Table des matières

Liste des tableaux :

Liste des Graphiques :

À ma famille, pour mes nombreuses heures d'absence durant cette scolarité.

Je voudrais exprimer toute ma reconnaissance à mon Directeur de recherche **Jean-Guy Blais**. Grâce à ses fonds de recherche, il m'a permis d'acquérir une license du logiciel Conquest v.3 (Wu, Adams et Wilson, 1997) qui a servi à la fois pour la simulation et la modélisation, et par sa disponibilité j'ai pu avancer dans ce travail.

Mes sincères remerciements à **Nathalie Loye** et **Michel Desmarais**, membres du jury. Leurs rétroactions ont contribué à l'amélioration de ce travail.

Mes sincères remerciements à tout le personnel du Département d'administration et fondement en éducation.

Contexte et problématique

La mesure est au centre des activités quotidiennes, mais avec des exigences particulières selon qu'elle se fait en sciences de la nature ou en sciences humaines. Dans les sciences exactes, les caractéristiques physiques des objets sont mesurées avec des instruments de mesure étalonnés, et une précision généralement connue à l'avance. Les mesures obtenues à différents lieux géographiques sont comparables puisque les instruments de mesure ainsi que les unités de mesure sont standards et ne souffrent d'aucune contestation.

Dans les sciences humaines, à l'exemple des sciences de l'éducation, l'on désire mesurer des caractéristiques inobservables chez un apprenant, appelées traits latents. Wagner et al. (2010) définissent un trait latent comme étant une variable qui ne peut pas être directement observée, et dont la présence ne peut qu'être détectée par son effet sur des variables observables.

Par exemple, on peut vouloir mesurer les connaissances à la suite des apprentissages pour les besoins de régulation ou de la certification, la sélection des apprenants dans certaines spécialités académiques, la promotion scolaire d'un niveau à un autre, etc. L'outil de mesure prend alors la forme d'un test ou d'un questionnaire, constitué d'un ensemble d'items. C'est donc à travers la manifestation extérieure sous la forme de la réponse ou de choix de l'apprenant aux questions qu'on lui pose que l'on mesure le trait latent.

Le contenu du test est en général adapté aux exigences de chaque société. Par exemple, les compétences exigées en matière de maîtrise de l'outil informatique chez l'enseignant dans un pays donné est fonction de la disponibilité de cet outil dans les salles de classe et de son degré d'intégration pédagogique dans le système éducatif national.

En plus du trait qui influence les choix ou les réponses de l'apprenant, il y a d'autres facteurs qui sont en interaction pendant le processus de mesure : ce sont entre autres le contenu de l'outil de mesure et son support (papier, ordinateur, etc.), les conditions de son administration, les habitudes et l'état d'esprit du répondant. L'on peut donc constater qu'à l'exception du niveau des traits de chacun des apprenants, tous les autres facteurs qui peuvent déterminer leur performance sont supposés avoir la même influence chez tous les apprenants, mais cela n'est pas toujours vrai car tous ne vivent pas les mêmes expériences quotidiennes.

Par exemple, les candidats qui ne sont pas familiers avec l'ordinateur peuvent avoir des performances moins bonnes lorsque le test est administré par ordinateur comparativement à ce que seraient ces performances si celui-ci avait été administré sur un support papier.

Il ressort donc de ce constat qu'il n'est pas toujours certain que l'on mesure seulement le trait latent visé ou que seul celui-ci détermine les choix des apprenants lorsqu'ils sont évalués. Mais, par pragmatisme, habitude ou par souci de parcimonie et de clarté, il est la plupart du temps considéré que l'on mesure un seul trait à la fois.

Le nombre de trait latents permettant d'expliquer le patron des réponses ou des choix des répondants est un indicateur de la dimension du test. Le concept de dimension réfère au nombre minimal de facteurs qui permettent d'expliquer les variations dans les données. En considérant que seuls les traits latents permettent d'expliquer ces variations, le nombre de traits mesurés est donc la dimension du test. Wang, Wilson et Adams (1996) ont traité la question de la dimensionnalité à deux niveaux dont celui de l'item et celui de l'outil de mesure (test ou questionnaire).

Lorsque l'outil de mesure comprend plusieurs sous-ensembles d'items qui sont supposés mesurer chacun un et un seul trait différent chez le candidat, on dira de l'outil qu'il est multidimensionnel. Sinon, il est unidimensionnel lorsque tous les items ne servent à mesurer qu'un seul trait. En rapport avec un item, celui-ci peut être unidimensionnel lorsqu'il ne sert qu'à la mesure d'un seul trait, et que la réponse à cet item n'est déterminée que par le niveau de présence de ce trait chez le répondant. Par contre, il est dit multidimensionnel lorsqu'il sert à la mesure de plusieurs traits à la fois, et que la réponse à cet item est déterminée par la présence et le niveau de ces traits chez le répondant.

Idéalement, l'on aimerait travailler avec un outil unidimensionnel et des items unidimensionnels, c'est-à dire permettant de mesurer un trait unique. Les trois raisons principales de cette préférence sont les suivantes: En premier lieu, il est plus facile de placer les scores à un test sur un même continuum pour permettre une meilleure appréciation des décisions à prendre sur la base du classement des répondants. Ce classement est rendu possible par le score au test que ces derniers obtiennent. Ce qui n'est pas facile à percevoir lorsque les individus sont placés dans un espace à plusieurs dimensions. En effet, il est difficile de trouver un indice combinant les différences observées entre eux sur chacune des dimensions du test, et permettant de les classer les uns après les autres. En deuxième lieu, le principe de parcimonie commande que l'on

choisisse comme modèle approprié pour mimer les données, celui qui est simple et qui produit des résultats adéquats. En troisième lieu, les modèles implémentés dans les logiciels courants d'analyse des données d'items sont basés sur l'hypothèse de l'unidimensionnalité. En plus, l'interprétation des modèles multidimensionnels est complexe et pourrait s'avérer difficilement utilisable pour le preneur de décisions.

Il existe des modèles unidimensionnels et multidimensionnels de mesure, parmi lesquels ceux de la famille de Rasch. On trouve dans cette famille trois modèles unidimensionnels qui sont appropriés pour les items dichotomiques dont les modèles de Rasch, binomial et de poisson et trois autres pour des items polytomiques à options ordonnées dont le modèle à échelle de notations (Andrich, 1978) qui est approprié lorsque tous les items du test ont le même nombre d'options ordonnées; le modèle du crédit partiel (Wright et Masters, 1982) qui s'applique lorsque le nombre d'options des items peut varier d'un item à l'autre et le modèle des préférences (Linacre, 1994) lorsque les choix sont ordonnés par ordre de préférence du candidat.

Les données sont souvent multidimensionnelles à cause des multiples éléments qui façonnent nos manières de répondre à des questions, mais les modèles de mesure que l'on utilise sont souvent unidimensionnels parce qu'on les connaît mieux, ils sont plus simples à utiliser, les interprétations sont plus facilement compréhensibles. Cela veut dire que l'on utilise des modélisations qui ne sont peut-être pas toujours bien adaptées à nos données. Cette étude vise ainsi à étudier la qualité des modélisations issues de l'application d'un modèle de mesure unidimensionnel comme le modèle de Rasch à des données multidimensionnelles, dont le patron ne peut être pourtant expliqué que par un seul facteur.

Question de recherche

Les raisons précédentes sur l'utilisation des modèles unidimensionnels justifient notre recherche dont la question générale est la suivante :

Dans quelles conditions le modèle de Rasch est-il robuste à la violation de l'hypothèse d'unidimensionnalité ?

La robustesse d'un modèle peut se définir comme étant son degré de sensibilité lorsque l'une des hypothèses fondamentales sur lesquelles il est bâti n'est plus respectée. Dans le cadre du modèle

de Rasch unidimensionnel ou du modèle logistique, cette sensibilité peut s'étudier en variant les paramètres items ou en modélisant des données multidimensionnelles avec le modèle de Rasch. Par exemple, on peut appliquer le modèle de Rasch ou le modèle logistique à un paramètre alors même que les items ont des paramètres de discrimination différents, ou alors le modèle logistique à 2 paramètres alors que le paramètre de pseudo-chance n'est pas nul pour tous les items du test. La violation de la propriété de l'invariance des estimés et la variation des paramètres items ont été traitées par Forsyth et al. (1981) et Fons (1986) respectivement.

Notre travail consiste à explorer les conditions d'application des modèles unidimensionnels de mesure en présence de plusieurs traits latents dans les données avec les modèles de Rasch. Le choix des modèles de Rasch est justifié par le fait que leur utilisation s'est très rapidement vulgarisée en sciences humaines au cours des quatre dernières décennies. Nous travaillerons donc sous l'hypothèse selon laquelle le modèle unidimensionnel de Rasch produit de bonnes mesures même lorsque les données sont multidimensionnelles.

En effet dans la pratique, le choix des items n'est pas souvent parfait car, il arrive que des items qui ne servent pas uniquement à mesurer le seul trait latent en étude soient présents dans l'outil de mesure. Il peut aussi arriver que des items qui mesurent un trait tout à fait différent de celui qui est visé fassent partie de l'outil. Cela a pour conséquence la production des mesures biaisées du trait latent en étude, conduisant par conséquent à des prises de décisions erronées parce que le modèle de mesure unidimensionnel utilisé pour produire ces mesures n'est plus approprié, étant donnée la présence de plusieurs traits latents dans les données.

Le chercheur doit faire un choix puisqu'il se pose la question de savoir s'il faudrait systématiquement faire usage des modèles multidimensionnels, mais que faire du principe de parcimonie?

Des études sur cette question dont Drasgow et Parsons (1983) ; Blais (1987) ; Way, Ansley et Forsyth (1988); Ackerman (1991, 1994) ; Cuesta et Muniz (1999) ; Kirisci et al. (2001) ; Walker et Beretvas (2003), puis Kahraman et Thompson (2011) que nous détaillons en revue de littérature n'ont pas de façon spécifique abordé le cas des modèles unidimensionnels de Rasch, en particulier avec des tests et items multidimensionnels.

Ces études ont en commun les deux points suivants:

- L'utilisation des données d'items dichotomiques. Même lorsque les items étaient à choix multiples au départ, les réponses ont été dichotomisées afin de produire les mesures
- L'estimation des paramètres s'est faite avec les variantes du modèle logistique à deux ou trois paramètres.

Une approche suffisamment documentée pour tester l'écart à l'unidimensionnalité avec le modèle de Rasch est basée sur le test du ratio de vraisemblance : le test sur items dichotomiques de Martin-Löf (1973) et ses extensions au modèle à items polytomiques (Kristensen et al., 2002; Kristensen et Kreiner, 2007), mais ce ratio de vraisemblance est une statistique globale, c'est à dire à l'échelle du test entier, donc ne permettant pas de prendre une décision ciblée à une partie du test.

Par contre, les statistiques d'ajustement au modèle de Rasch permettent d'identifier des items problématiques c'est-à-dire qui montrent un comportement différent de celui que prévoit le modèle de Rasch. Cette identification permet au chercheur d'approfondir les investigations en vue du retrait final de tels items du test, ou d'envisager leur reformulation. Cette statistique sera mise à profit dans notre étude car elle est centrale à la décision initiale d'explorer ou non l'existence d'une dimension supplémentaire dans les données.

Nous travaillerons avec les données d'items dichotomiques bidimensionnelles à structure simple c'est-à-dire que chacun des items sature sur une seule dimension, simulées selon un ensemble de situations que l'on peut rencontrer dans la pratique : d'une dimension fortement dominante à deux dimensions équivalentes en termes du nombre d'items. Le modèle de Rasch simple sera notre modèle de mesure, car il est approprié pour les items dichotomiques.

Questions spécifiques

Les questions spécifiques qui nous permettront de répondre à notre question de recherche, lesquelles seront étayées et justifiées au troisième chapitre, sont les suivantes :

i. En rapport avec les paramètres des items, quelle relation existe-t-il entre les valeurs simulées et les valeurs estimées par le modèle de Rasch?
ii. En rapport avec les paramètres des personnes, quelle relation existe-t-il entre les valeurs simulées et les valeurs estimées par le modèle de Rasch?
iii. En se référant sur les statistiques t d'ajustement pour retirer ou non un item du test, le modèle unidimensionnel de Rasch mesure-t-il une des dimensions connues du test?

Ce travail est réparti en cinq chapitres : le chapitre 1 traite du cadre théorique des modèles de la théorie des réponses aux items (TRI) et en particulier les modèles de Rasch. Le chapitre 2 se focalise sur les études empiriques en liaison avec la détection de la dimensionnalité d'un test. Bien que la détection de la dimensionnalité ne soit pas notre thème principal de recherche, il semble utile de consulter quelques recherches sur cette question, étant donné qu'elle reste d'actualité puisque les approches utilisées ne font pas l'unanimité des chercheurs. Le chapitre 3 traite de la robustesse de modèles unidimensionnels d'analyse des réponses aux items dans un premier temps, puis de celle du modèle de Rasch. Après avoir répondu à la question de savoir quel est l'intérêt d'étudier la robustesse des modèles unidimensionnels, le reste du chapitre brosse un état des lieux de cette question sur la base des études empiriques. Le chapitre 4 est dédié à la méthodologie. Il présente le plan de la simulation ainsi que le justificatif des choix qui ont été opérés pour cette simulation. Le chapitre 5 sera consacré aux analyses en vue d'évaluer la robustesse du modèle de Rasch en violation de l'hypothèse d'unidimensionnalité du test.

Chapitre 1:

Cadre théorique des modèles de mesure des réponses aux items

Nous commencerons ce chapitre par un bref exposé sur le concept de mesure d'une part, et sur le lien entre théorie et modèle. Nous présenterons ensuite trois familles de modèle de mesure: la théorie classique des tests (TCT), la théorie de la réponse à l'item (TRI) et la famille des modèles de Rasch. Il sera ensuite présenté successivement la comparaison entre la TCT et la TRI, un bref aperçu des méthodes d'estimation des paramètres dans les modèles, en particulier l'estimation par le maximum de vraisemblance qui est la plus fréquemment utilisée.

1.1. Les concepts de mesure et modèle de mesure en sciences sociales

Selon Bertrand et Blais (2004, p.18-19); il existerait deux définitions de la notion de mesure : la définition classique ou rigoureuse telle qu'utilisée dans les sciences exactes et affinée par Campbell; et la définition souple apparue dans la deuxième moitié du 20^{e} siècle sous l'impulsion des travaux de Thurstone, mais surtout de ceux de Stevens faisant suite aux rapports Ferguson (1938, 1940) qui remettaient en question la mesure en psychologie. La mesure peut être définie comme l'acte d'établir une forme de correspondance entre les observations et un concept théorique (Zeller, 2005). C'est aussi le processus qui consiste à lier les concepts abstraits aux indicateurs empiriques de ces concepts (Carmines & Woods, 2005). Le Petit Larousse Illustré (2012, p. 680) propose huit sens au concept de mesure, dont l'un étant *« l'action d'évaluer une grandeur d'après son rapport avec une grandeur de même espèce, prise comme unité et comme référence*».

Nous retenons la définition du Petit Larousse Illustré dans le cadre de notre travail dans une perspective comparative entre plusieurs grandeurs.

Les deux définitions de Zeller et Carmines & Woods suggèrent qu'il existe une fonction pour lier d'un côté les manifestations observables d'un concept, et de l'autre, le concept abstrait ou théorique encore appelé variable latente (Wagner, Tchatcher, & Piasta, 2010). Une telle fonction, lorsqu'elle existe est une formulation mathématique du comportement du concept théorique en rapport avec les observations et sera considérée comme notre modèle de mesure, à l'exemple des

modèles de la théorie des réponses à l'item et les modèles de Rasch que nous développerons au chapitre suivant.

1.1.1. Les exigences en mesure diffèrent selon le domaine d'application

Le processus de mesure a des exigences particulières selon qu'elle se fait en sciences exactes ou en sciences humaines.

En effet dans les sciences sociales et humaines, il n'existe pas d'instrument de mesure au contenu toujours consensuel pour évaluer les compétences ou les habiletés acquises par les apprenants. Par exemple, c'est la manifestation extérieure de l'habilité chez l'élève que l'on mesure à travers des stimuli élaborés par les spécialistes du domaine en étude. La question du changement à la fois de la définition et des standards en matière de mesure dès lors que le focus est mis sur la condition humaine est d'ailleurs posée par Bond et Fox (2007, p.1). Elle nous situe au cœur de la problématique de la mesure des caractéristiques inobservables et des instruments de mesure en sciences sociales et en sciences de l'éducation en particulier, c'est-à-dire la modélisation en sciences de l'éducation.

1.1.2. Modèles de mesure et dimensionnalité

Un modèle est une représentation simplifiée du monde réel dont le but est entre autre de décrire, prédire, expliquer ou contrôler des systèmes (Grégoire & Laveault, 1997). Il est formulé en rapport avec un cadre théorique général, ce dernier lie les variables observables (comme les scores au test et aux items) avec les variables inobservables et spécifie les détails des relations qui existent entre les concepts théoriques, au regard des hypothèses qui les gouvernent (Hambleton & Russel, 1993). Par exemple, en théorie classique des tests, la relation entre les concepts théoriques score vrai (T), score observé (X), et l'erreur (E), est $X = T + E$.

Les items et les modèles de la théorie de la réponse à l'item qui vérifient les postulats suivants sont ceux avec lesquels l'on souhaite travailler en psychométrie ou en sciences de l'éducation (Hambleton, Swaminathan, & Rogers, 1991; Wright & Mok, 2004):

- ✓ les caractéristiques des items ne dépendent pas du groupe des examinés
- ✓ les scores en rapport avec les aptitudes des candidats sont indépendants du test
- ✓ le modèle est bâti autour de l'item et non du test
- ✓ le modèle ne requiert pas des tests strictement parallèles pour évaluer sa fidélité

- ✓ le modèle génère la mesure de précision de chaque score sur la caractéristique d'intérêt
- ✓ le modèle produit des mesures linéaires.

La question de la modélisation de la mesure dans notre cadre renvoie à la fois à celle de l'outil de collecte de données, et à celle de l'évaluation ou de la détection de la présence du trait latent recherché et son niveau chez l'individu. Il est donc question d'élaborer des outils de collecte et des modèles de mesure appropriés qui répondent aux exigences des sciences sociales.

La dimensionnalité d'un instrument de mesure se réfère au nombre de traits latents dominants (Stout, 1987) censés être mesurés à travers le patron de réponses. La question de la dimensionnalité d'un test a été abordée entre autres par Hattie (1981, 1985); Nandakumar (1994); Blais et Laurier (1995); Linacre (1998); Smith E. (2002); Smith A. et al. (2008); Jasper (2010); Chou et Wang (2010) et Teol et al. (2011). Ces études seront détaillées au chapitre suivant.

Les modèles généralement utilisés dans les sciences sociales sont ceux de la théorie classique des tests (TCT), la théorie de la réponse à l'item (TRI) et les modèles d'analyse des classes latentes (ACL). A cause des limites de la TCT (Hambleton et al. (1991), Wilson et al. (2006) et Magno (2009)), les modèles de la TRI sont les plus utilisés de nos jours. Mais à cause de leur simplicité en termes d'interprétation, les modèles unidimensionnels sont les plus rencontrés dans le domaine appliqué, contrairement aux modèles multidimensionnels.

Nous présentons dans ce chapitre la théorie classique des tests (TCT) et les modèles de la théorie de la réponse à l'item (TRI). Ces modèles ont en commun le fait qu'ils servent à produire des mesures du trait latent sur une échelle continue, à travers ses manifestations extérieures sur d'autres variables. En plus, les modèles de la TRI permettent de produire des estimés des caractéristiques des items tels que l'indice de difficulté, de discrimination ou la pseudo-chance. Les modèles de mesure unidimensionnels et multidimensionnels les plus connus et utilisés de la TRI seront expliqués en détail dans la suite.

1.2. La théorie classique des tests (TCT)

1.2.1. Exposé du modèle de la théorie du score vrai

Les origines de cette théorie remontent aux travaux du psychologiste britannique Charles Spearman, entre 1907 et 1913, et de ceux d'autres auteurs comme Guilford (1936), Gulliksen (1950), Magnuson (1967) et Lord et Novick (1968) qui l'ont affiné dans sa forme actuelle (Crocker et Algina, 1986 ; Grégoire et Laveault, 1997).

Il existe plusieurs modèles de la théorie classique des tests parmi lesquels ceux basés sur la distribution du terme d'erreur, dont le plus connu est celui du score vrai. Dans cette approche, l'intérêt est mis sur le score global de l'individu suite à un test (de Ayala, 2009 ; p.5).

1.2.2. Propriétés du modèle de la TCT

Le modèle de la théorie classique des tests repose sur les propriétés suivantes (Crocker et Algina, 1986 ; Grégoire et Laveault, 1997 ; Bertrand et Blais, 2004 ; de Ayala, 2009) :

i. Le score observé d'un individu est la somme de son score vrai qui est une constante, et un terme d'erreur de mesure de ce dernier. Le score vrai est inobservable et c'est pour cette raison qu'il est estimé par le score observé à partir des réponses aux items.
ii. La moyenne des erreurs de mesure pour un individu est nulle lorsqu'on lui administre le test un très grand nombre de fois.
iii. La valeur attendue ou espérée pour le score observé est le score vrai, comme conséquence de la propriété (ii).
iv. La corrélation est nulle entre le score vrai et l'erreur de mesure pour un individu donné.
v. Lorsque deux tests différents sont administrés aux candidats et si en plus l'on suppose que les scores aux deux tests choisis proviennent de deux distributions indépendantes parmi les scores possibles, alors la corrélation est nulle entre les erreurs aux scores des deux tests.
vi. Les erreurs sur un test donné ne sont pas corrélées avec les scores vrais d'un autre test.
vii. Deux tests sont parallèles si et seulement si leurs scores vrais sont égaux et leurs erreurs de mesures sont égales.
viii. Deux tests sont considérés τ-équivalents lorsque leurs scores vrais diffèrent par une constante additive k : $T_1 = T_2 + k$.

1.2.3. Formulation et discussion

D'après le postulat i, pour un individu donné, son score observé (X) à un test est une somme de deux composantes dont son score vrai (T) et une composante aléatoire (E);

soit alors $X = T + E$ (1.1).

Toutefois, cette composante n'est connue que si T est connu, étant donné X. Or c'est précisément T que l'on souhaite découvrir pour chaque individu, en lui administrant des tests.

L'équation (1.1) montre que le score observé X est une variable aléatoire dont la distribution est celle de la composante E. Ses valeurs peuvent donc varier en fonction du nombre de fois que le même test est pris par le même individu et en fonction de la longueur de chacun des tests. Pour avoir une bonne estimation du score vrai T, il faut que le test soit administré un nombre infini de fois au même individu. En introduisant l'espérance mathématique $\in (.)$ à l'équation précédente, on obtient une estimation de T :

$$\epsilon(X) = \epsilon(T + E) = \epsilon(T) + \epsilon(E) = \text{T} + 0 = \text{T} \quad (1.2)$$

L'équation (1.2) montre que le score vrai d'un individu est la moyenne des scores observés lorsqu'il a passé le même test ou des tests parallèles sur un nombre infini de fois. Le score vrai d'une variable latente reste un concept statistique en sciences humaines et dont la valeur dépend du processus de mesure, contrairement aux données obtenues de la mesure sur les variables physiques en sciences exactes (Crocker et Algina, 1986 ; p.110).

Validité et fidélité du test

L'équation (1.1) et les hypothèses du modèle classique sont à la base des concepts de validité et de fidélité de l'instrument dont les définitions suivent :

La validité d'un test se réfère soit à la validation apparente, soit du contenu, d'un critère externe, d'un concept ou d'un modèle théorique (Grégoire & Laveault, 1997). La validation apparente du test est une évaluation superficielle des items dudit test par des juges non experts du domaine, contrairement à la validation du contenu. La validation du contenu d'un test permet entre autre d'apprécier la représentativité des items retenus par rapport au concept visé par la mesure, la qualité des instructions pour guider le processus de testing, la manière de présenter les stimuli, les modalités de réponses et leur notation, et le temps alloué pour compléter le test. La validité du test par rapport à un critère externe est dite concomitante lorsqu' elle se rapporte au degré de corrélation entre les scores au test et une mesure prise comme référence. Elle est dite prédictive

lorsqu' il s'agit de prédire une observation future à partir des scores au test. La validité conceptuelle quant à elle se réfère à l'évaluation du sens à attribuer aux scores à l'item ou à une échelle de mesure sur la base d'un modèle théorique.

Quant à la fidélité d'un instrument, elle se définit comme la corrélation entre les scores observés à deux tests parallèles (Crocker & Algina, 1986).

En utilisant les variances des deux scores aux tests parallèles, on trouve la formule suivante de la fidélité :

$$\rho(X, X') = {\sigma_T}^2/{\sigma_X}^2 \quad (1.3)$$

avec X et X' des scores observés à deux tests parallèles, ${\sigma_T}^2$ la variance du score vrai T et σ_X la variance du score observé X.

D'après Crocker & Algina et Grégoire & Laveault, la formule (1.3) donne trois interprétations possibles :

i. Elevée au carré, c'est la proportion de la variance du score observé qui est attribuable aux variations des scores vrais des candidats. On parle de la définition théorique du coefficient de fidélité.
ii. Elle s'interprète comme la corrélation entre scores observés à deux formes parallèles X et X' du test.
iii. Si l'on remplace dans cette équation X' par T, on obtient la corrélation entre les scores vrais et les scores observés. On parle alors de l'indice de fidélité.

Les méthodes qui permettent d'estimer le coefficient de fidélité d'un instrument sont classées en trois catégories:

- ✓ les méthodes de formes parallèles (tests alternatifs, test-retest, test-retest avec des tests alternatifs).
- ✓ la méthode de bissection ou split-half.
- ✓ les méthodes des covariances (coefficient alpha de Cronbach, les formules de Kuder Richardson KR 20 et KR 21, la formule de Hoyt).

Parmi les facteurs limitant la fidélité des résultats on peut citer : la difficulté du test, l'étendue des différences individuelles ou l'homogénéité du groupe, la limite de temps et la longueur du test.

1.2.4. Limites de la TCT

Les limites du modèle du score vrai sont en rapport non seulement avec la faisabilité des calculs des indices suggérés comme l'indice de fidélité, mais aussi avec des questions auxquelles auraient souhaité répondre les développeurs des tests et les analystes pour une meilleure orientation des décisions.

Ces limites sont entre autres (Hambleton, Swaminathan, & Rogers, 1991) :

- ✓ les caractéristiques du candidat et celles du test (difficulté et discrimination des items ; validité et fidélité du test) ne peuvent pas être séparées, l'un n'étant interprété que dans le contexte de l'autre. Par exemple, la difficulté de l'item d'un test est fonction des habiletés (scores vrais) des candidats, et inversement. Il est donc difficile de comparer des candidats qui prennent des tests différents, de même qu'il est difficile de comparer les items lorsqu'ils sont administrés à des populations différentes. Cela rend la tâche compliquée aux développeurs de tests qui ont besoin de capitaliser leurs expériences précédentes avec des items ou de constituer des banques d'items calibrés. Que dire par exemple de deux candidats qui ont le même score à deux tests mesurant la même habileté, mais dont les difficultés moyennes ne sont pas égales?
- ✓ Il est difficile en pratique d'avoir des tests parallèles, notion sur laquelle repose la définition de la fidélité considérée en TCT comme la corrélation entre les scores à deux tests parallèles, raison pour laquelle cet indice ne peut qu'être estimé par d'autres qui sont soit des bornes inférieures, soit alors des estimés de la fidélité avec des biais inconnus.
- ✓ L'erreur standard de mesure est supposée identique pour tous les candidats.
- ✓ La TCT est orienté sur le test, et non sur l'item. La TCT ne fournit pas d'information sur la façon dont les candidats répondent à un item donné, ce qui rend impossible de prévoir les performances (probabilité de bonnes réponses) des candidats face aux items. Il n'est pas possible de trouver la façon de discriminer les candidats à partir des items précis par exemple pour des besoins de sélection à l'entrée d'une profession.
- ✓ D'autres limites sont en rapport avec les tests adaptatifs, le développement des tests et l'identification des items biaisés.

La théorie de la réponse à l'item dont l'exposé est fait au point suivant, a été développée en tenant compte des limites de la TCT.

1.3. Modèles de la théorie de la réponse aux items (TRI)

Par rapport à la TCT, les modèles de la TRI permettent, en plus de produire des échelles de mesure, d'estimer les habiletés des candidats et les paramètres (difficulté, discrimination, pseudo-chance) des items. Les estimés des paramètres des items permettent de distinguer les items plus discriminants de ceux qui le sont moins.

Les modèles de la TRI les plus étudiés ont un paramètre habileté pour les personnes noté Θ, et trois paramètres pour les items dont le niveau de difficulté (b), sa capacité à discriminer entre candidats de faibles et fortes habiletés ou son indice de discrimination (a), le paramètre de pseudo-chance (c) qui est la probabilité qu'un candidat trouve par hasard la bonne réponse à l'item, peu importe son habileté. Ces paramètres peuvent être considérés comme des scalaires ou des vecteurs selon qu'on travaille en contexte unidimensionnel ou multidimensionnel.

La suite de cette partie est consacrée à l'exposé des deux postulats des modèles de la TRI et des deux hypothèses qui sont faites sur les données d'observation, et à la formulation mathématique des modèles unidimensionnels de la TRI communément rencontrés à savoir le modèle logistique à 3 paramètres (pour item dichotomique) noté ML3P duquel sont dérivés deux autres dont ML2P (2 paramètres) et ML1P (1 paramètre) ; et le modèle nominal de Bock (1972) pour item polytomique à échelle nominale.

