléter une expérience qui serait indépendante du programme de l'élève.

En somme, c'est l'objectif même du CFIM de construire un seul instrument pour tous ses programmes de formation et c'est dans cette perspective que la TRI et les objectifs de la recherche s'insèrent. Plus concrètement, la recherche se demande quels avantages offrent la TRI

dans ce contexte. Est-ce que l'appartenance d'un étudiant à un groupe influence la modélisation? Est-il possible de construire un seul instrument de recueil d'information pour tous les différents groupes ? Peut-il y avoir invariance d'un groupe à l'autre ? Comment est-ce que la TRI peut être utilisée dans ces circonstances ? Quels sont les avantages et les désavantages à l'utiliser ?

Dans le but d'éviter une présentation exhaustive des résultats et dans le but unique et arbitraire d'avoir un échantillon le plus adéquat possible, les programmes pour lesquels il y a eu le plus grand nombre de répondants sont retenus. Pour l'année 2000, ces deux programmes sont enseignement secondaire et éducation préscolaire et enseignement primaire. Pour 2001, ce sont les deux mêmes programmes, nous avons décidé d'utiliser une méthode de comparaison des données qui donne une unité de base pour la comparaison entre les données provenant des différents programmes. Comme unité de base, nous avons choisi les données recueillies auprès des étudiants du programme d'enseignement secondaire. La méthode de comparaison consiste à comparer ces données provenant du programme d'enseignement secondaire pour les années de distribution 2000 et 2001, pour la troisième et la quatrième du programme, et pour les versions 1 et 2 du questionnaire. Ces données sont aussi comparées à celles obtenues pour le programme d'éducation préscolaire et enseignement primaire. Ce schéma permet de couvrir plusieurs comparaisons particulièrement intéressantes au moment d'étudier la propriété d'invariance. Tout au long de l'analyse, l'appellation groupe de données (ou échantillon) réfère aux données recueillies auprès des étudiants d'un programme.

4.5.6 Limites de la recherche

Cette recherche possède certaines limites évidentes. La première limite découle de l'utilisation des logiciels informatiques. C'est donc dire que nos résultats dépendent des procédures utilisées par ces logiciels et, de plus, ils dépendent de la précision et de la justesse de la programmation du logiciel. La deuxième limite de cette recherche est la taille de notre échantillon. La troisième limite est le nombre d'items. Pour certaines des analyses, il se pourrait que la taille de l'échantillon ou le nombre d'items ne soient pas assez grands pour obtenir des estimations de paramètre qui soient consistantes. Quatrièmement, le schéma d'analyse se limite à explorer quelques liens parmi plusieurs liens possibles. Cinquièmement, il n'est pas certain que la présentation de l'ordre des items induira un FDI et conséquemment qu'il sera possible de répondre à la troisième question de recherche. Sixièmement, l'analyse faite du nombre de catégories de l'échelle ne permet pas de savoir si un plus grand nombre de catégories aurait été

préférable. Finalement, les résultats seront liés de près au modèle choisi et aux décisions prises quant à l'ajustement des données aux modèles. C'est pour ces raisons que les résultats seront difficilement généralisables suite à cette recherche qui demeure, nous le rappelons, exploratoire.

CHAPITRE V

ANALYSE DES DONNÉES

Les deux objectifs principaux de l'analyse des données sont premièrement d'explorer les caractéristiques de l'instrument à partir des données recueillies à l'aide de celui-ci et, deuxièmement, d'explorer la présence de fonctionnement différentiel des items dans ces données. Les résultats de la recherche sont présentés en regard de la démarche de l'analyse de la section 4.5. La première partie du chapitre permet de voir quels sont les items rejetés par le modèle et quels sont les items qui s'ajustent adéquatement au modèle. La deuxième partie étudie la pertinence du nombre de catégories à l'échelle de Likert. La troisième partie explore la présence de FDI dans les données et la quatrième partie étudie le respect de la propriété d'invariance.

5.1 RETOUR SUR LE SCHÉMA D'ANALYSE

Comme mentionné à l'alinéa 4.5.4, quatre variables qualitatives peuvent permettre de distinguer les données recueillies. Elles sont l'année de cueillette des données, la version du questionnaire, le programme et l'année scolaire des étudiants inscrits à l'un de ces programmes. Chacune de ces variables peut potentiellement influencer l'ajustement des données au modèle. Pour tenir compte de ces variables qualitatives, les résultats sont présentés tout en respectant le schéma expérimental de l'alinéa 4.5.5, c'est-à-dire que ce ne sont pas toutes les possibilités qui sont explorées.

5.2 L'AJUSTEMENT DES DONNÉES AU MODÈLE

La première étape de ce chapitre consiste à vérifier l'ajustement des données au modèle choisi, le *Rating Scale Model*. Pour ce faire, le logiciel CONQUEST utilise une statistique d'ajustement qui s'inspire de la statistique de Wright et Masters (1982). CONQUEST fournit quatre statistiques pour étudier l'ajustement : la moyenne des carrés et T (valeur standardisé) autant pour l'*infit* que pour l'*outfit*[x]. Bond et Fox (2001) font quelques propositions dans l'utilisation de ces statistiques produites par CONQUEST. Les propositions de Bond et Fox (2001) suggèrent que la moyenne des carrés pour l'*infit* devrait préférablement se situer entre 0,75 et 1,3 et que la valeur standardisée T pour *l'outfit* devrait préférablement se situer entre –2,0 et 2,0. Dans le cas d'un instrument qui utilise une échelle de référence comme celle de Likert, la moyenne des carrés peut varier entre 0,6 et 1,4 pour l'*infit*. Dans la mesure où un item ne se retrouve pas dans cet intervalle, il devrait être éliminé pour cause de manque d'ajustement au modèle. Toutefois, Smith et Schumacker (1998) soulignent que la moyenne des carrés peut être affectée par la grandeur de l'échantillon (plus petit que 500) et, à ce moment, une valeur critique de 1,3 (ou 1,4 pour le RSM) peut s'avérer beaucoup trop grande. La valeur standardisée T serait toutefois moins affectée et donc plus rigoureuse comme point de référence. L'ajustement des items est analysé à partir de ces propositions.

5.2.1 Exemple d'application : programme secondaire, année 2000

Le tableau 5.1 ci dessous est un exemple de sorties produites par CONQUEST pour l'échantillon du secondaire en 2000. La première colonne donne la valeur de l'estimation du paramètre δ. Ce paramètre donne l'emplacement de l'item sur le continuum, c'est-à-dire l'endroit sur la courbe où les CCI pour chaque catégorie se situent[xi]. Pour les catégories d'intersection, la première colonne correspond à la valeur de l'estimation de τ, c'est-à-dire l'endroit sur la courbe où il devient plus probable que la réponse du sujet franchisse une autre catégorie. Les autres colonnes donnent l'erreur de mesure et les statistiques d'ajustement.

x Pour un retour sur la définition de infit et outfit, le lecteur peut aller à l'alinéa 2.4.3

xi Revoir les alinéas 2.2.1 et 2.2.2 pour une explication complète de ce paramètre de location de l'item sur la courbe.

Tableau 5.1 Sorties du programme CONQUEST pour le programme secondaire année 2000

Item	δ	Erreur de mesure	*Infit*		*Outfit*	
			Somme des carrés	*T*	Somme des carrés	*T*
1	0.688	0.104	0.90	-1.0	0.89	-1.2
2	1.943	0.115	0.87	-1.4	0.81	-2.1
3	-0.136	0.105	1.08	0.8	1.07	0.7
4	0.257	0.104	1.45	4.0	1.45	4.2
5	0.688	0.104	1.40	3.6	1.40	3.9
6	0.753	0.104	1.33	3.0	1.33	3.4
7	0.361	0.104	0.99	-0.1	0.99	-0.1
8	-1.506	0.115	0.96	-0.4	0.97	-0.3
9	-0.762	0.108	0.84	-1.6	0.84	-1.7
10	-1.025	0.110	1.08	0.8	1.08	0.8
11	-0.473	0.107	0.69	-3.4	0.69	-3.6
12	-1.148	0.111	1.02	0.2	1.03	0.3
13	-0.048	0.105	0.63	-4.3	0.62	-4.5
14	0.148	0.104	1.15	1.5	1.14	1.5
15	-0.416	0.106	1.15	1.4	1.13	1.3
16	0.118	0.105	0.77	-2.4	0.77	-2.6
17	0.022	0.105	1.02	0.2	1.01	0.2
18	0.148	0.104	1.02	0.3	1.02	0.2
19	0.526	0.104	0.94	-0.6	0.94	-0.6
20	1.813	0.113	0.80	-2.1	0.79	-2.5
Catégorie	τ					
Intersection 1	-2.159	0.040	3.53	15.6	3.52	15.6
Intersection 2	0.190	0.035	2.18	8.9	2.26	9.5

Un premier aperçu des statistiques d'ajustement *infit* montre que les items 4, 5, 6, 11, 13, 16 et 20 ont des valeurs *outfit* inférieures à –2,0 ou supérieures à 2,0. Les valeurs *infit* pour les items 5, 6, 16 et 20 se situent dans la limite acceptable proposée par Fox et Bond. L'item 2 a une valeur hors de la limite (-2,1) pour le *outfit* mais acceptable pour le *infit*. Puisque notre échantillon est plus petit que 500, les valeurs *outfit* sont préférables comme critère de décision. Il faudrait alors éliminer les items 4, 5, 6, 11, 13 et 20 mais garder l'item 2 parce que son *infit T* est acceptable et que son *outfit T* est très près de la limite. En éliminant ces items, les valeurs pour l'estimation et les statistiques d'ajustement pourraient être quelque peu modifiées. C'est pourquoi il est préférable de reprendre la vérification de l'ajustement avec seulement les items restants.

Tableau 5.2 Sorties du programme CONQUEST pour le programme secondaire année 2000 sans les items 4, 5, 6, 11, 13, 16 et 20

			Infit		*Outfit*	
Item	δ	Erreur	Somme des carrés	*T*	Somme des carrés	*T*
1	0.695	0.106	1.00	0.1	1.00	0.0
2	1.991	0.117	0.96	-0.3	0.93	-0.7
3	-0.151	0.106	1.05	0.5	1.04	0.4
4						
5						
6						
7	0.361	0.106	0.99	-0.0	1.00	-0.0
8	-1.540	0.115	0.90	-1.0	0.91	-0.9
9	-0.788	0.109	0.87	-1.3	0.87	-1.4
10	-1.055	0.111	1.04	0.5	1.04	0.5
11						
12	-1.179	0.112	0.97	-0.3	0.98	-0.2
13						
14	0.140	0.106	1.24	2.2	1.23	2.3
15	-0.437	0.107	1.09	0.9	1.08	0.8
16						
17	0.011	0.106	1.01	0.1	1.00	0.1
18	0.140	0.106	1.05	0.6	1.04	0.5
19	0.529	0.105	0.99	-0.0	0.99	-0.0
20						
Catégorie	τ					
Intersection 1	-2.194	0.050	2.50	10.7	2.51	10.8
Intersection 2	0.143	0.044	1.57	4.9	1.65	5.5

Comme prévu, la valeur de l'estimation de δ a légèrement changé par rapport au tableau précédent. Ces petits changements dans les valeurs calculées par le logiciel font en sorte que l'item 14 ne s'ajuste plus au modèle avec des valeurs *T* de 2,2 et 2,3. Il faut donc reprendre une autre fois la procédure pour s'assurer que le reste des items ne sera pas influencé par l'absence de l'item 14. Le résultat est présenté au tableau 5.3. Encore une fois, les valeurs estimées par le logiciel ont été modifiées par ce changement (l'absence de l'item 14). Maintenant, tous les items s'ajustent au modèle. Bien que les données pour ces douze items s'ajustent maintenant au modèle, il faut aussi vérifier l'ajustement des paramètres des catégories d'intersection. Tel que les chiffres le suggèrent, les intersections s'ajustent très mal aux données avec des statistiques *T* variant de 5,1 à 10,1. La structure de l'échelle de référence est peut-être à revoir. La question des catégories de l'échelle de référence est traitée plus en longueur à la section 5.3.

Tableau 5.3 Sorties du programme CONQUEST pour le programme secondaire année 2000 sans les items 4, 5, 6, 11, 13, 14, 16 et 20

			Infit		*Outfit*	
Item	δ	Erreur	Somme des carrés	*T*	Somme des carrés	*T*
1	0.714	0.106	0.95	-0.5	0.94	-0.7
2	2.021	0.117	0.89	-1.1	0.87	-1.5
3	-0.147	0.107	1.05	0.5	1.04	0.4
4						
5						
6						
7	0.374	0.106	1.01	0.1	1.00	0.1
8	-1.565	0.116	0.97	-0.3	0.97	-0.2
9	-0.797	0.110	0.87	-1.3	0.87	-1.3
10	-1.070	0.112	1.06	0.6	1.05	0.5
11						
12	-1.197	0.113	1.01	0.2	1.02	0.3
13						
14						
15	-0.438	0.109	1.11	1.1	1.10	1.0
16						
17	0.018	0.107	1.07	0.7	1.06	0.7
18	0.150	0.106	1.03	0.4	1.03	0.3
19	0.545	0.106	1.09	0.9	1.09	1.0
20						
Catégorie	τ					
Intersection 1	-2.252	0.052	2.39	10.1	2.36	9.9
Intersection 2	0.163	0.046	1.59	5.1	1.66	5.6

Les figures 5.1 et 5.2 montrent la CCI de chacune des quatre catégories de l'échelle de Likert pour les items 1 et 8 respectivement. Ces figures permettent d'interpréter visuellement les résultats du tableau 5.3. Sur les figures, le paramètre δ situe les CCI sur le continuum thêta. Plus δ est petit, plus les CCI se situent à gauche sur le continuum et plus il est facile de franchir les catégories comme dans le cas de l'item 1 à la figure 5.1, comparativement à l'item 8 à la figure 5.2. Ensuite, le paramètre τ indique où les CCI se croisent, c'est-à-dire leurs intersections. Pour quatre catégories, les courbes se croisent trois fois, il a donc trois valeurs de τ. Pour obtenir l'endroit exact où se croisent les CCI sur le continuum, il faut additionner τ à δ (la position relative d'une courbe par rapport à l'autre et la position de toutes les courbes par rapport au continuum). Le tableau 5.3 donne les résultats estimés par CONQUEST[xii]. Pour l'item 1, les

xii L'addition des trois τ doit égaler 0 par définition avec le RSM, c'est pourquoi CONQUEST ne donne pas de valeur estimée pour le paramètre d'intersection de la troisième catégorie. Il suffit de faire le calcul pour arriver à 0.

trois τ sont : -2.252, 0,163 et 2,089. Ainsi, en les additionnant à δ (0,714), la figure 5.1 montre que pour l'item 1 les intersections sont à –1,538 (-2,252 + 0,714); 0,877 (0,163 + 0,714); et à 2,803 (2,089 + 0,714). Dans le cas du *Rating Scale Model*, les CCI se situeront toujours au même endroit l'une par rapport à l'autre, seulement δ peut alors varier. En somme, l'item 8 à la figure 5.2 nécessite un plus grand thêta pour franchir chacune des catégories de l'échelle de référence parce que la valeur estimée de δ pour cet item est plus grande que pour l'item 1. Ce que l'on peut voir en comparant les deux figures.