La présentation de ces modèles est justifiée parce que la simplification du modèle à trois paramètres permet de retrouver une formulation mathématique identique à celle du modèle de Rasch simple.

1.3.1. Les postulats

Les modèles de la théorie des réponses aux items sont basés sur deux postulats fondamentaux (Hambleton, Swaminathan, & Rogers, 1991) :

1) La performance d'un candidat à un test peut être prédite par un ensemble de facteurs
2) La relation entre les performances des candidats à un item et l'ensemble de leurs habiletés qui sous-tendent ces performances peut être décrite par une fonction monotone croissante appelée courbe ou fonction caractéristique de l'item.

Le deuxième postulat soutient donc que la probabilité de trouver la bonne réponse augmente avec le niveau de l'habileté du candidat.

1.3.2. Hypothèses sur les données

Deux hypothèses importantes sont faites sur les données, et expliquent les méthodes d'estimation des paramètres qui sont appliquées dans les programmes informatiques :

- ✓ L'hypothèse de l'unidimensionnalité du test signifie qu'une seule habileté est mesurée à travers les items du test.

Il semble important de relever que dans la pratique, plusieurs éléments associés à la fois aux items, aux candidats et au contexte ainsi que leurs interactions entrent en jeu (Bertrand et Blais, 2004, p.203) et donc il reste difficile de vérifier cette hypothèse, d'où la considération en pratique d'un facteur dominant pouvant expliquer les performances des candidats (Stout, 1987).

- ✓ L'hypothèse d'indépendance locale stipule que si l'on considère deux items quelconques du test, en maintenant constantes les habiletés qui expliquent les réponses aux items du test, la réponse d'un candidat à un des items donnés n'est pas influencée par sa réponse à l'autre item du test.

Cette hypothèse a pour conséquence que la probabilité d'avoir un vecteur type d'observations pour un candidat de niveau d'habileté donné est le produit des probabilités associées à chaque observation étant donné l'habileté.

En plus de ces deux hypothèses, il est aussi sous-entendu que la fonction caractéristique spécifiée reflète la vraie relation qui lie les habiletés des candidats aux réponses des items obtenues d'eux.

1.3.3. Les modèles logistiques de la TRI pour items dichotomiques

Les modèles logistiques de la TRI applicables aux items dichotomiques dont les résultats possibles sont l'échec (X=0) ou le succès (X=1) considèrent que la probabilité pour un candidat de donner la bonne réponse à un item dépend à la fois de son habileté et des caractéristiques de l'item. D'après le postulat 2, cette probabilité augmente avec l'habileté.

La formulation mathématique de ces modèles est :

$$P(X = 1|\theta) = c + (1 - c) * \frac{1}{1+e^{\alpha*(b-\theta)}} \quad (1.3)$$

L'équation (1.3) est celui du modèle logistique noté ML3P et proposé par Lord (1980) qui considère un paramètre (θ) pour les personnes et trois paramètres (a, b et c) pour les items.

Dans cette équation, on a posé

- ✓ c >=0 est l'indice de pseudo-chance de l'item.
- ✓ θ est le niveau d'habileté du candidat
- ✓ b est l'indice de difficulté. C'est le point du continuum des habiletés (encore appelé point d'inflexion), au niveau duquel la courbe caractéristique de l'item change de courbure. Lorsque le paramètre de pseudo-chance (c) est nul, la probabilité qu'a un individu d'habileté θ =b de donner la bonne réponse à l'item est 0,5 sinon, elle est de 0,5+0,5*c. La condition c>0 se traduit par une asymptote horizontale d'équation Y=c à la courbe caractéristique de l'item. Pour un item j donné, la probabilité de trouver la bonne réponse en tout point du continuum des habiletés est donnée par $p_j+c_j*(1-p_j)$, avec p_j la probabilité de trouver la bonne réponse lorsque $c_j=0$.
- ✓ α= d*a avec a>0, l'indice de discrimination (un item est plus discriminant qu'un autre lorsque son indice de discrimination est plus grand). Il est proportionnel à la pente de la courbe caractéristique de l'item lorsque le niveau d'habileté du candidat est égal au niveau de difficulté de l'item (θ =b)
- ✓ en rapport avec l'ogive normale, il est établi que d=1,7.

L'un des avantages des modèles de la TRI est de pouvoir positionner les individus et les items sur un même continuum. Donc, si l'item est trop difficile pour un candidat, cela se traduira par la différence b- θ élevée, et donc une probabilité plus petite pour le candidat de donner la bonne réponse. Si cette différence est négative cela se traduit par une probabilité plus élevée alors que si elle est nulle, cela signifie que la difficulté et l'habileté sont égales, soit alors que le candidat a une probabilité égale à 0,5 +0,5*c de réussir cet item.
En donnant des valeurs particulières aux paramètres a, b et c de l'item, les courbes caractéristiques changent aussi. Par exemple, si on augmente le paramètre difficulté b tout en maintenant les deux autres inchangés, la courbe caractéristique de l'item subit une translation vers la droite, montrant ainsi que le candidat a besoin de plus d'habileté pour trouver la bonne réponse, car le niveau de difficulté a augmenté. De manière analogue, en diminuant la valeur de b, la courbe glisse vers la gauche en conservant sa forme. L'item devient donc moins difficile.

Les modèles suivants sont des variantes du modèle unidimensionnel de la TRI obtenus en variant les paramètres a, b et c:

Le modèle logistique à un paramètre ou ML1P

Si l'on pose c=0 et α une constante, alors l'équation (1.3) devient

$$P(X = 1|\theta) = \frac{1}{1+e^{\alpha*(b-\theta)}} \qquad (1.4)$$

L'équation (1.4) montre que la probabilité de donner la bonne réponse est seulement fonction de la distance entre la difficulté de l'item et l'habileté du candidat. De plus, la conséquence immédiate d'avoir α constant est que les courbes caractéristiques des items ne se croisent pas.

Sur un plan représenté par les habiletés et les indices de difficultés en abscisse (X) et l'indice de pseudo-chance et les probabilités en ordonné (Y), la condition c=0 se traduit par une asymptote horizontale à la courbe caractéristique de l'item aux valeurs inférieures des habiletés des candidats sur l'axe des abscisses.

Le modèle logistique à deux paramètres noté ML2P (Birnbaum,1968)

Si nous supposons que seule la pseudo-chance (c) est nulle, c'est-à-dire que la probabilité est nulle pour un candidat d'habileté trop faible par rapport à la difficulté de l'item de trouver la bonne réponse, alors on obtient la formulation suivante du modèle logistique à deux paramètres :

$$P(X = 1|\theta) = \frac{1}{1+e^{\alpha*(b-\theta)}} \qquad (1.5)$$

À niveau égal de la difficulté, un item est plus discriminant que l'autre lorsque son indice de discrimination (a) est plus élevé, se traduisant par une pente plus forte de sa courbe caractéristique au point du continuum $lorsque\ \theta - b = 0$.

1.3.4. Modèles de la TRI pour items polytomiques :

Ces modèles s'appliquent aux items à choix multiples dont les réponses possibles constituent une échelle nominale. L'apport de ce type de modèle est d'établir une distinction parmi les candidats qui n'ont pas choisi la bonne option car avec le choix dichotomique, ceux-ci sont automatiquement considérés comme étant de faible habileté. Pourtant, certaines modalités de l'item peuvent être plus proches de la bonne réponse, demandant ainsi plus d'habileté pour les choisir que d'autres qui sont plus éloignées de la bonne réponse, soit requérant moins d'habileté (Bertrand et Blais, 2004). Ces modèles permettent donc de maximiser la précision des estimés des habiletés en capitalisant toute l'information contenue dans les réponses des examinés (Hambleton, Swaminathan, & Rogers, 1991).

On trouve des modèles paramétriques à items polytomiques dans la littérature dont entre autres:

- ✓ le modèle généralisé du crédit partiel de Muraki (1992) qui est une généralisation du modèle du crédit partiel appartenant à la famille des modèles de Rasch (de Ayala, 2009) comme nous le verrons dans la suite
- ✓ le modèle à réponse graduée de Samejima (1969) qui est approprié pour analyser les items de type Likert.
- ✓ le modèle nominal de Bock (1972)
- ✓ le modèle à choix ou à réponses multiples de Thissen et Steinberg (1984) pour les données nominales.

Pour la discussion sur l'utilisation de ces modèles l'on peut se référer à Crocker et Algina (1986), Hambleton et al. (1991), Grégoire et Laveault (1997), Bertrand et Blais (2004), et de Ayala (2009).

1.3.5. *Modèle multidimensionnel de la théorie de la réponse aux items*

Il existe deux types de modèles multidimensionnels dont ceux dits non compensatoires ou partiellement non compensatoires et ceux dits compensatoires (Reckase, 2009). Les modèles non compensatoires ont reçu peu d'intérêt (de Ayala, 2009), c'est pour cela que nous nous intéresserons aux modèles multidimensionnels compensatoires dans notre travail. Dans les modèles unidimensionnels, on a supposé que la réponse à un item n'est déterminée que par un seul trait du candidat.

Dans le cadre multidimensionnel, on suppose que la réponse à l'item est déterminée par plusieurs traits latents. De même on s'attend à ce que la difficulté de l'item soit un vecteur de même dimension que l'espace des traits latents. Le candidat ne peut donc plus directement se positionner sur le même et seul continuum avec l'item, puisque tous sont devenus des vecteurs ayant plusieurs composantes, sauf si l'on décide de les projeter sur une même droite. Dans un espace multidimensionnel, le pouvoir discriminant de l'item varie en fonction de l'angle entre celui-ci et la direction considérée (Reckase, 2009).

Dans la TRI, on trouve des modèles multidimensionnels pour items dichotomiques comme le modèle logistique multidimensionnel, et pour items polytomiques à l'exemple du modèle multidimensionnel généralisé du crédit partiel (MGCP) et le modèle multidimensionnel à réponse gradué (MRG). Ces modèles sont présentés dans ce qui suit. La présentation de ces modèles permet de comprendre la façon dont les traits sont supposés interagir entre eux pour déterminer un choix de réponse à un item lorsque ce dernier requièrent plusieurs traits chez le candidat.

Modèle logistique multidimensionnel à trois paramètres MLM3P (Reckase, 1985; Ackerman, 1996)

Des modèles multidimensionnels pour items polytomiques seront présentés dans la famille des modèles de Rasch à la section suivante.

Sans perte de généralité, considérons un espace de dimension 2, et posons $\theta = (\theta_1, \theta_2)$ le vecteur des habiletés du candidat au test, $\alpha=(\alpha_1,\alpha_2)=d*(a_1,a_2)$ avec $a=(a_1,a_2)$ le vecteur des indices de discrimination de l'item X aux deux dimensions, $b=(b_1, b_2)$ le vecteur des indices de difficulté de l'item X aux deux dimensions.

Alors la probabilité de trouver la bonne réponse à l'item X est donnée par :

$$P(X=1|\theta,\alpha,b) = P(X=1|\theta_1,\theta_2,\alpha_1,\alpha_2,b_1,b_2) = P(X=1)$$
$$= c + (1-c) * \frac{e^{\alpha_1(\theta_1-b_1)+\alpha_2(\theta_2-b_2)}}{1+e^{\alpha_1(\theta_1-b_1)+\alpha_2(\theta_2-b_2)}}) \quad (1.6)$$

L'équation (1.6) est le modèle compensatoire multidimensionnel à trois paramètres, car on a supposé que le paramètre de pseudo chance est non nul. Cette équation peut être généralisée à plus de deux dimensions (f>2), en utilisant les formules matricielles des vecteurs habiletés et paramètres items. On peut obtenir la formule simple suivante :

$$P(X=1) = c + (1-c) * \frac{e^{\alpha_1\theta_1+\alpha_2\theta_2+\cdots+\alpha_f\theta_f-(\alpha_1b_1+\cdots+\alpha_fb_f)}}{1+e^{\alpha_1\theta_1+\alpha_2\theta_2+\cdots+\alpha_f\theta_f-(\alpha_1b_1+\cdots+\alpha_fb_f)}}$$
$$= c + (1-c) * \frac{e^{\alpha_1\theta_1+\alpha_2\theta_2+\cdots+\alpha_f\theta_f-\gamma}}{1+e^{\alpha_1\theta_1+\alpha_2\theta_2+\cdots+\alpha_f\theta_f-\gamma}} \quad (1.7)$$

On peut constater au numérateur que l'expression se trouvant dans le symbole exponentiel (e) est une somme pondérée aux f dimensions. Cela signifie que si un candidat a une habileté faible par rapport à la difficulté de l'item sur une dimension mais une forte habileté par rapport à la difficulté sur l'autre dimension, il y a compensation de façon globale, mais cela dépend des indices de discrimination de l'item aux deux dimensions. L'une des conséquences est que des individus ayant des vecteurs d'habiletés différents, soit placés en divers endroits du plan Θ (f=2), peuvent avoir la même probabilité de trouver la bonne réponse à l'item.

Le développement de l'équation (1.7) permet de donner une forme scalaire aux paramètres difficulté (Δ_i) et discrimination (A_i) de l'item i dans le cadre multidimensionnel :

En posant $\gamma_i = \alpha_1 b_1 + \cdots + \alpha_f b_f$,

$$\Delta_i = -\gamma_i / A_i \quad (1.8)$$
$$A_i = \sqrt{\sum_{f=1}^{F}(\alpha_f)^2} \quad (1.9)$$

Un item est d'autant plus discriminant aux f dimensions que sa valeur A_i est élevée. Donc pour les items qui saturent sur une seule dimension, leur indice de discrimination est unidimensionnel puisqu'alors leurs saturations sont nulles sur les autres dimensions.

Pour comparer deux items selon leur capacité à discriminer, il faut s'assurer qu'ils discriminent sur la même direction dans l'espace des habiletés (de Ayala, 2009, p.281-285).

Modèle multidimensionnel généralisé du crédit partiel (Reckase, 2009)

Le modèle a été développé pour décrire les interactions entre les personnes et les items polytomiques.

Soit un item X_i donné ayant plusieurs options dont les scores assignés à un individu j vont de k= 0 à K_i, alors le nombre de choix possibles est de K+1. La probabilité pour que l'individu j dont le vecteur des habiletés est θ_j obtienne le score X_{ij}=k à l'item X_i dont le vecteur des paramètres de discrimination est a_i est donnée par la formule suivante :

$$P(X_{ij} = k|\theta_j) = \frac{e^{ka_i\theta'_j - \sum_{x=0}^{k}\tau_{ix}}}{\sum_{h=0}^{K_i} e^{ha_i\theta'_j - \sum_{x=0}^{h}\tau_{ix}}}$$

θ_j' est la transposé de la matrice θ_j, τ_{ix} est le paramètre de décalage de l'option x, et $\tau_{i0} = 0$.

Modèle multidimensionnel à réponse graduée (Muraki et Carlson, 1993)

Supposons que la réalisation d'une tâche requiert qu'un nombre d'étapes soit réalisés, et qu'atteindre l'étape k nécessite la réalisation avec succès de l'étape k-1. C'est le cas avec les modèles du crédit partiel tel que nous le verrons dans les modèles de Rasch. Le score le plus faible à l'item i est donnée par 0 alors que le score maximum est noté K_i. On considère que la probabilité de compléter avec succès au moins l'étape k est monotone croissante lorsque l'une des composantes du vecteur des habiletés θ_j augmente.

Cette formulation revient à créer une variable dichotomique sur l'échelle avec k comme valeur de coupure, cette variable prenant pour valeur 1 si le score obtenu par l'individu est supérieur ou égal à k, et 0 sinon. La probabilité de réaliser au moins k étapes est donnée par le modèle de l'ogive normale dans lequel le paramètre habileté est une combinaison linéaire des éléments du vecteur des habiletés dont les poids sont les paramètres de discrimination. La probabilité de réaliser avec succès k étapes est donc la différence entre la probabilité d'en réaliser k ou plus avec succès (notée $P^*(X_{ij} = k|\theta_j)$ pour θ_j fixé), et celle d'en réaliser k+1 ou plus avec succès.

$$P(X_{ij} = k|\theta_j) = P^*(X_{ij} = k|\theta_j) - P^*(X_{ij} = k+1|\theta_j)$$

Avec $P^*(X_{ij} = 0|\theta_j) = 1$ car réaliser la tâche au moins à l'étape zéro est une certitude pour toutes les personnes évaluées; et $P^*(X_{ij} = K_i + 1|\theta_j)$=0 puisque le travail ne requiert pas la réalisation de plus de K_i étapes.

La formule suivante est la forme de l'ogive normale du modèle :

$$P(X_{ij} = k|\theta_j) = \frac{1}{\sqrt{2\pi}} \int_{a'_i\theta_j + d_{i,k}+1}^{a'_i\theta_j + d_{ik}} e^{\frac{-t^2}{2}} dt$$

d_{ik} est le paramètre qui reflète le niveau de facilité avec laquelle l'individu pourra réaliser la tâche jusqu'à l'étape k. Il prend une valeur très grande quand il est facile d'obtenir le score k, mais très petite en valeur négative sinon. $d_{i0} = 0$ et lorsque le score est K_i+1, $d_{i,K_i+1} = -\infty$. Seules les valeurs d_{ik}allant de k= 1 à K_i sont estimées en pratique.

La formule de la probabilité d'obtenir le score k prend sa forme ogive normale selon la formule suivante :

$$P(X_{ij} = k|\theta_j) = \frac{1}{\sqrt{2\pi}} \int_{a'_i\theta_j + d_{i,k}+1}^{a'_i\theta_j + d_{ik}} e^{\frac{-t^2}{2}} dt \; - \; \frac{1}{\sqrt{2\pi}} \int_{a'_i\theta_j + d_{i,k}+1}^{a'_i\theta_j + d_{i,k+1}} e^{\frac{-t^2}{2}} dt$$

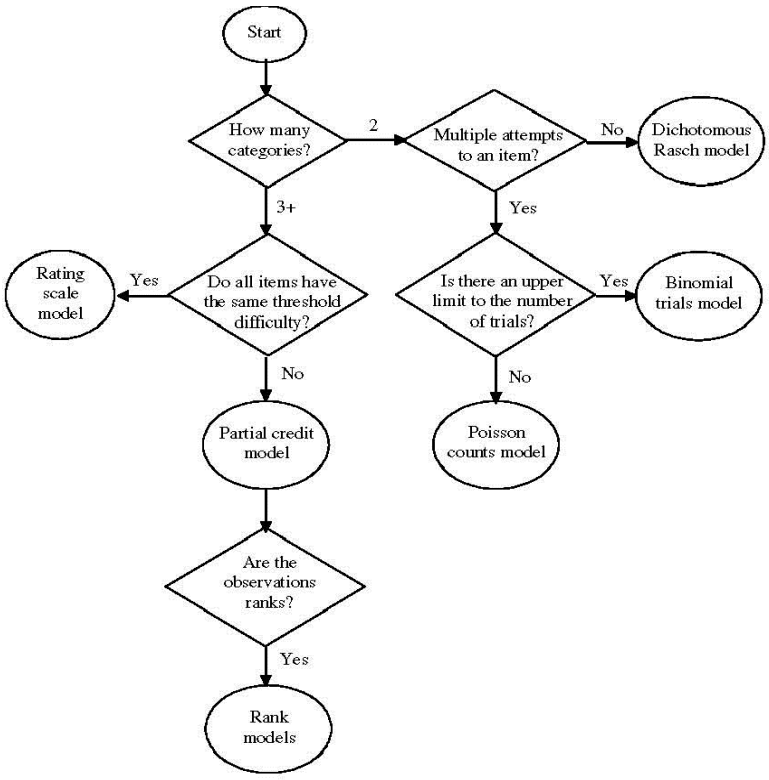
Les modèles multidimensionnels et les estimations des paramètres sont assez complexes comme l'attestent les formules 1.8 et 1.9. C'est l'une des raisons pour lesquelles on leurs préfère les modèles unidimensionnels car ils sont faciles à interpréter et leurs solutions sont autant simples à traduire en prise de décision.

La section qui suit est réservée aux modèles unidimensionnels et multidimensionnels qui appartiennent à la famille des modèles de Rasch.

1.4. Modèles de Rasch

1.4.1. Modèles unidimensionnels

Wright et Mok (2004) ont présenté six modèles qui appartiennent à la famille des modèles unidimensionnels de Rasch à partir de plusieurs critères dont en entrée le nombre de choix possibles pour répondre à l'item (voir graphique 1 ci-dessous).



Graphique 1. Classification des modèles unidimensionnels de Rasch

Source : Wright et Mok, 2004

On trouve dans la famille des modèles de Rasch 3 modèles appropriés pour les items dichotomiques et 3 autres pour des items polytomiques dont les options sont ordonnées. L'existence de la relation d'ordre entre les options signifie que certaines indiquent plus que d'autres, la présence faible ou forte de la caractéristique qui est mesurée.

Parmi les modèles pour items dichotomiques, on trouve

- ✓ le modèle de Rasch dichotomique
- ✓ le modèle binomial lorsque le nombre d'essais est fini
- ✓ le modèle de comptage de poisson lorsque le nombre d'essais est infini.

Les modèles polytomiques comprennent :

- ✓ le modèle à échelle de notations qui est approprié lorsque tous les items du test ont le même nombre d'options ordonnées
- ✓ le modèle du crédit partiel qui s'applique lorsque le nombre d'options des items peut varier d'un item à l'autre. C'est la forme non contrainte du modèle à échelle de notation

- ✓ le modèle des préférences lorsque les choix sont ordonnés par ordre de préférence du candidat.

Le modèle de Rasch pour item dichotomique est celui sur lequel repose notre étude. Les modèles à échelle de notation et de crédit partiel sont abondement utilisés en sciences de l'éducation et sont des extensions du modèle de Rasch simple aux items polytomiques. Chacun de ces trois modèles est exposé en détails ci-dessous.

Modèle de Rasch dichotomique

Le modèle de Rasch pour les items dichotomiques est identique sur le plan mathématique au modèle logistique à un paramètre vu dans la section précédente.

$$p(X = 1|\theta) = p(X) = \frac{\exp(\theta - b)}{1+\exp(\theta - b)} \quad (1.10)$$

Ce modèle est retrouvé à partir des modèles à items polytomiques lorsque l'on réduit le nombre d'options à deux.

Modèle à échelle de notation ou Rating scale model en anglais (Andrich, 1978)

Les modèles à échelle de notation sont appropriés pour les données des items de type Likert et celles de l'évaluation des performances. Les items de type Likert offrent un support pour la réponse consistant en en une série de catégories ordonnées qui varient par exemple de l'option ''totalement en désaccord'' à ''totalement en accord''.

Il est supposé que les options, qui sont ordonnées, sont séparées l'une de l'autre par une série de seuils aussi ordonnés et dont chacun (sur le continuum de la caractéristique mesurée) sépare deux options adjacentes de l'item. En plus d'avoir le même format, c'est-à dire les mêmes options, chaque item doit avoir le même nombre de valeurs de décalage τh, et ceux-ci sont identiques à tous les items. La position du seuil de rang h (ou *treshold*) entre les options h-1 et h de l'item j de niveau de difficulté b_j est donnée par $\delta_{jh} = b_j + \tau_h$ (De Ayala, 2009, p. 179-180). Les seuils ne se localisent donc pas forcément aux mêmes endroits pour tous ces items sur le continuum, car ils dépendent des niveaux de difficultés (b_j) de chacun.

La probabilité qu'a un individu d'habileté θ de passer Xj seuils ou mieux de choisir une catégorie de l'item j de paramètre de difficulté bj est:

$$p(X_j/\theta, b_j, \tau) = \frac{\exp\left[\sum_{h=0}^{X_j}(\theta-(b_j+\tau_h))\right]}{\sum_{h=0}^{m}\exp(\sum_{h=0}^{k}(\theta-(b_j+\tau_h)))} = \frac{\exp\left[-\sum_{h=0}^{X_j}\tau_h + x_j(\theta-b_j)\right]}{\sum_{h=0}^{m}\exp(-\sum_{h=0}^{k}\tau_h - k(\theta-b_j))} \quad (1.11)$$

avec τ le vecteur de m seuils de l'item j tel que $\sum_h \tau_h = 0$ et $X_j = \{0,1,2,\dots,m\}$

Modèle de crédit partiel ou Partial credit model en anglais (Wright et Masters, 1982)

Masters (1982) avait proposé une façon de modéliser les données d'items polytomiques ordonnés qui consiste à décomposer les réponses possibles en une série de paires formées par des catégories ou des scores catégoriels adjacents, et à appliquer successivement le modèle dichotomique à chacune des paires. Le modèle de crédit partiel est la version du modèle à échelle de notation sans contrainte sur le nombre identique de choix ou d'options possibles à tous les items. Ce modèle s'applique dans les conditions suivantes :

- ✓ des crédits partiels peuvent être donnés aux réponses partiellement correctes
- ✓ la demande cognitive sur le répondant est hiérarchique à chacun des items
- ✓ chacun des items requiert une séquence de tâches à compléter
- ✓ les réponses aux items sont ordonnées, et chacun des items a son nombre de seuils ou paramètres de transitions entre deux options de réponses adjacentes.

Le modèle de crédit partiel permet donc de donner des crédits différents pour des niveaux distincts de la réalisation d'une tâche donnée. Lorsque la tâche est entièrement réalisée, le candidat reçoit le crédit complet, mais reçoit une portion dudit crédit lorsqu'elle n'est pas complétée. Ainsi, pour une tâche scindée en deux sous-tâches dont la deuxième ne peut être réalisée que si la première est réussie, le candidat qui réalise la première sous-tâche reçoit le crédit de 1, celui qui réalise les deux reçoit un crédit de 2, et celui qui ne réalise aucune des sous-tâches reçoit un crédit de 0.

Le nombre de crédits possibles peut varier d'un item à l'autre, c'est-à-dire qu'un item du test peut avoir 3 crédits possibles, alors que d'autres en ont 4 ou 5, etc. Lorsqu'un item admet m modalités possibles, alors le nombre de crédit qu'un candidat peut recevoir est de m+1, car le minimum est 0 lorsqu'il ne réussit à aucune des opérations attendues. Le nombre équivalent de seuils au-delà desquels pourrait se situer l'habileté du candidat est de m-1.

Soit X un item ayant m+1 crédits possibles $j = \{0.,1,2,\dots m\}$, alors la probabilité que le candidat obtienne le crédit j=0,1,2,…m à l'item X est donnée par l'équation (1.13) :

$$p(X = j/\theta, b_j) = \frac{exp\left[\sum_{h=0}^{j}(\theta - b_h)\right]}{\sum_{k=0}^{m} exp(\sum_{h=0}^{k}(\theta - b_h))} \qquad (1.12)$$

avec

$\sum_{h=0}^{0}(\theta - b_h) = 0$; $\sum_{j=0}^{m} p(X = j|\theta_i) = 1$ (la somme des probabilités sur tous les crédits possibles à l'item X, pour un individu i d'habileté θ est égale à l'unité); et b_h est le paramètre de transition entre les modalités adjacentes h-1 et h de l'item X. Il est encore appelé le paramètre de difficulté d'étape (*step difficulty parameter*). C'est le lieu du continuum de la caractéristique mesurée au niveau duquel la probabilité qu'un candidat se place dans l'une des deux options adjacentes (h versus h-1) est égale : $(X = h - 1/\theta = b_h) = p(X = h/\theta = b_h))$.

Toutefois, les paramètres b_h ne sont pas ordonnés forcément et dépendent de la difficulté de la sous-tâche ou de l'opération h de l'item. Ils doivent donc être interprétés comme un ensemble de paramètres de l'item, soit donc dans le contexte de ce dernier. Il est aussi supposé que les items discriminent les personnes au même degré (paramètre de discrimination constant pour tous les items du test). Cette hypothèse est relaxée dans le cadre du modèle généralisé de crédit partiel.

Modèle des préférences ou Ranks model en anglais (Linacre, 1994)

Le modèle de préférence est utile lorsqu'il est demandé aux individus de dévoiler leurs préférences sur un ensemble de biens, en les classant par ordre de préférence, plutôt que de les noter par exemple. Ce modèle a été développé comme alternative au modèle à échelle de notation. La notation dans ce dernier est remplacée par le rang de l'objet dans la hiérarchie des préférences.

Modèle binomial (Wright et Masters, 1982)

Lorsqu'une épreuve n'a que deux issues possibles qui sont l'échec ou le succès, et qu'elle doit être répétée un grand nombre de fois (n), le nombre de succès (m) suit une loi Binomiale. Dans le cadre de la théorie de la réponse à l'item, cela revient à supposer qu'un candidat devra passer plusieurs fois un item donné.

Modèle de poisson (Wright et Masters, 1982)

La loi de poisson est encore appelée la loi des évènements rares. Lorsque le nombre d'essais de la loi binomiale est infini, et que la probabilité de succès à un essai est assez petite, alors on peut approximer le modèle binomial par celui de Poisson.

1.4.2. Les modèles multidimensionnels de Rasch

Nous simulerons les données multidimensionnelles avec le logiciel Conquest, qui est spécialisé pour les modèles de Rasch et dans lequel le modèle multidimensionnel de Rasch (MRCML) est pris en charge. Ce dernier est flexible et permet de retrouver les autres modèles de Rasch par manipulation de deux matrices de données. Le MRMCL est exposé dans ce qui suit.

Le modèle logistique multinomial à coefficients aléatoires (MRCML)

En général, lorsque l'instrument de mesure est multidimensionnel et que la distribution des items entre les dimensions est connue, l'on a deux choix possibles pour estimer les paramètres : soit on estime les paramètres par dimension en considérant tous les items à la fois, auquel cas on obtient des scores par dimension, sauf que des informations différentielles des performances sur les dimensions individuelles sont inexploitées. Une autre façon de procéder serait de faire des estimations consécutives, c'est-à-dire considérer chacune des dimensions indépendamment des autres pour produire les estimations.

L'avantage de cette méthode est la production des estimés des habiletés en autant de fois que le nombre de dimensions, ainsi que les erreurs standards des scores par dimension. Un inconvénient est qu'on ne tient pas compte d'une corrélation potentielle entre les dimensions, car on a supposé qu'elles sont orthogonales entre elles.

Le modèle logistique multinomial à coefficients aléatoires, ou Multidimensional Random Coefficients Multinomial Logit Model (MRCML), a été développé par Wang, Wilson et Adams (1996) pour pallier aux insuffisances des deux approches précédentes. C'est une généralisation des modèles unidimensionnels et multidimensionnels de la famille des modèles de Rasch.

Ce modèle est basé sur deux matrices qui peuvent être manipulées par le chercheur à savoir la matrice de scores notée B, et la matrice de construction (design matrix) notée A. Les deux matrices servent à donner la spécification de la forme fonctionnelle du modèle à partir de la distribution des items en rapport avec les paramètres des items (A) et les dimensions (B). Le

MRCML est adapté pour la multi dimensionnalité entre les items et la multi dimensionnalité intra items. L'estimation des paramètres prend en compte les corrélations entre les dimensions, et les estimés obtenus sont plus stables que dans les approches consécutive ou unidimensionnelle composite.

Formulation

Supposons I items indexés par $i = 1,2,3, \dots I$, ayant chacun $K_i + 1$ catégories de réponses possibles ($k = 1,2, \dots K_i$). Considérons la variable aléatoire X_{ik} tel que

$X_{ik} = \begin{cases} 1 \; si \; la \; réponse \text{ à } l'item \; i \; est \; dans \; la \; catégorie \; k \\ 0 \; sinon \end{cases}$,

Ce modèle de dimension D peut s'écrire, à l'échelle de la catégorie par :

$$P(X_{ik} = 1; A, B, \varepsilon|\theta) = \frac{\exp(b_{ik}\theta + a'_{ik}\varepsilon)}{\sum_{i=1}^{K_i} \exp(b_{ik}\theta + a'_{ik}\varepsilon)} \quad (1.19) \text{ avec}$$

$\theta \; le \; vecteur \; colonne \; des \; habiletés \; de \; dimension \; D$,

$\varepsilon \; est \; le \; vecteur \; des \; paramètres \; des \; items \; et \; leurs \; catégories.$

a_{ik} et b_{ik} (score obtenu lorsque l'option k de l'item i est choisie en rapport avec une dimension spécifique) sont des éléments des matrices A et B respectivement.

La probabilité de choisir la bonne option pour un item dichotomique est la suivante :

$$P(U_i = 1|a_i, d_i, \theta) = \frac{\exp(a_i\theta\prime + d_i)}{1+\exp(a_i\theta\prime + d_i)} \quad (1.20)$$

Les moyennes des paramètres items et population de ce modèle et les variances des paramètres théta (θ) sont estimées par la méthode du maximum de vraisemblance marginale.

Par exemple, si l'on suppose disposer des données d'items polytomiques à cinq options 0,1,2,3 et 4, et d'un instrument de dimension D, alors les probabilités que le répondant n choisisse l'une des options de l'item i de niveau de difficulté δ_i et disposant de 4 paramètres de difficulté d'étape τ_{jd} (j=1,2,3,4 est l'indice des paramètres de difficulté d'étape) étant donné son habileté (θ_{nd}) à la dimension d sont les suivantes:

$$P(0) = {}^{1}/_{\gamma}$$

$$P(1) = {}^{\exp(\theta_{nd} - \delta_i - \tau_{1d})}/_{\gamma}$$

$$P(2) = {}^{\exp(2\theta_{nd} - 2\delta_i - \tau_{1d} - \tau_{2d})}/_{\gamma}$$

$$P(3) = {}^{\exp(3\theta_{nd} - 3\delta_i - \tau_{1d} - \tau_{2d} - \tau_{3d})}/_{\gamma}$$

$$P(4) = {}^{\exp(4\theta_{nd} - 4\delta_i - \tau_{1d} - \tau_{2d} - \tau_{3d} - \tau_{4d})}/_{\gamma} = {}^{\exp(4\theta_{nd} - 4\delta_i)}/_{\gamma}$$

Car on pose $\tau_{4d} = -(\tau_{1d} + \tau_{2d} + \tau_{3d})$ pour des contraintes d'identification des paramètres et

$$\gamma = 1 + \exp(\theta_{nd} - \delta_i - \tau_{1d}) + \exp(2\theta_{nd} - 2\delta_i - \tau_{2d}) + \exp(3\theta_{nd} - 3\delta_i - \tau_{1d} - \tau_{2d} - \tau_{3d}) + \exp(4\theta_{nd} - 4\delta_i)$$

La difficulté de l'item sur une dimension d quelconque se divise donc en la somme de sa difficulté moyenne et la difficulté d'étape.

En manipulant les matrices A et B, les différents modèles de Rasch unidimensionnels ou multidimensionnels peuvent être retrouvés. Par exemple, les probabilités dans le cadre du modèle multidimensionnel du crédit partiel prennent la formulation suivante pour un individu quelconque.

$$P(X_i = k|\theta) = \frac{\exp(\sum_{l=1}^{m}(\theta_l - b_{ilk})w_{ilk})}{\sum_{r=0}^{K_i}\exp(\sum_{l=1}^{m}(\theta_l - b_{ilk})w_{ilk}} \quad (1.21)$$

b_{ilk} est le paramètre difficulté de l'item i sur la dimension l pour le score catégoriel k et w_{ilk} est un poids du score prédéfini pour l'item i en rapport avec la dimension l et le score catégoriel k.

1.4.3. Les modèles de Rasch et la mesure de la dimensionnalité

L'évaluation de l'écart à l'unidimensionnalité d'un test avec le modèle de Rasch peut être étudiée à partir des statistiques d'ajustements (infits et outfits). En plus, des représentations graphiques appelées cartes de Wright sont générées et mettent ensemble les items et les personnes sur un même continuum. Ils donnent ainsi les positions des personnes en fonction des habiletés et celles de chacun des items en fonction du niveau de difficulté. Cette facilité donne la possibilité de juger de l'appariement entre les items et le groupe d'individus en étude. Mieux, cette option

permet de savoir si dans l'ensemble, les items sont de difficulté plus ou moins élevée par rapport aux habiletés de la population étudiée.

Toutefois, le fait d'avoir des items de niveau de difficulté plus élevé que celui des habiletés de la population pourrait faire ressortir une dimension illusoire, c'est-à-dire que la dimension et la difficulté sont confondues (Drasgrow et Parsons (1983) ; Akerman (1991)).

Le rapport de performance entre les modèles multidimensionnels et unidimensionnels se fait communément à partir des indices connus comme la déviance (G^2), le critère d'information d'Akaïke (AIC) ou le critère d'information de Bayes (BIC). Un exposé sur la formulation et les interprétations des statistiques d'ajustement infit et outfit sera donné au chapitre trois, sur leur contribution à l'étude de la robustesse des modèles unidimensionnels de mesure.

1.5. Comparaison entre la TCT, la TRI et le modèle de Rasch

Notre choix d'aborder la robustesse des modèles unidimensionnels dans le cadre du modèle de Rasch, et pas de celui de la TCT est justifié par le fait que plusieurs recherches à l'exemple de celles de Wilson et al. (2006), et celle de Magno (2009) ont révélé que la modélisation de Rasch produit des résultats qui prennent en compte les limites de la TCT.

Wilson et al. (2006) ont comparé les deux approches TCT et Rasch à travers trois groupes de critère dont i) le choix du modèle ; ii) l'évidence de la fidélité, l'évidence du calcul du coefficient de fidélité et son erreur de mesure et iii) l'évidence sur la validité y compris celle basée sur le contenu de l'instrument, le processus de réponse, la structure interne et autres variables. Bien que les résultats aient montré qu'il y a des aspects techniques similaires entre les deux approches, l'étude a conclu que le modèle de Rasch performe mieux que la TCT pour tous les aspects en rapport avec la construction et l'évaluation de l'instrument.

(Magno, 2009) a comparé la TCT et la TRI en partant de deux tests de 70 items polytomiques en chimie, administrés à deux échantillons de 109 et 110 étudiants sur les 3 aspects suivants : i) la difficulté des items, ii) la consistance interne de chaque test et iii) les erreurs de mesure.

Des résultats de cette étude, il ressort que:

- ✓ les estimés des paramètres difficulté des items par la TRI ne changent pas à travers les deux échantillons comme le sont ceux obtenus avec la TCT
- ✓ les estimés des paramètres difficulté de la TRI sont plus stables à travers les deux questionnaires que ne le sont ceux obtenus avec la TCT

- ✓ la consistance interne est plus stable à travers les 2 échantillons avec la TRI qu'avec la TCT
- ✓ les erreurs produites par la TRI sont faibles par rapport à celles qui sont produites par la TCT.

D'autre part, les modèles de la TRI sont falsifiables dans ce sens qu'un modèle donné peut ne pas être approprié pour un jeu de données particulières. En effet, les statistiques produites par les modèles de la TRI au niveau de chacun des items et chacune des personnes, en combinaison avec les statistiques globales sur le test permettent de confirmer si les données s'ajustent ou non au modèle de mesure appliqué.

Par rapport à la TCT, les habiletés des candidats sont indépendantes du test, et les caractéristiques des items du test ne dépendent pas du groupe de candidats. Ce sont donc des paramètres invariants dans le cadre de la TRI. Les modèles de la TRI permettent de générer des erreurs standards des estimés des habiletés de chacun des candidats et items, et non une erreur globale comme dans la TCT.

Enfin, des similitudes conceptuelles au niveau des paramètres difficulté et discrimination des items entre la TCT et la TRI peuvent être notées dont (Reckase, 2009) :

- ✓ Paramètre difficulté : En TCT, la proportion de personnes qui donnent la bonne réponse à un item est considérée comme un indicateur de son niveau de difficulté, souvent noté p. Un item sera dit difficile lorsque cette proportion est faible, et facile quand elle est élevée. Dans la TRI, la difficulté de l'item est notée par le symbole b. Pour déterminer la fonction de lien entre p et b, la fonction de distribution de la variable θ , $f(\theta)$ doit être connue.

On obtient alors la relation suivante :

$$p = \int_{-\infty}^{+\infty} P(X = 1/a, b, c) f(\theta)\, d\theta \quad (1.22)$$

Avec p la probabilité de donner la bonne réponse à l'item X dont les paramètres sont a, b et c. Les deux paramètres p et b sont liées par une relation inverse et non linéaire, c'est-à-dire que lorsque b augmente, p diminue.

- ✓ Paramètre discrimination

La discrimination est la capacité de l'item du test à différencier les candidats de faible habileté, de ceux à forte habileté. En TCT, elle est calculée à partir de la corrélation bisérielle par point entre les réponses à l'item et le score total au test. En TRI, le paramètre discrimination de l'item est noté par la lettre a.

Sous certaines hypothèses, on trouve la relation suivante entre les deux théories en rapport avec ce paramètre :

$$a = \frac{\rho}{\sqrt{1-\rho^2}} \quad (1.23)$$

Avec ρ la saturation de l'item au premier facteur commun des items du test.

1.6. Méthodes d'estimation des paramètres

L'estimation est le processus par lequel une quantité inconnue appelée paramètre se voit attribuer une valeur numérique sur la base des informations recueillies auprès d'un échantillon représentatif d'une population donnée (Besbeas, 2010). Lorsque la valeur estimée est un nombre, on parle d'estimation par point. Sinon l'estimation est dite par intervalle lorsque le processus produit une plage de données encore appelée intervalle de confiance, qui est supposé contenir la vraie valeur du paramètre en considérant une marge d'erreur α, généralement fixée à 5%.

La précision de l'estimation d'un paramètre θ donnée est mesurée par le biais B qui est la différence entre la vraie valeur du paramètre et sa valeur estimée $\hat{\theta}$, mais surtout par l'erreur quadratique moyenne EQM(.) qui prend la forme d'une somme du biais au carré et de la variance de l'estimateur.

$$\text{EQM}(\theta, \hat{\theta}) = \text{E}(\theta - \hat{\theta}(X))^2 = \text{E}^2(\hat{\theta}(X) - \theta) - \text{V}(\hat{\theta}(X)) = \text{B}^2(\hat{\theta}, \theta) - \text{V}(\hat{\theta}(X))$$

Dans le cadre des modèles de mesure des réponses aux items, les problèmes d'estimation ou d'adéquation peuvent surgir au niveau des paramètres personnes (habiletés), et des paramètres items (difficulté, discrimination, paramètre de pseudo-chance).

Il existe trois techniques d'estimation des paramètres dont les techniques paramétriques, les techniques semi-paramétriques et les techniques non paramétriques. Les techniques paramétriques sont basées sur la vraisemblance, et donc impliquent que soit connue la distribution de probabilité des variables étudiées. On peut citer l'estimateur de vraisemblance et l'estimateur Bayésien. Les techniques semi paramétriques sont moins contraignantes. On trouve dans cette catégorie la méthode des moments généralisés (GMM), l'estimation par les moindres écarts absolus, les méthodes par densité du noyau, la régression partiellement linéaire etc.

Dans les techniques non paramétriques, les paramètres ne sont pas figés, et le but est de trouver une formulation satisfaisante des données. On peut citer l'estimation de densité et l'estimation de la fonction de régression.

Selon Green (2005. P.410), lorsqu'il n'y a pas de problème de spécification du modèle paramétrique, le meilleur estimateur paramétrique surpasse généralement le meilleur estimateur semi paramétrique, et celui-ci surpasse le meilleur estimateur non paramétrique. Dans le cadre de notre travail, les modèles de réponses aux items sont basés sur la fonction de densité logistique, d'où notre intérêt pour les estimateurs paramétriques.

Nous présentons trois méthodes d'estimation des paramètres rencontrées dans les logiciels courants de la théorie de la réponse à l'item et de la modélisation de Rasch: La méthode d'estimation par le maximum de vraisemblance (EMV) et deux variantes d'estimation itératives des paramètres dont celle du maximum de vraisemblance jointe (EMVJ) et la méthode du maximum de vraisemblance marginale (EMVM) sont successivement exposées dans ce qui suit.

1.6.1. Estimation par le maximum de vraisemblance (EMV)

L'estimation d'un paramètre par la méthode du maximum de vraisemblance suppose la connaissance de sa distribution de probabilité. Dans notre cas, nous cherchons à estimer les paramètres des items et ceux des personnes. L'EMV ne permet pas d'estimer les deux ensembles de paramètres à la fois, c'est pour cela qu'il faut connaître l'un des deux pour estimer l'autre.

Considérons un test constitué de J items dichotomiques. La probabilité de donner la bonne réponse à un item j à choix dichotomique $X_j = \{0,1\}$ de difficulté b_j pour un individu i d'habileté θ_i dans le cadre du modèle de Rasch est :

$$p(X_{ij} = 1|\theta_i, b_j) = \frac{1}{1 + e^{-(\theta_i - b_j)}} = p_{ij}(X)$$

$$p(X_{ij} = 0|\theta_i, b_j) = 1 - p_{ij}(X)$$

Considérons $\underline{X}$ un vecteur de J réponses 0 ou 1 de cet individu. Alors, la fonction de vraisemblance du vecteur d'observations de cet individu s'écrit :

$$L_i(\underline{X}|\theta_i, b) = \prod_{j=1}^{J} {p_j}^{x_{ij}} (1 - p_j)^{(1-x_{ij})} \quad (1.24)$$

La fonction de vraisemblance de N observations aux J items est alors donnée par l'équation suivante :

$$L(\underline{X}|\theta, b) = \prod_{i=1}^{N} \prod_{j=1}^{J} {p_j}^{x_{ij}} (1 - p_j)^{(1-x_{ij})} \quad (1.25)$$

Avec θ le vecteur des N habiletés et b le vecteur des J difficultés.

Si nous supposons connus les paramètres difficultés des items, l'estimation des paramètres habiletés des individus passe par plusieurs étapes :

- ✓ Calculer les probabilités p_j pour chacun des items sur le continuum des habiletés
- ✓ Calculer la probabilité du vecteur d'observations, ou sa vraisemblance par multiplication des probabilités de la précédente étape, en exploitant l'hypothèse de l'indépendance des observations pour chaque point habileté sur le continuum
- ✓ On choisit alors la valeur de l'habileté pour lequel le vecteur a la plus forte vraisemblance.

Les erreurs standards sur les estimés peuvent aussi être obtenues par la formule suivante :

$1/\sigma_j = \sqrt{I_j(\theta_i)}$ avec $I_j(\theta_i) = \omega^2 p_j(1 - p_j)$ l'information apportée par l'item j dont le paramètre de discrimination est ω, dans l'estimation de l'habileté de l'individu i. L'information totale apportée par le test pour estimer une habileté θ est la somme des quantités d'information apportée par chacun des items du test $I(\theta) = \sum_{j=1}^{J} I_j(\theta)$

Un intervalle de confiance du paramètre θ au seuil α est donnée par $IC: \hat{\theta} \mp z_{1-\alpha}\sigma(\hat{\theta})$

Dans la pratique, l'on ne dispose ni des estimés des paramètres items, ni ceux des personnes. Des variantes les plus rencontrées de la méthode du maximum de vraisemblance qui permettent d'estimer ces paramètres à la fois sont l'estimation par le maximum de vraisemblance jointe ou inconditionnelle et l'estimation par le maximum de vraisemblance marginale (voir de Ayala, 2009).

1.6.2. Estimation par le maximum de vraisemblance jointe ou inconditionnelle (EMVJ)

Cette méthode est itérative et permet d'estimer à la fois les paramètres des items et ceux des personnes en trois étapes, car les scores brutes constituent des statistiques suffisantes de mesure à la fois pour les items et les personnes (de Ayala, 2009). L'algorithme souvent utilisé dans cette

méthode est celui de Newton-Raphson. La fonction à maximiser est le logarithme de la fonction de vraisemblance de l'équation (1.25).

Etape 1 : cette étape consiste à estimer les paramètres des items à partir des estimés provisoires des paramètres des personnes. Étant donné le nombre moins élevé des items par rapport aux personnes, il est conseillé de procéder d'abord à l'estimation des paramètres des items, puisque l'on dispose de plus d'information pour procéder ainsi.

Etape 2: les estimés des paramètres précédemment obtenus sont traités comme connus et mis à contribution pour estimer les paramètres des personnes indépendamment les uns des autres.

Etape 3 : les estimés des paramètres personnes ayant été améliorés à l'étape 2, ces derniers sont de nouveau utilisés pour ajuster les paramètres des items. Cette boucle entre les deux étapes est répétée jusqu'à ce que l'écart entre les estimés consécutifs du même paramètre soit inférieur au critère de convergence fixé par le chercheur.

Comme avantage, cette méthode prend en compte des valeurs prédéfinies des paramètres s'il en existe. Toutefois, elle produit des biais d'estimation lorsque le test contient 15 items ou moins (Lord, 1986) ou lorsque l'échantillon est de taille assez faible. L'on note aussi le fait que l'on se trouve obligé d'estimer un ensemble de paramètres en fonction d'un autre non connu en réalité (problème de paramètres structurels). D'autre part, les estimations se faisant ensemble, lorsqu'un item ne se comporte pas conformément au modèle, il doit être retiré, et cela exige une nouvelle calibration des items et des personnes. La méthode d'estimation par le maximum de vraisemblance marginale permet de surmonter les deux dernières limites.

1.6.3. Estimation par le maximum de vraisemblance marginale(EMVM)

Cette méthode permet d'estimer uniquement les paramètres des items, alors que les paramètres des personnes sont ensuite estimés avec une autre méthode comme la méthode du maximum de vraisemblance ou l'approche Bayésienne à posteriori (de Ayala, 2009).

Supposons un vecteur de réponses à un test de longueur J. Alors la probabilité pour un individu (dont le paramètre de localisation est θ) d'obtenir un tel vecteur, avec ϑ la matrice des paramètres items est :

$$P\left(\underline{X}\middle|\theta,\vartheta\right) = \prod_{j=1}^{J} {p_j}^{x_j} \left(1-p_j\right)^{(1-x_j)}$$

L'introduction de l'information sur les personnes (sans avoir à estimer leurs paramètres) en vue d'estimer les paramètres des items se fait par le biais de la fonction de distribution d'un

échantillon aléatoire représentatif de la population $g(\theta|\delta)$ où δ contient les paramètres localisation et échelle des personnes. L'estimation des paramètres items est obtenue à partir de la fonction suivante :

$$P(X) = \int_{-\infty}^{+\infty} \prod_{j=1}^{J} {p_j}^{x_j} (1 - p_j)^{(1-x_j)} g(\theta|\delta) \, \mathrm{d}\theta$$

En termes d'avantages, les valeurs manquantes sur un échantillon de données peuvent être gérées même lorsque l'information est insuffisante en vue de produire des estimés acceptables pour les personnes. Cette méthode est très utilisée dans la modélisation des items à deux ou trois paramètres, et supporte aussi les modèles multidimensionnels de Rasch.

L'un des inconvénients de l'EMVM est qu'elle impose une fonction de distribution sur les paramètres des personnes, généralement la distribution normale uni variée ou multi variée (selon que le modèle de mesure est unidimensionnel ou multidimensionnel).

Toutefois, les deux méthodes d'estimations EMVJ et EMVM produisent des résultats semblables (Linacre, 1999).

Conclusion

Dans ce chapitre, nous avons abordé le cadre théorique à travers un exposé sur les modèles de la TCT, de la TRI et de Rasch. Les limites de la TCT ainsi que les apports de la TRI et du modèle de Rasch par rapport à la TCT ont été mises en exergue. Les similitudes en termes de formulation des paramètres de discrimination et difficulté des items ont été établies. Des modèles unidimensionnels et multidimensionnels de la TRI et de Rasch y ont aussi été exposés. Comme méthodes d'estimation des paramètres, le focus a été mis sur la méthode du maximum de vraisemblance, la méthode du maximum de vraisemblance jointe et la méthode du maximum de vraisemblance marginale car ces méthodes sont utilisées dans la plupart des logiciels, parmi lesquels Conquest qui est notre logiciel de travail. Les chapitres 2 et 3 qui suivent traiteront respectivement de la revue de littérature sur la question de la dimensionnalité d'un test, et sur celle de la robustesse des modèles unidimensionnels de mesure des réponses aux items.

Chapitre 2:

La dimensionnalité de l'instrument de mesure

Ce chapitre se focalise sur les études empiriques en liaison avec la détection de la dimensionnalité d'un test. Bien que la détection de la dimensionnalité ne soit pas notre thème principal de recherche, il semble utile de consulter quelques recherches sur cette question, étant donné qu'elle reste d'actualité puisque les approches utilisées ne font pas l'unanimité des chercheurs. Ce concept est au cœur de notre problématique, puisqu'il s'agit d'explorer la robustesse d'un modèle de mesure unidimensionnel, mais avec des données produites par un instrument de collecte qui reflète plusieurs dimensions.

La détermination de la dimensionnalité se fait à deux niveaux : au niveau de chaque item et à celui de l'instrument de mesure entier. La combinaison de la dimensionnalité entre item et questionnaire donne donc trois cas de figure rencontrés en pratique qui sont (Wang, Wilson, & Adams, 1996):

- ✓ Un questionnaire unidimensionnel ne contenant que des items unidimensionnels permettant de ne mesurer que le trait latent visé par le questionnaire
- ✓ Un questionnaire multidimensionnel ne contenant que des items unidimensionnels, c'est à dire que chaque item ne mesure qu'un seul des traits visés par le questionnaire et seul celui-ci détermine la réponse à l'item. C'est la multi dimensionnalité entre les items
- ✓ Un questionnaire multidimensionnel avec des items unidimensionnels et multidimensionnels, donc il y a présence d'items mesurant plus d'un des traits visés par l'ensemble du questionnaire, et les réponses à ces items particuliers ne sont déterminées que par ces traits qu'ils sont censés mesurer. C'est la dimensionnalité de l'instrument et des items qui le composent ou multi dimensionnalité intra items.

La section qui suit est centrée sur la revue de littérature au tour des méthodes généralement utilisées pour détecter la dimensionnalité d'un instrument de mesure.

Les résultats obtenus dans les études sur la question de la détermination de la dimension d'un instrument de mesure ont certaines limites qui ne permettent pas toujours leur généralisation.

Nous listons certaines de ces limites ci-dessous, puis nous nous attarderons sur la définition

opérationnelle de la dimensionnalité avant de présenter une sélection parmi des études qui ont été réalisées.

2.1 Les limites potentielles à la généralisation des résultats

Les résultats d'analyse de la dimensionnalité d'un même jeu de données peuvent être différents pour plusieurs raisons (Kirisci, Hsu, & Yu, 2001) :

- ✓ différence de méthodes d'estimation utilisées dans les logiciels. Par exemple les logiciels BILOG et MULTILOG utilisent la méthode du maximum de vraisemblance marginale, RUMM2030 utilise entre autre la méthode du maximum de vraisemblance jointe, alors que Conquest 3 utilise la méthode du maximum de vraisemblance marginale et celle du maximum de vraisemblance jointe. Toutefois, avec un nombre élevé d'observations, ces méthodes convergent vers les mêmes résultats.
- ✓ avec une même méthode d'estimation, les options et les algorithmes peuvent être différents d'un logiciel à l'autre.
- ✓ les processus générateurs de données multidimensionnelles peuvent être différents même lorsque ce sont des données de simulations en particulier si le modèle utilisé pour générer les données n'est pas le même dans deux études différentes.
- ✓ la définition et la structure retenue de la multi dimensionnalité avec des données simulées peut varier selon deux approches de la dimensionnalité: certaines études considèrent plusieurs dimensions dominantes avec des degrés de corrélations entre elles, tandis que d'autres considèrent une dimension dominante et plusieurs dimensions mineures.

À ces éléments, on peut ajouter deux autres raisons dont la distribution des paramètres dans les simulations qui varie selon les expériences, l'écart entre les valeurs souhaitées et les valeurs obtenues des paramètres par simulation (Blais, 1987). Cette contrainte a pour conséquence de restreindre le champ pratique des comparaisons des résultats entre logiciels sauf au sacrifice de l'uniformité des données.

2.2 Définition opérationnelle de la dimensionnalité

Selon Nandakumar (1991), Il existe deux variantes opérationnelles de la dimensionnalité dont: la dimensionnalité au sens traditionnel (McDonald, 1986) qui est le nombre exacte de dimensions

d'un instrument, alors que la dimensionnalité essentielle (Stout, 1987) s'intéresse au nombre de dimensions dominantes.

Les méthodes de détermination de la dimensionnalité, ou mieux de l'écart à l'unidimensionnalité d'un ensemble d'items ont varié non seulement avec le temps grâce au développement de nouvelles théories en mesure et en analyse des données, mais surtout grâce au développement de logiciels informatiques qui permettent de nos jours de tester la robustesse des différentes méthodes d'estimation, comme nous le verrons dans les études qui suivent.

2.3 Des études empiriques sur la dimensionnalité

Selon le type de données disponibles, les méthodes de détection de l'écart à l'unidimensionnalité généralement rencontrées font références aux indices statistiques produits dans le cadre des modèles théoriques d'analyse des données se rapportant à la théorie classique des tests, la théorie des réponses aux items, et les modèles de Rasch. Toutes sont des méthodes basées sur des techniques d'estimation paramétriques, thème qui a été abordé au chapitre 1. On rencontre aussi des méthodes non paramétriques à l'exemple de :

- ✓ DETECT (dimensionality Evaluation to Enumerate Contributing Traits) développée par Kim (1994), Zhang et Stout (1999)
- ✓ DIMTEST (Test of essential dimensionality (Stout, 1987, 1990)
- ✓ MSP (Mokken scaling program) développée par Molenaar et Sijstma (2000)
- ✓ TETRAD concu par Glymour (1982, 2000)
- ✓ HCA/CCPROX (Hierarchical cluster analysis with proximity matrix développée par Roussos, Stout et Marden (1998)).

La mesure du nombre de dimensions présentes dans les données à travers ces modèles est basée sur des indices statistiques souvent propres à chacun d'eux. Cette section est consacrée aux études de Hattie (1981, 1985); Nandakumar (1994); Blais et Laurier (1995); Linacre (1998); Smith E. (2002); Smith A. et al. (2008); Jasper (2010); Chou et Wang (2010) et Teol et al. (2011) qui ont contribué à la définition et la mise en œuvre pratique du concept de dimensionnalité, et à faire ressortir sa complexité au travers des données empiriques.

Hattie (1981, 1985) avait produit une revue de littérature des méthodologies qui existaient à l'époque de ses travaux pour détecter l'écart à l'unidimensionnalité d'un ensemble d'items. En

tout, 87 indices sur lesquels il est recommandé de baser l'appréciation de la dimensionnalité ont été recensés, et classés en 5 catégories:

- ✓ les indices basés sur le patron des réponses aux items (*answer patterns* en anglais)
- ✓ les indices basés sur la fidélité de l'instrument
- ✓ les indices basés sur l'analyse factorielle
- ✓ les indices basés sur l'analyse en composantes
- ✓ les indices basés sur les modèles de variables latentes qui de nos jours font notamment référence aux modèles de réponse à l'item.