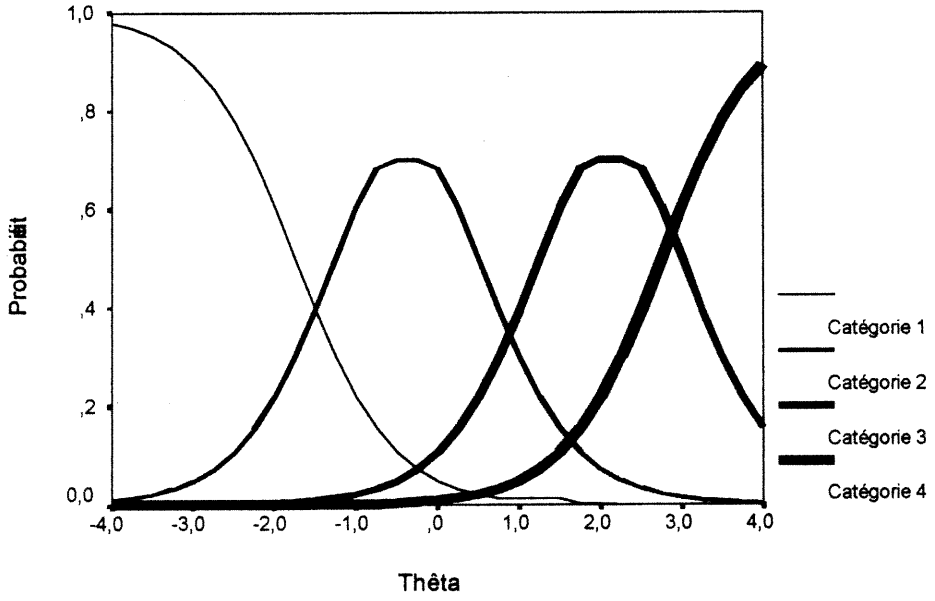

Figure 5.1 Courbes de réponses des catégories pour l'item 1 du programme secondaire en 2000

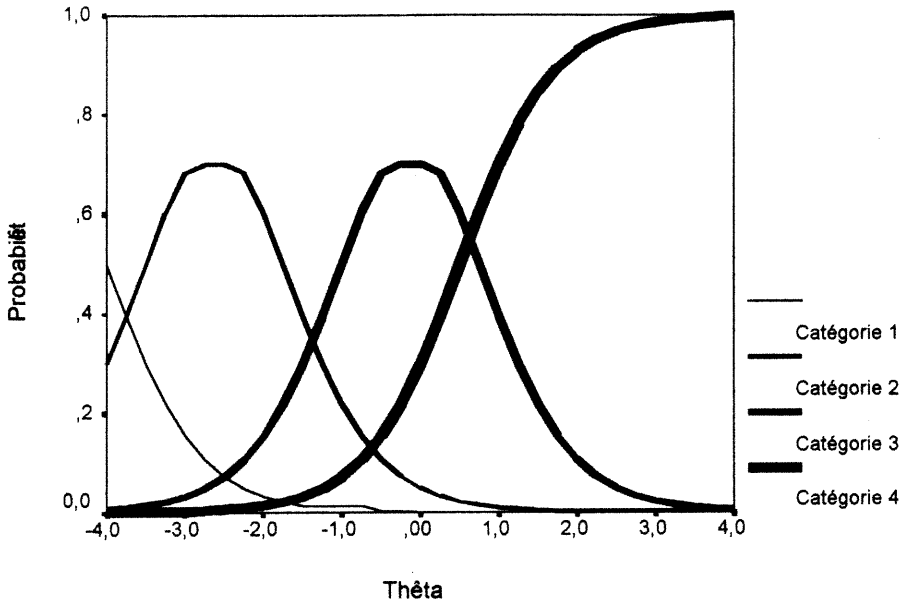

Figure 5.2 Courbes de réponses des catégories pour l'item 8 du programme secondaire en 2000

Cet exemple a permis de présenter comment interpréter l'analyse de l'ajustement des données au modèle pour ce chapitre. Maintenant que la procédure est précisée, elle peut être reprise selon le programme, l'année de distribution, l'année scolaire et la version du questionnaire. À noter qu'il n'est pas question dans cette première étape de savoir si la valeur de l'estimation du paramètre de la difficulté est différente pour un item en fonction des variables qualitatives, mais seulement de savoir pour quels items les données s'ajustent adéquatement au modèle et de savoir si ces items diffèrent d'un groupe d'étudiants à l'autre.

5.2.2 Comparaison de l'ajustement selon l'année de cueillette des données

Pour augmenter le niveau de comparabilité, l'année de cueillette des données est comparée à partir d'un même groupe de données (enseignement secondaire) pour une même version (version 1) du questionnaire pour chacune des deux années scolaires du programme.

Tableau 5.4 Valeur estimée de δ pour les items qui s'ajustent au modèle entre 2000 et 2001

	Secondaire Troisième		*Secondaire Quatrième*	
Item	2000	2001	2000	2001
1	0.562	1.262	0.919	1.431
2		2.556	2.303	2.767
3		0.113	-0.406	0.365
4		0.885		0.867
5		0.391		
6		0.437		
7	0.450	0.667	0.277	1.008
8	-1.720	-1.690	-1.540	-1.185
9	-0.852		-0.842	0.068
10	-1.454	-1.581	-0.790	-0.971
11		-0.408		0.217
12	-1.559	-1.690	-0.946	
13		0.113		
14	0.052	0.529	0.226	
15	-0.333	0.067	-0.635	0.401
16	0.193	0.529	-0.001	1.008
17	-0.279	-0.360	0.302	0.328
18	-0.264	0.529	0.582	1.503
19	0.452	0.987	0.659	0.973
20	1.786	2.613	2.187	
Catégorie Intersection 1				0.474

Le tableau 5.4 donne la valeur estimée pour le paramètre δ des items pour lesquels les données s'ajustent au modèle. Ce tableau montre que les données qui s'ajustent au modèle sont différentes d'une année de cueillette à l'autre. Il y a tout de même dix items pour lesquels les données s'ajustent toujours d'une année de cueillette à l'autre pour les deux années scolaires. Pour les autres items, l'ajustement des données varie. C'est pour la troisième année du secondaire en 2001 que l'ajustement des données des items est le meilleur avec 19 items sur 20, donc 95 %. D'un autre côté, pour la troisième secondaire en 2000, les données de seulement 65 % des items s'ajustent. Pour le groupe de données de la troisième année du secondaire en 2000 et celui de la quatrième année en 2001, ce sont les mêmes étudiants qui ont répondu (parce qu'ils sont passés de la troisième à la quatrième). Toutefois, les données s'ajustent différemment pour sept items du questionnaire. Néanmoins, dans l'ensemble, pour un certain nombre d'items, les données s'ajustent et, pour le paramètre catégorie d'intersection, les données s'ajustent rarement.

Il faut noter que certaines valeurs des estimations sont très différentes d'une année à l'autre ou d'une colonne du tableau à l'autre. Par exemple, pour l'item 1, la valeur de l'estimation du paramètre δ varie de 0,562 à 1,431 entre le groupe de troisième année et celui de quatrième. Pour l'item 18, la valeur estimée passe de –0,264 à 1,503. Ceci suggère qu'en plus de ne pas être les données des mêmes items qui s'ajustent, la valeur de l'estimation de δ pour ces items ne se situe pas au même endroit sur le continuum thêta d'une année de cueillette à l'autre. Comme cette section a pour but unique de vérifier l'ajustement, ces remarques sont mises de côté pour l'instant. Cette question de la variance dans l'estimation sera abordée à la section 5.6.

5.2.3 Comparaison de l'ajustement selon la version du questionnaire en 2001

À l'alinéa précédent, l'année de cueillette des données est isolée pour les besoins de la comparaison. À cet alinéa, c'est plutôt l'ajustement des différentes versions du questionnaire qui est vérifié. Pour isoler cette variable, l'échantillon du programme de l'enseignement secondaire avec ces deux années scolaires (troisième et quatrième) demeure l'unité de comparaison. Comme les versions varient seulement en 2001, l'ajustement est vérifié pour 2001. Le tableau 5.5 permet de constater que l'ensemble des données s'ajuste très bien au modèle pour l'échantillon de la troisième année du secondaire pour les deux versions du questionnaire. Seules les données de l'item 9 pour la version 1 ne se sont pas ajustées au modèle. Toutefois, pour l'échantillon de quatrième secondaire, l'ajustement des données au modèle est beaucoup plus variable entre les deux versions. Le tableau montre une différence dans l'ajustement des données pour les items 4 à 6, 11 à 14 et pour l'item 20 d'une version à l'autre. Il y a, au total, six items pour lesquels les données s'ajustent différemment au modèle selon la version du questionnaire pour l'échantillon de quatrième année. Le groupe pour lequel les données s'ajustent le moins a tout de même 70 % des items avec un ajustement favorable des données, ce qui est appréciable Pour le paramètre catégorie d'intersection, encore très peu des données s'ajustent.

Tableau 5.5 Valeur estimée de δ pour les items qui s'ajustent au modèle pour les deux versions du questionnaire de 2001 pour les deux années du secondaire

	Secondaire Troisième		*Secondaire Quatrième*	
	δ		δ	
Item	Version 1	Version 2	Version 1	Version 2
1	1.262	0.394	1.431	0.556
2	2.556	1.716	2.767	2.462
3	0.113	0.188	0.365	0.092
4	0.885	1.078	0.867	
5	0.391	0.188		
6	0.437	0.257		
7	0.667	0.667	1.008	-0.054
8	-1.690	-1.417	-1.185	-1.153
9		-0.586	0.068	-0.233
10	-1.581	-1.182	-0.971	-0.851
11	-0.408	-0.229	0.217	
12	-1.690	-1.338		-1.077
13	0.113	0.188		-0.233
14	0.529	0.050		0.483
15	0.067	0.119	0.401	-0.017
16	0.529	-0.089	1.008	0.268
17	-0.360	-0.513	0.328	-0.377
18	0.529	-0.229	1.503	-0.167
19	0.987	0.394	0.973	1.032
20	2.613	1.643		2.606
Catégorie Intersection 2		0.186	0.474	

5.2.4 Comparaison de l'ajustement pour deux programmes différents

Jusqu'ici, les tableaux précédents montrent que l'ajustement des données au modèle varie non seulement avec l'année de cueillette des informations mais aussi avec la version du questionnaire. Les tableaux étudiés jusqu'ici laissent aussi croire que les données recueillies pour l'année scolaire influencent l'ajustement de celles-ci au modèle. Il serait maintenant intéressant de voir si l'ajustement est différent pour l'échantillon de deux programmes différents. Pour augmenter la comparabilité, nous avons choisi d'utiliser la version 1 du questionnaire de 2001 comme unité de base.

Tableau 5.6 Valeur estimée de δ pour les items qui s'ajustent au modèle pour les deux années scolaires de la version 1 du questionnaire de 2001

	Secondaire		*Primaire*	
	δ		δ	
Item	Troisième	Quatrième	Troisième	Quatrième
1	1.262	1.431	0.363	0.585
2	2.556	2.767	1.013	
3	0.113	0.365	-0.212	0.846
4	0.885	0.867	-0.367	-0.020
5	0.391		-0.881	-0.491
6	0.437		-0.675	-0.901
7	0.667	1.008	0.631	
8	-1.690	-1.185	-2.360	-2.373
9		0.068		
10	-1.581	-0.971	-1.892	-1.418
11	-0.408	0.217	-1.173	-0.531
12	-1.690		-2.179	-1.027
13	0.113		-0.881	-0.136
14	0.529		0.631	0.771
15	0.067	0.401	-0.367	0.285
16	0.529	1.008	-0.057	
17	-0.360	0.328	-0.675	-0.058
18	0.529	1.503	-0.624	0.808
19	0.987	0.973	0.299	1.486
20	2.613		0.739	1.799
Catégorie Intersection 2		0.474		

Le tableau 5.6 indique que, dans l'ensemble, les données des échantillons de troisième et de quatrième année s'ajustent de façon assez semblable pour les programmes primaire et secondaire. Toutefois, entre les groupes d'étudiants de quatrième année des deux programmes, les données de plus de dix items ne s'ajustent pas de la même façon entre le primaire et le secondaire. Il est aussi possible de remarquer que le nombre d'items, pour lesquels les données des groupes d'étudiants de quatrième et de troisième années du programme primaire s'ajustent, est semblable. Encore une fois, pour au moins 70 % des items, les données s'ajustent au modèle et les données s'ajustent très mal au paramètre d'intersection des catégories. Essentiellement, il semble bien que le programme influence lui aussi le nombre d'items pour lesquels les données s'ajustent.

5.2.5 Comparaison de l'ajustement pour les variables année scolaire et version du questionnaire sans spécification du programme ou de l'année de cueillette

Les comparaisons faites jusqu'ici permettent de généraliser suffisamment pour dire que toutes les variables qualitatives peuvent jouer un rôle dans l'ajustement des données au modèle. Mais avant de passer à la section suivante, il serait intéressant d'explorer comment se comporte l'ajustement des données en ne tenant pas compte de l'année scolaire ni du programme, c'est-à-dire en évitant de procéder à une isolation délicate des variables qualitatives, en regroupant donc un plus grand nombre d'étudiants.

Tableau 5.7 Valeur estimée de δ pour les items qui s'ajustent au modèle pour toutes les données de 2000 et 2001

	Année		*Version du questionnaire 2001*	
Item	2000	2001	Version 1	Version2
1		0.644	0.864	0.280
2	1.831			1.381
3			0.300	
4			0.343	
5			0.014	-0.101
6			0.004	-0.078
7	0.201	0.506	0.666	0.246
8	-2.257		-1.712	-1.851
9	-1.120			
10	-1.329	-1.360	-1.306	-1.317
11				
12	-1.666	-1.291		-1.304
13				
14		0.479	0.641	0.337
15			0.119	
16				
17	-0.120	-0.348	-0.132	-0.612
18	-0.016	0.181	0.601	-0.320
19	0.667	0.827	0.890	0.670
20			1.858	1.478

Comme le tableau 5.7 le suggère, en regroupant plusieurs groupes d'étudiants, le nombre d'items pour lesquels les données s'ajustent diminue. En 2000, les items s'ajustent pour seulement 45 % des items comparativement à 40 % en 2001 et ce ne sont pas les mêmes items. En 2001, la situation est moins alarmante, avec 70 % pour la version 1 et 65 % pour la version 2. Ces nombres sont tout de même inférieurs à ceux obtenus dans les analyses précédentes. Donc, la variabilité dans le nombre d'items avec des données qui s'ajustent semble être plus grande si

plusieurs des variables qualitatives (programme, année scolaire, etc) sont considérées simultanément.

En résumé, l'ensemble des données s'ajuste de façon différente selon la version du questionnaire, le programme, l'année du programme et l'année de cueillette des questionnaires. Plusieurs raisons peuvent probablement expliquer l'écart entre ces ajustements, la première étant peut-être le nombre restreint de sujets pour faire l'estimation de δ. Une autre raison qui peut influencer l'ajustement des données au modèle serait un fonctionnement différentiel des items pour les sous-groupes. Toutefois, il faut rappeler que la théorie de la réponse à l'item devait permettre de remédier à ces problèmes de variabilité dans l'estimation. Il semble que, non seulement les items varient dans la valeur de leur estimation, mais ce ne sont pas les données des mêmes items qui s'ajustent au modèle d'un sous-groupe de données à l'autre. Finalement, les données pour le paramètre d'intersection des catégories s'ajustent très mal au modèle peu importe le groupe de données utilisées. Cette question est justement abordée à la section suivante.

5.3 MODIFICATION DU NOMBRE DE CATÉGORIES

La valeur estimée pour les statistiques d'ajustement pour le paramètre d'intersection des catégories varient autour de 8-9 pour la première intersection et autour de 5-6 pour la deuxième. À quelques reprises seulement les données se sont ajustées adéquatement pour ce paramètre. Il faudrait donc trouver une façon de réaménager les catégories pour obtenir un meilleur ajustement. La solution proposée par Bond et Fox (2001) s'appuie sur les suggestions de Linacre (1999). Pour améliorer la qualité des catégories de l'échelle, Bond et Fox (2001) suggèrent de joindre deux catégories pour en former une qui possède de meilleures propriétés. Dans une analyse post facto avec seulement quatre catégories, il y a peu d'option logique qui s'offre. Parallèlement à ces suggestions, il faut s'assurer d'apporter des modifications qui ont du sens en regard des objectifs de la cueillette de données.

5.3.1 Aperçu de la distribution des données dans les différentes catégories

La distribution des données dans les catégories est un point de départ pour vérifier les changements qui pourraient être apportés à l'échelle de référence. Elle permet aussi de vérifier une condition de Linacre (1999) qui propose de maintenir un minimum de dix observations par

catégorie. Pour l'ensemble de nos groupes de données, la distribution pour un item ressemble à un des quatre exemples du tableau 5.8. Les distributions présentées sont celles des données pour l'échantillon du programme d'enseignement secondaire, année 2000, et elles sont représentatives de tous les autres programmes ou sous-groupes de données.

Tableau 5.8 Exemples de distribution des données selon les catégories

Catégorie	*Nombre de réponses dans la catégorie*			
	Exemple 1	Exemple 2	Exemple 3	Exemple 4
1	18	0	2	16
2	76	13	70	122
3	85	105	95	53
4	20	79	32	7
Pourcentage de rejet par le modèle	55 %	33 %	40 %	40 %

Dans le tableau 5.8, peu importe l'exemple ou la forme de la distribution, souvent les deux catégories du milieu obtiennent souvent le plus grand nombre de réponses. De plus, lorsque le nombre de réponses dans la catégorie 3 est plus grand que dans la catégorie 2, le nombre de réponses dans la catégorie 4 est toujours plus grand que dans la catégorie 1, et vice versa. Pour le reste, les variations sont attribuables au degré d'accord donné à l'item. Pour les items plus neutres, comme à l'exemple 1, la distribution semble normale et les catégories extrêmes obtiennent un nombre appréciable de réponses. Toutefois à mesure qu'un item est davantage apprécié (bon degré d'accord), certaines catégories sont de moins en moins utilisées. Comme pour l'exemple 2, les catégories 1 ou 2 ne contiennent presque pas de réponses tandis que les deux autres se divisent assez bien la majorité des réponses. D'un autre côté, l'exemple 4 indique que, malgré des réponses favorables, on n'obtient pas beaucoup de réponses dans la première catégorie (1). Quant à l'exemple 3, il démontre que, quelquefois, le score extrême permettra de juger du degré d'accord à l'item. Pour certains exemples, seulement les deux catégories du centre sont nécessaires ou permettent de nuancer les données recueillies par le questionnaire Mais dans d'autres cas, les catégories extrêmes permettent de nuancer les données (exemples 2 et 3). Pour les quatre exemples, trois catégories extrêmes ne réussissent pas à obtenir le minimum de dix observations.

Selon Linacre (1999), l'idéal est de pouvoir diviser l'échelle en deux, et puis les catégories devraient encore pouvoir se rediviser en deux avec assez d'observations dans chaque subdivisions. C'est le cas de l'exemple 1. Mais pour les autres distributions, seulement un côté

pourrait être divisé une deuxième fois et même pour l'exemple 4, aucun des deux côtés ne pourrait être divisé une deuxième fois.

5.3.2 Facteurs à prendre en considération dans la création de nouvelles catégories

La partie précédente décrit les données de notre échantillon et devrait inspirer la modification des catégories de l'échelle de référence de départ. Selon les propositions de Linacre (1999) et Bond et Fox (2001), il faudrait avoir des catégories ayant au moins dix observations, distribuées assez normalement (c'est-à-dire pas de grande irrégularité), qui regroupent plus d'une catégorie et qui ont du sens. Cette dernière proposition est la plus contraignante. Pour que les nouvelles catégories aient du sens, il n'y a qu'une seule possibilité, celle de joindre les catégories extrêmes et de faire seulement deux catégories au total. La réunion des deux catégories du milieu n'a pas de sens parce que toutes les observations se retrouveraient dans cette catégorie et les concepteurs ne voulaient pas de catégorie neutre au départ. De plus, ces deux catégories sont généralement celles qui permettent de distinguer l'attitude des candidats. De plus, il serait peu pratique de réunir seulement un côté (c'est-à-dire faire 1134) parce que, pour certains items, il faudrait plutôt avoir 1244 selon le degré d'accord apporté à l'item. Une seule modification possible des données est alors sensée, celle de réunir les deux catégories de chaque extrême et d'obtenir des données dichotomiques.

Puisque la première version du questionnaire ne comprend que quatre catégories, le choix est restreint. Puisqu'il est impossible d'inventer de nouvelles catégories avec nos données, le concepteur ne peut jamais savoir si un plus grand nombre de catégories est mieux pour son instrument. Voilà pourquoi une seule autre combinaison de catégorie est possible. Il est important de noter que cette nouvelle forme de données implique un changement de modèle vers le modèle dichotomique de Rasch. Avec seulement deux catégories, il n'y a plus de paramètre d'intersection des catégories. Il sera intéressant de voir comment les données s'ajustent au nouveau modèle, c'est-à-dire à ces nouvelles catégories.

5.4 AJUSTEMENT DES DONNÉES AUX NOUVELLES CATÉGORIES

Encore une fois, ce n'est pas la différence qui existe dans la valeur de l'estimation des paramètres qui est étudiée à ce moment, mais bien la différence dans l'ajustement des données au modèle.

5.4.1 Comparaison de l'ajustement au modèle dichotomique selon l'année de distribution

Tableau 5.9 Valeur estimée de δ pour les items qui s'ajustent au modèle DICHOTOMIQUE pour le secondaire en 2000 et 2001

	Secondaire Troisième		*Secondaire Quatrième*	
	δ		δ	
Item	2000	2001	2000	2001
1	-0.464	1.772	1.268	0.524
2			2.706	3.459
3			-0.608	0.185
4		0.460	0.063	0.565
5	0.485	-0.110	0.682	0.565
6	0.433	0.191	0.847	0.799
7	0.268	0.600	0.214	0.799
8		-2.792	-2.511	-1.821
9	-1.311		-1.258	-0.424
10	-2.257	-2.584	-0.773	-1.703
11	-1.301		-0.773	-0.116
12		-2.225	-1.193	-0.916
13	-0.480	-0.211	0.063	0.412
14	-0.023	0.191	0.367	0.631
15		-0.311	-0.662	0.336
16	0.190	0.812	-0.190	0.352
17	-0.614	-1.046	0.316	0.035
18	-0.584	0.191	0.904	0.524
19	0.485	0.812	0.904	0.027
20	1.992			

À la lecture du tableau 5.9, il semble que le changement de catégorie ait permis à un plus grand nombre d'items de s'ajuster au modèle. La situation est particulièrement évidente pour les données de l'échantillon de la quatrième année scolaire pour le secondaire. Toutefois, pour le cas des données du groupe de troisième secondaire, il n'y a pas vraiment plus d'items pour lesquels les données s'ajustent au modèle. Pour l'instant, on ne peut pas conclure que les données s'ajustent mieux à ce nouveau modèle.

5.4.2 Comparaison de l'ajustement pour le secondaire selon la version du questionnaire en 2001 pour le modèle dichotomique

Tableau 5.10 Valeur estimée de δ pour les items qui s'ajustent au modèle pour les deux années scolaires de la version 1 du questionnaire de 2001

	Secondaire Troisième		*Secondaire Quatrième*	
	δ		δ	
Item	Version 1	Version 2	Version 1	Version 2
1	1.772	0.138	0.524	0.365
2		1.912	3.459	2.294
3		-0.014	0.185	0.067
4	0.460	1.074	0.565	0.129
5	-0.110	-0.323	0.565	0.686
6	0.191	-0.014	0.799	0.851
7	0.600	0.594	0.799	-0.187
8	-2.792	-2.792	-1.821	-1.441
9		-0.818	-0.424	-0.513
10	-2.584	-1.610	-1.703	-1.550
11		-0.483	-0.116	-0.430
12	-2.225	-2.429	-0.916	-1.136
13	-0.211	-0.323	0.412	
14	0.191	-0.014	0.631	0.686
15	-0.311	-0.014	0.336	-0.093
16	0.812	-0.323	0.352	0.129
17	-1.046	-1.187	0.035	-0.596
18	0.191	-0.818	0.524	-0.416
19	0.812	0.289	0.027	1.386
20		1.912		

En examinant le tableau 5.10, on observe que, pour une deuxième fois, il semble y avoir plus d'items pour lesquels les données s'ajustent au modèle dichotomique qu'au RSM. Dans ce cas-ci, les données de presque tous les items s'ajustent pour les différents groupe d'étudiants. Seul l'échantillon de la version 1 du programme de troisième secondaire a plus de deux items (5) pour lesquels les données ne s'ajustent. Bien que les données s'ajustent pour plusieurs items, il reste tout de même une certaine différence dans l'ajustement des données selon la version. Par exemple, pour le groupe d'étudiants du programme de troisième année, les données de cinq items s'ajustent différemment, ce qui est semblable au RSM. Mais, au total, avec les nouvelles catégories, il y a plus d'items pour lesquels les données s'ajustent avec le modèle dichotomique.

5.4.3 Comparaison de l'ajustement pour les deux années scolaires selon le programme

Tableau 5.11 Estimation de δ pour les items qui s'ajustent au modèle pour les deux années scolaires de la version 1 du questionnaire de 2001

	Secondaire		*Primaire*	
	δ		δ	
Item	Troisième	Quatrième	Troisième	Quatrième
1	1.772	0.524		0.574
2		3.459	1.837	
3		0.185	0.033	0.490
4	0.460	0.565	-0.274	-0.499
5	-0.110	0.565	-0.912	-0.939
6	0.191	0.799	-0.480	-1.430
7	0.600	0.799	0.774	
8	-2.792	-1.821	-1.673	
9		-0.424	-1.902	
10	-2.584	-1.703	-1.902	
11		-0.116		-0.939
12	-2.225	-0.916	-0.520	-1.430
13	-0.211	0.412	-0.108	-0.332
14	0.191	0.631	1.006	0.407
15	-0.311	0.336	-0.171	-0.166
16	0.812	0.352	0.238	-0.034
17	-1.046	0.035	-1.144	-0.332
18	0.191	0.524	-0.801	
19	0.812	0.027	0.506	1.722
20			1.389	1.272

D'après les données du tableau 5.11, il y a ici aussi des différences dans les données qui s'ajustent au modèle selon l'année scolaire ou le programme. Entre quatre et huit items de différences selon le cas. Dans l'ensemble, plus de données s'ajustent au modèle à l'exception du groupe d'étudiants de la quatrième année du primaire qui ressemble à l'ajustement moyen du RSM avec environ 70 % des items pour lesquels les données s'ajustent.

En général, les données recueillies pour les différents échantillons s'ajustent mieux au modèle dichotomique, c'est-à-dire s'ajustent peut-être mieux avec seulement deux catégories au lieu de quatre. Toutefois, le changement de modèle n'a pas permis d'obtenir plus de stabilité dans l'ajustement d'un groupe de données à l'autre. L'année de distribution, l'année scolaire, la version du questionnaire et le programme semblent toujours affecter l'ajustement des items au modèle. La modification des catégories a, au moins, réglé le problème de l'ajustement des catégories puisqu'elles n'existent plus avec ce nouveau modèle.

5.5 FONCTIONNEMENT DIFFÉRENTIEL DES ITEMS SELON LES VARIABLES

Une des causes possibles de cette variabilité est peut-être un fonctionnement différentiel des items d'un échantillon à l'autre, d'une version du questionnaire à l'autre, etc. Cette partie du chapitre se divise en deux. La première partie essaie de voir si le biais de mesure que nous avons tenté d'induire dans les données a créé un FDI entre les différents items du questionnaire. Un exemple d'analyse permet de comprendre sur quelle base la détection du FDI est évaluée. Dans une deuxième partie, une courte analyse de la relation entre le FDI et l'ajustement des données au modèle aide à inférer sur le lien qui peut exister entre les deux. Comme la recherche s'attarde à la construction d'un instrument de mesure d'un point de vue purement quantitatif, le but n'est pas de faire une analyse des items pour analyser la présence de biais de l'item, mais d'explorer l'influence potentielle que peut avoir le FDI dans la construction d'un instrument de mesure.

5.5.1 Exemple d'analyse du fonctionnement différentiel des items avec POLYSIBTEST

Dans cette partie, le premier objectif est de vérifier si la modification de l'ordre de présentation des items a produit un FDI pour certains items. L'emphase est donc mise sur les deux versions du questionnaire de 2001 puisque c'est seulement pour l'année de cueillette 2001 que l'ordre de présentation des questions a été modifié. L'établissement de l'ordre des questions ne permet pas clairement d'identifier quels items pourraient être influencés par ce changement; pour cette raison, la démarche d'analyse est exploratoire. Comme Stout et Roussos le proposent dans le manuel d'instructions du logiciel POLYSIBTEST, une démarche exploratoire se fait en trois temps. Une première étape consiste à vérifier le FDI pour chaque item où le reste des items est considéré comme bloc de référence; ensuite l'analyse est reprise sans les items considérés comme ayant un FDI à la première étape; finalement, une dernière analyse est effectuée en identifiant clairement tous les items rejetés comme potentiellement porteurs de FDI et les autres items comme bloc de référence.

L'hypothèse nulle de la statistique SIBTEST utilisée par le logiciel POLYSIBTEST stipule que la différence entre les valeurs estimées de δ pour deux items est 0. L'hypothèse alternative stipule que cette différence est différente de 0 peu importe quel groupe elle favorise. Puisqu'il n'y a pas de groupe de référence stable d'une comparaison à l'autre, les groupes de données ne sont pas étiquetés R ou F. De toute façon, l'objectif est seulement de savoir s'il y a FDI, peu

importe le groupe favorisé. Le tableau qui suit est un exemple de sortie obtenue avec le logiciel POLYSIBTEST.

Tableau 5.12 Exemple de sorties avec POLYSIBTEST

Item suspect	*Différence de δ*	*Erreur de mesure*	*Valeur p au test*	*Proportion items éliminés*		*Différence pour le sous-groupe de référence*
				Dans le groupe R	*Dans le groupe F*	
1	0.016	0.046	0.724	.02	.08	-0.14
2	0.075	0.043	0.079	.02	.07	-0.15
3	-0.197	0.055	0.000	.02	.09	-0.12
4	-0.054	0.054	0.315	.03	.09	-0.13
5	0.051	0.056	0.364	.02	.07	-0.15
6	0.097	0.057	0.087	.02	.06	-0.15
7	-0.044	0.049	0.368	.03	.08	-0.14
8	-0.076	0.042	0.070	.02	.07	-0.13
9	-0.064	0.042	0.134	.02	.07	-0.13
10	0.110	0.045	0.015	.02	.09	-0.15
11	-0.003	0.041	0.943	.02	.08	-0.14
12	-0.026	0.046	0.577	.03	.08	-0.14
13	0.001	0.040	0.982	.02	.07	-0.14
14	-0.063	0.053	0.236	.02	.09	-0.14
15	-0.158	0.052	0.002	.02	.07	-0.12
16	0.012	0.042	0.765	.04	.08	-0.14
17	0.166	0.048	0.001	.03	.10	-0.16
18	0.008	0.048	0.871	.03	.09	-0.14
19	0.013	0.047	0.786	.03	.08	-0.14
20	0.026	0.042	0.548	.02	.09	-0.14

Pour les sorties du logiciel POLYSIBTEST, la valeur la plus importante est celle de la quatrième colonne, la valeur de la probabilité associée au test statistique. Une valeur inférieure à 0,05 suggère le rejet de l'hypothèse nulle et la présence d'une différence significative entre les valeurs estimées pour les items. Dans ce cas ci, il y aurait un FDI pour les items 3, 10, 15 et 17. C'est une différence dans la valeur estimée pour δ d'environ 0,1 qui est significative dans ce cas-ci (première colonne).

5.5.2 Exploration du FDI pour les deux versions de 2001

Les deux tableaux 5.13 et 5.14 indiquent quels items ont un FDI pour les groupes de données des deux versions, des deux programmes et selon l'année scolaire. Il ne semble pas y avoir de constance dans la détection de FDI, c'est-à-dire que, selon l'année, le programme ou l'année scolaire, ce ne sont pas les mêmes items pour lesquels on détecte du FDI. Par exemple, pour les

données des deux versions du programme d'enseignement secondaire, ce sont les items 1, 16 et 18 pour lesquels il y a du FDI. Pour les données des deux versions du programme primaire, c'est seulement l'item 4. Puis, pour les données des deux versions des programmes pris ensemble, ce sont les items 2, 10, 16 et 18. En plus, au tableau 5.14, les items qui fonctionnent différemment pour les données des deux versions du programme secondaire ne sont pas les mêmes en troisième et quatrième. Alors d'où provient le FDI dans les données entre les deux versions au secondaire ? Comme il n'y a pas de constance dans la détection du FDI, il est difficile de comprendre comment le changement de l'ordre des questions affecte un item précis.

Tableau 5.13 Items identifiés avec du FDI pour l'année 2001

Item	*Entre version 1 et version 2*	*Entre Secondaire version 1 et 2*	*Entre Primaire version 1 et 2*
1		X	
2	X		
3			
4			X
5			
6			
7			
8			
9			
10	X		
11			
12			
13			
14			
15			
16	X	X	
17			
18	X	X	
19			
20			

Tout de même, il reste intéressant d'explorer la nature des items pour lesquels il y a eu détection de FDI. L'ordre des items présenté à la section 4.2.1 est : 2-20-19-1-5-6-14-7-18-4-16-17-15-3-13-11-9-10-12-8. Si on étudie les items 16 et 18 par exemple, ils font partie du groupe des items du milieu où l'influence de l'ordre de présentation des items est logiquement difficile à établir. Entre autre, peu de FDI était attendu pour ces items. Il se pourrait que l'ordre de présentation des items soit responsable du FDI. Pour les tableaux 5.13 et 5.14, les données pour lesquelles les items devraient probablement démontrer du FDI (19-1-5-6 ou 10-12-8) le font dans seulement 8 cas sur 49, environ 16 %, et principalement pour les items 12 et 8. Conséquemment, il n'est pas possible d'affirmer que l'ordre de présentation des items est responsable du FDI détecté

d'une version à l'autre pour les données des deux programmes et des deux années scolaires présentées à cet alinéa. Il n'y a pas assez de constance dans les items. En conclusion, la modification apportée à l'ordre des items n'a pas réussi à induire un FDI constant.

Tableau 5.14 Items identifiés avec du FDI pour l'année 2001 selon l'année scolaire

Item	*Entre secondaire troisième version 1 et 2*	*Entre secondaire quatrième version 1 et 2*	*Entre primaire troisième version 1 et 2*	*Entre primaire quatrième version 1 et 2*
1				
2	X			X
3			X	
4				X
5				
6	X			
7		X		
8		X		X
9				
10				
11				
12	X		X	X
13	X			
14	X			
15	X		X	
16				
17				
18				
19				
20	X		X	X

5.5.3 Exploration de la variation dans l'ajustement en regard du FDI

Pour les items qui démontrent du FDI, un lien peut-il être fait avec l'ajustement des données au modèle ? La présence de FDI entre deux items influence-t-elle l'ajustement des données au modèle ? Le tableau 5.15 montre quels items s'ajustent différemment d'un groupe d'étudiants à l'autre et s'il y a présence de FDI pour ces mêmes items. De tous les items de ce tableau, seulement l'item 2 entre les deux versions qui incluent tous les étudiants montre un FDI et une différence entre l'ajustement des items. C'est très peu.