Hattie a simulé des données par la méthode Monte Carlo pour vérifier la pertinence de certains de ces indices pour détecter l'écart à l'unidimensionnalité. Il a finalement recommandé que soient utilisés les indices qui sont basés sur la taille des résidus issus des modèles de variables latentes à deux ou trois paramètres.

On comprend donc, d'après Hattie, que certaines approches comme l'analyse factorielle qui fait partie de la classe des modèles de la théorie classique, seraient moins pertinentes que celles des modèles de variables latentes pour la détection de la dimensionnalité des données associées aux réponses des items.

Nandakumar (1994) avait fait une comparaison des performances de trois méthodes de détection de la dimensionnalité: DIMTEST, la procédure de Holland et Rosenbaum (1986), et l'analyse factorielle non-linéaire. Elle utilise des résultats à 7 tests simulés dont 3 sont unidimensionnels et 4 sont bidimensionnels, et des résultats à 8 tests réels dont 4 unidimensionnels et 4 sont bidimensionnels. Entre les dimensions, la corrélation est fixée à 0,3 et 0,7. Elle aboutit à la conclusion selon laquelle DIMTEST fait bien la distinction entre les tests unidimensionnels et bidimensionnels dans les deux contextes en présence d'une corrélation forte (0,7) ou faible (0,3) entre les dimensions. L'analyse factorielle non-linéaire et la procédure de Holland et Rosenbaum (1986) identifient l'absence de l'unidimensionnalité seulement lorsque la corrélation entre les dimensions est faible (0,3) en présence des données tant réelles que simulées.

Blais et Laurier (1995) se sont intéressés aux aspects méthodologiques de la procédure non paramétrique de détermination de la dimensionnalité DIMTEST. Ils ont recommandé que le construit inclut la théorie de l'apprentissage et la sélection d'un contenu qui prenne en compte le

processus mental approprié. Ils ont aussi utilisé plusieurs approches pour évaluer la dimensionnalité de trois sous-ensembles d'items à choix multiples, calibrés à partir du modèle logistique à trois paramètres. Ces trois sous-ensembles qui constituent le test (administré par ordinateur) de langue française à l'entrée dans les universités et collèges des étudiants anglophones au Canada sont: la compréhension d'un paragraphe court, le choix d'une réponse appropriée dans une mise en contexte et la partie consistant à compléter le vide par le mot approprié.

Les items à ces trois sous tests sont stockés dans trois banques différentes et le score final de l'étudiant est la combinaison des scores aux trois sous-tests comprenant chacun 50 items dans la version papier-crayon. Il y est mis à profit la contribution d'un expert pour classer les items de chaque catégorie en sous-thématique, afin d'utiliser certaines approches d'évaluation.

Les auteurs ont mis à contribution les modèles suivants :

- ✓ Les modèles d'équations structurelles avec le logiciel LISREL. Ils ont contribué à confirmer la théorie qui sous-tend la répartition ou l'association des items dans les trois banques, mais pas la dimensionnalité de chacun de ces 3 ensembles d'items.
- ✓ L'analyse factorielle avec information complète avec le programme TESFACT. Elle a été appliquée au test entier avec les résultats de ceux qui ont répondu aux trois ensembles d'items, puis à chacun des 3 groupes d'items séparément en vue d'explorer la dimensionnalité dans chaque groupe. L'unidimensionnalité du premier sous-groupe d'items (paragraphe) a été confirmé, trois facteurs ont été suggérés pour le troisième sous ensemble d'items alors que le second ensemble posait des problèmes d'interprétation.
- ✓ L'approche non paramétrique (DIMTEST) qui exige la prise en compte de l'opinion d'un expert en contenu. L'analyse factorielle préalable pour diviser l'ensemble d'items en trois classes n'a pas été concluante, par rapport à la répartition de l'expert. L'unidimensionnalité du premier sous-ensemble (paragraphe) d'items a été confirmée, le troisième ensemble d'items a été séparé en grammaire et vocabulaire, au contraire du sous-test 2 (mise en situation) qui est resté sans un statut précis.
- ✓ L'approche basée sur la théorie des réponses aux items avec BILOG. Cette dernière a contribué à l'identification des items à retirer du sous-test sur la compréhension de paragraphe, et n'a pas confirmé d'unidimensionnalité pour les deux autres sous-tests.

Comme enseignement, chaque méthode apporte une information complémentaire qui peut être utilisée pour améliorer le contenu et la répartition des items du test.

Linacre (1998)

L'étude conduite par Linacre avait pour but de comparer trois méthodes de détection de la multi dimensionnalité à partir des résidus du modèle de Rasch. L'analyse de la structure des résidus se justifie par le fait que, lorsque le modèle unidimensionnel de Rasch est appliqué aux données, la composante unidimensionnelle existante est extraite et il ne reste que des résidus normalement distribués et dont la matrice de covariance prévisible ne dégage aucun lien entre eux. Les trois types de forme de résidus à tester pour la détection d'une dimension additionnelle à partir de l'analyse en composantes principales (ACP) sont :

- ✓ Les résidus bruts définis ou la différence entre la valeur observée et celle prédite par le modèle
- ✓ Les résidus standardisés
- ✓ Les résidus logit *ou logit residual en anglais* (résidu brut divisé par sa variance pour un item donné).

Des données dichotomiques bidimensionnelles sur 1190 vecteurs d'observations ont été simulées couvrant 3 domaines : 100 items de mathématiques, 25 en lecture et 50 en vocabulaire. Les paramètres difficulté aux deux dimensions varient entre -2 et 2 logits, et les paramètres habiletés proviennent de la loi normale centrée réduite N (0,1). Pour simuler les données sur le vocabulaire, il a été fait usage de la plus faible habileté entre celle en mathématique et celle en lecture de chaque individu. Le modèle de Rasch a alors été appliqué aux données ainsi obtenues avec le programme BIGSTEPS, l'analyse en composantes principales (ACP) a été appliquée aux trois types de résidus.

Une deuxième simulation de 1000 vecteurs de données sur deux dimensions de 50 items chacune, dont les difficultés sont uniformément distribuées sur l'intervalle (-2,2), a été effectuée avec les deux habiletés suivant une normale N(0,1), mais avec une corrélation de 0,9 entre elles.

Il en ressort que l'ACP des résidus standardisés produit de meilleurs résultats que les résidus non standardisés, et l'ACP de ces derniers produisant de meilleurs résultats que l'ACP sur les résidus logit.

Smith E. (2002) a comparé l'efficacité des statistiques d'ajustement du modèle de Rasch à échelle de notation avec les valeurs propres issus de l'ACP des résidus pour inférer l'existence ou non des items de l'instrument qui font appel à d'autres habiletés en plus de celle visée. Il conclut de l'existence d'un écart à l'unidimensionnalité du test lorsque la première valeur propre de l'ACP des résidus est supérieure à 1,5. Ce résultat s'est avéré limité et difficilement généralisable, car la deuxième valeur propre d'une ACP sur les données unidimensionnelles peut prendre des valeurs plus importantes que le seuil de 1,5 utilisé par Smith dans son étude (Raîche, 2005), et celle-ci dépend à la fois de la taille de l'échantillon et de la longueur de l'instrument.

Smith A. et al. (2008) se sont intéressés à l'influence de la taille de l'échantillon sur les statistiques d'ajustement du modèle de Rasch avec des données polytomiques issues de deux questionnaires (*Patient health questionnaire*, et *Hospital anxiety and depression scale*). Des échantillons constitués par le tirage avec remise de tailles variant entre 25 et 3200 ont été obtenus sur l'ensemble des réponses de 4072 malades atteints du cancer qui ont rempli les deux questionnaires. Le modèle de crédit partiel a été mis à contribution pour les analyses, lesquelles ont produit les statistiques t (*outfit* et *infit*), ainsi que les carrés moyens (*mean squares statistics*). Ils ont conclu que les carrés moyens (*mean squares statistics*) sont plus stables que les statistiques t d'ajustement (fit statistics) lorsque la taille de l'échantillon varie.

Fabian et Jasper (2010) proposent un arbre de décision qui permet de tester à la fois la dimensionnalité et la structure d'un test de performance constitué d'items à réponses binaires. L'application en est faite sur les données du test pré-emploi en mathématiques des 16 ans et plus en Allemagne START-M *Mathematics*, sur un échantillon de 1554 individus. Les auteurs adoptent la définition de Mcdonald (1986), qui définit la dimensionnalité d'un test de performance comme le nombre de facteurs qui sont nécessaires pour rendre compte des relations de corrélations entre les items.

Il aborde aussi la question de la structure du test qui peut être simple (Thurstone, 1947), c'est-à-dire que pour chaque variable, la saturation est plus importante sur un seul facteur et négligeable sur les autres facteurs. Elle est dite complexe lorsque certaines variables ont des saturations aussi importantes sur plusieurs facteurs à la fois, rendant ainsi difficiles la détection et l'interprétation des facteurs.

Il se sert de DIMTEST (Stout, 1987; Namdakumar & Stout, 1993) comme méthode (non paramétrique) d'exploration de la dimensionnalité et de DETECT[1] (Zhang & Stout, 1999) comme méthode confirmatoire pour valider la structure (simple ou complexe) du test. La combinaison de ces deux méthodes permet d'aboutir à la sélection des modèles d'analyses tels que: la théorie des réponses aux items (IRT), l'analyse factorielle non linéaire et les modèles d'équations structurelles (SEM). DIMTEST permet de déterminer si la structure de la matrice de covariance justifie l'hypothèse de l'unidimensionnalité ou non, sur la base d'une statistique T. Sous DETECT, deux indices permettent d'inférer l'unidimensionnalité ou la qualité de la multi dimensionnalité (faible, modérée, forte), tandis que l'indice rmax permet d'inférer qu'une structure est simple ou complexe. L'application de cet arbre de décision sur les données réelles provenant du test START-M a permis de relever l'importance de maîtriser les bases théoriques du trait que le test est supposé mesurer, car pour être utile, la séparation des items par DETECT en présence de plusieurs dimensions doit trouver une justification théorique. Les auteurs ont enfin suggéré de tester cet arbre décisionnel à partir des données issues de la simulation.

Chou et Wang (2010) suggèrent trois statistiques du khi carré pour tester l'hypothèse de l'indépendance des résidus à partir de la matrice de corrélation des résidus standardisés obtenus par le modèle de Rasch, au lieu de faire l'ACP comme suggéré par Smith E. et en conclure à la multi dimensionnalité lorsque la première valeur est supérieure à 1,5, la valeur de référence. Les auteurs posent deux critères à vérifier simultanément pour considérer une valeur de référence d'une statistique:

En premier lieu, la distribution d'échantillonnage de ladite statistique doit être très stable sous plusieurs conditions comme la longueur du test et la taille de l'échantillon.

En deuxième lieu, la distribution d'échantillonnage de la statistique sous l'hypothèse nulle (égalité à la valeur référence) doit être très distincte lorsque l'hypothèse alternative est vraie.

Des données dichotomiques ont été simulées selon cent soixante conditions. Dans chacune des conditions, la simulation a été répliquée cent fois, soit cent ensembles de données. La longueur du test variait de 10, 20, 40 et 60 et la taille de l'échantillon était de 100, 200, 500, et 1 500. La différence entre la difficulté moyenne des items et l'habileté moyenne des individus a été fixée à 0, puis à -1 (items relativement faciles pour les répondants).

[1] *Dimensionality Evaluation to Enumerate Contributing Traits*

Les analyses ont été faites avec le modèle de Rasch, mais les données ont été générées pour tous les items par 2 modèles différents dont le modèle de Rasch et le modèle logistique à trois paramètres. La corrélation entre les deux dimensions était fixée à 0 (indépendance), 0,5 (corrélation modérée) et 0,8 (forte corrélation).

Le calcul de la matrice des corrélations des résidus standardisés a permis de constater que son espérance n'est pas nulle sous Ho (unidimensionnalité). Chou et Wang soustraient donc des valeurs hors diagonale de cette matrice l'espérance trouvée qui est d'environ -1/(I-1) avec I la longueur du test ou le nombre d'items du questionnaire, avant de calculer les trois statistiques d'indépendance multi variée qui sont la statistique de Bartlett (1950, 1954), la statistique de Brien et al. (1984) et la statistique de Steiger (1980a, 1980b).

La comparaison entre ces trois statistiques a montré que la statistique de Steiger pour tester l'indépendance multivariée de la matrice de variance covariance des résidus du modèle de Rasch donne des résultats satisfaisants.

Teol et al. (2011) ont mis à profit les moindres carrés partiels et les équations structurelles pour évaluer la dimensionnalité de la version malaysienne du *Consumer Ethnocentrism Scale* (CETSCALE). Les auteurs évaluent la dimensionnalité du test CETSCALE (*Consumer Ethnocentrism Scale*) auprès d'un échantillon de 398 jeunes âgés entre 16 et 30 ans. Le test CETSCALE a été conçu par Shimp et Sharma en 1987 aux États-Unis. Il est composé de 17 items associés à une échelle de type Likert à 5 modalités. Cette version a été adaptée dans le contexte de l'étude en Malaisie, dont le but était non seulement de tester la dimensionnalité de cet ensemble de 17 items, mais aussi de mesurer la perception des consommateurs vis-à-vis des produits locaux (10 items).

L'unidimensionnalité du CETSCALE a été confirmée, de même que la relation entre le score moyen à ce test et la perception du consommateur vis-à-vis les produits locaux grâce au modèle d'équations structurelles.

Conclusion

L'on peut donc constater que le développement des méthodes de détection de l'unidimensionnalité reste un champ fertile pour la recherche. En l'absence d'une approche consensuelle de détermination du nombre de dimensions, il reste toujours difficile de vérifier que

les réponses à tous les items d'un questionnaire ou sous-questionnaire ne sont déterminées que par une seule variable latente. Au-delà de cet aspect du problème, les méthodes existantes de détermination du nombre de dimensions mesurées par un instrument sont fonction de la définition de la dimensionnalité utilisée, des modèles de mesure, des logiciels et des options des méthodes d'estimations ou de détermination de la dimensionnalité qui y sont implémentées. L'approche abondement utilisée pour confirmer l'unidimensionnalité consiste à faire une analyse en composante principale des résidus du modèle unidimensionnel comme c'est le cas avec le modèle de Rasch (Linacre, 1998 ; Smith, 2002). Mais des études récentes ont montré que cette approche a des limites (Chou et Wang, 2010). C'est compte tenu de l'absence d'une approche unifiée de détermination de la dimension d'un test que nous nous intéressons dans le prochain chapitre à la question de la robustesse des modèles unidimensionnels lorsque l'hypothèse de l'unidimensionnalité est violée.

Chapitre 3:
Robustesse des modèles unidimensionnels de mesure

Étudier les individus selon leur score sur un continuum unidimensionnel lorsque ceux-ci sont décrits avec des items dont les réponses nécessitent plusieurs habiletés peut être dommageable ou approximatif, et sujet à des biais souvent assez importants. Par exemple, plusieurs individus ayant le même score unidimensionnel ne se positionnent pas forcément au même endroit dans le plan, si l'on considère que le score unidimensionnel obtenu avec le modèle de mesure unidimensionnel est une combinaison linéaire des réponses aux items bidimensionnels, c'est à dire une combinaison des scores aux deux dimensions.

La multi dimensionnalité peut se trouver absorbée dans le score unidimensionnel qui confond donc des individus aux caractéristiques pourtant bien distinctes. Par exemple, les tests adaptatifs par ordinateur fonctionnent avec des algorithmes de choix d'items basés sur des méthodes de calibration unidimensionnelles. Si ce calibrage n'est pas fait sous cette condition, c'est-à-dire malheureusement que des items de la banque ne mesurent pas un seul et même trait, la conséquence est que deux individus peuvent se voir proposer deux tests de contenus et longueur différents mesurant des traits latents différents, pourtant ils devraient mesurer un seul et même trait (Way, Ansley, & Fortsyth, 1988).

Lorsque les items contenus dans la banque ne mesurent pas le même trait, il peut arriver que deux individus semblables dans l'espace multidimensionnel décrit par les items de la banque soient évalués avec des tests qui ne sont pas parallèles ou identiques, ou que deux individus ayant des scores totalement différents sur chacune des dimensions de départ se trouvent placés au même endroit sur le continuum unidimensionnel comme s'ils avaient les mêmes habiletés (Akerman, 1991).

Ce chapitre traite de la robustesse du modèle unidimensionnel et comprend trois sections. L'intérêt d'une telle étude est présenté dans la section qui suit. La deuxième section est consacrée dans un premier temps à la revue de littérature sur la robustesse des modèles unidimensionnels de la TRI suivie par une synthèse des études présentées, et ensuite à celle sur la robustesse du modèle de Rasch. Dans cette dernière partie, une place de choix est réservée aux statistiques

d'ajustements (t) au modèle de Rasch, étant donné leur importance pour notre troisième question de recherche.

3.1 Intérêt de l'étude de la robustesse des modèles unidimensionnels

Il y a trois raisons fondamentales pour lesquelles l'on préfèrerait l'analyse avec les modèles unidimensionnels:

En premier lieu, il est plus facile de placer les résultats d'une évaluation sur un même continuum pour permettre une meilleure appréciation des décisions à prendre sur la base du classement des répondants. Ce classement est rendu possible par le score que ces derniers obtiennent suite à une évaluation, mais celui-ci n'est pas facile à percevoir lorsque les individus sont placés dans un espace à plusieurs dimensions, et les décisions à prendre devront tenir compte de la position des individus sur chacune des dimensions. En deuxième lieu, le principe de parcimonie commande que l'on choisisse comme modèle approprié pour mimer les données, celui qui est simple et qui produit des résultats adéquats. En troisième lieu, les modèles implémentés dans les logiciels d'analyse des données d'items sont basés sur le postulat de l'unidimensionnalité, c'est-à-dire qu'un ensemble d'items donnés du questionnaire mesure une seule dimension, et que la propriété d'invariance tient. Comme nous l'avons vu au chapitre 2, des méthodes d'estimation des paramètres utilisent entre autre cette propriété pour le calcul de la fonction de vraisemblance.

Ces trois raisons justifient donc notre recherche sur les conditions de la robustesse des modèles unidimensionnels, en particulier dans la famille des modèles de Rasch.

Dans la section qui suit, nous allons aborder cette question à travers des études empiriques existantes qui traitent de la robustesse des modèles unidimensionnels.

3.2 Études empiriques sur la robustesse des modèles unidimensionnels

Pour appliquer un modèle unidimensionnel à des données multidimensionnelles, il faut que tous les items contenus dans le test mesurent le même construit et de la même façon (c'est-à dire que les saturations des items sur la dimension sont toutes positives), et que les paramètres de difficulté des items soient confondus avec la dimension additionnelle Reckase(1990). Nous présentons des études de robustesse du modèle logistique en premier lieu, puis les tests récents sur l'unidimensionnalité du modèle de Rasch

3.2.1 Modèle unidimensionnel de la théorie des réponses aux items (TRI)

Nous nous proposons de présenter les études faites sur la robustesse du modèle unidimensionnel par Drasgow et Parsons (1983); Blais (1987); Way, Ansley et Forsyth (1988); Ackerman (1992, 1994); Cuesta et Muniz (1999); Kirisci et al. (2001); Walker et Beretvas (2003), puis Kahraman et Thompson (2011).

Drasgow et Parsons (1983) ont étudié la robustesse de l'estimation des paramètres items et personne sous violation de l'hypothèse d'unidimensionnalité avec la version du logiciel LOGIST utilisant la méthode du maximum de vraisemblance pour estimer les paramètres. Cinq ensembles de données sur 50 items à structure simple ont été simulées à la fois sur 1000 observations (sans paramètre de pseudo-chance) et 1500 observations (avec paramètre de pseudo-chance de 0,15 aux items de rang pair et 0,20 aux items de rang impair) avec le programme IMSL (5 facteurs constitués respectivement de 15, 5, 10, 10 et 10 items; suivant 5 niveaux de corrélation entre les facteurs : corrélation totale entre les facteurs; de 0,68 à 0,90; 0,46 à 0,6; de 0,25 à 0,39; de 0,02 à 0,14), à partir du modèle d'analyse factorielle hiérarchique proposé par Schmid et Leiman (1957). Un facteur général de second ordre permettait de contrôler la corrélation entre les facteurs communs du premier ordre.

Sur la base de la racine carrée de la moyenne des carrés des différences des estimés (par le ML2P pour les données sans pseudo-chance, et ML3P pour les données avec pseudo-chance) des paramètres items et personnes, l'étude révèle les cinq points suivants: 1)-la dimension ayant plus de variables (15) semble être assimilée à celle que capte le programme lorsque la corrélation entre les facteurs est trop faible ; 2)-les moyennes des carrés des différences augmentent plus vite dans les conditions ou les paramètres de pseudo-chance sont considérés ; 3)-les modèles unidimensionnels de la théorie des réponses aux items (TRI) sont appropriés pour décrire les données provenant de modèles multidimensionnels lorsqu'il existe une dimension dominante qui est suffisamment prépondérante ; 4)-LOGIST est robuste aux violations mineures de l'unidimensionnalité, ou lorsque la corrélation entre dimensions est modeste (entre 0,46 et 0,60, c'est-à-dire les items sont assez hétérogènes entre les dimensions) ; 5)-lorsque plusieurs items ont des niveaux de difficultés trop élevés, cela pourrait induire la multi dimensionnalité.

Comme limite, on peut noter que travailler en présence des items ayant une structure simple n'est pas évident en pratique car en général les saturations d'un item peuvent être significatives sur

plus d'un facteur à la fois. De plus, l'étude est faite avec cinq dimensions, ce qui s'éloigne des conditions pratiques car au préalable, la validation du contenu du test en rapport avec le trait visé doit être faite au moins avant et après la phase pilote de la collecte des données. Idéalement, on devrait obtenir un outil du test disposant des items en majorité servant à mesurer un même trait.

Blais (1987) s'était penché sur la sensibilité du modèle unidimensionnel avec des données dichotomiques bidimensionnelles simulées par la version du modèle logistique multidimensionnel à deux paramètres, utilisé par Doody-Bogan et Yen en 1983. Le schéma expérimental comprend 55 configurations (p.56) de 30 items à choix multiples et 500 vecteurs d'observations dans trois types de situations généralement rencontrés en pratique : une configuration très proche de l'unidimensionnalité (corrélation de 0,8 entre les dimensions), une situation bidimensionnelle (corrélation de 0,2 entre les dimensions), et une troisième situation dans laquelle domine la confusion (dimensions simultanément en opposition, ayant «des influences contraires sur le processus de mesure»: corrélation de -0,5).

La méthode du maximum de vraisemblance marginale a été utilisée pour estimer les paramètres avec le programme LOGIST. La génération des données s'est faite en deux étapes dont celle de la génération des paramètres suivant une loi uniforme sur l'intervalle (-2 ; 2) pour les habiletés et les paramètres difficultés des items, (0 ; 1,17) pour les paramètres discrimination tel que : $0,5=<a_1=<1,7$ et $0,0=<a_2=<0,5$; et $0,5=<a2$, $a1=<1,7$. Les données dichotomiques représentant 500 candidats sur 30 items à choix multiples ont ensuite été obtenues via une version modifiée du programme DATAGEN qui est écrit en FORTRAN 4.

Des écarts ont été observés entre les valeurs attendues des paramètres et leurs valeurs obtenues par simulations (p. 62). En rapport avec la stabilité des paramètres générés, il ressort que les paramètres difficultés des items sont très stables, contrairement aux paramètres de discrimination dont la stabilité est modérée.

En conclusion, le modèle logistique est sensible à la dimensionnalité au niveau des habiletés, contrairement à celle qui est introduite au niveau de la difficulté des items. L'ajout graduel d'une deuxième dimension (sur la base de la corrélation entre les paramètres habiletés) peut provoquer soit une contraction, soit une extension de la distribution des estimations du paramètre difficulté et une baisse générale de celles du paramètre discrimination (p.133). Compte tenu de cette dernière implication sur la variation des paramètres des items, on est emmené à penser qu'une

situation idéale pourrait se produire avec des items disposant un seul paramètre comme c'est le cas avec le modèle de Rasch.

Way, Ansley et Forsyth (1988) ont étudié les effets de l'estimation unidimensionnelle des paramètres des modèles compensatoires et non compensatoires bidimensionnels. Le logiciel utilisé pour la production des mesures avec le modèle logistique à 3 paramètres est LOGIST. La simulation (cinq réplications) est faite avec 60 items à structure simple sur deux dimensions, et sur 2000 observations.

Les résultats suivants ont été obtenus : Pour le modèle non compensatoire, la discrimination moyenne du test est la moyenne des discriminations moyennes des items aux deux dimensions, alors que pour le modèle compensatoire, c'est la somme des discriminations moyennes des items aux deux dimensions. Pour la difficulté moyenne des items sur le continuum unidimensionnel, il est apparu que celle-ci est plus grande que la difficulté moyenne des items à la première dimension pour le modèle non compensatoire, mais pour le modèle compensatoire, la valeur de la difficulté moyenne est la moyenne des difficultés moyennes des items aux deux dimensions.

Ces résultats montrent que la combinaison des paramètres aux différentes dimensions qui résulte de l'utilisation du modèle unidimensionnel n'est pas toujours stable même avec le même ensemble de données : somme versus moyenne.

Ackerman (1992, 1994) avait étudié la sélection d'items d'une banque contenant des items bidimensionnels dans le cadre des tests adaptatifs par ordinateur. Il en ressort que les items ayant des niveaux de difficultés aux deux dimensions qui sont similaires à celui estimé par le modèle unidimensionnel ont un paramètre de discrimination plus grand et donc susceptible d'être sélectionnés pour être administrés. Dans la deuxième étude en 1994, il relève deux points importants à retenir avant toute interprétation du score unidimensionnel issu de données multidimensionnelles : non seulement il faut savoir quelle est la combinaison des dimensions initiales qui est mesurée le long du continuum unidimensionnel, mais aussi il faut s'assurer que tous les individus sont évalués selon la même combinaison. Lorsque le modèle unidimensionnel est utilisé pour estimer les habiletés alors que le modèle multidimensionnel est approprié, l'estimé de l'habileté qui en résulte est une combinaison linéaire des habiletés qui auraient pu être

obtenues si le modèle multidimensionnel était utilisé. De plus, si la difficulté et la dimension se confondent, la combinaison linéaire n'est pas constante le long du continuum unidimensionnel.

Cuesta et Muniz (1999) avaient étudié la robustesse des modèles logistiques à la violation du postulat de l'unidimensionnalité. Deux ensembles de données bidimensionnelles ont été simulés à partir du modèle logistique multidimensionnel à deux paramètres, avec le programme M2GEN2, sous les contraintes suivantes : Les deux tests comprennent 40 items dichotomiques dont 20 saturent sur la première dimension, et les 20 autres saturent sur la deuxième dimension pour le premier test, et 25 saturent sur la première dimension, et les 15 autres saturent sur la deuxième dimension pour le deuxième test. Pour chacun des tests, les données simulées contiennent 300 et 1000 vecteurs d'observations sur les 40 items dichotomiques, et la corrélation entre les dimensions qui sont calibrées sur une loi normale centrée réduite, comprend cinq niveaux dont 0,05, 0,30, 0,60, 0,90 et 0,95.

Dans les deux conditions de tests, pour simuler les paramètres habiletés, 25 items ont été contraints de discriminer plus sur une dimension, et 15 sur l'autre. Les logiciels d'analyse sont SPSS pour calculer le coefficient Alpha de Cronbach en particulier, alors que les paramètres estimés ont été obtenus avec le logiciel BILOG.

Les racines des moyennes des carrés des différences ainsi que les coefficients de Pearson ont servi de critères de comparaison entre les paramètres multidimensionnels et leurs correspondants unidimensionnels. Selon cette étude, les estimations obtenues avec le modèle logistique à deux paramètres (2PLM) sont robustes.

En particulier, le paramètre difficulté est moins affecté par la violation de l'unidimensionnalité, en comparaison aux autres paramètres. En effet, les résultats montrent que le paramètre discrimination et le paramètre habileté sont influencés par le niveau de corrélation entre les dimensions et le type de la multi dimensionnalité des données.

La taille d'échantillonnage passant de 300 à 1000 n'a pas eu d'effet sur la précision des estimés des paramètres. Par contre cette dernière s'améliore lorsque le coefficient de corrélation entre les deux dimensions augmente. Pour un item quelconque, l'indice de discrimination d'un item sur sa composante de la deuxième dimension est toujours plus proche de celui obtenu avec le modèle unidimensionnel, que ne l'est sa composante sur la première dimension. Toutefois, leur moyenne

est proche de l'estimé unidimensionnel, et c'est l'indice de discrimination multidimensionnel qui lui est le plus distant.

Lorsque la corrélation entre les dimensions diminue, les estimés unidimensionnels des paramètres discrimination des items sont plus corrélés avec leurs paramètres discrimination sur la première dimension, et le sont avec ceux sur la deuxième dimension au détriment de la première, lorsque la corrélation entre dimension augmente.

Il a aussi été observé à la fois de fortes corrélations entre le paramètre difficulté unidimensionnel et le paramètre difficulté multidimensionnel (d) d'une part, et avec la distance (D) entre le point de discrimination maximale et l'origine du plan formé par les deux dimensions.

Pour la deuxième étude, il est ressorti que l'estimé unidimensionnel du paramètre discrimination est plus proche de la moyenne de ce paramètre aux deux dimensions. Sa corrélation avec l'indice multidimensionnel de discrimination est aussi élevée. La corrélation du paramètre difficulté avec D ou d est très élevée. Au sujet des habiletés, l'écart est plus faible entre les valeurs unidimensionnelles estimées et les habiletés à la première dimension, qu'avec la seconde. Aussi, les valeurs unidimensionnelles sont fortement corrélées avec l'habileté moyenne.

Les deux principales limites de cette étude sont les suivantes : Le ratio 25 :15 pourrait ne pas être suffisant pour inférer une dimension dominante, car 15 items peuvent être suffisants pour constituer une dimension indépendante et à part entière.

En plus du fait que nous avons des données dichotomiques qui ne se prêtent pas aux manipulations des variables continues, l'utilisation des valeurs propres pour juger de l'unidimensionnalité est un problème comme nous l'avions signalé précédemment (Chou et Wang, 2010). On se serait attendu à ce que même dans le cas de 20 items par dimension, et compte tenu du cas de très faible corrélation (0,05), que les 3 valeurs propres reportées indiquent clairement une bi dimensionnalité plutôt qu'une seule dimension.

Comme dans l'étude de Blais (1987), étudier la robustesse du modèle unidimensionnel avec des items ayant plus d'un paramètre est plus ardue et pourrait s'avérer improbable que si elle est conduite avec des items à un seul paramètre, la difficulté puisque le paramètre discrimination obtenu par application du modèle unidimensionnel sur des données bidimensionnelles semble imprévisible.

Kirisci et al. (2001) avaient étudié la robustesse des estimations des paramètres items produites avec les logiciels BILOG, MULTILOG et XCALIBRE lorsque les hypothèses d'unidimensionnalité et de normalité de l'habileté sont violées. Cette expérimentation a été conduite, autant que faire se peut dans des conditions identiques pour les analyses à travers les trois logiciels.

Les données ont été simulées à partir de la variante unidimensionnelle d'une part, et tridimensionnelle d'autre part du modèle logistique compensatoire à trois paramètres proposé par Hattie en 1981 dans les conditions suivantes : Trois variantes de la distribution normale centrée réduite de l'habileté ont été retenues dont l'une parfaitement normale : a) coefficient d'asymétrie=coefficient d'aplatissement=0 ; b) coefficient d'asymétrie=0,75 et d'aplatissement=0 ; c) coefficient d'asymétrie=0 et coefficient d'aplatissement=-1. La corrélation entre les habiletés dans le cadre tridimensionnel a été fixée à 0,6 pour garantir une forte corrélation.

Les paramètres items suivent tous la distribution uniforme avec les spécifications suivantes : $.4<a<2$; $-2<b<2$; $.2<c<.3$. Trois vecteurs différents ont été simulés pour les paramètres discrimination et difficultés des items dans le cadre multidimensionnel, mais un seul pour le paramètre de pseudo-chance. La déviation à la normale a été obtenue par la fonction puissance de Fleishman (1978) (p.149). Le test comprend 40 items et 1000 observations. Il a donc été créé 6 conditions dont 10 réplications dans chaque cas, sous lesquelles les trois logiciels ont été mis à contribution.

La comparaison des résultats entre les trois logiciels s'est faite par la racine de la moyenne des carrés des erreurs pour deux raisons : en premier lieu, elle possède des propriétés différentielles lorsque les scores sont normalisés par la transformation logarithmique- deuxièmement, elle est généralement utilisée dans les études de récupération des paramètres.

La méthode du maximum de vraisemblance a été mise à contribution pour estimer les paramètres items pour BILOG et MULTILOG. Cette option n'est pas disponible dans XCALIBRE. L'estimation a priori de l'habileté a été faite dans BILOG et XCALIBRE, MULTILOG n'ayant pas cette fonctionnalité lorsque les données items sont en input. Il est cependant important de noter que ces deux caractéristiques qui sont implémentées chacune dans deux des trois logiciels, limiteraient finalement les comparaisons des résultats entre les trois logiciels.

Il ressort en particulier que l'augmentation de la taille de l'échantillon ne contribue pas toujours à améliorer l'estimation du paramètre thêta (habileté de la personne), mais cela est possible en augmentant la longueur du test- la robustesse de l'hypothèse de l'unidimensionnalité est fonction du logiciel, et non de la distribution de thêta, car en effet tous les logiciels n'utilisent pas les mêmes méthodes d'estimations- lorsque dans les données il y a une seule dimension dominante avec des dimensions faibles supplémentaires, le modèle unidimensionnel de la théorie des réponses aux items peut être appliqué (mais il faut un pourcentage minimal de la variance expliquée par le premier facteur)- si la corrélation entre les dimensions est très forte ou supérieure à 40%, on peut aussi appliquer le modèle unidimensionnel- lorsque la corrélation entre les dimensions varie assez ou qu'elle est faible (inférieure à 40%), alors il faut utiliser le modèle multidimensionnel.

Walker et Beretvas (2003) avaient travaillé avec les données sur 63 533 et 65 279 étudiants de 4^{e} et 7^{e} secondaire respectivement obtenues suite à l'administration en 1998 du test *Washington Assessment of Student Learning* (WASL).

Leur but premier est d'étudier l'effet du classement des étudiants des niveaux 4 et 6 du secondaire sur un continuum unidimensionnel sur la base du sous-test en mathématique qui est réputé bidimensionnel et comprend 24 ou 30 items à choix multiples en rapport avec l'habileté générale en mathématiques; et 16 items à réponse ouverte qui sont supposés mesurer l'habileté à communiquer en mathématiques.

Le second but est de comparer le positionnement des étudiants sur le continuum unidimensionnel d'avec leur positionnement sur le modèle bidimensionnel. Le calibrage des paramètres items des deux modèles a été fait avec le logiciel NOHARM dont la version disponible à cette date ne fonctionne qu'avec les données dichotomiques.

Pour rendre possible le calibrage, tous les scores aux questions ouvertes ont été dichotomisés. Les paramètres items ont été estimés par le modèle d'ogive normal, l'équivalent du modèle logistique à 3 paramètres dans le cas unidimensionnel. Toutefois, le logiciel MULTILOG a d'abord été mis à contribution pour estimer les asymptotes inférieures ou paramètres de pseudo-chance des items à choix multiple, puisque ces paramètres doivent être fixés lorsque le calibrage doit se faire avec NOHARM.

Compte tenu des études déjà conduites sur ce sujet, il est donc postulé deux dimensions, avec les 16 items à réponses ouvertes saturés seulement sur la deuxième dimension, mais tous les items du test y compris ceux à réponses ouvertes sont saturés à la première. Ceci suppose donc une corrélation entre les deux facteurs.

Les estimés des habiletés des deux modèles ont ensuite été obtenus grâce au programme 2D-EAP de Luecht (1992). Les résultats obtenus sont les suivants : sur la base de l'indice RMSDE, le modèle multidimensionnel est plus adéquat que le modèle unidimensionnel, mais la différence (0,01) du RMSDE est faible.

Le test sur les fonctions d'information peut aussi servir de validation au côté du RMSDE, couplé avec les nuages de coquilles (*Clam shell plots* en anglais). La corrélation entre les dimensions est de 0,61 au secondaire 4 et de 0,81 au secondaire 7. Quatre classes dans lesquelles sont affectés les étudiants suivant les deux modèles.

Le croisement entre les deux confirme la différence minime entre les deux modèles aux deux niveaux scolaires. Quatre-vingt-dix pourcent des étudiants du secondaire 4 et 84% au secondaire 7 ont été affectés aux classes d'habileté (1, 2, 3 et 4) identiques par les deux modèles. Toutefois, malgré ces résultats, il faut noter que l'habileté obtenue suite à l'application du modèle unidimensionnel est l'habileté générale en mathématique- les étudiants qui avaient de faibles scores sur la composante communication en mathématiques avaient de fortes chances d'être placés dans une classe à faible habileté en mathématiques sur le continuum unidimensionnel, contrairement à leur classement d'après le modèle multidimensionnel. Ce qui est révélé par l'analyse des 10% et 16% des étudiants qui ont été placés dans des classes différentes selon les deux modèles.

Les auteurs retiennent le modèle multidimensionnel malgré qu'il soit plus complexe que le modèle unidimensionnel au vu des résultats obtenus, comme il est identifiable et interprétable dans le cas étudié.

Il ressort des résultats de cette étude que l'analyse des données multidimensionnelles aux fins de classement des étudiants avec le modèle unidimensionnel conduit à surclasser sur le continuum unidimensionnel les étudiants qui ont de bonnes performances sur la deuxième dimension et déclassent ceux qui n'y ont pas de bonnes performances.

Le diagnostic quant à l'aide à apporter aux étudiants ne serait pas efficace étant donné que le modèle unidimensionnel ne capte que la dimension prépondérante qui est la connaissance

générale en math. Bien qu'étant très habile en mathématiques, un étudiant serait identifié comme ayant des difficultés en mathématiques alors que le diagnostic approprié est celui de la difficulté à communiquer les mathématiques.

Kahraman et Thompson (2011) ont conduit une étude dont le but était de comparer les résultats des estimations des paramètres obtenues à partir de deux procédures de projection des items multidimensionnels sur une droite composite. L'une est analytique et l'autre empirique. Ces deux modèles, dits modèles de projection de la TRI ont été développés pour paramétrer les projections unidimensionnelles des items multidimensionnels d'un test.

En effet, dans la pratique, lorsqu'un test est composé de plusieurs sous-tests qui sont censés mesurer des dimensions différentes, il arrive qu'un des sous-tests n'ait pas assez d'items pour fournir des mesures fiables et significatives. Cette étude a montré que l'information collatérale provenant des items appartenant aux autres sous-tests peut être mise à profit pour améliorer la fidélité et la validité des sous-tests sous représentés, et ceux-ci peuvent produire de l'information pour diagnostiquer les forces et faiblesses des candidats à un test.

Les données étudiées sont dichotomiques et obtenues par simulation des réponses à 20 items bidimensionnels (dont 10 items ayant des paramètres difficultés plus élevés sur une des deux dimensions, et faibles sur l'autre). Les paramètres discrimination et localisation ont été calibrés à partir des études antérieures. La simulation de 1000 observations est répliquée 30 fois avec le modèle logistique à 2 paramètres (ML2P). Deux logiciels sont utilisés dont Mplus pour analyser les données bidimensionnelles, et Splus version 7.0 pour les autres calculs. La méthode du maximum de vraisemblance marginale a été mise à contribution pour les estimations des paramètres.

Cette étude utilise deux méthodes dont l'une est analytique et l'autre est empirique: Suivant la méthode analytique, l'estimation sur le continuum unidimensionnel des paramètres discrimination, difficulté et localisation des items multidimensionnels est faite à partir des équations équivalentes à celles vues au chapitre 2. La procédure de la méthode empirique se déroule en deux étapes comme suit : Supposons N=N1+N2 des items bidimensionnels dont N1 saturant le plus sur la première dimension, et N2 sur la deuxième dimension. 1) les items sont calibrés en fonction de la dimension principale qu'ils sont censé mesurer. Dans notre cas, les N1 premiers et N2 derniers items sont considérés séparément pour être calibrés avec un modèle

unidimensionnel, 2) les items qui ne sont pas supposés mesurer prioritairement la dimension en étude mais pourraient apporter une information sur celle-ci sont ajoutés un à un puis calibrés successivement, tout en conservant les estimés déjà obtenus du premier ensemble d'items. Les items supplémentaires sont calibrés en maximisant une fonction de vraisemblance.

Les résultats de l'étude sont les suivants: 1)-Les estimés obtenus des paramètres de localisation et discrimination sont sujets à des biais, mais la procédure analytique a l'avantage de la précision ; 2)-En rapport avec les vraies valeurs des paramètres bidimensionnels, la procédure analytique produit des estimés extrêmes lorsque l'item est trop facile ou a une discrimination nulle en rapport avec l'une des dimensions ; 3)-La procédure empirique pourrait conduire à des estimés faibles ou élevés du paramètre discrimination lorsque le sous ensemble d'items de la dimension est (respectivement) trop facile ou ne discrimine pas assez ; 4)-Des items conservent un pouvoir de discrimination dans leur dimension, mais perdent suffisamment celui de l'autre dimension ; 5)-Lorsque les items se trouvent plus en adéquation ou plus alignés avec le continuum composite de référence du test, ils perdent leur pouvoir de discrimination sur leur dimension originale au profit de l'autre, mais discriminent le plus sur la droite composite de référence du test.

Cela pourrait signifier que de tels items requièrent une combinaison équitable des habiletés aux deux dimensions ; 6)-L'indice de discrimination d'un item par rapport à une des droites composites diminue lorsque l'angle entre l'item et cette droite augmente ; 7)-Les mesures de fidélité au niveau des dimensions individuelles de départ s'améliorent avec l'information collatérale due à l'introduction graduelle des N2 ou N1 items supplémentaires lorsque ceux-ci contribuent à mesurer la caractéristique visée. Cela peut aider à fournir des informations aux parties prenantes sur les forces et faiblesses des candidats.

Cette étude a fourni une application concrète des formules du modèle de projection de la TRI. Cette approche est semblable à celle que nous comptons explorer avec le modèle de Rasch, mais à la différence que nous ne ferons pas une d'étude comparative entre deux modèles de mesure.

3.2.1.1 Synthèse des études sur la TRI

Le tableau 1 ci-dessous est un résumé des paramètres des études précédentes sur la robustesse.

Les deux questions essentielles auxquelles ces études ont tenté de répondre sont en général orientées vers le rapport entre les estimés des paramètres et leurs valeurs réelles obtenues par la simulation, et l'identification du trait latent mesuré par le modèle lorsque les données réelles sont

utilisées dans l'étude. Dans notre travail, nous aborderons aussi la question de l'ajustement des scores des items, en plus de deux citées précédemment. Les constats suivants se dégagent de la revue de littérature :

- ✓ l'essentiel des études de robustesse des modèles unidimensionnels tourne en particulier autour des modèles multidimensionnels de la théorie des réponses aux items (TRI).
- ✓ la simulation des données avec les items à deux paramètres posent des défis complexes en particulier lorsqu'on s'intéresse à la relation entre l'indice de discrimination estimé par le modèle unidimensionnel et les indices de discrimination multidimensionnels simulés. Comme révèlent les études de Blais(1987), Way et al.(1988), Akerman (1991, 1994) et Cuesta et Muniz (1999), les paramètres unidimensionnels obtenus sont des combinaisons linéaires des paramètres multidimensionnels, mais ces combinaisons peuvent être instables le long du continuum unidimensionnel. Pour avoir le plein contrôle des résultats, il semble indispensable de maîtriser le comportement des paramètres items, en particulier celui de la discrimination. Une exploration de la robustesse des modèles précédents avec le paramètre difficulté seulement pourrait aboutir à des résultats différents et stables.
- ✓ dans les études utilisant des données simulées, l'on a pu constater l'absence du contrôle de l'égalité entre la difficulté moyenne du test et l'habileté moyenne des personnes (Test Person Matching (TPM) =0) comme souligné par Chou et Wang (2010) car les items trop difficiles peuvent induire une dimension additionnelle illusoire (Drasgrow et Parsons (1983) ; Akerman (1991)).
- ✓ le fait que les lois de probabilité des paramètres items et personnes diffèrent comme dans les études de Blais (1987), Way et al. (1988) et Kirisci et al. (2001) ne contribuent pas à garantir l'une des conditions idéales qui est la nullité du TPM.
- ✓ les logiciels de simulation des données multidimensionnelles et ceux avec lesquels les modèles de mesure unidimensionnel sont implémentés diffèrent. Dans plusieurs cas, les données simulées ont été dichotomisées afin d'être pris en charge par le logiciel.

Tableau 1

Synthèse des Conditions des Études sur la Robustesse du Modèle Logistique

Études **Paramètres de l'étude**	Drasgow et Parsons (1983) ;	Blais (1987)	Way, Ansley et Forsyth (1988)	Cuesta et Muniz (1999)	Kirisci et al. (2001)	Walker et Beretvas (2003)	Kahraman et Thompson (2011)
Logiciel ou modèle utilisé pour la simulation	ISML	Logiciel : I.M.S.L. (1982) Modèle : MLM2P (Doody-Bogan et Yen, 1983)	Voir Ansley et Forsyth (1985a)	M2GEN2	ML1Pet MLC3P* (Hatie, 1981)	pas de simulation	pas de simulation
Lois de probabilité des paramètres simulés	Modèle d'analyse factorielle hiérarchique	Loi uniforme pour les paramètres habileté et items	la loi bi normale centrée réduite pour l'habileté, mais la distribution uniforme pour les paramètres items	la loi normale centrée réduite pour le paramètre habileté	la loi normale centrée réduite pour le paramètre habileté, mais la distribution uniforme pour les paramètres items	pas de simulation	pas de simulation
Logiciel utilisé pour la mesure	LOGIST	LOGIST	LOGIST	SPSS, BILOG	BILOG, MULTILOG, XCALIBRE	NOHARM, MULTILOG, 2D-EAP (Luecht, 1992)	Mplus, Splus 7.0
Modèle de mesure	ML2P et ML3P	ML2P	ML3P	ML2P	ML3P	ML3P	ML2P
Nombre total d'items du test	50	30 CM	60	40	40	40 et 46	20
Taille de l'échantillon	1000 et 1500	500	2000	300 et 1000	1000	63533 et 65279	1000
Format des items à l'estimation	dichotomiques	dichotomiques	dichotomique	dichotomique	dichotomiques	dichotomiques	dichotomiques
Structure des items ou type de la multi dimensionnalité	simple	simple	simple	simple	simple	Dimensionnalité inter-items	Dimensionnalité intra-items
Nombre de dimensions	5	2	2	2	3	2	2
corrélation entre dimensions ?	oui	oui	oui	oui	oui	-	-

*Modèle logistique compensatoire à trois paramètres

En pratique, d'autres modèles sont aussi utilisés pour modéliser les données d'items ou pour aider au développement des outils de collecte de qualité. C'est le cas des modèles de Rasch, dont l'application est de plus en plus répandue en médecine et en sciences sociales depuis quelques décennies. En plus, l'offre s'est diversifiée en termes de logiciels d'application des modèles de Rasch ou de simulation de données s'ajustant aux modèles de Rasch, créant ainsi les possibilités de mener des études de robustesse semblables à celles que nous venons d'explorer.

Nous allons explorer la famille des modèles de Rasch, en travaillant avec des données d'items dichotomiques obtenues par simulation avec le modèle bidimensionnel, auxquelles sera appliqué le modèle de Rasch unidimensionnel simple. Contrairement aux simulations faites dans les études précédentes, le modèle de Rasch considère le niveau de difficulté comme le seul paramètre item. Trois des tests généralement pratiqués pour tester la validité de l'hypothèse d'unidimensionnalité des données lorsque le modèle de Rasch est utilisé pour produire les mesures sont ci-dessous présentés.

3.2.2 Robustesse du modèle de Rasch unidimensionnel.

L'avantage du modèle de Rasch réside dans le fait qu'il produit des estimations de meilleure précision même avec de petits échantillons, contrairement aux modèles logistiques. L'étude de la robustesse du modèle de Rasch peut être abordée sous plusieurs angles parmi lesquels la violation des hypothèses sur les paramètres items, de la propriété d'invariance des estimés ou la violation de l'hypothèse de l'unidimensionnalité. Plusieurs facteurs en pratique justifient une telle démarche (Forsyth, 1981): en premier lieu, les tests standards de compétences sont généralement construits sur la base d'un tableau de spécification de contenu par processus, ce qui impliquerait donc plusieurs facteurs dans les données, même si un seul facteur dominant en émerge qui explique une forte proportion de la variance dans les données. En deuxième lieu, on utilise souvent un même stimulus pour plusieurs items du test. Par conséquent la propriété d'indépendance locale n'est pas toujours respectée. En plus, de tels items pourraient avoir un pouvoir de discrimination distinct l'un de l'autre. Enfin l'usage des items à choix multiples peut emmener des répondants à faire des choix de bonnes réponses au hasard.

Forsyth et al. (1981) avaient traité la question de l'invariance des estimations des paramètres items et personnes par le modèle de Rasch. Cette approche de l'étude de la robustesse avait été

abordée à travers un même test (Iowa tests of Educational development) administré à des étudiants de niveau scolaires différents (9 à 12) en Iowa. Le modèle de Rasch s'était avéré robuste car les estimés des paramètres items étaient invariants à travers les groupes d'étudiants. De même les estimés des paramètres personnes étaient restés invariants à travers les sous-groupes d'items du test.

Fons (1986) avait étudié la sensibilité du modèle de Rasch à la violation des hypothèses sur les paramètres items. Avec le modèle de Rasch, Il est supposé que tous les items ont un paramètre discrimination identique et égal à 1, et le paramètre de pseudo-chance est nul. Il suffit de faire une modélisation de Rasch avec des données de réponses aux items dont le paramètre discrimination n'est pas unique pour tous les items, ou le paramètre de pseudo-chance n'est pas du tout nul pour tous (surtout avec certains items à choix multiples dont les réponses sont ensuite dichotomisées).
Les données avaient été simulées selon que le paramètre de pseudo-chance ou discrimination n'est pas nul pour tous les items, ou que les deux paramètres à la fois ne respectent pas la condition d'application du modèle de Rasch. Les paramètres discrimination et de pseudo-chance suivaient une distribution uniforme, alors que les paramètres difficulté et personnes suivaient la distribution de Gauss. La formule ci-dessous de la probabilité de donner la bonne réponse étant donné les paramètres items et personnes avait permis de générer 50 matrices des observations dichotomiques dans chacune des combinaisons nombre d'items-nombre de personnes suivantes (10:25), (25:50), (25:100), (25:500) et (50 :500).

$$P(X = 1|\theta) = c + (1 - c) * \frac{1}{1+e^{\alpha*(b-\theta)}}$$

Où c est le paramètre de pseudo-chance, b le paramètre difficulté de l'item et θ le paramètre habileté de la personne.

La modélisation de Rasch avait été appliquée à ces matrices de données obtenues malgré que le paramètre de pseudo-chance n'était pas nul, ou le paramètre de discrimination n'était pas identique et égal à l'unité pour chacun des items.
Les mesures de la précision des estimés utilisées sont:

- ✓ la corrélation entre les paramètres items et leurs estimés

- ✓ la corrélation entre les paramètres personnes et leurs estimés
- ✓ les erreurs standards des estimés des paramètres items et personnes
- ✓ la racine carrée de la moyenne des carrés des différences entre les paramètres et leurs estimés

Avec le critère de la corrélation, le modèle de Rasch était robuste même en présence d'un paramètre de discrimination hétérogène, sauf quand ce dernier prenait des valeurs extrêmes. Aussi, la corrélation diminuait lorsque la pseudo-chance était introduite, en particulier lorsque ce paramètre prenait des valeurs élevées. Mais globalement, l'impact sur la corrélation n'était pas important. Il n'y avait pas d'effet amplifié sur la corrélation lorsque ces deux paramètres étaient introduits simultanément.

En rapport avec le critère de l'erreur standard, l'étude avait montré que le modèle de Rasch restait valide même lorsque le paramètre discrimination était hétérogène. En plus, des valeurs élevées de la pseudo-chance conduisaient à des erreurs standards plus petites du trait latent et des paramètres items et personnes, sans que ceci soit corrigible par la longueur du test. Enfin, sous violation des deux hypothèses, il avait été constaté que l'erreur standard était surtout déterminée par la pseudo-chance. Il semble donc possible d'estimer l'erreur standard d'un trait latent sans faire d'hypothèse sur sa distribution lorsqu'il y a assez d'observations et le paramètre de pseudo-chance est nul.

Les valeurs du biais et de la racine carrée de la moyenne des carrés des écarts étaient plus fortes en présence de la pseudo-chance, que ce soit pour les personnes ou les items. Toutefois, l'ordre entre les difficultés des items était maintenu bien que la différence entre ces paramètres et leurs estimés était plus élevée. Dans cette étude, il est recommandé la mise à contribution du biais et de la racine carrée de la moyenne des carrés des écarts lorsque l'intérêt du chercheur est mis sur les valeurs exactes des paramètres.

L'étude de la robustesse par la violation de l'hypothèse de l'unidimensionnalité a été abordée à travers des tests exposés ci-dessous sur les statistiques globales. Les trois approches suivantes sont généralement utilisées pour tester l'unidimensionnalité des données de réponses aux items (Brentani, 2007) :

- ✓ la première approche est basée sur le test a priori, à l'exemple de l'analyse factorielle linéaire ou non linéaire, le test de Martin-Löf (1973) pour items dichotomiques et ses extensions aux items polytomiques.
- ✓ La deuxième approche est basée sur le test a posteriori, à l'exemple des statistiques d'ajustements infits et outfits et l'analyse en composante principale des résidus du modèle de Rasch.
- ✓ La troisième est basée sur des procédures non paramétriques comme DIMTEST (Stout, 1987).

Les premiers tests sur la violation de l'axiome d'unidimensionnalité avec les modèles de Rasch sont apparus au cours des années 1970, soit une décennie après le développement du modèle de Rasch en 1960. Parmi eux on peut citer :

- ✓ le test utilisant la technique du splitter-item de Molenaar (1983)
- ✓ le test Q 2 de Van den Wollenberg (1979, 1982)
- ✓ le test du ratio de vraisemblance (Anderson, 1973)
- ✓ le test de Martin-Löf (1973)
- ✓ le test de Fisher-Scheiblechner (1970)
- ✓ le test de Wright et Panchapakesan (1969)

Tous sont construits autour des statistiques globales et requièrent que les items soient a priori partitionnés en deux groupes représentant chacun une dimension du test. Cette contrainte qui nécessite du chercheur une parfaite connaissance du contenu est la source de nombreuses imperfections (Wollenberg, 1982). Nous présentons dans cette section trois tests dont celui de Martin-Löf (1973) qui est considéré comme le plus efficace parmi ceux proposés au cours des années 1970 (Verhelst, 2001), et deux autres plus récents dans la littérature, qui sont des extensions du premier. Tous trois ont été mis en œuvre pour tester l'écart à l'unidimensionnalité avec des données simulées à partir des modèles de Rasch.

3.2.2.1 **Le test de Martin-Löf (1973)**

Le test de Martin-Lof (1973) sur les données d'items dichotomiques est présenté, et suivront ses extensions aux données à items polytomiques.

Dans son article, **Andersen (1973**) a résumé les travaux de Martin-Löf qui n'étaient disponibles qu'en suédois, et qui avaient été dispensés dans une série de cours à l'université de Stockholm.

Supposons un test composé de n= n_1 + n_2 items dont n_1 et n_2 saturent seulement sur la première et sur la deuxième dimension respectivement.

L'hypothèse à tester est celle selon laquelle tous ces n items saturent sur une seule et même dimension, ce qui justifierait l'utilisation du modèle de Rasch pour générer des mesures sur un seul trait. Ce test est basé sur le ratio de vraisemblance dont le numérateur est la vraisemblance obtenue lorsque les n items sont mis ensemble et supposés mesurer le même trait, et le dénominateur est le produit des vraisemblances obtenues avec les deux groupes d'items de tailles n_1 et n_2.

En effet, si l'hypothèse de l'unidimensionnalité tient, alors le produit des deux vraisemblances conditionnelles est égale à la vraisemblance totale L= L_1* L_2. Dans ce cas, le ratio $\lambda = \frac{L}{L1*L2}$ est sensiblement égal à 1. Ce ratio prend des valeurs entre 0 et 1. Il s'éloigne de 1 lorsque l'hypothèse d'unidimensionnalité ne tient pas. La statistique $Z = -2log\lambda$ suit un khi carré. Sur la base du niveau du test z_0, on pourra accepter ou rejeter l'hypothèse nulle selon que Z est inférieur ou supérieur à z_0. Deux faiblesses de ce test résident dans le fait que sa puissance dépend de la répartition a priori des items entre les deux groupes par le chercheur. Ce qui n'est pas toujours garanti en particulier lorsque l'étude est à but exploratoire et le chercheur ne dispose pas d'éléments suffisants pour prendre des décisions éclairées en rapport avec le contenu du test.

En deuxième lieu, la valeur critique (p-value en anglais) sur laquelle se base le test conduit à des décisions biaisées, car on suppose toujours une distribution du trait latent en étude, qui pourrait être différente de sa vraie distribution. En effet sous l'hypothèse nulle (unidimensionnalité), la distribution s'écarte asymptotiquement du khi carré lorsque la taille de l'échantillon n'est pas trop élevée (Christensen et al., 2002). L'approche Monte Carlo proposée par Kristensen et Kreiner (2007) permettant d'obtenir une meilleure estimation de la valeur critique est exposée dans le paragraphe suivant.

3.2.2.2 Extensions du test de Martin-Löf aux modèles de Rasch à items polytomiques

Kristensen et al. (2002) ont proposé une extension du test de Martin-Löf aux modèles de Rasch à items polytomiques. Deux tests ont été proposés par ces auteurs dont l'un est basé sur la fonction de vraisemblance marginale multidimensionnelle (MML) et l'autre sur la fonction de vraisemblance conditionnelle multidimensionnelle (CML). Tous s'appuient sur le ratio de vraisemblance entre l'unidimensionnalité et une multi dimensionnalité fixée par le chercheur.

Pour évaluer la distribution asymptotique des statistiques de ces tests ainsi que leur puissance, des données bidimensionnelles ont été obtenues par simulation avec successivement 200, 500 et 1000 observations; et une corrélation entre les deux dimensions fixée à 0,43, 0,8 et 0,9. Le test est à structure simple et comprend neuf items polytomiques dont les valeurs possibles vont de 0 à 4, répartis sur trois dimensions constituées chacune de trois items. En vue de tester si les items forment une seule dimension ou non, le test du ratio de vraisemblance proposé par Andersen (1973) a été mis à contribution.