Tableau 5.15 Liste des items qui diffèrent dans l'ajustement au modèle et détection de FDI pour ces items en 2001

Groupe d'étudiants (échantillon)	*Liste des items qui diffèrent dans leur ajustement entre les versions 1 et 2*	*Présence de FDI pour l'item*
Tous les étudiants	2	Oui
	3	Non
	4	Non
	12	Non
	15	Non
Secondaire troisième	9	Non
Secondaire quatrième	4	Non
	11	Non
	12	Non
	13	Non
	14	Non
	20	Non

5.6 VÉRIFICATION DE L'HYPOTHÈSE D'INVARIANCE DANS L'ESTIMATION DE δ

À quelques reprises, l'estimation de δ pour les données d'un programme à l'autre donne des résultats qui varient considérablement. Cette variation porte à croire que l'hypothèse d'invariance ne serait pas confirmée par les données de cette recherche. Dans la méthodologie, nous avons établi que la meilleure façon de déterminer s'il y a une variation dans la valeur de l'estimation est d'explorer l'ordre de grandeur des valeurs. En plus, seuls les items retenus par le modèle sont comparés parce que la présence de données qui ne s'ajustent pas au modèle peut influencer la valeur de l'estimation et, conséquemment, l'ordre. Une comparaison de l'ordre pour chacune des variables qualitatives de la recherche donne un indicateur de la présence d'invariance dans l'estimation des paramètres de l'item. Toutefois, comme le chapitre IV l'indique, c'est un indicateur qui sert à défaut d'avoir recensé une méthode scientifique éprouvée et appropriée aux besoins de cette recherche.

Tableau 5.16 Ordre de l'estimation de la valeur de δ pour quatre groupes d'étudiants

Item	*Ordre de grandeur de l'estimation de δ*							
	Secondaire		*Primaire*		*Secondaire 3 – 2000*		*Secondaire 4 – V1 - 2001*	
	RSM	*Dicho*	*RSM*	*Dicho*	*RSM*	*Dicho*	*RSM*	*Dicho*
1	11/12	12/12	16/16	12/16	12/13	7/14	12/14	12/19
2	12			16			14	19
3	6						6	8
4			9	6			8	15
5		9	5	3		12		14
6		10	6	5		11		17
7	9	8	13	11	10	10	11	18
8	1		1		1		1	1
9	4	2	4	2	4	2	3	4
10	3	1	3		3	1	2	2
11		3	7	4		3	4	5
12	2		2	1	2			3
13			8			6		11
14		7		13	8	8		16
15	5		10	7	5		7	9
16		5		10	9	9	10	10
17	7	4	11	9	6	4	5	7
18	8	6	12	8	7	5	13	13
19	10	11	14	14	11	13	9	6
20			15	15	13	14		

Le tableau 5.16 ci-haut donne l'ordre de la valeur de l'estimation des items pour différents groupes d'étudiants. En prenant les groupes pour le programme secondaire et primaire et en regardant uniquement les items pour lesquels les données s'ajustent avec le RSM, les items 1, 19 et 7 sont toujours ceux avec la plus grande valeur estimée. Les items 8, 9, 10 et 12 sont aussi toujours ceux avec la plus petite valeur estimée. L'ordre des items serait identique pour ces deux programmes avec la séquence suivante : 8, 12, 10, 9, 15, 17, 18, 7, 19, 1. Il y aurait donc invariance dans l'ordre des valeurs estimées dans le cas spécifique du RSM pour ces deux programmes. Cette invariance est plus facile à voir à la figure 5.3. Comme la figure le montre, les lignes ne se croisent jamais, ce qui signifie que l'ordre de l'estimation du paramètre des items ne varie pas. Dans la mesure où les lignes ne se croisent pas, il y a invariance dans l'ordre de la valeur du paramètre estimé. Aussi, plus les lignes sont droites, moins la valeur de l'estimation du paramètre s'est modifiée d'un groupe à l'autre. Il est intéressant de voir sur la figure 5.3 que les lignes les plus droites sont celles qui se situent dans le milieu du continuum. Alors peut-être que les items aux extrémités de l'échelle sont celles dont la valeur varie le plus d'un groupe de données à l'autre. Mais pour l'instant, ce n'est qu'une observation.

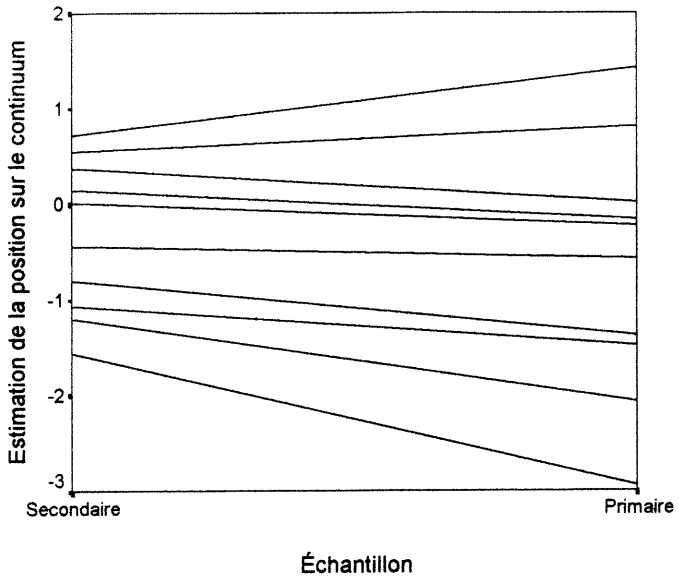

Figure 5.3 Graphique de la variation de l'ordre des valeurs estimées pour δ sur le continuum pour les groupes des programmes secondaire et le primaire en 2000 avec le *Rating Scale Model*

Il ne faut toutefois pas oublier que le paramètre τ ne s'ajuste pas au modèle pour ces deux groupes. Conséquemment, il n'est pas possible d'affirmer que l'ordre d'estimation de ce paramètre est invariant, ce qui cause problème. Dans ces circonstances, il est intéressant de voir comment se comporte l'ordre des valeurs estimées pour le modèle dichotomique qui, lui, permet un ajustement des catégories. Pour le modèle dichotomique, le tableau 5.16 plus haut indique que les items 19 et 20 sont ceux dont la valeur est la plus grande dans les deux groupes (secondaire et primaire). Mais pour les items 1,17 et 18, l'ordre ne se maintient pas comme il est possible de le voir à la figure 5.4. Cette fois, certaines lignent se croisent sur la figure. Plus de sept lignes se croisent. Encore une fois, les lignes les plus droites sont celles situées au milieu, c'est-à-dire près de 0. La valeur estimée de certains paramètres varie jusqu'à 1,888 de différence (item 6). Sur la figure, cet item croise plusieurs lignes. Pour le modèle dichotomique, la valeur de l'estimation des paramètres varie, du moins pour quelques items. La question est de savoir qui serait le mieux entre un modèle où les données ne s'ajustent pas bien et un modèle dont l'ordre de grandeur de la valeur des estimations n'est pas invariant. Il faut aussi se demander si ces deux échantillons représentent en fait des populations différentes.

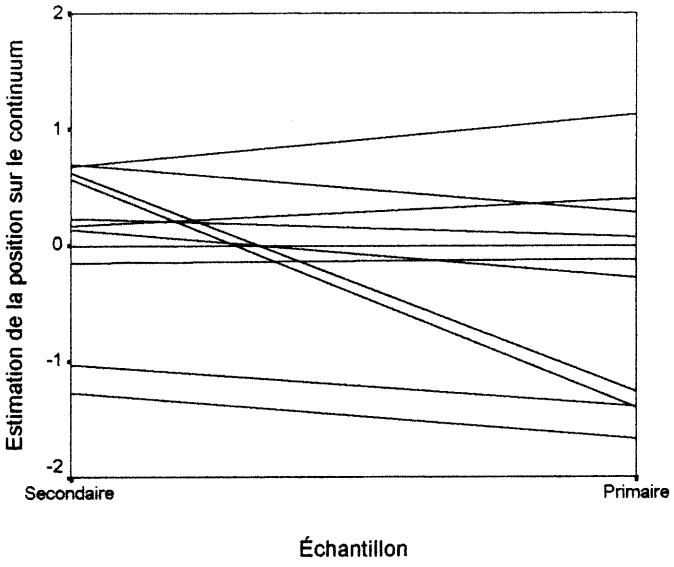

Figure 5.4 Graphique de la variation de l'ordre des valeurs estimées pour δ sur le continuum pour les groupes des programmes secondaire et le primaire en 2000 avec le modèle dichotomique de Rasch

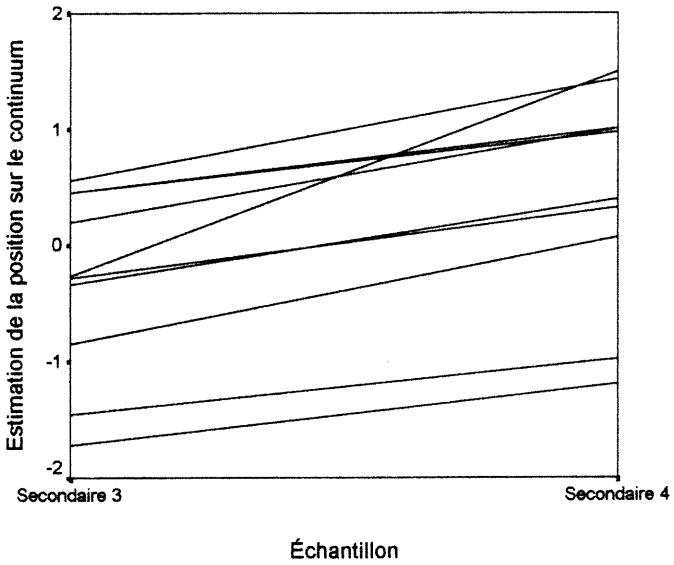

Figure 5.5 Graphique de la variation de l'ordre des valeurs estimées pour δ sur le continuum pour les groupes des programmes secondaire 3 en 2000 et le secondaire 4 en 2001 avec le *Rating Scale Model*

L'exemple de la figure 5.5 est intéressant parce que ce sont, grosso modo, les mêmes étudiants qui composent ces deux groupes, c'est-à-dire qu'ils ont passé de la troisième année à la quatrième. Tout de même, pour le RSM, la figure 5.5 montre que l'ordre de la valeur des

estimations varie beaucoup puisque plusieurs lignes se croisent sur le graphique. Conséquemment, pour un groupe d'étudiants semblable d'une année à l'autre avec le *Rating Scale Model*, il ne semble pas y avoir d'invariance dans la valeur de l'estimation du paramètre de l'item.

Deux des trois figures ci-haut indiquent qu'il n'y aurait pas d'invariance en se fiant à l'indicateur utilisé. Néanmoins, la démarche d'analyse exige de vérifier cette affirmation plus rigoureusement. Le tableau 5.17 indique si l'ordre de valeur estimée du paramètre δ pour les items est invariant pour différents groupes d'étudiants pour le RSM. Les cases du tableau donnent une indication négative (NON) lorsqu'au moins une ligne croise les autres lignes. Une indication positive (OUI) signifie que toutes les lignes sont parallèles sur les figures, qu'aucune ligne ne se croise. Cette façon (oui ou non) de trancher dès qu'une ligne entrecoupe une des autres lignes n'est pas idéale, mais c'est celle que nous avons choisie. Évidemment, il pourrait arriver que seulement une ligne entrecoupe légèrement les autres. À ce moment-là, on pourrait peut-être évaluer l'invariance selon chaque cas, mais nous avons plutôt choisi de trancher sans discrimination pour donner un portrait quantitatif sommaire.

Comme le tableau l'indique, pour les combinaisons autres que celles des figures plus hautes, il y a toujours au moins une ligne qui croise les autres. Ce tableau est seulement pour le RSM, mais il suggère que la propriété d'invariance apparaît rarement d'un groupe à l'autre. Les données ont démontré de l'invariance dans la valeur estimée du paramètre seulement pour les données des programmes du secondaire et du primaire (dans notre premier exemple plus haut).

Tableau 5.17 Vérification de la propriété d'invariance en regard des variables qualitatives pour le RSM de Rasch

	Année de cueillette				*Version du questionnaire*				*Programme scolaire*			
	2000		2001		V1		V2		Secondaire		Primaire	
	Sec3	Sec4	Sec3	Sec4	Sec3	Sec4	Sec3	Sec4	Sec3	Sec4	Pri3	Pri4
Année de cueillette												
2000-Sec 3		Non	Non	Non								
2000-Sec 4				Non								
Version												
V1-Sec3							Non	Non				
V1-Sec4							Non	Non				
Programme												
Secondaire 3											Non	Non
Secondaire 4											Non	Non

CHAPITRE VI

DISCUSSION DES RÉSULTATS

Les résultats présentés au chapitre précédent doivent maintenant être remis dans le contexte de nos questions de recherche. L'essentiel de la discussion de ce chapitre porte sur cette remise en contexte des résultats de la recherche. La première section de ce cinquième chapitre fait un retour sur les questions de la recherche et sur la façon dont elles seront étudiées dans ce chapitre. La deuxième section regroupe l'information qui permet de répondre à la première question secondaire de la recherche. La troisième section offre des éléments de réponse à la deuxième question secondaire de la recherche. La quatrième section reprend l'analyse faite pour ces deux questions et remet celles-ci dans le contexte de la question principale de la recherche.

6.1 RETOUR SUR LES QUESTIONS DE LA RECHERCHE

Au chapitre I, dans la problématique à la page 6, nous avons spécifié notre question de recherche. Cette question est :

Quels sont les avantages et les désavantages à utiliser les modèles de la théorie de la réponse à l'item dans un processus de validation d'un instrument de cueillette d'information ?

Pour répondre à cette question, il est préférable de commencer par répondre aux deux questions spécifiques suivantes :

Dans quelle mesure le *Rating Scale Model* de Rasch permet-il de construire une échelle de mesure à partir d'un instrument de cueillette d'information ?

Dans quelle mesure la présence d'un biais de mesure influence-t-elle l'utilisation du *Rating Scale Model* dans cette construction ?

Les réponses à ces questions s'insèrent dans le cadre conceptuel de ce travail de recherche. C'est-à-dire que la discussion des résultats doit tenir compte des chapitres 2 et 3. Le chapitre 2 présente les fondements de base de la TRI et le chapitre 3 fait le lien entre la TRI et le processus de validation d'un instrument de cueillette d'information. À la lecture de chapitre, il faut tout de même retenir que les résultats d'un processus de validation d'un instrument ne s'atteignent pas en une seule analyse. Ces résultats sont plutôt le fruit d'efforts continus. Conséquemment, les arguments apportés à la suite de la présentation des résultats doivent être nuancés pour tenir compte de la réalité de cette recherche. Le troisième chapitre explique de façon linéaire la manière dont nous avons présenté et analysé les données. La discussion est conduite dans la même direction.

6.2 PREMIÈRE QUESTION SECONDAIRE DE LA RECHERCHE

Pour répondre à cette question, il faut se demander si le RSM s'ajuste aux données recueillies, si le nombre de catégories est approprié pour l'échelle de référence et si les valeurs des estimations des paramètres pour les différents groupes sont invariantes. Ces trois indices donnent des arguments de réponse à la question. Ils donnent de l'information sur la façon de mettre au point un instrument de mesure et permettent de décrire les conséquences que peut avoir cette mise au point sur la révision d'un instrument. Avec cette forme de modélisation, et comme les chapitres II et III l'expliquent, l'objectif final de ce processus d'analyse est de construire une échelle qui possède les propriétés de la mesure fondamentale en sciences sociales.

6.2.1 Analyse de l'ajustement des données aux RSM et conséquences pour l'instrument

Les résultats présentés au chapitre 5 suggèrent que l'ajustement des données au RSM est possible pour certains items. La figure 6.1 donne le pourcentage d'items pour lesquels les données s'ajustent au modèle pour les différents groupes d'étudiants en regard de leurs caractéristiques qualitatives. Dans l'ensemble, la figure montre que le pourcentage d'items est de 70 % en moyenne. Ceci veut dire que l'instrument original conserve la plupart de ces items après l'élimination des items pour lesquels les données ne s'ajustent pas. Dans ces circonstances, très peu d'information est perdu au niveau de la qualité de l'information. Comme le montre la figure, le pourcentage d'items varie avec le groupe d'étudiants. Sur l'abscisse, il y a plusieurs échantillons d'étudiants. Pour les trois premiers échantillons à gauche sur le tableau, les groupes de données incluent des étudiants de tous les programmes selon l'année de cueillette et la

version du questionnaire. Pour le reste des échantillons, les étudiants sont ceux d'un programme spécifique en fonction d'une année de cueillette; ensuite, en fonction de l'année scolaire et, finalement, en fonction de la version du questionnaire. Pour les groupes d'étudiants les plus hétérogènes, le pourcentage peut descendre jusqu'à 45 %, ce qui veut dire qu'il faudrait éliminer plus de la moitié des items. Pour les autres groupes, le pourcentage se situe autour de 70 %. Il y a une variation d'un groupe d'étudiants à l'autre, mais ce n'est pas clair que cette variation soit attribuable à l'homogénéité des groupes. Dans l'ensemble, on peut tout de même dire qu'une année scolaire s'ajuste mieux, ou au moins également, aux deux années scolaires prises ensemble. Ceci laisse croire que l'hétérogénéité pourrait jouer un rôle. Ces résultats semblent indiquer qu'il est possible d'ajuster un bon nombre d'items au RSM. C'est-à-dire qu'au moins 45 % des items ont des données qui s'ajustent au modèle avec une moyenne d'environ 70 %. Dans ces circonstances, l'instrument révisé maintiendrait la plupart de ces items après une première validation des données par le RSM.

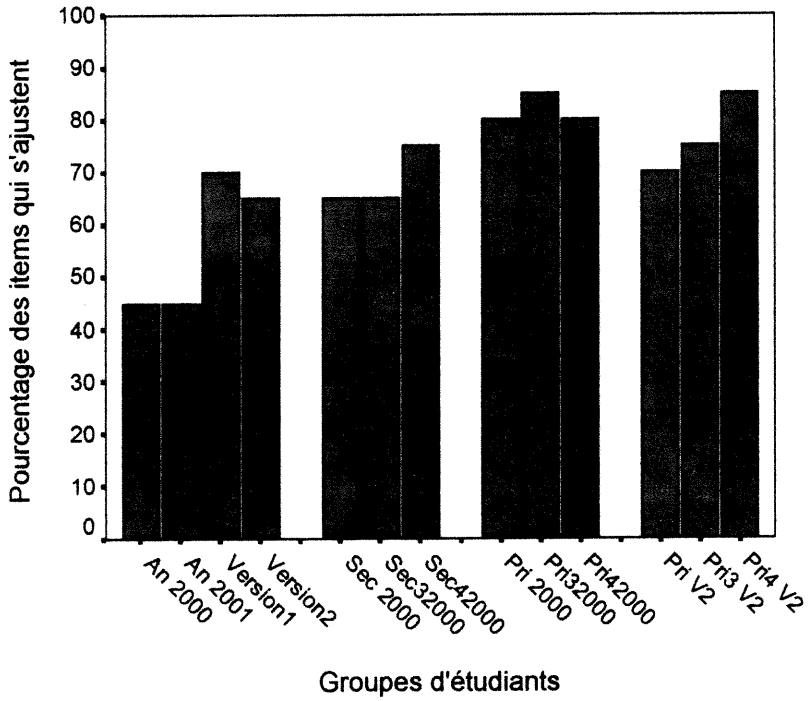

Figure 6.1 Pourcentage des items qui s'ajustent au RSM selon les variables qualitatives

L'élimination de certains items peut avoir des conséquences importantes sur l'information obtenue avec l'instrument. Parmi les items éliminés, certains ont peut-être une importance capitale aux yeux des concepteurs et il y a toujours le dilemme de savoir ce qui prime entre le modèle et les données. À titre d'exemple, prenons les données de la version 1 de 2001 pour le

secondaire en quatrième année. Pour cette version, les données des items 5, 6, 12, 13, 14 et 20 ne s'ajustent pas. Parmi ces items, deux interrogent les élèves par rapport aux parents et un par rapport aux différences ethniques et culturelles. Pour ce même exemple, les items 5 et 6 concernent la langue écrite et orale. Pour ces deux items, les données s'ajustent moins bien dans le cas des étudiants du programme d'enseignement secondaire que dans le cas de l'éducation préscolaire et primaire. À ce moment-là, il y a peut-être une raison externe (une deuxième dimension par exemple) qui influence les réponses à ces items pour les étudiants du programme secondaire et qui explique le manque d'ajustement des données. Un autre désavantage de l'ajustement est que ce ne sont pas les données des mêmes items qui s'ajustent d'un groupe d'étudiants à l'autre comme nous l'avons vu tout au long du chapitre précédent. Ceci ne veut toutefois pas dire que le modèle ne permet pas d'améliorer l'information obtenue. Au contraire, il permet de soulever la question et d'améliorer la composition de ces items. Il serait ensuite possible de remettre à l'essai ces items modifiés suite à l'analyse de l'ajustement. En somme, l'élimination de certains items provoquerait une diminution du nombre d'items et, conséquemment, une diminution de la quantité d'information recueillie. D'un autre côté, la validité de l'instrument est améliorée parce qu'il ne contient plus d'items pour lesquels les données ne s'ajustent pas au modèle. Le contexte entourant la distribution de l'instrument peut donner une indication de ce qui est préférable entre éliminer des items ou maintenir une grande quantité d'information. Mais rappelons qu'il serait probablement préférable d'amener des modifications aux items pour lesquels il n'y a pas d'ajustement et de reprendre ensuite l'analyse de l'ajustement des données. D'ailleurs, le processus pourrait prendre quelques années avant d'obtenir un instrument qui possède toutes les caractéristiques et toutes les qualités désirables.

Bien que la figure 6.1 indique que les données des items s'ajustent assez bien au modèele, ce n'est pas le cas pour les paramètres d'intersection des catégories. Pour les analyses présentées au chapitre cinq, seulement 10 % des catégories ont des données qui s'ajustent au RSM, ce qui veut dire presque jamais. Il y aurait donc un problème dans la structure de l'échelle. Le manque d'ajustement indique qu'il faudrait aussi réévaluer l'échelle dans la mesure où l'on souhaite maintenir le RSM pour modéliser nos données.

6.2.2 Le nombre de catégorie et les conséquences d'un changement de ce nombre

Les statistiques d'ajustement du chapitre 5 pour les catégories indiquent qu'il y a un problème avec la structure de l'échelle de référence de l'instrument initial. Pour contrer ce problème, les

deux catégories des extrêmes ont été rassemblées pour en former une seule. Suite à ce changement, c'est un nouveau modèle qui devait être ajusté. Le figure 6.2 donne le pourcentage d'items pour lesquels les données s'ajustent à ce nouveau modèle en comparaison du RSM. La figure indique que le pourcentage d'items est plus élevé pour le modèle dichotomique pour certains groupes d'étudiants et moins pour d'autres. Parmi les groupes comparés à la figure 6.2, il ne semble pas que le modèle dichotomique permette d'ajuster beaucoup plus d'items que le RSM dans l'ensemble. Le pourcentage semble dépendre du groupe et il n'y a pas de trame commune entre les groupes d'étudiants. La différence majeure est que dans le modèle dichotomique, les données n'ont pas besoin d'être ajustées aux paramètre d'intersection de par sa définition. À ce moment-là, en changeant les catégories et en ajustant les données au modèle dichotomique, le problème de l'ajustement des catégories serait résolu. Dans ce cas, le modèle dichotomique serait plus approprié. Dans un processus à long terme, il faudrait peut-être essayer d'ajuster le RSM en incluant un plus grand nombre de catégories à l'échelle de référence ou en modifiant le libellé des catégories. Toutefois, il ne faut pas oublier que la statistique d'ajustement est une statistique du khi-carré. Comme nos groupes de sujets sont déjà petits (autour de 100 sujets pour certains), il ne reste déjà plus beaucoup de données dans chacune des cases du tableau de contingence de la statistique khi-carré, ce qui peut peut-être influencer les résultats obtenus et l'ajustement des données aux catégories.

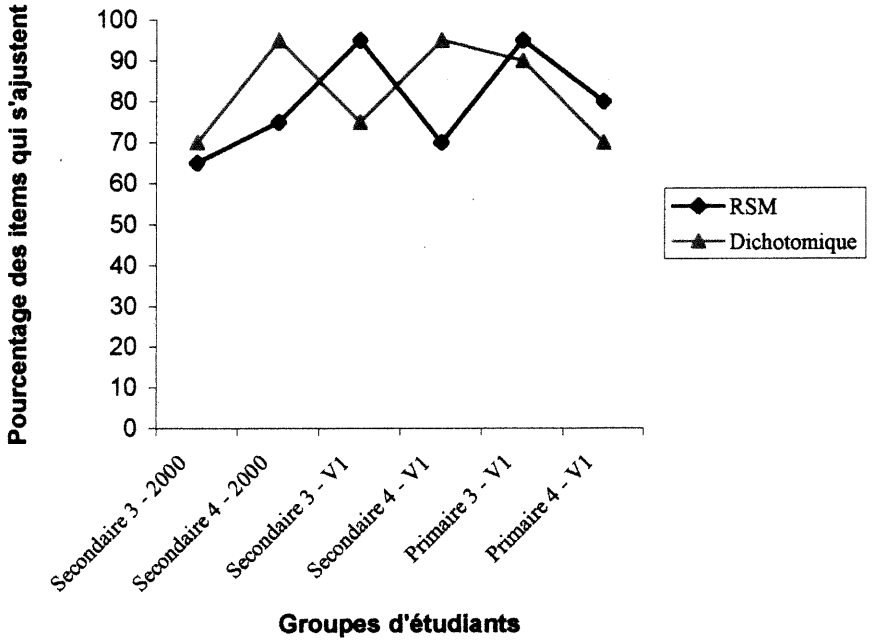

Figure 6.2 Comparaison du pourcentage des items qui s'ajustent au modèle dichotomique de Rasch et au RSM

Cette suggestion ne veut pas dire que le modèle dichotomique soit sans problème. Il faut encore éliminer les items pour lesquels les données ne s'ajustent pas au modèle. Mais en plus, ce nouveau modèle oblige d'avoir seulement deux catégories de réponse à l'échelle, ce qui n'était pas nécessairement l'intention du concepteur. Les réponses obtenues se limitent alors à « d'accord » ou « en désaccord ». Les catégories ne permettent plus de nuancer parmi ceux qui répondent une ou l'autre de ces deux réponses. Donc, en plus d'avoir à éliminer des items, le nombre de catégories ne correspond plus exactement à ce qui était souhaité au début. Mais, encore une fois, pour les avantages que peut apporter la TRI dans la validation de l'instrument, le concepteur peut décider d'ignorer ces désavantages. Il se pourrait aussi que deux catégories soient préférables dans ce cas-ci. En plus, rien n'empêche de modifier l'instrument avec un nombre différents de catégories à l'échelle. Dans ce sens, la TRI est très utile parce qu'elle permet d'étudier le nombre de catégories de l'échelle de référence vers un équilibre appréciable entre les catégories et le modèle choisi pour comprendre les données.

6.2.3 La vérification de l'hypothèse de l'invariance

Van der Linden définit la mesure implicite à partir de deux conditions : l'obtention d'une échelle linéaire et la présence d'invariance dans l'estimation des paramètres. Le propos de ce texte n'est pas de vérifier l'existence de la première condition. La section 2.2.2 réfère à quelques auteurs qui ont étudié cette condition. C'est l'existence de la deuxième condition qui est vérifiée ici. Pour cette recherche, nous avons supposé que, dans le mesure où le modèle de Rasch s'ajuste aux données, la présence de l'hypothèse de l'invariance devrait se confirmer par l'étude des données. C'est à partir d'une étude de la variation de l'ordre de grandeur de la valeur de l'estimation des paramètres des items que le chapitre précédent analyse celle-ci. À partir de cet indice, nous avons choisi d'inférer sur la présence de cette propriété.

Les résultats de la section 5.6 soulignent que ce n'est pas dans tous les cas que l'ordre de grandeur dans l'estimation des paramètres de l'item est invariant. Pour les deux exemples présentés au chapitre précédent, seulement les deux groupes des étudiants pour les programmes secondaire et primaire en 2000 avec le RSM présentaient de l'invariance dans les données. Dans le cas du modèle dichotomique, ces deux mêmes groupes ne présentaient pas d'invariance dans la valeur de l'estimation de ses données. Pour toutes les autres combinaisons de groupes, au moins un item ne se situait relativement pas au même endroit sur le continuum par rapport aux

autres. Il faut encore une fois remettre ces résultats en perspective. Dans certains de ces cas, la façon de trancher est peut-être radicale, tel que mentionné à la section 5.6.

Ce manque d'invariance d'un groupe à l'autre peut être attribuable à certains facteurs comme la taille des groupes ou l'homogénéité de ces groupes les uns par rapport aux autres. La question est alors de savoir si le questionnaire est en mesure de produire des valeurs estimées stables pour toute la population étudiante des programmes de formation des maîtres. Il serait plus sensé de croire qu'il existe peut-être plusieurs populations et que plusieurs modélisations différentes doivent être réalisées pour chacune des populations. Mais là n'est pas l'objectif. Au contraire, le but et la pertinence d'utiliser cette forme de modélisation est de construire un questionnaire unique qui possède des propriétés métriques intéressantes.

Pour l'instant, comme l'analyse faite de l'invariance ne permet pas de conclure en faveur de cette dernière, il faudrait probablement apporter des modifications à la structure de l'instrument ou redéfinir les échantillons visés par le questionnaire. Bien évidemment, une méthode plus sophistiquée comme indicateur de la présence d'invariance aurait peut-être donné des résultats différents. Nos résultats se basent uniquement sur l'ordre de grandeur de la valeur des estimations et ne sont que des inférences faites à partir de cet ordre.

6.2.4 Réponse à la première question secondaire de la recherche

Dans un processus de validation d'un instrument de cueillette d'information, le *Rating Scale Model* permet d'identifier les items qui ne s'ajustent pas, donc qui demanderaient une révision; il permet de faire ces mêmes vérifications et ajustements pour les catégories de l'échelle et, finalement, de s'interroger sur l'homogénéité de la population visée et, conséquemment, sur la relation entre les items et les caractéristiques de la population. À cette étape du processus de validation, la TRI ne permet toutefois pas de construire une échelle de mesure au sens propre. C'est-à-dire qu'elle ne permet pas d'obtenir des scores à l'échelle qui soient linéaires et invariants comme la TRI suggère qu'elle peut le faire à cette étape ci du processus de validation. Peut-être, qu'après une révision des items, il sera possible d'obtenir ce type de score. Pour répondre de façon plus directe à la question, le *Rating Scale Model* permet d'entreprendre une démarche qui pourrait éventuellement mener à la construction d'un instrument de mesure.

6.3 DEUXIÈME QUESTION SECONDAIRE DE LA RECHERCHE

La deuxième question de la recherche s'intéresse à l'effet que peut avoir la détection de FDI sur l'ajustement des données au modèle ou sur la propriété d'invariance. Pour répondre à cette question, il faut voir s'il existe un lien entre les items pour lesquels les données s'ajustent différemment au modèle du groupe d'étudiants à l'autre et la détection de FDI, et entre les items avec des valeurs estimées invariantes et la détection FDI. Mais avant de passer à l'analyse, il faut comprendre que les résultats du chapitre 5 ne permettent pas de conclure que nous ayons réussi à induire un FDI attribuable à l'ordre de présentation des items. De fait même, l'analyse faite dans cette section s'interroge sur l'effet de la présence de FDI sans connaître la raison qui justifie celle-ci. Il se pourrait très bien que ce FDI soit justifié par une différence d'opinion entre ces deux groupes ou par un biais de mesure non identifié. Comme le but de cette recherche n'est pas de clarifier cette question, nous allons tout de même procéder à l'analyse, mais en sachant que l'influence de la modification de l'ordre de présentation des items n'a pas pu être établie dans le cadre de cette recherche à partir de l'information présentée à la section 5.5.2.

6.3.1 Lien entre le FDI et l'ajustement des données au modèle

Les résultats du chapitre 5 et la figure 6.3 suggère qu'il n'y pas de lien important à faire entre la détection de FDI et le nombre d'items pour lesquels les données s'ajustent différemment au modèle d'un échantillon à l'autre. La figure compare le nombre d'items où les données s'ajustent différemment au modèle et combien de ces items ont été étiquetés avec du FDI. Cette comparaison peut donner un indice du lien à faire entre les deux. À regarder la figure, il ne semble pas que les items pour lesquels les données s'ajustent différemment d'un modèle à l'autre soient les mêmes items pour lesquels il y a eu détection de FDI. Sinon, les bandes de la figure seraient de la même grandeur. Donc, il est difficile d'établir que le manque d'ajustement dans le modèle est une conséquence de la présence d'un FDI. Mais encore, ce ne sont que des résultats partiels puisque nous ne connaissons pas vraiment la raison de cette présence de FDI.