Une limite à cette étude est le nombre trop faible d'items (3) par dimension, par rapport à la vie pratique ou l'on a parfois besoin de plus de neuf items en vue de couvrir une seule dimension. Les estimations par la vraisemblance marginale multidimensionnelle ont été produites dans le logiciel Conquest, alors que les estimations par la vraisemblance conditionnelle ont été obtenues avec le programme OPLM.

Kristensen et Kreiner (2007) ont proposé la méthode Monte Carlo pour une meilleure estimation de la valeur critique, sans distribution préalable du trait latent, laquelle est en général supposée être gaussienne. Les données bi dimensionnelles aux items polytomiques (dont 4 sur la première dimension, et 3 sur la deuxième) ont été simulées avec pour critère la corrélation entre les dimensions de 0,75, 0,8, 0,85 et 0,9 sur 250, 500, 750 et 1000 observations. Les traits latents sont supposés suivre la loi normale bivariée de moyenne nulle.

Toutefois, les auteurs relèvent que la détection de l'écart à l'unidimensionnalité est moins évidente lorsque les dimensions sont trop corrélées, et la puissance du test augmente avec le nombre d'observations. Cela revient encore à déduire que les tests basés sur le khi carré a priori seraient moins utiles sauf lorsqu'il s'agit de détecter la multi dimensionnalité globale en présence d'un échantillon de taille assez élevée et une faible corrélation entre les dimensions. La faiblesse de cette étude réside sur le nombre assez faible d'items par dimension et sur le ratio d'items par dimension (4 versus 3). Ce qui n'est pas habituel dans la pratique où en général on a une dimension dominante avec suffisamment d'items et une autre résiduelle que l'on ne souhaitait pas avoir dans les données. De plus, on pourrait se retrouver dans la pratique avec des dimensions faiblement corrélées, mais cet aspect n'a pas été abordé dans cette étude. Il aurait donc été intéressant d'étudier la puissance de l'approche Monte Carlo en présence d'une dimension fortement dominante.

Les deux tests ci-dessus comme celui de Martin-Löf ont en commun le fait que le test du ratio donne un diagnostic global sur la dimension du test à retenir, puisque les statistiques d'ajustement de deux dimensions concurrentes sont comparées entre elles. Il ne permet pas d'identifier les items que l'on aurait souhaité retirer si l'on veut travailler avec une seule des dimensions puisqu'on ne connaît pas les saturations des items sur chacune d'elles. En pratique, le chercheur souhaiterait retirer les items problématiques, les reformuler ou confirmer ses appréhensions sur certains des items. Les statistiques outfits du modèle de Rasch permettent de mettre en évidence de tels items du test qui méritent un second regard par les spécialistes du contenu avant de prendre une décision finale quant à leur inclusion dans le test.

3.2.2.3 Contribution des statistiques outfit et infit du modèle de Rasch

La formulation du modèle de Rasch pour item dichotomique montre que la probabilité de trouver la bonne réponse ne dépend que de la distance entre le paramètre personne et le paramètre item. Elle augmente lorsque le paramètre personne augmente, celui de l'item restant inchangé.

$p(X = 1|\theta) = p(X) = \frac{\exp(\theta - b)}{1+\exp(\theta - b)}$

En général, le critère permettant d'identifier un vecteur (personne ou item) de réponses qui ne s'ajuste pas au modèle de Rasch est la statistique t dont la valeur critique absolue est 2. La formulation mathématique et les interprétations qu'on donne à cette statistique sont exposées dans cette section.

Supposons un test composé de I items dichotomiques dont les valeurs sont 0 (mauvaise réponse) et 1 (bonne réponse), et qui a été administré à N individus.

Posons $\varepsilon_{in} = O_{in} - E_{in}$ l'erreur de prévision entre la valeur de la réponse observée et celle prédite à l'item i pour la personne n. E_{in} est obtenue à partir de l'équation définissant la formulation du modèle de Rasch précédente. Les valeurs possibles sont donc des probabilités comprises entre 0 et 1, et par conséquent, les erreurs sont comprises entre -1 et 1.

ε_{in} prend de petites valeurs lorsque les valeurs prédites par le modèle sont proches des valeurs observées. Par contre, un grand écart signifie que la valeur prédite est loin d'être conforme aux attentes du modèle de Rasch.

Pour chacune des paires constituées d'un individu et un item donnés, E_{in} prend la valeur 0,5 lorsque le paramètre de l'item est égal à celui de l'individu. Dans ce cas, le résidu ε_{in} prend la valeur 0,5 si la réponse observée est correcte (égale à 1) et -0,5 si elle est incorrecte (égale à 0).

Les valeurs résiduelles négatives sont associées aux réponses observées qui sont incorrectes alors que celles qui sont positives correspondent aux réponses observées qui sont correctes. Les résidus qui se trouvent hors de la plage $[-0{,}75; +0{,}75]$ sont considérés comme non conformes pour le modèle de Rasch (Bond et Fox, 2007). Les valeurs standardisés de ces résidus z_{in} sont utilisées pour le calcul de deux statistiques d'ajustement au modèle de Rasch, dites outfit et infit, généralement notées t. La statistique z_{in}^2 suit un khi deux à un degré de liberté, si l'on suppose que les résidus sont des bruits blancs.

Formulation de la statistique outfit

Pour un item i donné, la statistique d'ajustement outfit est la moyenne de la somme des carrés des erreurs standardisées:

$$outfit_i = \sum_{n=1}^{N} \frac{1}{N} * z_{in}^2$$

Pour un individu n donné, l'expression équivalente est

$$outfit_n = \sum_{i=1}^{I} \frac{1}{I} * z_{in}^2$$

Formulation de la statistique infit

L'estimation des paramètres items et personnes est sujet à des erreurs. La faiblesse de la statistique outfit est qu'elle est sensible aux valeurs extrêmes. Si l'on prend en compte la variance des estimés, la statistique infit est obtenue en divisant par la somme des variances, la somme des carrés des résidus pondérée par leurs variances.

Pour un item i donné, la statistique d'ajustement infit est:

$$infit_i = \sum_{n=1}^{N} \frac{1}{\sum_n v_{ni}} * v_{ni} * z_{in}^2$$

Pour un individu n donné, l'expression équivalente est :

$$infit_n = \sum_{i=1}^{I} \frac{1}{\sum_i v_{ni}} * v_{ni} * z_{in}^2$$

Ces statistiques sont rapportées par les logiciels de modélisation sous forme de rapport de statistique de khi carré sur le degré de liberté. Elles prennent en conséquence des valeurs

positives dont l'espérance est 1, mais augmentent avec la taille de l'échantillon ou le nombre d'items.

Une valeur de la statistique infit ou outfit de 1+x signifie qu'il y a 100*x fois plus de variations entre les observations et les prédictions du modèle dans les données que celles qu'on obtiendrait si les données et le modèle de Rasch étaient parfaitement compatibles. Ceci arrive lorsque les réponses sont hasardeuses comme lorsqu'une personne ayant de fortes habiletés donne des réponses fausses aux items faciles, ou qu'une personne de faible habileté trouve les bonnes réponses aux items difficiles. Par contre une valeur y de ces statistiques inférieure à 1 signifie qu'il y a 100*(1-y) moins de variations dans les observations que celles que prédirait le modèle (Bond et Fox, 2007).

En pratique, on préfère travailler avec la forme standardisée des statistiques outfits et infits, généralement désignés par la lettre t car elles sont plus robustes et stables. Elles suivent approximativement la loi de Student qui est bien connue, d'espérance zéro et dont les valeurs vont de $-\infty$ à $+\infty$.

Ainsi lorsque les données sont conformes au modèle de Rasch ces versions normalisées ont pour espérance 0 et pour variance 1.

En conséquence (Bond et Fox, 2007, p.240):

- ✓ les items et personnes dont les valeurs de t sont inférieures à -2 ou supérieures à 2 sont considérées comme étant moins compatibles avec le modèle de Rasch au seuil de 5%.
- ✓ Les valeurs de t négatives indiquent qu'il y a moins de variations dans les réponses que celles que prévoit le modèle, c'est-à dire que les réponses sont trop prévisibles. C'est le cas lorsqu'un individu réussit à tous les items faciles, et en donne de mauvaises aux items difficiles.
- ✓ Lorsque t est inférieur à -2 cela signifie que les réponses sont trop prévisibles, il y a très peu de variations que prévu par le modèle et les données s'ajustent trop au modèle. Pour les items, cela indique la violation du postulat de l'indépendance locale entre eux.
- ✓ Les valeurs positives de t indiquent des réponses hasardeuses.
- ✓ Lorsque t est supérieur à 2, on considère que les réponses sont trop hasardeuses (performances ayant trop de bruit dans les données), il y a beaucoup de variations ou de perturbations et donc les données s'ajustent très mal au modèle.

Étant donné que nous travaillons avec des données simulées dont le processus de génération est bien connu, pour un item, si la statistique infit t est supérieur à 2, cela devrait signaler non pas des réponses hasardeuses sur cet item à travers tout l'échantillon des personnes, mais surtout que l'item ne mesure pas le même trait latent que les autres. On pourra alors confronter ceux des items ayant un t supérieur à 2 avec leur dimension d'origine pour déterminer si l'ensemble des items restant forment une des deux dimensions de départ, constituant ainsi la réponse à notre troisième question de recherche.

Nous allons donc mettre à profit les statistiques t d'ajustement dans notre étude de la robustesse du modèle de Rasch. Compte tenu des études existantes dans notre revue de littérature, la réponse à notre question générale de recherche peut être construite au tour des trois questions spécifiques ci-dessous, dont les justifications suivent:

i. En rapport avec les paramètres des items, quelle relation existe-t-il entre les valeurs simulées et les valeurs estimées par le modèle de Rasch?
ii. En rapport avec les paramètres des personnes, quelle relation existe-t-il entre les valeurs simulées et les valeurs estimées par le modèle de Rasch?
iii. En se référant sur les statistiques t d'ajustement pour retirer ou non un item du test, le modèle unidimensionnel de Rasch mesure-t-il une des dimensions connues du test?

Les deux premières sous questions permettront de déterminer les relations statistiques entre les paramètres personnes et items (respectivement) estimés par le modèle de Rasch, et leurs valeurs simulées. Ces relations permettront de tirer des conclusions sur la capacité à reproduire les paramètres des items et des personnes du modèle de Rasch dans chacune des conditions de la simulation. Les combinaisons des paramètres personnes, ces derniers qui sont bidimensionnels au départ, seront déterminés dans chacune des situations de la simulation, puisque la modélisation de Rasch produit des estimés à partir d'un modèle unidimensionnel, et on aimerait vérifier si ces estimés sont malgré tout adéquats étant donné qu'une forme de multi dimensionnalité a été introduite par simulation.

La troisième question trouve sa justification dans le fait qu'un test est bâti pour mesurer un ou des traits bien définis. Les mesures produites doivent donc être statistiquement valides dans ce sens qu'elles doivent se rapporter au trait visé et non un autre. Il est donc important de vérifier

que les mesures produites sont interprétables et conformes à la visée de la recherche. Or les statistiques t permettent de guider le choix des items à retenir pour produire des mesures adéquates.

En effet, une fois que la validation de la structure des items et du contenu du test par des experts du domaine dans lequel on veut produire des mesures est achevée, la validation statistique qui se fait sur la base des données réelles à collecter est une étape permettant de conforter la démarche vers la production des mesures valides. Pour vérifier que les items sélectionnés mesurent ensemble un seul et même trait, plusieurs approches sont utilisées, à l'exemple de l'analyse en composantes principales ou des équations structurelles comme nous l'avions vu au chapitre 2 dans lequel la mesure de la dimensionnalité a été abordée. Avec le modèle de Rasch, la statistique d'ajustement infit t est produite pour chacun des items du test qui sont supposés mesurer un même trait, et permet d'identifier ceux des items dont le patron des réponses est suffisamment distinct de celui prédit par le modèle. De tels items ont des statistiques d'ajustement standardisées supérieures à 2 en valeur absolue.

En combinant les réponses à ces trois questions, l'on pourra alors identifier celles des conditions dans lesquelles le modèle de Rasch unidimensionnel s'appliquerait à un test quand bien même les items le composant sont associés à différentes dimensions. Dans ces conditions, on pourra alors se fier essentiellement sur les statistiques t d'ajustement pour affiner la composition du test.

Le chapitre qui suit décrit la méthodologie qui nous permettra de répondre aux questions spécifiques sous-jacentes à notre question principale.

Chapitre 4:
Approche méthodologique

Introduction

La question principale que nous posons dans notre recherche est celle de savoir si le modèle unidimensionnel de Rasch est robuste lorsque le postulat de l'unidimensionnalité du test est violé. Les modèles de Rasch permettent d'estimer les paramètres des items et les paramètres des personnes (habiletés). Ils donnent aussi la possibilité de placer à la fois les individus et les personnes sur un même continuum, et sur un même graphique appelée carte de Wright pour apprécier la difficulté globale de l'ensemble des items par rapport à l'habileté globale des personnes.

Étudier la robustesse des modèles unidimensionnels revient donc à étudier les écarts entre les estimations de ces paramètres et leurs valeurs connues, lorsque tous les items du test ne saturent pas sur une seule dimension (en regard de la simulation qui sera faite dans ce travail), ou que certains des items saturent sur plusieurs dimensions à la fois. Les études vues au chapitre précédent ont mis à contribution la racine carré de la moyenne des carrés des différences ou des rapports linéaires entre paramètres simulés et estimés comme critère d'évaluation de la robustesse du modèle unidimensionnel de la TRI.

Quant aux modèles de Rasch, les statistiques globales sont à la base des tests énumérés. Dans cette étude, nous nous intéressons à la modélisation de la relation entre les paramètres simulés avec un modèle de Rasch bi-dimensionnel, et leurs estimés par un modèle de Rasch unidimensionnel.

En effet, les études de Blais (1987), Cuesta et Muniz (1999) ont montré que la difficulté des items est bien reproduite par le modèle logistique à deux paramètres avec un test multidimensionnel à structure simple. Il ressort aussi des autres études que les paramètres personnes générés par le modèle logistique sont des combinaisons linéaires des correspondants aux différentes dimensions. Pour répondre donc à notre question principale, les trois sous questions suivantes doivent être étudiées avec le modèle de Rasch unidimensionnel comme modèle de mesure, lorsque le test est multidimensionnel à structure simple:

i. En rapport avec les paramètres des items, quelle relation existe-t-il entre les valeurs simulées et les valeurs estimées par le modèle de Rasch?
ii. En rapport avec les paramètres des personnes, quelle relation existe-t-il entre les valeurs simulées et les valeurs estimées par le modèle de Rasch?
iii. En se référant sur les statistiques t d'ajustement pour retirer ou non un item du test, le modèle unidimensionnel de Rasch mesure-t-il une des dimensions connues du test?

Les deux premières questions ont aussi été abordées dans les études vues au chapitre 3, mais avec des critères de décision différents comme la racine carrée de la moyenne des carrés des écarts et des modèles de mesure différents comme le modèle logistique a un ou deux paramètres, ou les statistiques globales comme dans le cas du test de Martin-Löf.

En effet, pour exploiter la racine carrée de la moyenne des carrés des écarts pour un item multidimensionnel, il faut calculer celle-ci autant de fois que la dimension. Par exemple, pour des items bidimensionnels, il faut faire ce calcul deux fois avant d'identifier la dimension à laquelle les paramètres estimés (unidimensionnels) se rapprochent le plus, à partir des coefficients de régression. On pourrait envisager en effet de faire la régression multiple entre d'un côté le vecteur multidimensionnel des paramètres items simulés (variables explicatives), et le vecteur unidimensionnel des paramètres items estimés d'un autre côté, ce dernier étant le vecteur d'observations de la variable à expliquer.

Une manière généralement rencontrée dans la construction de la multi dimensionnalité est l'introduction d'une forme de corrélation entre les facteurs qui sont censés expliquer le patron des observations. Les moindres carrés ordinaires (MCO) ne sont donc plus applicables comme méthode d'estimation avec le modèle linéaire multiple car l'indépendance entre les variables explicatives qui est une hypothèse fondamentale de l'estimation par les MCO n'est plus respectée.

Dans notre cas nous mettons à profit les résultats de trois sources dont d'une part les résultats de régressions multiples contrairement à la régression une à une entre le paramètre estimé et chacun des paramètres simulés correspondant aux différentes dimensions comme on l'a vu dans la littérature; et d'autre part les résultats de la régression ridge lorsque il existe une forme de corrélation entre les variables explicatives; et enfin les statistiques t d'ajustement au modèle de Rasch pour répondre à la troisième question.

Il est donc indispensable de connaître les vraies valeurs des paramètres des items et des personnes pour pouvoir apprécier la performance du modèle de mesure utilisé. En pratique pour les items, cela est possible lorsqu'ils ont été calibrés au cours des études antérieures comme dans le cadre des tests dont l'administration est assistée par ordinateur. Pour les personnes, cela suppose que leurs habiletés ont déjà été estimées dans des tests parallèles pour servir de repères.

4.1 Type de recherche et justification

Notre recherche est quantitative et à but à la fois confirmatoire et exploratoire.

Elle est exploratoire car l'on ne sait pas d'avance comment les relations entre les paramètres simulés et estimés varient suivant les conditions de la simulation, ce qui est l'un des objectifs de cette étude. Elle est confirmatoire car nous connaissons la structure exacte du test en termes de saturation de chacun des items sur les deux dimensions, et l'un des objectifs de cette étude est de déterminer laquelle des deux dimensions de départ est surtout reproduite par le modèle unidimensionnel de Rasch, sur la base des statistiques t d'ajustement. Nous travaillons dans le cadre théorique de la modélisation statistique des réponses aux items. En particulier nous étudions la robustesse du modèle unidimensionnel pour modéliser des données multidimensionnelles de réponses aux items.

4.2 Visées de la recherche

Comme d'autres recherches précédentes sur la robustesse des modèles de la théorie de la réponse à l'item, notre recherche s'inscrit de façon spécifique dans le cadre exploratoire de la robustesse des modèles de Rasch en particulier. Elle vise à éclairer les utilisateurs de ces modèles sur le type de combinaison des paramètres aux différentes dimensions du test tel qu'utilisé par ce modèle lorsque l'unidimensionnalité des données n'est pas certaine, et suivant plusieurs scénarios. Ces résultats donnent les ingrédients à partir desquels certaines conclusions ne peuvent être dégagées sans le risque d'engendrer des décisions biaisées.

4.3 Les participants

Nous ne développerons pas d'outil de collecte, et nous ne collecterons pas de données auprès des individus. Comme notre recherche suppose que nous disposions des données réelles pour pouvoir comparer les estimations des paramètres, nous ferons comme si nous disposions à la fois d'un

questionnaire et des données réelles. Pour cela nous allons simuler des valeurs des paramètres des 30 items et de 500 personnes, ainsi que des réponses dichotomiques aux 30 items.

4.4 Les conditions de la simulation

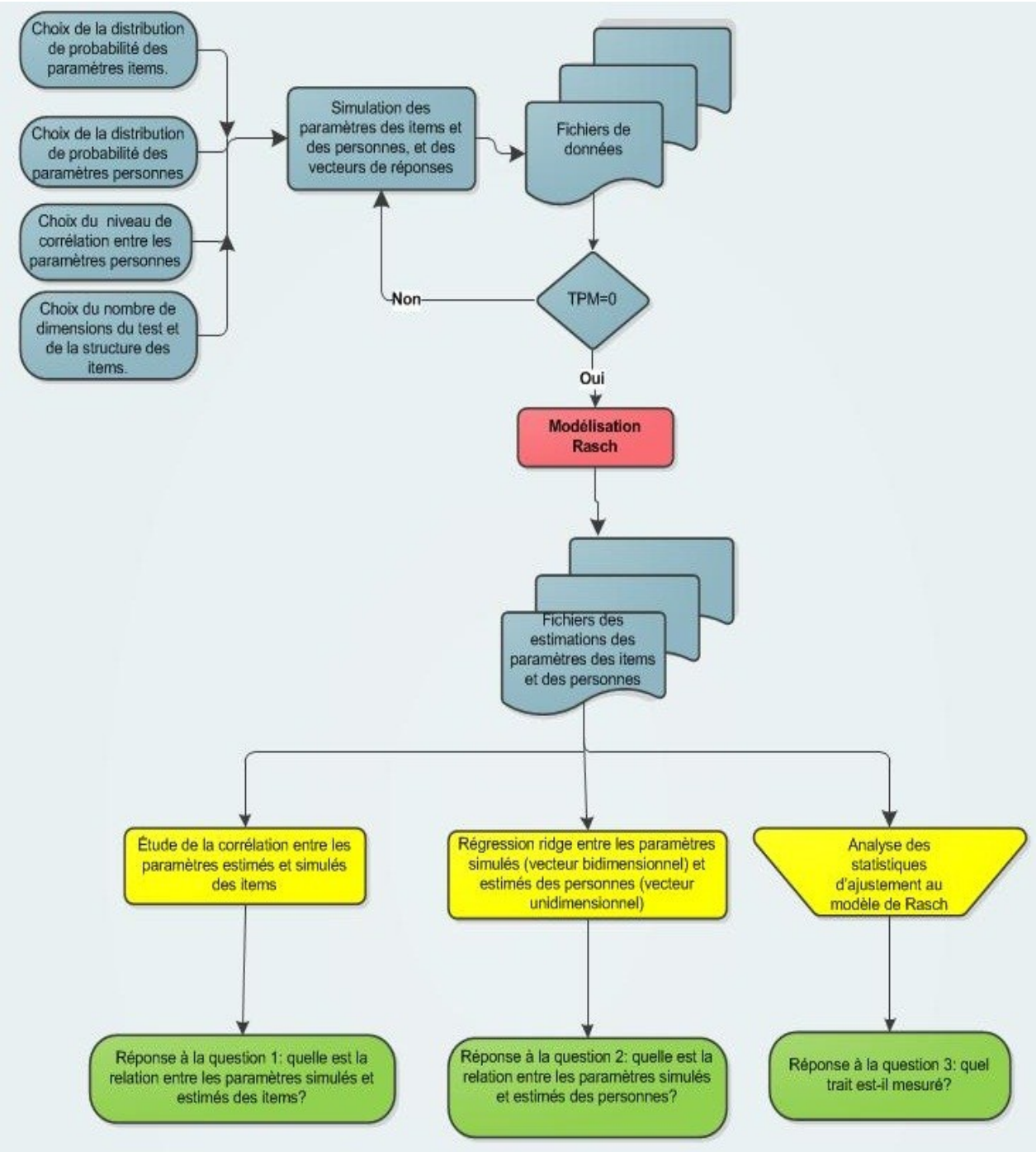
Les conditions de notre simulation tiennent compte des variables suivantes supposées être sous notre contrôle :

- ✓ les modèles de simulation des données
- ✓ les modèles de mesure
- ✓ le nombre d'observations (taille de l'échantillon)
- ✓ le nombre d'items (longueur du test)
- ✓ la distribution de probabilité des paramètres items
- ✓ la distribution de probabilité des paramètres personnes
- ✓ le type de dimensionnalité
- ✓ le nombre de dimensions
- ✓ la répartition des items par dimension
- ✓ le niveau de la corrélation entre les dimensions

Les données modélisées ont été obtenues par simulation à l'aide du logiciel Conquest v.3 (Wu, Adams et Wilson, 1997).

Ce logiciel permet non seulement de modéliser les données de réponses aux items avec les modèles unidimensionnels et multidimensionnels de Rach, mais aussi de simuler des réponses aux items pour des tests unidimensionnels, ou aux tests multidimensionnels (multi dimensionnalité inter-items). Le logiciel offre aussi la possibilité de choisir un degré de corrélation entre les dimensions du test dans le cadre d'un test multidimensionnel.

Le graphique 2 ci-dessous résume en quatre étapes le processus qui a conduit à inférer des réponses aux trois questions de recherche.



Graphique 2. Schéma de la simulation et de la modélisation des données.

En entrée, en plus du choix du nombre de dimensions du test qui est de deux dans notre cas, trois autres critères doivent être fixés pour générer les données appropriées :

- ✓ la distribution de probabilité des paramètres items est choisie entre la loi normale et la loi uniforme, puisque ce sont les deux choix possibles dans la version du logiciel Conquest à notre disposition.
- ✓ La distribution de probabilité des paramètres personnes peut être la normale uni variée ou multi variée, ou la loi uniforme uni variée.
- ✓ Dans le cas de la distribution normale multi variée, la corrélation entre les traits est prise en compte dans la commande.

Comme on l'a vu au chapitre deux, on rencontre la multi dimensionnalité soit au niveau de l'item, soit alors au niveau du test. Par construction, nous avons choisi un test bi dimensionnel à structure simple, c'est à dire que chacun des items sature uniquement sur une seule des deux dimensions. Ce qui est équivalent de dire que les réponses aux items du test requièrent des personnes deux types d'habiletés bien distinctes. Par conséquent, le vecteur des paramètres aux 30 items et celui des paramètres des habiletés des personnes sont bidimensionnels. Une habileté permet donc de justifier les réponses aux items de la dimension dominante et une autre habileté pour les items de la dimension résiduelle.

Par exemple, lorsque le test est constitué des 25 premiers items saturant sur la dimension 1 et 5 autres sur la dimension 2, chacun des items est un vecteur se présentant comme suit:

$I_i = (b_i, 0)$ pour les 25 premiers items de la dimension 1, i=0,…, 25

$I_i = (0, b_i)$ pour les 5 items de la dimension 2, avec i=26,.., 30, et b_i le paramètre difficulté de l'item i.

Quant aux personnes, chacune est représentée par un vecteur de paramètres habiletés ayant deux dimensions dont la première composante est son niveau d'habileté sur le premier trait requis pour répondre aux items de la dimension 1, et la deuxième composante, son niveau d'habileté sur le deuxième trait qui est requis pour répondre aux items de la dimension 2. On obtient alors le vecteur d'habiletés suivant pour l'individu j= 1, …, 500 : $\Theta_j = (\theta_{j1}, \theta_{j2})$ avec θ_{j1} son habileté sur le premier trait et θ_{j2} son habileté sur le second trait.

Les données sont ainsi multidimensionnelles car le test l'est à partir des items saturant sur deux dimensions distinctes, et les réponses à ces items nécessitent aussi en chacune des personnes des habiletés distinctes, mais qui peuvent être corrélées comme ce sera le cas dans ce travail.

Dans notre cas, nous avons simulé les paramètres items et les paramètres personnes selon la loi normale réduite uni variée et bi variée respectivement, et 500 vecteurs de données dichotomiques sur 30 items répartis sur deux dimensions selon trois critères dont :

- ✓ trois ratios de la répartition des items entre les deux dimensions dont 25 :5, 20 :10 et 15 :15
- ✓ quatre degrés de corrélations entre les deux dimensions sur les habiletés dont l'indépendance complète entre les deux dimensions (une corrélation nulle), une corrélation faible de 0,2, modérée de 0,6 et forte de 0,8 successivement.

Dans chacune des douze combinaisons, les données ont été simulées avec dix réplications en vue dégager des tendances qui, lorsque mises ensemble permettent de répondre à nos trois questions de recherche. Pour chacune des 12 situations, nous avons ainsi obtenu 10 fichiers de 500 réponses dichotomiques aux 30 items, de 1000 (500 sur chacune des deux dimensions) paramètres personnes, et de 30 paramètres items soit au total 360 fichiers de données à gérer pour la suite des traitements.

Avec le logiciel Conquest v.3, la simulation des paramètres personnes en vue d'obtenir des réponses aux items se rapportant aux tests multidimensionnels n'est possible que si la loi suivie par ces paramètres est la normale multi variée. C'est la raison pour laquelle nous avons travaillé avec la loi normale uni variée pour les paramètres items. La forme réduite a été retenue pour les deux types de paramètres en vue de s'assurer que les items ne soient pas ni difficiles, ni faciles pour les personnes. La condition sur l'appariement entre personnes et items (TPM ou test person matching égal à zéro) était ainsi assurée puisque les paramètres des personnes et des items sont contraints d'avoir une moyenne nulle.

La modélisation des données avec le modèle unidimensionnel de Rasch est la deuxième phase à partir de laquelle sont obtenus les estimés des paramètres des items et des personnes ainsi que les statistiques d'ajustement au modèle uni dimensionnel de Rasch, sur la base des 120 matrices de données dichotomiques obtenues par simulation. Ces estimés sont ensuite exploités au cours de la quatrième phase en vue de collecter les arguments qui permettent de répondre aux trois sous-questions de recherche. L'étude de la corrélation entre les paramètres simulés et estimés permet de répondre aux deux premières questions, alors que la troisième question trouve sa réponse dans l'exploitation des statistiques d'ajustement.

Nous donnons ci-dessous les raisons du choix des paramètres de la simulation.

4.4.1 Modèles de simulation et de mesure

Les études de robustesse de la TRI que nous avons explorées jusque-là ont étudié des variantes du modèle logistique à 2 ou 3 paramètres comme modèle de mesure, et les données simulées ont été générées par des variantes du modèle logistique multidimensionnel, soit le modèle factoriel et les modèles compensatoires.