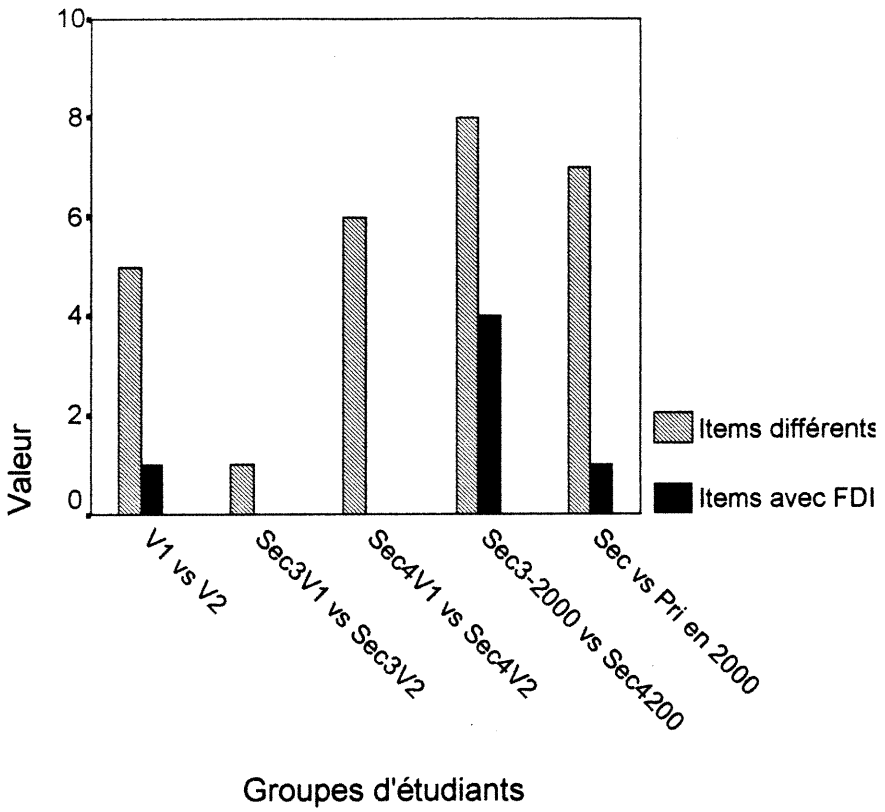

Figure 6.3 Nombre d'items où les données s'ajustent différemment entre deux groupes d'étudiants et nombre d'items avec un FDI entre ces deux mêmes groupes

6.3.2 Lien entre le FDI et la propriété d'invariance

Le deuxième impact possible de la présence de FDI sur l'utilisation du RSM est au niveau de la propriété d'invariance. L'idée est encore de vérifier s'il y a un lien à faire entre la variation dans l'ordre de grandeur, les valeurs estimées de δ et la présence de FDI. Pour ce faire, nous allons reprendre quelques figures du chapitre 5, mais cette fois-ci, nous allons indiquer quels items ont été étiquetés avec du FDI. Pour rendre la comparaison plus intéressante, nous allons présenter la seule combinaison qui a démontré de l'invariance avec deux combinaisons qui n'ont pas démontré d'invariance. Sur les figures plus bas, les items avec détection de FDI sont en gras. Il faut voir si ces items (lignes) croisent les autres items plus souvent que les items sans détection de FDI. Dans l'éventualité où les items avec FDI croisent plus souvent que les autres lignes, ce serait un indice d'un lien potentiel entre la détection de FDI et l'invariance dans l'estimation des paramètres.

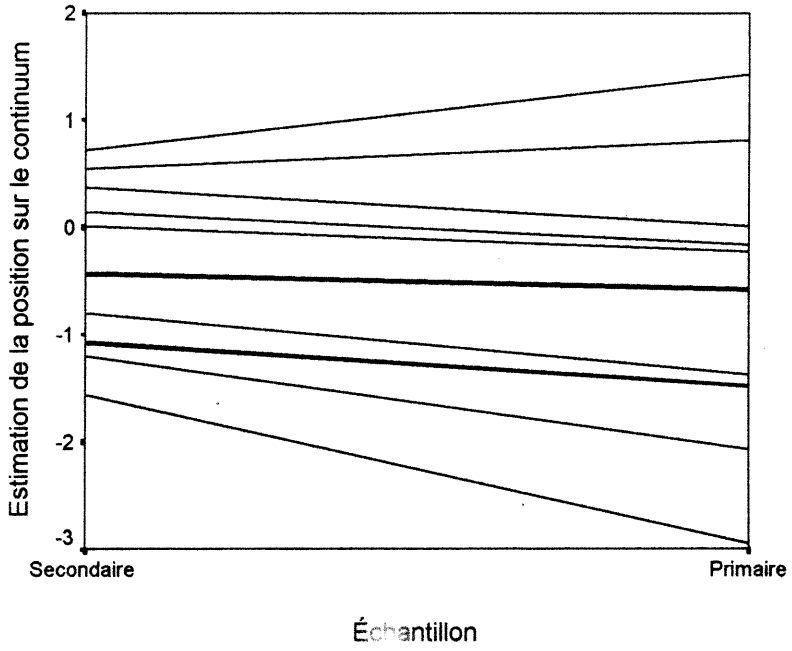

Figure 6.4 Comparaison de l'invariance et du FDI pour le secondaire et le primaire en 2000

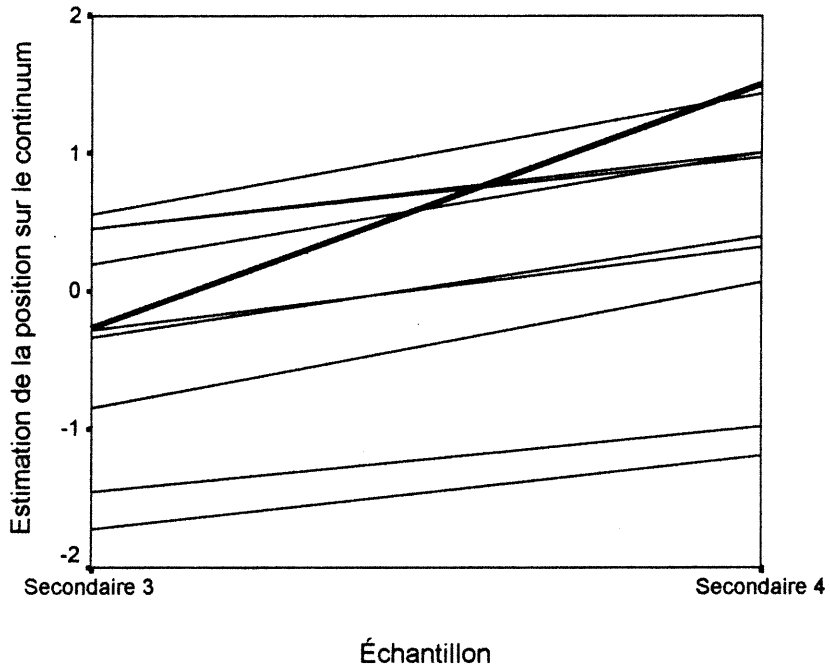

Figure 6.5 Comparaison de l'invariance et du FDI pour le secondaire 3 en 2000 et le secondaire 4 en 2001

Sur la figure 6.4, deux items (en gras) sont identifiés avec présence de FDI et, pourtant, ces deux items n'ont pas affecté l'ordre des valeurs estimées pour les items. Ces deux lignes sont assez parallèles au reste des autres lignes ce qui suggère qu'elles se comportent comme les autres, malgré la présence de FDI. Sur la figure 6.5, un seul item a un FDI, et ce dernier croise quatre autre lignes. Il affecte donc clairement l'ordre. En plus, c'est clairement l'item le moins parallèle aux autres. Mais d'autres items se croisent sur la figure, ce n'est donc pas le seul item responsable de l'invariance entre les deux groupes. À partir de la réflexion sur ces deux figures, nous ne pouvons pas affirmer que le FDI soit responsable du manque d'invariance entre les différents groupes. Ou plutôt, il n'y a pas d'indice en ce moment qui permette de le suggérer. Évidemment, une étude plus approfondie du FDI retracé dans cette recherche pourrait peut-être permettre d'éclaircir davantage la question comme nous le mentionnons depuis le début de cette section.

6.3.3 Réponse à la deuxième question secondaire de la recherche

Il n'est pas vraiment possible de répondre à cette question de recherche par manque d'information pertinente. Puisque la recherche n'a pas réussi à contrôler suffisamment l'insertion du biais de mesure, le FDI dans cette recherche est resté incompris. C'est pourquoi il ne faut pas trop avancer de réponse à cette question. Pour des fins exploratoires, il ne semble toutefois pas qu'il existe un lien entre les items qui donnent des valeurs significatives pour l'étude du FDI avec la statistique comptabilisée par le logiciel POLYSIBTEST, l'ajustement des données et la propriété d'invariance. Mais ce sont des résultats partiels et incomplets.

Jusqu'ici, ce chapitre a répondu aux deux questions secondaires de la recherche. La réponse à ces deux questions donne tous les éléments nécessaires pour répondre maintenant à la question principale de la recherche.

6.4 QUESTION PRINCIPALE DE LA RECHERCHE

Dans l'introduction de ce texte, lorsque la problématique a été développée, l'utilisation de la théorie de la réponse à l'item avait comme objectif premier de résoudre les problèmes associés à la théorie classique des tests et d'explorer comment la qualité de l'instrument pouvait être améliorée. En acceptant les présupposés théoriques de la TRI, nous devrions avoir les avantages suivants : les caractéristiques des items ne dépendent pas des groupes; les scores ne dépendent

pas de l'instrument; un modèle qui s'exprime au niveau des items et non du test; et un score qui possède les caractéristiques d'une mesure fondamentale.

En somme, à ce point du processus de validation d'un instrument de cueillette d'information, notre étude de la TRI ne permet pas de confirmer tous les avantages théoriques de celle-ci. Certaines données ne s'ajustent pas au modèle, le nombre de catégories de l'échelle n'est pas approprié et il n'y pas vraiment d'invariance dans les valeurs estimées entre les différents groupes de sujets. Tout de même, la TRI permet d'accomplir certaines choses intéressantes dans un processus de validation. Par exemple, elle permet d'identifier quels items doivent être modifiés pour améliorer l'ajustement des données au modèle. Un deuxième avantage est dans l'analyse des catégories. Puisque les données ne s'ajustent presque jamais aux catégories de l'échelle, la TRI soulève directement ce problème. D'ailleurs, elle donne une piste de réflexion pour les concepteurs de l'instrument. Il est peut-être préférable de commencer avec plus de catégories que moins pour se donner plus de flexibilité dans l'analyse du nombre idéal de catégories à avoir pour l'échelle de référence. Dans un processus de validation à long terme, la TRI peut permettre de sauver du temps et des efforts parce qu'elle cible certains des problèmes du questionnaire de façon pointue dès la première étape. En plus, par l'entremise de l'étude de la propriété d'invariance, la TRI donne une indication sur l'homogénéité des données, des groupes. Comme une population assez homogène devrait donner une estimation invariante des valeurs des paramètres selon la TRI, son absence remet en question l'homogénéité des groupes. Il faudrait alors revoir certains des items qui n'ont peut-être pas réussi à généraliser l'expérience des étudiants des différents groupes. Comme l'objectif demeure de construire une échelle unique, il faudrait modifier les items pour qu'ils reflètent l'homogénéité de ces groupes. Il est aussi possible de se penser qu'il serait peut-être préférable d'analyser les données recueillies pour les groupes d'étudiants par programme et de traiter celles-ci comme provenant de populations différentes.

Le premier désavantage à utiliser la TRI est peut-être plus instrumental que pratique. C'est-à-dire que nous avons supposé que dans la mesure où le modèle s'ajuste aux données, les caractéristiques de la TRI seraient présentes dans les données. Il y a aussi le choix du modèle. À ce moment-là, il faudrait par exemple rejeter le modèle et non les items qui ne s'ajustent pas à ce modèle. Alors, si le RSM de Rasch n'était pas approprié, il y aurait peut-être révision d'items qui sont adéquats pour la situation. Un autre désavantage est que, dans certains cas, le RSM exige de réviser un grand nombre d'items. À ce moment-là, il faut se demander si tous ces

changements n'affectent pas trop le contenu des informations tirées de ce nouvel instrument final. Le dernier désavantage est que le processus de validation qui mène à la construction d'un instrument de mesure peut être long. Il y aura sûrement plusieurs analyses à faire et plusieurs distributions différentes des questionnaires avant d'en arriver là. Donc, d'une année à l'autre, les résultats peuvent être plus difficilement comparables. En plus, l'instrument final peut s'avérer fort différent de ce qui était convenu au départ. Il ne faudrait pas sacrifier la qualité du contenu pour la qualité du construit bien que l'équilibre soit sûrement possible.

CONCLUSION

En conclusion, il y aurait certains avantages non négligeables à utiliser la TRI dans un processus de validation d'un instrument de cueillette d'information. Le plus évident est que la validité du construit s'en verrait améliorée à long terme par l'obtention de données qui posséderaient les caractéristiques de la mesure implicite. C'est le résultat final qui justifie encore le plus l'utilisation de la TRI même si la validité du construit de l'instrument s'améliore toujours à moyen terme. Bien sûr, la TRI permet de détecter certains problèmes avec précision ce qui améliore constamment la validité du construit, mais c'est dans la mesure où nous acceptons les fondements de la TRI que ce processus est intéressant. Si les conditions d'application de la TRI s'avéraient ne pas être respectées, les décisions prises pour améliorer le construit devraient, elles aussi, être remises en question et, par le fait même, tout le processus de validation.

En ce qui concerne l'impact de la présence d'un biais de mesure sur le processus de validation, la recherche n'a pas réussi à conclure grand chose à ce sujet. Nous n'avons pas réussi à insérer un biais de mesure dans les données, donc l'effet de sa présence reste incertain. Il ne semble pas que l'ordre de présentation des items ait affecté clairement les réponses. Il nous est donc impossible de conclure à l'amélioration ou à la détérioration de cette source de validité à partir de nos résultats. L'impact du FDI sur la modélisation est tout aussi inconcluant.

Les conclusions de la recherche s'insèrent dans un cadre méthodologique qui avait certaines limites. De celles-ci, la première limite de la recherche concerne le programme informatique. Le logiciel CONQUEST s'est avéré assez facile d'utilisation et a sûrement permis de faire une estimation juste de la valeur des paramètres du modèle. Il se pourrait toutefois qu'un autre logiciel informatique donne des résultats différents pour les mêmes données. Il serait intéressant d'explorer cette possibilité. En plus, il y a certaines sorties que CONQUEST ne permet pas d'obtenir tel le score moyen des candidats à l'item ou l'indice de fidélité de la valeur d l'estimation des paramètres et de l'ordre de grandeur, ce qui limite certaines analyses. Dans ces circonstances, il est conseillé de choisir un logiciel qui fournit ces informations tel RUMM2010. La deuxième limite de la recherche est dans la grandeur des échantillons. Avec des groupes d'étudiants aussi peu nombreux que 50, il est possible que ce nombre soit trop petit pour certaines des analyses. Troisièmement, la recherche n'a pas exploré toutes les combinaisons possibles de sous-groupes, mais s'est limitée à couvrir un certain nombre de possibilités. Puisque les résultats obtenus à partir du schéma d'analyse sont assez uniformes, rien ne porte à croire

qu'une enquête plus exhaustive apporterait de l'information qui viendrait contredire les résultats de la recherche, mais ce n'est pas impossible. Quatrièmement, la modification de l'ordre de présentation des questions n'a pas réussi à introduire un biais de mesure dans les réponses. Il a été impossible de répondre clairement à la deuxième question secondaire de la recherche. Il faudrait peut-être réessayer en modifiant la composition des items ou en introduisant un autre biais qui pourrait être vérifié à partir d'une analyse confirmatoire au lieu d'exploratoire. Cinquièmement, il existe d'autres modèles de la TRI qui auraient pu être utilisés comme par exemple le *Graded Response Model*. Il serait intéressant de reprendre l'analyse avec ce modèle ou un autre. Finalement, lorsqu'un item ne s'ajustait pas aux données, il pourrait être pertinent d'étudier le nombre de dimensions et/ou l'indépendance locale. Il serait peut-être adéquat de procéder à des vérifications de ce genre.

Dans le cadre de recherches futures, plusieurs angles d'analyse pourraient être envisagés. Premièrement, il serait intéressant de progresser dans le processus de validation avec ce même questionnaire et de voir où il peut nous mener à partir de modifications apportées à la lumière de cette recherche. Deuxièmement, il serait aussi intéressant de reprendre la même recherche avec plusieurs logiciels informatiques différents. Troisièmement, il va de soi que faire cette recherche avec de plus grands échantillons serait aussi intéressant. Mais d'un autre côté, ce questionnaire touche une population complète et il faut continuer à étudier les caractéristiques de la modélisation avec la théorie de la réponse à l'item dans ces circonstances.

RÉFÉRENCES

Adams, R. et Wright, B. D. (1994). When does misfit make a difference ?. Dans : Dans : Wilson, M. (ed.). Objective Measurement : Theory into Practice Vol.2. Ablex Publishing Corp. ISBN : 0-89381-843-1.

Aiken, L.R. (1997). *Questionnaires and inventories : Surveying Opinions and Assessing Personality.* New-York : J.Wiley.

American Psychological Association, American Educational Research Association et National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing.* Washington, DC : American Psychological Association.

Andersen, E.B. (1972). The numerical solution of a set of conditional estimation equations. *Journal of the Royal Statistical Society, Series B, 32,* 283-301.