Quant aux études de la robustesse du modèle de Rasch, les tests développés sont dits globaux dans ce sens qu'ils permettent de réfuter ou non l'unidimensionnalité des données, mais ne permettent pas d'identifier les items du test qui seraient la source de l'inadéquation au modèle de Rasch unidimensionnel comme modèle de mesure. Pourtant en pratique, une telle information permet d'affiner le test en vue de produire des mesures valides en rapport avec le trait étudié. Il est généralement recommandé de réévaluer l'insertion dans le test, des items dont les statistiques t d'ajustement sont supérieurs à 2 en valeur absolue. Cela signifie qu'ils doivent être reformulés, ou simplement retirés du test final lorsqu'ils s'avèrent moins pertinents.

Nous avons vu que ces statistiques sont en général couplées avec l'analyse en composante principale des résidus pour déterminer si une deuxième dimension est présente dans les données, mais cette option est limitée du fait que le critère fondé sur la valeur propre ne fait pas l'unanimité (Raîche, 2005;Chou et Wang, 2010). Nous travaillons donc sous l'hypothèse que le critère sur la statistique t d'ajustement produit par le modèle de Rasch est l'unique élément qui guide la décision de retirer ou de conserver un item dans le test.

Nous avons simulé des données d'items dichotomiques formant ensemble une bi dimensionnalité inter items (Wang, 1996) avec le modèle de Rasch. Ce choix est justifié par le fait que nous souhaitons travailler sous des conditions idéales dans lesquelles le processus générateur des données est maîtrisé. Dans notre cas, nous voulons étudier la robustesse du modèle de Rasch, donc il serait préférable de simuler des données bidimensionnelles à partir de ce modèle, ce qui garantit la validité de l'utilisation du modèle unidimensionnel de Rasch comme modèle de mesure. Le logiciel utilisé pour la simulation et l'analyse des données simulées est Conquest v.3 (Wu, Adams et Wilson). L'avantage ici est d'utiliser un même logiciel pour la simulation et l'analyse, car l'on n'a pas besoin de transformer les données pour accommoder un autre logiciel.

4.4.2 Nombre d'items et d'observations

Nous travaillons avec 30 items, ce qui est la moyenne pour un test supposé mesurer deux dimensions, et fixerons le nombre d'observations à 500 comme moyenne des autres études avec deux dimensions (Drasgow et Parsons (1983); Blais (1987); Cuesta et Muniz (1999); Kirisci et al. (2001); Kahraman et Thompson (2011).

4.4.3 Saturation des items aux facteurs (structure des items)

Les items seront répartis en deux groupes dont ceux saturant sur la première dimension uniquement et ceux saturant sur la deuxième dimension uniquement. La première dimension sera considérée comme étant la dimension dominante et donc aura toujours le plus grand nombre d'items qui y saturent uniquement, par rapport à la deuxième.

4.4.4 Distribution de probabilité des paramètres items et personnes

Les paramètres personnes multidimensionnels ne peuvent être simulés que sous l'hypothèse de la loi normale multi variée avec le logiciel Conquest qui est à notre disposition. Compte tenu de cette limite pratique du logiciel, pour avoir les chances d'obtenir des items et personnes de niveau de difficulté et habileté moyennes semblables garantissant ainsi l'appariement entre les personnes et les items, la loi normale centrée réduite a été retenue. Toutefois, l'on devra veiller à ce qu'il y ait une correspondance entre la difficulté du test et les habiletés des personnes (*test-person matching* (TPM) en anglais). Ce qui signifie que le test ne doit être ni facile, ni difficile pour l'échantillon des personnes en étude, et revient donc à vérifier l'égalité TPM = 0 (Chou et Wang, 2010). Cette condition est satisfaite par le choix de la distribution de probabilité identique des paramètres items et personnes.

4.4.5 Corrélation entre les facteurs

Nous travaillerons dans des conditions de corrélation nulle (0,0), faible (0,4), modérée (0,6) et forte (0,8). Des conditions semblables ont aussi été expérimentées dans des études revues dans la littérature.

4.5 Analyse des données

Le modèle de Rasch simple est le modèle de mesure puisque les items sont dichotomiques, mais le modèle de simulation est un modèle de Rasch bidimensionnel afin de constituer un test qui mesure deux traits différents.

L'indice communément utilisé pour apprécier l'écart global est la racine de la somme des carrés des écarts à la moyenne (Drasgow, 1983; Cuesta et al., 1999).

Dans notre étude, pour chacun des paramètres items et personnes, nous modélisons la relation entre les valeurs estimées et les valeurs simulées des paramètres items et personnes. La régression linéaire multiple est mise à profit lorsque la corrélation est nulle ou trop faible (0,2) entre les habiletés simulées des personnes qui sont présentées sous forme d'un vecteur bidimensionnel que nous considérons être le vecteur de deux variables explicatives (voir graphique III qui suit). Lorsqu'une forme de corrélation entre les habiletés est présente, la régression ridge sera utilisée car elle est appropriée en régression linéaire multiple lorsque les variables explicatives sont corrélées.

Nous accorderons ensuite une attention particulière aux statistiques d'adéquation (infit) issues du modèle de mesure pour apprécier l'émergence de l'une des deux dimensions de départ. En effet, à partir du vecteur des observations, le modèle unidimensionnel produit des estimés des paramètres items et personnes sur un même continuum, peu importe si plusieurs habiletés déterminent ces observations. Les vecteurs bidimensionnels sont donc transformés en vecteurs unidimensionnels de sorte qu'en bout de ligne les personnes ont une seule habileté au lieu de deux, et les items saturent sur une seule dimension peu importe leur distribution originale sur deux dimensions.

Les résultats de ces deux approches nous permettront de répondre aux trois sous questions de recherche.

Conclusion

L'exploration des études existantes sur la robustesse du modèle unidimensionnel nous a permis de justifier notre choix de travailler avec le modèle de Rasch, et de définir les paramètres de la simulation des données. Le chapitre suivant est consacré aux analyses nous permettant de trouver les réponses aux trois sous-questions, soit finalement à la question principale de notre recherche.

Chapitre 5 :

Analyse des résultats de la modélisation

Ce chapitre comprend trois sections. Les résultats de la modélisation avec le modèle de Rasch unidimensionnel seront présentés dans les deux premières sections. Ceux-ci permettront de répondre aux deux premières questions spécifiques à savoir quelle relation existe-t-il entre les valeurs simulées et les valeurs estimées des paramètres items et personnes (respectivement) par le modèle de Rasch?

Les réponses à ces deux questions sont déduites de l'analyse des régressions biaisées entre les paramètres simulés des personnes et des items et leurs estimés respectifs par le modèle de Rasch unidimensionnel. La troisième section est consacrée à l'étude de la dimension qui ressort de la modélisation des données avec le modèle de Rasch unidimensionnel, sur la base des statistiques t d'ajustement.

Le graphique 3 ci-dessous montre la transformation que subit chacun des vecteurs des paramètres items et personnes: Par simulation, chacun des items et chacune des personnes est décrit(e) par deux paramètres. Les matrices b et Θ sont des représentations de l'ensemble des 30 items et des 500 personnes respectivement. À partir de la confrontation entre la difficulté de chacun des items et le niveau de l'habileté requise pour trouver la bonne réponse à un item donné qui est présent chez l'individu, on obtient la matrice des observations dichotomiques dans notre cas, constituée de 500 lignes et 30 colonnes. C'est à partir de cette matrice des observations que la modélisation de Rasch permet d'obtenir les vecteurs unidimensionnels des estimés des paramètres des items et des personnes $\hat{b}$ et $\hat{\theta}$, puisque dans la pratique, on ne dispose que des manifestations du trait latent à travers les réponses aux items, qui sont regroupées ici dans la matrice des observations O. En bout de ligne donc, à partir des estimés des paramètres, tous les items et les personnes sont projetés sur une droite dont on ignore l'équation.

On est alors en droit de se poser la question de savoir quelle relation existe entre les valeurs en sortie des paramètres items et personnes, et les valeurs correspondantes qui ont permis d'obtenir

la matrice O des observations en entrée. Le processus correspondant à cette description est représenté à travers le graphique 3 ci-dessous.

a) Structure des vecteurs bidimensionnels des paramètres simulés des items et des personnes, et de la matrice des observations

$$b = (b_{i1}, b_{i2}) = \begin{bmatrix} b_{1,1} & 0 \\ b_{2,1} & 0 \\ \vdots & \vdots \\ b_{25,1} & 0 \\ 0 & b_{26,1} \\ \vdots & \vdots \\ 0 & b_{30,2} \end{bmatrix} \quad \Bigg\} \quad O = \begin{pmatrix} o_{1,1} & o_{1,2} & \cdots & o_{1,30} \\ \vdots & \vdots & \ddots & \vdots \\ o_{500,1} & o_{500,2} & \cdots & o_{500,30} \end{pmatrix}$$

$$\theta = (\theta_{j,1}, \theta_{j,2}) = \begin{bmatrix} \theta_{1,1} & \theta_{1,2} \\ \vdots & \vdots \\ \theta_{500,1} & \theta_{500,2} \end{bmatrix}$$

b) Production des vecteurs unidimensionnels des paramètres estimés des 30 items et des 500 personnes, et des statistiques d'ajustement infit t par le modèle de Rasch.

$$O = \begin{pmatrix} o_{1,1} & o_{1,2} & \cdots & o_{1,30} \\ \vdots & \vdots & \ddots & \vdots \\ o_{500,1} & o_{500,2} & \cdots & o_{500,30} \end{pmatrix} \quad \Bigg\}$$

Modélisation de Rasch

$$\hat{b} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \vdots \\ \hat{b}_{30} \end{bmatrix}; \quad \hat{t}_I = \begin{bmatrix} \hat{t}_1 \\ \hat{t}_2 \\ \vdots \\ \hat{t}_{30} \end{bmatrix}$$

$$\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_{500} \end{bmatrix}; \quad \hat{t}_J = \begin{bmatrix} \hat{t}_1 \\ \hat{t}_2 \\ \vdots \\ \hat{t}_{500} \end{bmatrix}$$

Graphique 3. Processus d'estimation des vecteurs unidimensionnels des paramètres items et personnes dans le cas du ratio 25:5 items sur deux dimensions.

5.1. Nature de la relation entre les paramètres items simulés et estimés par le modèle unidimensionnel de Rasch

Nous avons vu les deux résultats empiriques suivant : D'après l'étude de Way et al. (1988), avec les modèles compensatoires de la TRI, la difficulté moyenne estimée par le modèle logistique à 3 paramètres sous violation de l'hypothèse de l'unidimensionnalité est la moyenne des difficultés moyennes des items aux deux dimensions, alors que la discrimination moyenne estimée est la somme des discriminations moyennes de l'item aux deux dimensions. Aussi, selon l'étude menée par Akerman (1991), les items bidimensionnels les plus susceptibles d'être sélectionnés dans un test adaptatif sont ceux qui ont des niveaux de difficultés aux deux dimensions qui soient similaires à celui qui est estimé par le modèle unidimensionnel.

Avec le modèle de Rasch, nous allons explorer les corrélations entre les paramètres difficultés simulés des items et leur correspondants qui ont été estimés. Nous disposons donc dans chacune des conditions de la simulation, d'une part du vecteur bidimensionnel constitué des paramètres difficultés simulés et d'autre part du vecteur unidimensionnel des estimations de ces paramètres par le modèle de Rasch puisque les 30 items sont à structure simple, chacun ne disposant que d'un seul paramètre difficulté, l'autre étant par conséquent nul.

Les régressions effectuées entre ces deux vecteurs ont produit des coefficients de corrélation de Pearson trop élevés et avoisinant l'unité. Ce résultat montre bien que le niveau de difficulté des items est bien reproduit par le modèle de Rasch unidimensionnel bien que le test soit constitué de deux dimensions, quel que soit le degré de corrélation et le ratio des items entre ces deux dernières. Un résultat similaire avait aussi été obtenu dans les études réalisées par Blais (1987) et Cuesta et Muniz (1999).

Résultat d'analyse 1**:**

> Étant donné la structure simple des items du test, et compte tenu des conditions idéales dans lesquelles les réponses ont été obtenues (par simulation), nous pouvons déduire que la difficulté de l'item d'un test multidimensionnel à structure simple ne serait pas affectée même lorsque le modèle de Rasch unidimensionnel est mis à contribution pour la production des mesures. Toutefois, dans la pratique des résultats différents peuvent être obtenus en rapport avec la position des items dans le test, la longueur de celui-ci et le temps alloué à son administration. En effet sous la contrainte du temps et lorsque le test est long et que les items sont regroupés par dimension, le

candidat qui opte pour répondre aux items en continu l'un après l'autre, ou que le test est ainsi conçu (en ligne par exemple ou un test adaptatif), des items peuvent sembler difficiles à cause de la fatigue ou des réponses hasardeuses du fait du manque de temps vers la fin de la période de test. Une stratégie est de faire un mélange d'items de manière à ce que l'ordre de présentation des items soit indépendant de la dimension mesurée.

5.2. Nature de la relation entre les paramètres personnes simulés et estimés par le modèle unidimensionnel de Rasch

Un des résultats de l'étude conduite par Akerman (1994) montre que les estimés des habiletés des individus par le modèle de Rasch unidimensionnel en violation de l'hypothèse d'unidimensionnalité sont des combinaisons linéaires non stables des habiletés réelles sur chacune des dimensions du test. Il importe donc de savoir non seulement la combinaison linéaire entre les dimensions qui est utilisée dans l'estimation, mais aussi il faut s'assurer que cette combinaison est unique le long du continuum. Dans cette étude, le vecteur des paramètres habiletés simulés est de dimension deux alors que celui des paramètres estimés a une seule dimension. Nous avons mis à profit la régression linéaire multiple entre ces deux vecteurs d'observations, avec pour variable à expliquer le paramètre estimé, et pour variables explicatives les paramètres simulés sur les deux dimensions.

La formulation de ce modèle est le suivant :

$$Y = a + \beta 1 * X1 + \beta 2 * X2 + \xi$$

avec Y la valeur du paramètre estimé, X_1 et X_2 le paramètre des personnes à la première et deuxième dimension respectivement, et ξ un bruit blanc; β_1 et β_2 sont les paramètres à estimer.

En posant $X=(X_1, X_2)$, la solution des moindres carrés ordinaires est non biaisée et de variance minimale lorsque les conditions (i) à (iii) ci-dessous sont remplies. Elle est donnée par le vecteur $\beta^{mco}= (X'X)^{-1}X'Y$

i. X_1 et X_2 sont des vecteurs non corrélés
ii. Le terme d'erreur ξ a une distribution normale

iii. Homocédasticité des erreurs, c'est-à dire que les termes d'erreur ξ ont la même variance.

Dans les conditions de simulation impliquant une corrélation non nulle entre les deux dimensions, la première condition n'est plus vérifiée, ce qui introduit donc un biais dans l'estimation des paramètres. La régression ridge a été utilisée pour tenir compte de la colinéarité entre les deux variables explicatives X_1 et X_2.

La solution de la régression ridge est donnée par l'expression:

$$\beta_{\text{ridge}} = (X'X + \lambda I) - 1X'Y.$$

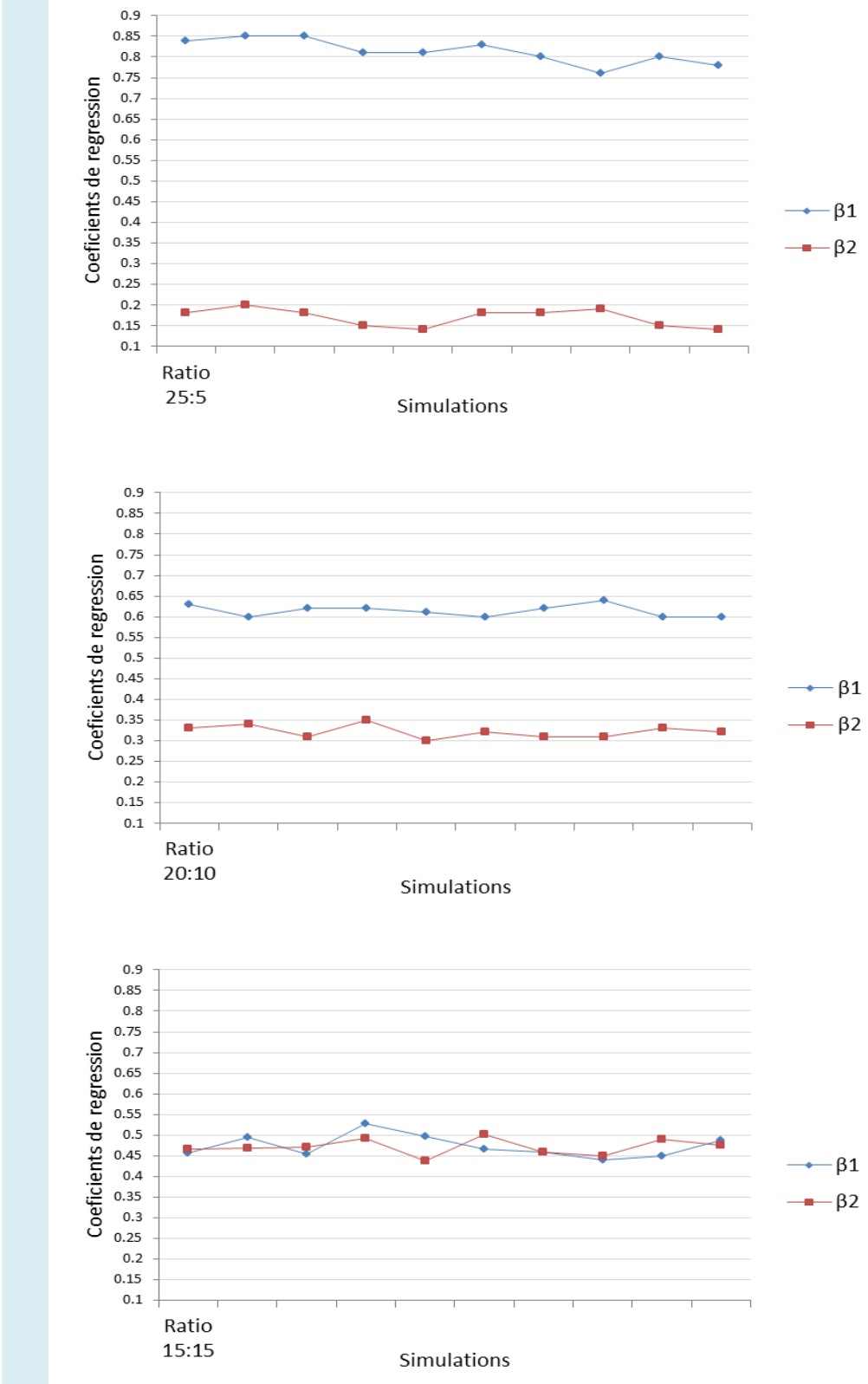
I est la matrice unité et λ est un scalaire tel que :

- ✓ lorsque λ tend vers zéro, l'estimateur β^{ridge} tend vers celui des moindres carrés,
- ✓ lorsque λ tend vers plus l'infini, β^{ridge} tend vers zéro.

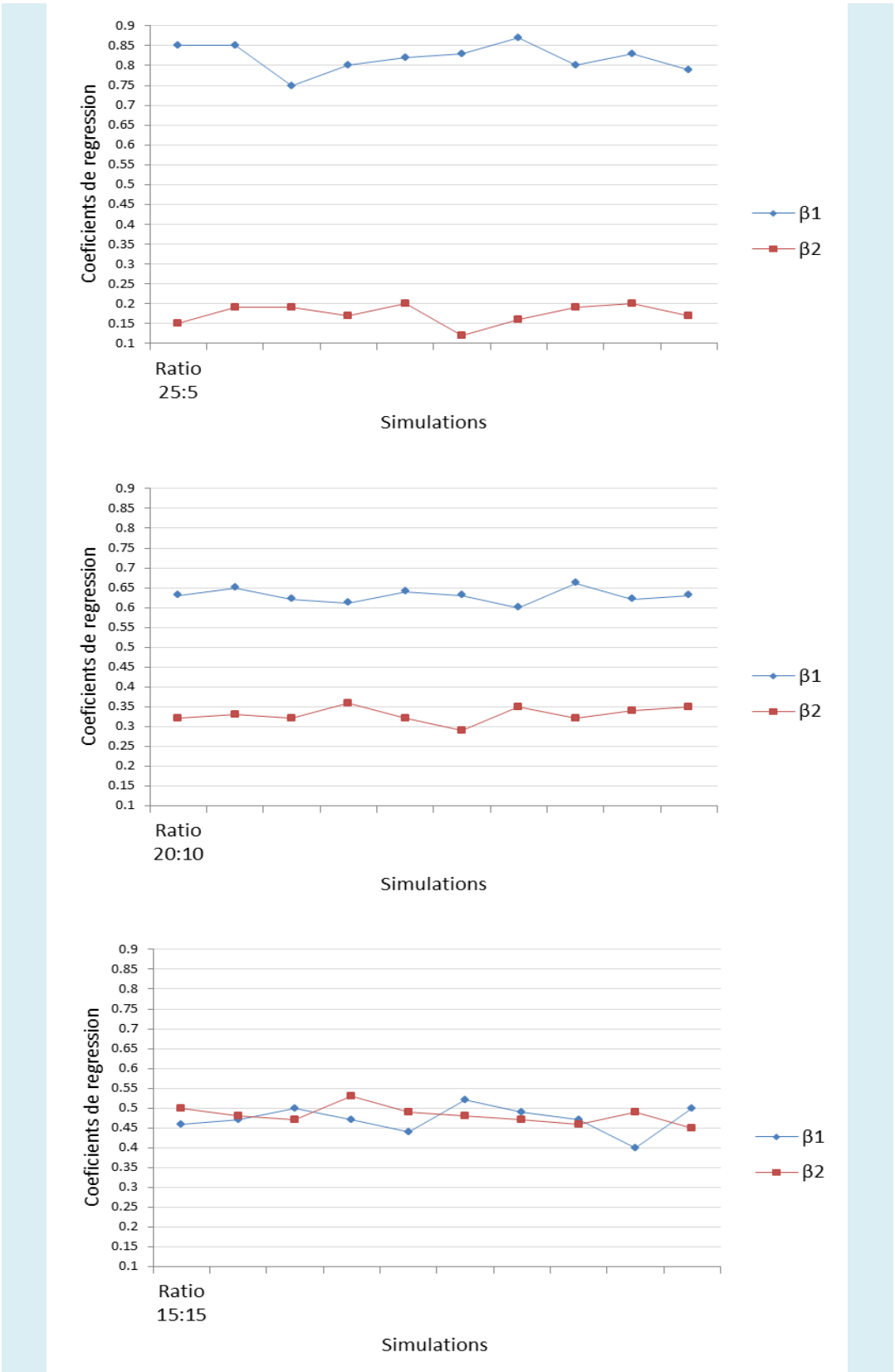
Ces régressions ont été faites dans le logiciel R grâce au package Ridge proposé par Cule et al.(2012) lorsque r>0. Le paramètre λ est choisi par une procédure automatique.

Les graphiques 4, 5, 6 et 7 suivants montrent les variations des poids des paramètres habiletés des personnes sur chacune des deux dimensions lorsque le paramètre habileté estimé par le modèle de Rasch unidimensionnel (Y) est régressé sur le vecteur bidimensionnel des habiletés simulées $X=(X_1, X_2)$, avec β_1 le poids de la dimension dominante et β_2 celui de la dimension résiduelle.

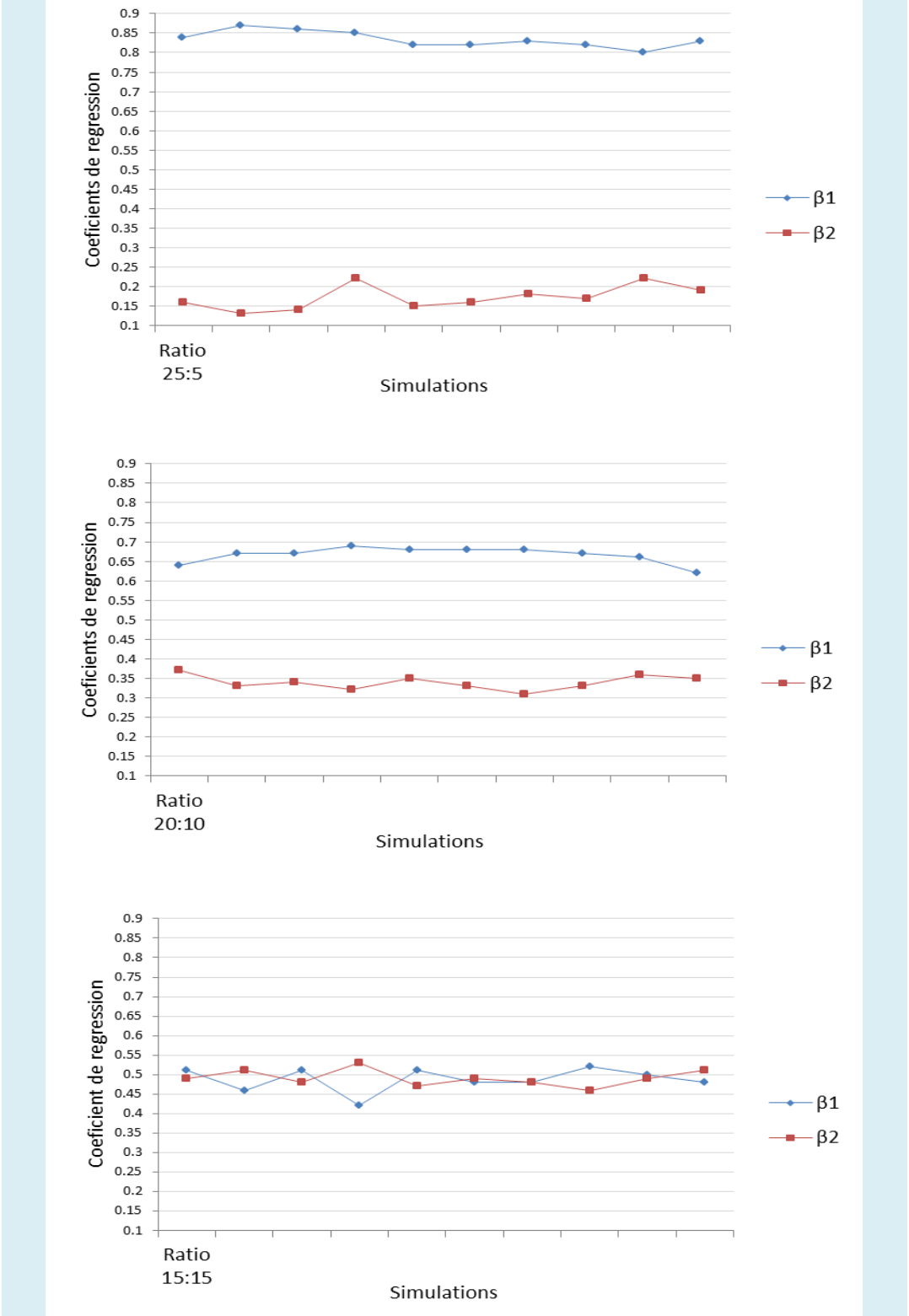
Le tableau 5.1 est un résumé des valeurs minimales et maximales des coefficients de régression obtenus soit par la régression linéaire multiple (r=0) soit par la régression ridge (r>0).



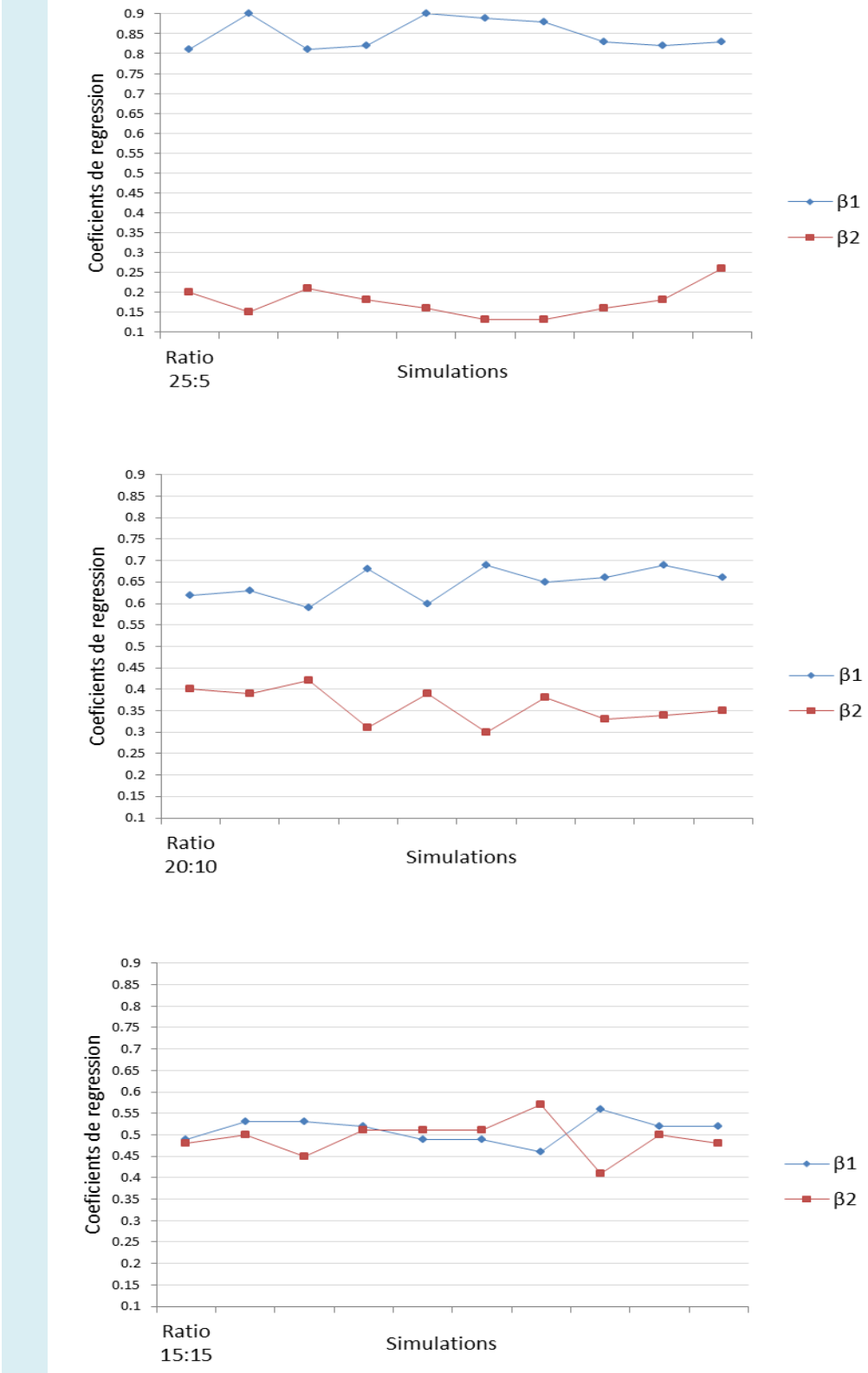
Graphique 4. Variation des coeficients de regression des habiletes aux deux dimensions selon le ratio d'items lorsque la correlation entre les deux dimensions est nulle.