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andrich, D. (1978b). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581-594.

Andrich, D. (1978c). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2, 3,* 449-460.

Andrich, D. (1988). *Rasch models for measurement.* (Sage University Paper Series on Quantitative Applications in the social sciences, no. 07-068). Newbury Park, CA : Sage.

Andrich, D., Lyne, A., Sheridan, B. et Luo, G. (1997). *RUMM : Rasch Unidimensional Measurement Models.* Perth, Australia : RUMM Laboratory.

Angoff, W.H. (1988). Validity : An Evolving Concept. Dans : Wainer, H. et Braun, H.I. *Test Validity.* New Jersey : Lawrence Erlbaum Associates.

Ansley, T.N. & Forsyth, R.A. (1985). An examination of the characteristics of unidimensional I.R.T. parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9,* 37-48.

Baker, F.B. (1987). Methodology review : item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement, 11,* 111-141.

Bartholomew, D.J. (1996). *The Statistical Approach to Social Measurement.* London : Academic Press, Inc.

Bejar, I.I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17,* 283-296.

Berk, R.A. (1982). *Handbook of Methods for Detecting Item Bias.* Baltimore, MD : John Hopkins University Press.

Birnbaum, A. (1968). Some latent trait models. Dans : *Statistical theories of mental test scores.* Reading Mass. : Addison-Wesley.

Blais, J.G. (1987). Effets des la violation du postulat d'unidimensionalité dans la théorie des réponses aux items. Thèse de doctorat non publié. Université de Montréal.

Blais, J.G. & Ajar, D. (1992). Théorie des réponses aux items et modélisation. *Mesure et Évaluation en Éducation, 14,* 5-18.

Blais, J.G. & Laurier, M.D. (1997). La détermination de l'unidimensionalité de l'ensemble des scores à un test. *Mesure et Évaluation en Éducation, 20, 1,* 65 -90.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bock, R.D. et Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters : application of an EM algorithm. *Psychometrika, 46,* 443-459.

Bock, R.D., Gibbons, R. et Muraki, E.J. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12,* 261-280.

Bond, T.G., Fox, C.M. (2001). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences.* Mahwah, New Jersey : Lawrence Erlbaum Associates.

Breithaupt, K. et Zumbo, B.D. (2000). *Testing for Sample Invariance of IRM Item Parameter Estimates : A case study on seal data.* (Paper No. ESQESS-2000-1). Vancouver, B.C. : University of British Columbia. Edgeworth Laboratory for Quantative Educational and Social Science.

Camilli, G. et Shepard, L.A. (1994). *Methods for Identifiying Biased Test Items.* Measurement Methods for the Social Sciences Series. Newbury Park, CA : Sage.

Campbell, N.R. (1928). *An account of the principles of of measurement and calculation.* London : Longmans, Green & Co.

Chang, H.H., Mazzeo, J. et Roussos, L. (1996). Detecting DIF for polytomously scored items : An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 33,* 333-353.

Chen, W.H. & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22,* 265-289.

Dong, H.K. & al. (1983). An empirical investigation of sample-free calibration claim of the Rasch model. Technical Report. Ball Foundation : Research and development of human potential.

Dorans, N.J. et Holland, P.W. (1993). DIF Detection and Description : Mantel-Haenzel and Standardization. Dans : Holland, P.W. et Wainer, H. *Differential Item Functioning*. Hillsdale, New Jersey : Lawrence Erlbaum Associates.

Dorans, N.J. et Kulick, E. (1983). *Assessing Unexpected Differential Item Performance of Female Candidates on SAT and TSWE Forms Administered in December, 1977 : An Application of the Standardized Approach* (ETS Research Report Rep. No. RR-83-9). Princeton, New Jersey : Educational Testing Service.

Dorans, N.J. et Kulick, E. (1986). Demonstrating the Utility of the Standardization Approach to Assessing Unexpected Differential Item Performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355-368.

Doogy-Bogan, E. & Yen, W.M. (1983). Detecting multidimensionalty and examining its effect on vertical equating with the three-parameter logistic model. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Montréal.

Downing, S.M. et Haladyna, T.M. (1997). Test Item Development : Validity Evidence from Quality Assurance Procedures. *Applied Measurement in Education,10,* 61-82.

Drasgow, F & Parsons, C.K. (1983). Application of unidimensional psychological item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189-199.

Embretson, S.E. & Hershberger, S.L. (ed). (1999). The new rules of measurement : What every psychologist and educator should know.

Embretson, S.E. & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, New Jersey : Lawrence Erlbaum Associates. Mahwah, New Jersey : Lawrence Erlbaum Associates.

Engelhard, George Jr. (1992). Historical views of invariance : Evidence from the measurement theories of Thorndike, Thurstone, and Rasch. *Educational and Psychological Measurement, 52, 2,* 275-291.

Fan, X. (1998). Item response theory and classical test theory : an empirical comparison of their item/person characterisitics. *Educational and Psychological Measurement, 58, 3,* 357-381.

Fan, X. (1999). *Assessing the effect of model-data misfit on the invariance property of IRT parameter estimates.* Texte présenté dans le cadre de la rencontre annuelle de l'American Educational Association. Montréal, Canada.

Fishbein, M. (1967). Attitude and the Prediction of Behavior. Dans : *Readings in Attitude Theory and Measurement*. New-York : John Wiley & Sons, Inc.

Fisher, G. H. (1995). Derivations of the Rasch model. Dans : G.H. Fisher et I.W. Molenaar (eds), *Rasch models : foundations, recent developments, and applications*. New-York : Springer-Verlag.

Fisher, G.H. (1998). *LPCM-WIN*. Minneapolis, MN : Assessment Systems Corp.

Flowers, C.P., Oshima, T.C. et Raju, N.S. (1999), A Description and Demonstration of the Polytomous-DIFT Framework. *Applied Psychological Measurement, 23, 4,* 309-326.

Fraser, C. (1988). NOHARM II : A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, N.S.W. : Université de Nouvelle-Angleterre, Centre des Études Behaviorales.

Gustafsson, J.E. (1981). *An introduction to Rasch's measurement model.* Princeton, N.J. : Eric Clearinghouse on Tests, Measurements, and Evaluation.

Haladyna, T.M. et Downing, S.M. (1989). A Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education, 2,* 37-50.

Hambleton, R.K. et Traub, R.E. (1973). Ananlysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology, 26,*273-281.

Hambleton, R.K. et Swaminathan, H. (1985). *Item response theory : principles and applications.* Boston : Klwer Nijhoff Publishing.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory.* Measurement Methods for the Social Sciences Series. Newbury Park, CA : Sage.

Hambleton, R.K. et Rogers, H.J. (1989). Detecting Potentially Biased Test Items : Compararison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education, 24,4,* 313-334.

Hattie, J.A. (1985). Methodology review : assessing unidimensionality of test and items. *Applied psychological measurement, 9,* 139-164.

Henning, G. (1988). The influance of test and sample dimensionality on latent trait person ability and item difficulty calibration. *Language Testing, 5,* 83-99.

Holland, P.W. et Thayer, D.T. (1988). Differential Item Performance and the Mantel-Haenzel Procedure. Dans : Wainer, H. et Braun, H.I. *Test Validity.* New Jersey : Lawrence Erlbaum Associates.

Humphreys, L. (1984). A theoritical and empirical study of the psychometric assessment of psychological test dimensionality and bias (ONR Research Proposal). Washington, DC : Office of Naval Research.

Karabatsos, G. (2000). A critique of the Rasch residual fit statistics. *Journal of Applied Measurement, 1(2),* 152-176.

Lane, D.S. et al. (1987). The Effects of Knowledge of Item Arrangement, Gender, and Statistical and Cognitive Item Difficulty on Test Performance. *Educational and Psychological Measurement, 47,* 865-879.

Laurencelle, L. (1998). *Théorie et techniques de la mesure instrumentale.* Québec : PUQ.

Laveault, D. et Grégoire, J. (1997). *Introduction aux théories des tests en sciences humaines*. Bruxelles : DeBoeck Université.

Levine, M. et Williams, B. (1991). *Non-parametric item response function estimation strategies*. Texte présenté dans le cadre de l'ONR Conference on Model-Based Measurement, Princeton, New-Jersey.

Linn, R.L. (1990). Has Item Response Theory Increased the Validity and Achievement Test Scores ? *Applied Measurement in Education,3,* 115-141.

Linn, R.L. (1993). The Use of Differential Item Functioning Statistics : A Discussion of Current Practice and Future Implications. Dans : Holland, P.W. et Wainer, H. *Differential Item Functioning*. Hillsdale, New Jersey : Lawrence Erlbaum Associates.

Linacre, J.M. (1995). *The effect of misfit on measurement*. Texte présenté à l'occasion de la Eight International Objective Measurement Workshop, Berkeley, California.

Linacre, J. M. (1999). Understanding Rasch measurement : estimation methods for Rasch measures. *Journal of Outcome Measurement, 3,* 382-405.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph, 7.*

Lord, F.M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, Mass. : Addison-Wesley.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. New Jersey : Erlbaum Associates.

Luce, R.D. & Tukey, J.W. (1964). Simultaneous conjoint meaurement : A new type of fundamental measurement. *Journal of Mathematical Psychology, 1,* 1-27.

Ludlow, L.H. et Haley, S.M. (1995). Rasch model logits : interpretation, use, and transformation. *Educational and Psychological Measurement, 55,* 967-975.

Maranon, P.P., Barbero Garcia, M.I. et Costas, C.S, (1997). Identification of Nonuniform Differential Item Functioning : A Comparison of Mantel-Haenszel and Item Response Theory Ananlysis Procedures. *Educational and Psychological Measurement, 57, 4,* 559-568.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Masters, G.N. et Wright, B.D. (1996). The partial credit model. Dans : van der Linden, W.J. et Hambleton, R.K. (1996). *Handbook of modern item response theory*. New York : Springer.

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34,* 100-117.

McDonald, R.P. & Mok, M.M.C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research, 30,* 23-40.

McKinley, R.L. et Reckase, M.D. (1982). *The use of a general Rasch model with mutidimensional item response.* Research report 82-1. Iowa City, IA : The American College Testing Program.

Meijer, R.R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement,18,* 311-314.

Meijer, R.R. et Sijtsma, K. (1999). A review of methods for evaluating the fit of item score patterns on a test. Research Report 99-01. AE Enschede, the Netherlands : Faculty of Educational Science and Technology, Université de Twente.

Messick, S. (1989). Validity. Dans : Linn, R.L. (Ed.). *Educational Measurement.* New-York : Macmillan.

Messick, S. (1993). *Foundations of Validity : Meaning and Consequences in Psychological Assessment.* Research Report. New Jersey : Educational Testing Service.

Miller, T.R. et al. (1992). *A Comparison of Three Methods for Identifying Non-Uniform DIF in Polytomously Scored Test Items.* Texte présenté dans le cadre de la rencontre de la Psychometric Society Meeting, Colombus OH.

Miller, T.R. et Spray, J.A. (1993). Logistic Discriminant Function Analysis for DIF Identification of Polytomously Scored Items. *Journal of Educational Measurement, 30,* 107-122.

Millsap, R.E. et Everson, H.T. (1993). Methodology Review : Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement, 17, 4,* 297-334.

Mislevy, R.J. et Bock, R. D. (1990). *BILOG-3; Item analysis and test scoring with binary logistic models* [Programme informatique]. Mooresville, IN : Scientific Software.

Molenaar, I.W. (1995). Estimation of item parameters. Dans : G.H. Fisher et I.W. Molenaar (eds), *Rasch models : foundations, recent developments, and applications.* New-York : Springer-Verlag.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59-71.

Muraki, E. (1992). A generalized partial credit model : Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Muraki, E. (1993). *Implementing Item Parameter Drift and Bias in Polytomously Item Response Models.* Texte présenté dans le cadre de la rencontre annuelle du National Council on Measurement in Education, Atalanta GA.

Muraki, E. et Bock, R.D. (1993). *PARSCALE : IRT based test scoring and item analysis for graded open-ended exercises and performance tasks.* Chigago : Scientific Software Int.

Newman, D.L. et al. (1988). Effect of Varying Item Order on Multiple-Choice Test Scores : Importance of Statistical and Cognitive Difficulty. *Applied Measurement in Education, 1,* 89-97.

Osterlind, S.J. (1983). *Test Item Bias*. (Sage University Paper Series on Quantitative Applications in the social sciences, no. 07-030). Beverly Hills, CA : Sage.

Özçelik, D.A. & Berboglu, G. (1991). Contributions of the Rasch model to objectivity in measurement. *Studies in Educational Evaluation, 17,* 167-188.

Potenza, M.T. et Dorans, N.J. (1995). DIF Assessment for Polytomously Scored Items : A Framework for Classification and Evaluation. *Applied Psychological Measurement, 19, 1,* 23-37.

Rasch, G. (1960). Probabilistic model for some intelligence and attainment tests. Copenhaugen : Danish Institute for Educational Research.

Raju, N.S., Drasgow, F. et Slinde, J.A. (1993). An empirical Comparison of the Area Methods, Lord's Chi-Square Test, and the Manytel-Haenszel technique for Assessing Differential Item Functioning. *Educational and Psychological Measurement, 53,* 301-314.

Resckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests : results and implications. *Journal of Educational Statistics, 4,* 207-230.

Roskam, E. et Jansen, P.G.W. (1984). A new derivation of the Rasch model. Dans : G.H. Fisher et I.W. Molenaar (eds), *Rasch models : foundations, recent developments, and applications*. New-York : Springer-Verlag.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, N.17.*

Samejima, F. (1994). Nonparametric estimation of the plausibility functions of the distractors of vocabulary test items. *Applied Psychological Measurement, 18,* 35-51.

Samejima, F. (1996). The graded responde model. Dans : van der Linden, W.J. et Hambleton, R.K. (1996). *Handbook of modern item response theory*. New York : Springer.

Schuman, H. et Presser, S. (1996). *Questions and Answers in Attitude Surveys. Expirements on Question Form, Wording, and Context.* San Diego, CA : Sage Publications.

Shealy, R.T. et Stout, W.F. (1993). An Item Response Theory Model for Test Bias and Differential Item Functioning. Dans : Holland, P.W. et Wainer, H. *Differential Item Functioning*. Hillsdale, New Jersey : Lawrence Erlbaum Associates.

Sijtsma, K. et Junker, B.W. (1996). A survey of theory and methods for invariant item ordering. *British Journal of Mathematical Psychology,49,* 79-105.

Sijtsma, K. et Hemker, B.T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika, 63,* 183-209.

Sijtsma, K. et Hemker, B.T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics, 25,* 391-415.

Sireci, S.G. , Thissen, D. & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

Smith, R.M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1(2)*, 199-218.

Smith, R.M., Schumacker, R.E. & Bush, M.J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2(1)*, 66-78.

Shepard, L.A., Camilli, G. & Williams, D.M. (1984).Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*, 93-128.

Spector, P.E. (1992). *Summated rating scale construction : an introduction.* (Sage University Paper Series on Quantitative Applications in the social sciences, no. 07-082). Newbury Park, CA : Sage.

Stout, W. (1987). A non-parametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 55*, 293-326.

Stout, W., Nandakumar, R., Junker, B., Chang, H. & Steidinger, D. (1991). DIMTEST and TESTSIM, programs for dimensionality testing and test simulation. Université de l'Illinois à Urbana-Champaign, Département de Statistique.

Thissen, D. (1991). *MULTILOG user's guide : Multiple categorical item analysis and test scoring using item response theory.* Chigago : Scientific SoftwareInt.

Thissen, D. et Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.

Thissen, D. et Steinberg, L. (1986). A toxonomy of item response models. *Psychometrika, 51*, 567-577.

Thissen, D., Steinberg, L. et Wainer, H. (1988). Use of Item Response Theory in the Study of Group Differences in Trace Lines. Dans : Wainer, H. et Braun, H.I. *Test Validity.* New Jersey : Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L. et Wainer, H. (1993). Dectection of Differential Item Functioning Using the Parameters of Item Response Models. Dans : Holland, P.W. et Wainer, H. *Differential Item Functioning.* Hillsdale, New Jersey : Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L. et Mooney, J.A. (1989). Trace lines for test-lets : a use of multiple-categorical-response models. *Journal of Educational Measurement, 26*, 247-260.

Thompson, T.D. & Pommerich, M. (1996). Examining the sources and effects of local dependence. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, New-York , NY.

van der Linden, W.J. (1994). Fundamental measurement and the fundamentals of Rasch measurement. Dans : Wilson, M. (ed.). Objective Measurement : Theory into Practice Vol.2. Ablex Publishing Corp. ISBN : 0-89381-843-1.

van der Linden, W.J. et Hambleton, R.K. (1996). *Handbook of modern item response theory.* New York : Springer.

Van der Maren, J.M. (1999). *Recherche appliquée en éducation.* Bruxelles : DeBoeck Université.

Wainer, H. et Thissen, D (2001). *Test Scoring.* United States : Lawrence Erlbaum Associates.

Wainer, H. et Thissen, D. (2001). True score theory : the traditional method. Dans : Wainer, H. et Thissen, D. *Test Scoring.* United States : Lawrence Erlbaum Associates.

Whitely, S.E. & Dawis, R.V. (1974). The nature of objectivity with the Rasch model. *Journal of Educational measurement, 11, 2,* 163-178.

Wilkie, C.J. (1999). Case Study : The Process of Constructing and Validating an Attitude Survey. *Research and Teaching in Developmental Education, 15,2,* 65-78.

Wilson, D., Wood, R. & Gibbons, R.D. (1987). *TESTFACT : Test scoring, item statistics and factor analysis.* Mooresville, Dans : Scientific Software, Inc.

Wingersly, M.S., Barton, M.A. et Lord, F.M. (1982). *LOGIST user's guide.* Princeton, N.J. : Educational Testing Service.

Wright, B.D. et Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23-48.

Wright, B.D. et Stone, M.H. (1979). *Best test design : Rasch measurement.* Chigago : Mesa Press.

Wright, B.D. et Masters, G.N. (1982). *Rating scale analysis.* Chicago : MESA Press.

Wright, B.D. et Linacre, J.M. (1991). *Winsteps Rasch Measurement Computer Program.* Chigago : MESA Press.

Wu, M.L., Adams, R.J., Wilson, M.R. (1998). *Conquest : Generalised Item Response Modeling Software.* Melbourne, Australia : Australian Council for Educational Research.

Yan, J.W. (1997). Examining local item dependence effects in a large-scale science assessment by a Rasch partial credit model. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Chigaco.

Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

Yen, W.M. (1993). Scaling performance assessments : strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

Zieky, M. (1993). Practical Questions in the Use of DIF Statistics in Test Development. Dans : Holland, P.W. et Wainer, H. *Differential Item Functioning*. Hillsdale, New Jersey : Lawrence Erlbaum Associates.

Zumbo, B.D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF) : Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores.* Ottawa, ON : Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Donoghue, J.R. et Grima, A. (1993). Assessment of Differential Item Functioning for Performance Tasks. *Journal of Educational Measurement, 30,* 233-251.

Zwick, R. et Thayer, D.T. (1996). Evaluating the Magnitude of Differential Item Functioning in Polytomous Items. *Journal of Educational and Behavioral Statistics, 21,* 187-201.

Zwick, R., Thayer, D.T. et Mazzeo, J. (1997). *Describing and Categorizing DIF in Polytomous Items : Final Report* (Research Report No. 97-05, GRE Report No. 93-10P). Princeton, New-Jersey : Educational Testing Service.

Zwinderman, A.H. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement, 19,* 369-375.

APPENDICE A

VERSIONS DU QUESTIONNAIRE D'ENQUÊTE SUR L'ATTITUDE

Cet appendice présente un exemplaire du questionnaire utilisé dans le cadre de cette recherche. Bien que seule la deuxième section ait fait l'objet de notre étude, le questionnaire en entier est présenté. Les deux versions du questionnaire utilisé pour cette recherche ne diffèrent que par l'ordre de présentation des items à la section II PRÉPARATION À L'ENSEIGNEMENT. Cet appendice ne présente pas les deux versions, mais se contente plutôt d'insérer la deuxième version de la section 2 à la suite de sa première version. Donc, à la page 114, il y a la première version de la section 2 de l'instrument. Elle est identifiée VERSION A dans la partie supérieure droite; et à la page 115, il y a la deuxième version identifiée VERSION B avec l'ordre de présentation des items modifié à la lumière de l'analyse faite à la section 4.2.1 du chapitre 4.

QUESTIONNAIRE D'ENQUÊTE
AUPRÈS DES ÉTUDIANTS ET DES ÉTUDIANTES
DE LA FORMATION INITIALE DES MAÎTRES

3e et 4e années

Nous nous adressons à vous en tant qu'étudiants ou étudiantes du Centre de formation initiale des maîtres (CFIM) de la Faculté des sciences de l'éducation. Nous cherchons à recueillir des informations sur votre programme d'études. Votre participation à cette enquête est primordiale pour les personnes touchées par la qualité de la formation en enseignement.

Nous vous remercions de bien vouloir répondre à ce questionnaire d'évaluation; il vous suffira de dix à quinze minutes.

Il n'y a pas de bonnes ou de mauvaises réponses; l'important est de répondre en fonction de votre expérience personnelle. Nous vous assurons que vos réponses ne seront utilisées qu'aux fins de cette étude et qu'elles seront traitées dans la plus grande confidentialité.

QUELQUES DIRECTIVES AVANT DE RÉPONDRE AU QUESTIONNAIRE

1) Pour chaque question, répondez :

- en encerclant le chiffre qui correspond à la réponse que vous choisissez,
- ou en écrivant une brève réponse, s'il y a lieu.

2) À moins d'avis contraire, n'indiquez qu'une seule réponse par question.

NOUS VOUS REMERCIONS DE VOTRE COLLABORATION.

Pour de plus amples informations sur la présente enquête, veuillez communiquer avec :

Monsieur Jean-Guy Blais, vice-doyen à la gestion et au développement	343-7844
Monsieur Michel Thérien, directeur du CFIM et vice-doyen aux études de 1er cycle	343-6652

A.1. PERCEPTION GÉNÉRALE DE LA FORMATION

Indiquez votre degré d'accord ou de désaccord avec les énoncés suivants :

		Tout à fait d'accord	Plutôt d'accord	Plutôt en désaccord	Tout à fait en désaccord
1.	De manière générale, je suis satisfait(e) de l'ensemble de la formation (cours, encadrement, stages).	1	2	3	4
2.	De manière générale, je suis satisfait(e) des cours que j'ai suivis à la Faculté des sciences de l'éducation.	1	2	3	4
3.	De manière générale, je suis satisfait(e) du support pédagogique que j'ai reçu de mes professeurs à la Faculté des sciences de l'éducation.	1	2	3	4
4.	Mon programme d'études m'a amené(e) à faire des liens entre la pédagogie, la didactique et la pratique.	1	2	3	4
5.	Mon programme d'études m'a permis de prendre contact avec les contenus que j'aurai à enseigner.	1	2	3	4
6.	Mon programme d'études m'a permis d'examiner des façons d'établir des relations avec les parents.	1	2	3	4
7.	Mon programme d'études m'a permis d'enrichir ma culture générale.	1	2	3	4
8.	Mon programme d'études m'a sensibilisé(e) à l'importance de l'éthique professionnelle.	1	2	3	4
9.	Mon programme d'études m'a sensibilisé(e) à l'importance de la formation continue.	1	2	3	4
10.	Mon programme d'études a répondu à mes attentes de futur enseignant ou enseignante.	1	2	3	4

A.2. PRÉPARATION À L'ENSEIGNEMENT (VERSION A)

Cette section a pour objet de connaître ce que vous pensez de votre degré de préparation pour la réalisation de certaines tâches. Vous devez répondre en tenant compte de la phrase d'introduction suivante :
Je considère que mon programme d'études m'a préparé(e) adéquatement pour .....

		Tout à fait d'accord	Plutôt d'accord	Plutôt en désaccord	Tout à fait en désaccord
11.	Identifier les contenus difficiles à faire apprendre aux élèves	1	2	3	4
12.	Répondre aux demandes des parents	1	2	3	4
13.	Construire des outils pour l'évaluation sommative (contrôles, examens, etc.)	1	2	3	4
14.	Maîtriser les contenus que j'enseignerai en conformité avec les programmes du ministère de l'Éducation	1	2	3	4
15.	Corriger la langue écrite des élèves	1	2	3	4
16.	Corriger la langue orale des élèves	1	2	3	4
17.	Intervenir individuellement auprès des élèves à risque d'échouer	1	2	3	4
18.	Planifier le déroulement d'activités d'apprentissage	1	2	3	4
19.	Adapter mes activités d'enseignement aux caractéristiques des élèves	1	2	3	4
20.	Établir les règles de fonctionnement de la classe	1	2	3	4
21.	Motiver les élèves à s'engager dans leur apprentissage	1	2	3	4
22.	Respecter les différences ethniques ou culturelles des élèves	1	2	3	4
23.	Identifier les points forts et les points faibles des élèves	1	2	3	4
24.	Établir des relations avec les membres de l'équipe-école	1	2	3	4
25.	Construire des outils pour l'évaluation formative (exercices, devoirs, etc.)	1	2	3	4
26.	Aider les élèves à développer leurs méthodes de travail	1	2	3	4
27.	Sanctionner les problèmes de discipline chez les élèves	1	2	3	4
28.	Sensibiliser les élèves aux situations de discrimination qui existent entre eux	1	2	3	4
29.	Orienter les élèves vers les services d'aide appropriés	1	2	3	4
30.	Établir des relations avec les parents	1	2	3	4

A.3. PRÉPARATION À L'ENSEIGNEMENT (VERSION B)

Cette section a pour objet de connaître ce que vous pensez de votre degré de préparation pour la réalisation de certaines tâches. Vous devez répondre en tenant compte de la phrase d'introduction suivante :

Je considère que mon programme d'études m'a préparé(e) adéquatement pour .....

		Tout à fait d'accord	Plutôt d'accord	Plutôt en désaccord	Tout à fait en désaccord
11.	Répondre aux demandes des parents	1	2	3	4
12.	Établir des relations avec les parents	1	2	3	4
13.	Orienter les élèves vers les services d'aide appropriés	1	2	3	4
14.	Identifier les contenus difficiles à faire apprendre aux élèves	1	2	3	4
15.	Corriger la langue écrite des élèves	1	2	3	4
16.	Corriger la langue orale des élèves	1	2	3	4
17.	Établir des relations avec les membres de l'équipe-école	1	2	3	4
18.	Intervenir individuellement auprès des élèves à risque d'échouer	1	2	3	4
19.	Sensibiliser les élèves aux situations de discrimination qui existent entre eux	1	2	3	4
20.	Maîtriser les contenus que j'enseignerai en conformité avec les programmes du ministère de l'Éducation	1	2	3	4
21.	Aider les élèves à développer leurs méthodes de travail	1	2	3	4
22.	Sanctionner les problèmes de discipline chez les élèves	1	2	3	4
23.	Construire des outils pour l'évaluation formative (exercices, devoirs, etc.)	1	2	3	4
24.	Construire des outils pour l'évaluation sommative (contrôles, examens, etc.)	1	2	3	4
25.	Identifier les points forts et les points faibles des élèves	1	2	3	4
26.	Motiver les élèves à s'engager dans leur apprentissage	1	2	3	4
27.	Adapter mes activités d'enseignement aux caractéristiques des élèves	1	2	3	4
28.	Établir les règles de fonctionnement de la classe	1	2	3	4
29.	Respecter les différences ethniques ou culturelles des Élèves	1	2	3	4

30.	Planifier le déroulement d'activités d'apprentissage	1	2	3	4

A.4. LES STAGES

Indiquez votre degré d'accord ou de désaccord avec les énoncés suivants :

		Tout à fait d'accord	Plutôt d'accord	Plutôt en désaccord	Tout à fait en désaccord
31.	Je suis satisfait(e) de la façon dont mes stages ont été organisés (choix de l'école, gestion des dossiers, etc.)	1	2	3	4
32.	Je suis satisfait(e) de l'accueil que j'ai reçu dans mes milieux de stage.	1	2	3	4
33.	Je suis satisfait(e) de l'encadrement de mes maîtres associés.	1	2	3	4
34.	Je suis satisfait(e) de l'encadrement de mes superviseurs de stage à l'université.	1	2	3	4
35.	Je suis satisfait(e) de mes apprentissages d'enseignant lors de mes stages.	1	2	3	4
36.	Je suis satisfait(e) de la façon dont mes stages ont été évalués (évaluation globale).	1	2	3	4
37.	Mes stages m'ont permis de transférer le contenu de mes cours dans la pratique.	1	2	3	4
38.	Mon programme d'études m'a bien préparé(e) à effectuer mes stages.	1	2	3	4
39.	Dans l'ensemble, mes stages étaient placés à des moments opportuns de ma formation.	1	2	3	4
40.	Le nombre d'heures consacrées aux activités de stage était suffisamment élevé.	1	2	3	4

A.5. ÉTUDES ET GESTION DU TEMPS

Indiquez votre degré d'accord ou de désaccord avec les énoncés suivants :

Durant l'année académique 2000-2001, ...

		Tout à fait d'accord	Plutôt d'accord	Plutôt en désaccord	Tout à fait en désaccord
41.	J'ai eu l'impression d'être dépassé(e) par ma charge de travail d'étudiant(e).	1	2	3	4
42.	J'ai eu du temps pour échanger avec des collègues étudiant(e)s.	1	2	3	4
43.	J'ai manqué de temps pour ma vie personnelle.	1	2	3	4
44.	J'aurais pu obtenir de meilleurs résultats si j'avais consacré plus de temps à mes études.	1	2	3	4
45.	Je donne la priorité à mes études en tout temps.	1	2	3	4

46. Durant l'année académique 2000-2001, avez-vous occupé un ou des emplois?

Oui _____ 1

Non _____ 2

47. Si oui, quel(s) emploi(s) avez vous occupé(s)?

_____

_____

_____

_____

48. Si vous avez occupé un ou des emplois durant les trimestres d'automne et d'hiver, quel nombre d'heures (en moyenne) avez-vous travaillé par semaine?

- 1 à 5 hres/sem. 1
- 6 à 10 hres/sem. 2
- 11 à 15 hres/sem. 3
- plus de 15 hres/sem. 4

A.6. POURSUITE DES ÉTUDES

Avez-vous pensé à poursuivre vos études au 2e cycle en éducation après avoir complété votre baccalauréat?

49. *Encercler une des réponses proposées.*

1. Oui, je continuerai mes études immédiatement après le baccalauréat.

2. Oui, j'aimerais poursuivre immédiatement mais ma situation personnelle ne me le permet pas (emploi, famille, etc.).

3. Oui, mais pas immédiatement. Je préfère prendre quelques années d'expérience (2 à 3) avant de poursuivre mes études.

4. Oui, mais plus tard, dans cinq ou six ans.

5. Non, à l'heure actuelle je ne pense pas du tout à poursuivre mes études au 2e cycle en éducation.

50. Quelles sont les raisons qui motivent votre réponse à la question précédente ?

_____

_____

_____

_____

_____

A.7. COMMENTAIRES SUR VOTRE PROGRAMME D'ÉTUDES

51. Les quelques lignes qui suivent vous permettent d'exprimer votre opinion sur différentes facettes de votre programme d'études. Vous pouvez aborder les aspects de votre choix, par exemple, les améliorations qui pourraient être apportées à la formation et au programme lui-même.

A.8. RENSEIGNEMENTS GÉNÉRAUX

52. Votre âge _____ ans

53. Votre sexe
Féminin 1
Masculin 2

٥٤. Votre programme d'études

Baccalauréat en enseignement secondaire _____ 1
Baccalauréat en éducation préscolaire et enseignement primaire _____ 2
Baccalauréat en orthopédagogie _____ 3
Baccalauréat en enseignement de l'éducation physique et santé _____ 4
Baccalauréat en enseignement du français langue seconde _____ 5

55. En quelle année êtes-vous? Troisième année 1
Quatrième année 2

٧. VOLONTAIRES RECHERCHÉ(E)S

56. Afin de compléter les informations recueillies à l'aide de ce questionnaire, nous pourrions faire appel à des volontaires pour participer à des entrevues de groupe. Si vous êtes intéressé(e) à participer à des échanges sur votre formation et votre programme d'études, nous vous prions de bien vouloir nous laisser vos coordonnées.

Nom : _____

Numéro de téléphone :
(incluant le code régional) _____

Courriel : _____

MERCI BEAUCOUP DE VOTRE COLLABORATION.

APPENDICE B

DIRECTIVES DONNÉES AUX SUPERVISEURS DES SÉMINAIRES

B.1. FEUILLE DE DIRECTIVE REMISE AU SUPERVISEUR DES SÉMINAIRES

INFORMATION SUR LA DISTRIBUTION DES QUESTIONNAIRES

Les questionnaires dans cette enveloppe ont pour but d'obtenir des informations auprès des étudiants sur les différents programmes du CFIM. Il est très important que tous les étudiants remplissent ce questionnaire. La plupart d'entre eux ont déjà rempli ce questionnaire l'an dernier et sont familiers avec la procédure. En conséquence, il n'est pas nécessaire pour vous d'expliquer longuement les motifs de cette enquête. L'information essentielle se retrouve sur la première page du questionnaire. Les étudiants auront besoin d'environ 15 minutes selon la longueur des commentaires écrits qu'ils fourniront.

PROCÉDURES À SUIVRE

- Distribuer les questionnaires aux étudiants au moment où vous le jugerez opportun;
- Donnez le temps nécessaire pour remplir le questionnaire;
- Recueillir les questionnaires;
- Insérez tous les questionnaires dans l'enveloppe;
- Remettre les questionnaires au local A-223.7 dans la boîte prévue à cet effet.

Nous vous remercions grandement de votre participation.