Graphique 5. Variation des coeficients de regression des habiletes aux deux dimensions selon le ratio d'items lorsque la correlation entre les deux dimensions est 0,2.



Graphique 6. Variation des coeficients de regression des habiletes aux deux dimensions selon le ratio d'items lorsque la correlation entre les deux dimensions est 0,6.



Graphique 7. Variation des coeficients de regression des habiletes aux deux dimensions selon le ratio d'items lorsque la correlation entre les deux dimensions est 0,8.

Tableau 2

Variations des Coefficients de la Régression en Fonction de la Corrélation entre les Dimensions et le Ratio d'Items

Correlation entre les dimensions	Ratio des items sur les deux dimensions	Coefficient de regression				Coefficient de correlation
		β_1		β_2		
		Minimum	Maximum	Minimum	Maximum	
r=0	15:15	0,44	0,52	0,43	0,50	(0,70;0,76)
	20:10	0,60	0,64	0,30	0,35	(0,70;0,76)
	25:5	0,76	0,85	0,14	0,20	(0,76;0,81)
r=0,2	15:15	0,40	0,52	0,45	0,53	0,75
	20:10	0,60	0,66	0,29	0,36	0,80
	25:5	0,75	0,87	0,12	0,20	0,82
r=0,6	15:15	0,42	0,52	0,46	0,53	
	20:10	0,62	0,69	0,31	0,35	
	25:5	0,80	0,87	0,13	0,22	
r=0,8	15:15	0,46	0,53	0,48	0,57	
	20:10	0,59	0,69	0,30	0,42	
	25:5	0,81	0,90	0,13	0,21	

Pour tirer des éléments de réponses à notre question, nous considérons à la fois le ratio d'items entre les deux dimensions et le niveau de corrélation entre ces dernières. Lorsqu'il n'y a pas de corrélation entre les deux dimensions, ou qu'elle est faible (r=0), le modèle linéaire multiple a été mis à contribution pour déterminer la relation entre les paramètres estimés par le modèle de Rasch simple, et ceux simulés. Dans les cas de corrélation moyenne ou forte (r=0,6 ou 0,8), la régression ridge a été effectuée pour estimer les coefficients de régression. Étant donné la violation de l'hypothèse d'indépendance des variables explicatives, le modèle linéaire n'était plus approprié. En général, le paramètre estimé est une combinaison linéaire des correspondants simulés.

Les résultats spécifiques suivants se dégagent des graphiques 4, 5 et 6, et du tableau précédents :

- ✓ Lorsque les deux dimensions sont perpendiculaires (r=0), le modèle linéaire explique entre 70% et 76% des variations dans les données en présence d'une dimension

faiblement dominante. Il explique entre 76% et 81% de ces variations lorsqu'il y a une dimension fortement dominante (25 :5). Tous les coefficients sont très statistiquement significatifs, de même que le modèle global. On note aussi les éléments suivants (graphique 4) :

- Le paramètre habileté estimé tourne autour de la simple moyenne des paramètres simulés sur les deux dimensions lorsque celles-ci ont un nombre identique d'items (ratio 15:15).
- Le paramètre habileté simulé sur la dimension dominante a un poids de plus en plus important lorsque son nombre d'items augmente. Il varie de 0,6 à 0,64 contre 0,3 à 0,35 sur la dimension résiduelle pour un ratio de 20 contre 10; et de 0,76 à 0,85 contre 0,14 à 0,20 lorsque le ratio est de 25 contre 5.

✓ Lorsque la corrélation est assez faible entre les deux dimensions (r=0,2), les résultats sont semblables à ceux obtenus pour r=0 ci-dessus (graphique 5). Le modèle linéaire explique 75% des variations dans les données lorsque les deux dimensions ont le même nombre d'items (absence de dimension dominante). Il explique respectivement 80% de ces variations en présence d'une dimension dominante, et 82% lorsqu'il y a une dimension fortement dominante. Tous les coefficients sont très statistiquement significatifs, de même que le modèle global. On note aussi les éléments suivants :

✓ Lorsque la corrélation entre les deux dimensions est r=0,6 (graphique 6)

- Le paramètre estimé se situe autour de la moyenne des deux paramètres simulés au ratio 15:15.
- Le paramètre sur la dimension dominante a un poids allant de 0,62 à 0,69 contre 0,31 à 0,35 lorsque le ratio est de 20 contre 10, et de 0,8 à 0,87 contre 0,13 à 0,22 lorsque le ratio est de 25 contre 5.

✓ Lorsque la corrélation entre les deux dimensions est r=0,8 (graphique 7)

- Le paramètre estimé se situe autour de la moyenne des deux paramètres simulés au ratio 15:15.
- Le paramètre sur la dimension dominante a un poids variant entre 0,59 et 0,69 contre 0,3 à 0,42 au ratio de 20 contre 10.

- Le paramètre sur la dimension dominante a un poids variant entre 0,81 et 0,90 contre 0,13 à 0,21 au ratio de 25 contre 5.

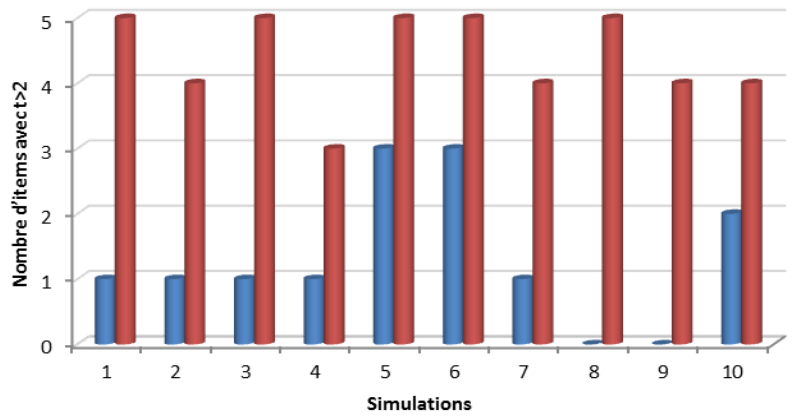
Résultat 2 :

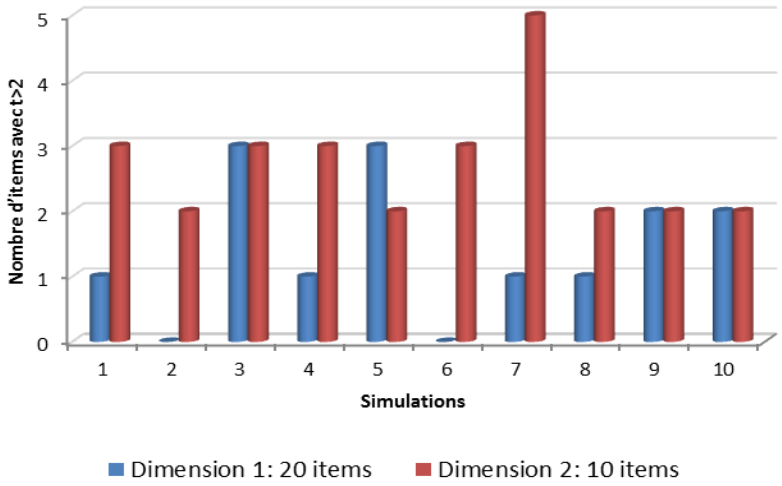
La dimension dominante a un poids qui tend vers l'unité lorsque le niveau de corrélation entre les deux dimensions augmente simultanément avec le nombre d'items qui y saturent uniquement. Les poids les plus élevés de cette dimension sont obtenus lorsque r=0,8 et le ratio d'items de 25 contre 5. Par conséquent dans la pratique, dans une situation de forte corrélation entre les deux dimensions, parmi lesquelles une est fortement dominante, le modèle de Rasch unidimensionnel serait indiqué par principe de parcimonie.

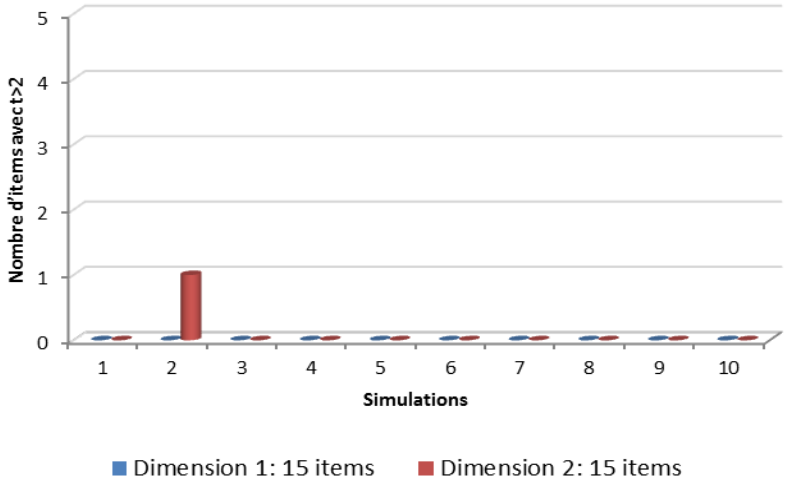
5.3. Dimension mesurée par le modèle unidimensionnel de Rasch

La modélisation des données avec le modèle de Rasch permet d'obtenir non seulement les estimés des paramètres, mais aussi les statistiques d'ajustement au dit modèle en vue, entre autre d'aider le chercheur à prendre une décision sur la validité des mesures générées ou alors la révision des items choisis pour mesurer le trait en étude. La statistique d'ajustement au modèle est produite pour chacun des items et chacune des personnes à partir des résidus entre la réponse observée et celle qui est estimée. Pour un item donné elle est obtenue en tenant compte des résidus aux réponses des N personnes ayant complété le test. La version standardisée de cette statistique (t) est couramment utilisée pour apprécier si un item a un comportement conforme au modèle de Rasch. Sa distribution de probabilité est la loi de Student. De ce fait, au risque de se tromper à 5%, un item dont la valeur de t est comprise entre -2 et 2 est considérée conforme alors qu'à l'extérieur de cet intervalle, il est considéré non conforme au modèle de Rasch, et par conséquent nécessite soit un retrait (item mesurant un autre trait que celui en étude, item trop facile ou trop difficile), soit alors une reformulation.

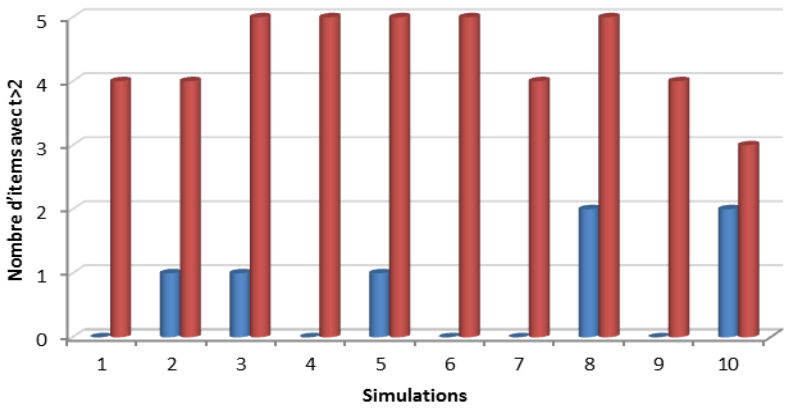
Dans notre étude, nous travaillons dans des conditions idéales car nous connaissons les distributions des paramètres des items et des personnes, la structure des items et leur saturation sur les deux dimensions, l'appariement entre les personnes et les items (TPM=0). Le critère sur t peut donc être utilisé avec un minimum de risque de se tromper.Les résultats obtenus des 120 fichiers de données simulées sont résumés aux graphiques 8, 9, 10 et 11 ci-dessous, tandis que le tableau 5.2 présente la moyenne d'items ayant une statistique t en dehors de la plage de valeurs acceptables.

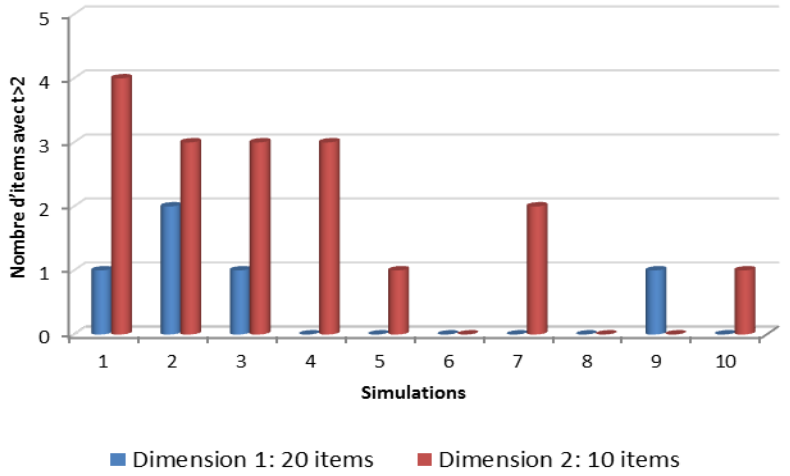


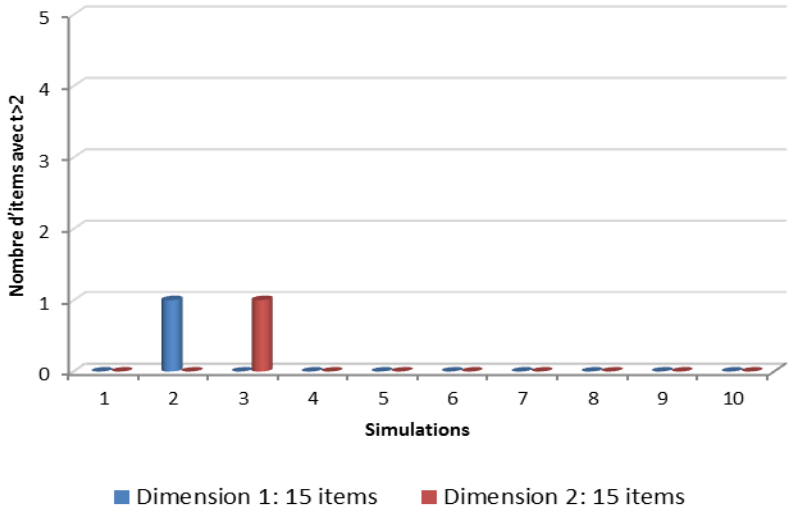




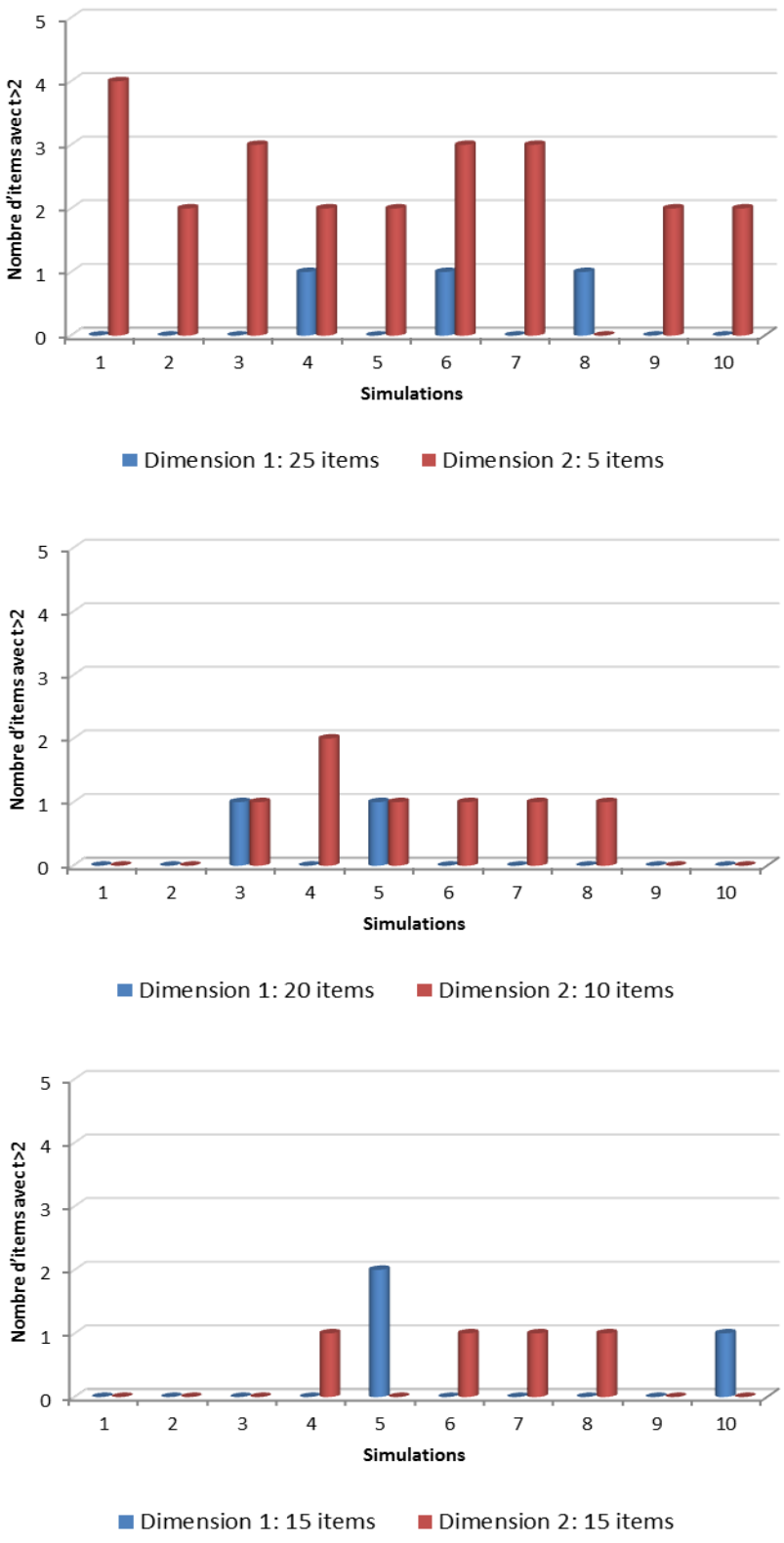
Graphique 8. Nombre d'items dont la statistique d'ajustement t est hors de la plage ⊢]-2,2⊣ [selon le ratio d'items lorsque la correlation entre les deux dimensions est nulle.



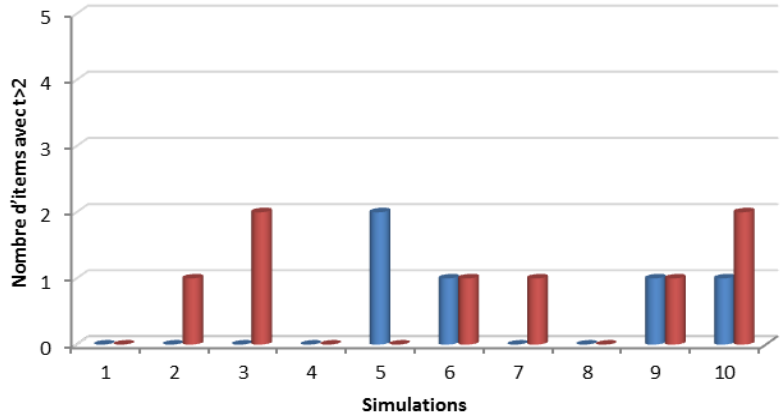


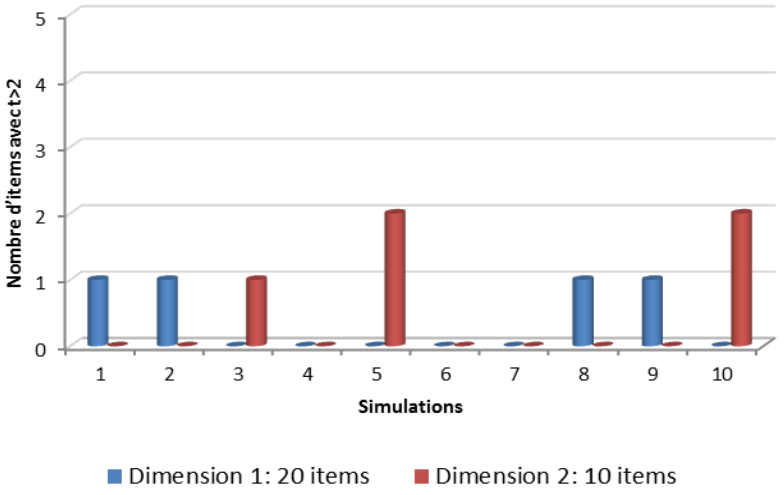


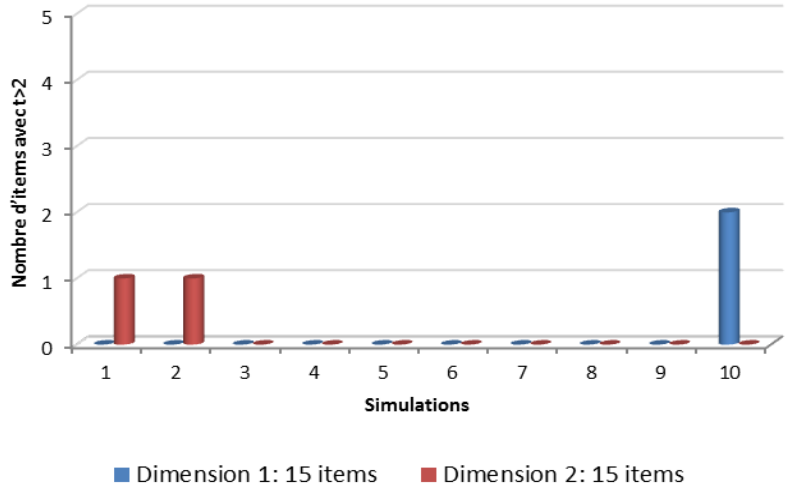
Graphique 9. Nombre d'items dont la statistique d'ajustement t est hors de la plage $]-2,2[$ selon le ratio d'items lorsque la correlation entre les deux dimensions est 0,2.



Graphique 10. Nombre d'items dont la statistique d'ajustement t est hors de la plage $]-2,2[$ selon le ratio d'items lorsque la correlation entre les deux dimensions est 0,6.







Graphique 11. Nombre d'items dont la statistique d'ajustement t est hors de la plage $]-2,2[$ selon le ratio d'items lorsque la correlation entre les deux dimensions est 0,8.

Tableau **3**

Nombre Moyen d'Items dont la Valeur Absolue de la Statistique d'Ajustement t au Modèle de Rasch est Supérieure à 2

Corrélation entre les deux dimensions	Ratio 25:5		Ratio 20:10		Ratio 15:15	
	Dim1=25	Dim2=5	Dim1=20	Dim2=10	Dim1=15	Dim2=15
r=0,0	1,3	4,4	1,4	2,7	0	0,1
r=0,2	0,7	4,4	0,5	1,7	0,1	0,1
r=0,6	0,3	2,3	0,2	0,7	0,3	0,4
r=0,8	0,5	0,8	0,4	0,5	0,2	0,2

Il ressort des graphiques 8, 9, 10 et 11, et du tableau 5.2 les résultats suivants:

- ✓ Lorsque les deux dimensions sont non corrélées ou perpendiculaires, en moyenne un item sur la dimension dominante (25 ou 20 items) a un t en dehors de la plage de référence de données acceptables $[-2,2]$ pour la conformité au modèle de Rasch. Cette moyenne est de zéro lorsque les deux dimensions ont chacune 15 items.
- ✓ Lorsque le ratio est de 25 contre 5 items sur les dimensions dominante et secondaire respectivement, en moyenne, 4 items sur 5 de la dimension secondaire sont identifiés avec des t en dehors de la plage acceptable lorsque la corrélation est soit nulle, soit assez faible (0,2). Cette moyenne chute à 2 lorsque la corrélation est de 0,6, et à 1 en présence d'une très forte corrélation entre les dimensions (0,8).
- ✓ Quand le ratio est de 20 items sur la dimension dominante contre 10 sur la dimension secondaire, en moyenne 3 items sur 10 de la dimension secondaire sont identifiés avec des t en dehors de la plage acceptable lorsque la corrélation est soit nulle, contre 2 lorsque la corrélation est assez faible (0,2). Cette moyenne chute à 1 lorsque la corrélation est de 0,6 ou en présence d'une très forte corrélation entre les deux dimensions (0,8).
- ✓ Lorsque les deux dimensions ont le même nombre d'items, tous sont considérés comme mesurant un même trait car il est rare qu'un des items ait un t en dehors de la plage $[-2,2]$.

Les trois conclusions suivantes peuvent être tirées des analyses précédentes.

Résultat 3**:**

> Lorsque les deux dimensions sont non corrélées, le modèle de Rasch est très robuste à la violation de l'hypothèse d'unidimensionnalité en présence d'une dimension fortement dominante. 4 sur 5 des items de la dimension résiduelle sont identifiés comme problématiques par les statistiques d'ajustement, et doivent donc être retirés des données afin de produire des mesures valides. C'est donc la situation idéale pour le développeur des tests, ce qui montre que l'étape de la validation du contenu reste capitale, car seule la situation de forte dominance garantit que les mesures produites se rapportent à un trait particulier.

Résultat 4:

> Le modèle de Rasch est moins robuste lorsque la dominance est faible (ratio de 20 contre 10 items). Seulement 3 sur 10 items de la dimension secondaire sont identifiés lorsque la corrélation est nulle, contre 2 sur 10 en présence d'une faible corrélation.

Résultat 5:

> Lorsque la corrélation entre les deux dimensions est modérée ou forte le modèle de Rasch n'est pas robuste à la violation de l'hypothèse d'unidimensionnalité des données. Le trait mesuré est une combinaison des deux traits car les items de la dimension secondaire sont rarement identifiés.

On peut remarquer que la statistique t d'ajustement n'est pas parfaite car même lorsque les deux dimensions sont perpendiculaires, il y a au moins un item en moyenne de la dimension dominante qui risque d'être retiré du test si l'on se fie seulement sur cette statistique.

Conclusion

Dans ce chapitre, nous avons exploité les résultats des régressions multiples et les statistiques d'ajustement au modèle de Rasch pour tirer des éléments de réponses à notre question de recherche. Au regard des résultats dans chacune des conditions de la simulation, certaines similitudes apparaissent à l'exemple de la confusion lorsque les dimensions ont le même nombre d'items. On a pu constater des résultats satisfaisants de la modélisation lorsqu'existe idéalement une dimension fortement dominante dans les données.

Conclusion générale

Cette étude a permis de relever l'importance de l'étape de la validation du contenu lorsqu'on développe des tests. Les développeurs des tests doivent donc suffisamment explorer les outils à leur disposition pour s'assurer que les items retenus pour produire des mesures avec le modèle unidimensionnel de Rasch sont suffisamment corrélés pour mesurer un seul trait. Lorsque sur le plan théorique, deux traits sont fortement corrélés, il pourrait être justifié de les considérer comme mesurant ensemble un même trait de niveau supérieur, car le modèle de Rasch les distingue difficilement en particulier lorsque la différence du nombre d'items aux deux dimensions est faible.

Une situation dans laquelle le modèle de Rasch se révèle inefficace est celle où le test comprend un nombre équitable d'items sur deux dimensions différentes peu importe le niveau de corrélation, car le modèle de Rasch ne distingue pas les items pouvant être problématiques sur l'une ou l'autre des deux dimensions. C'est une situation de confusion ou de fourre-tout, que l'on aimerait idéalement éviter dans la réalité.

En combinant les résultats des graphiques 4 à 11 et des tableaux 5.1 et 5.2, il ressort en que :

En présence de deux dimensions égales en termes du nombre d'items, quel que soit le niveau de corrélation entre celles-ci, les paramètres des personnes estimés par le modèle de Rasch unidimensionnel est la simple moyenne des valeurs réelles des paramètres correspondant aux deux dimensions. Les statistiques t d'ajustement ne sont pas un critère sur lequel le chercheur devrait se baser pour approfondir l'identification d'une dimension supplémentaire dans les données, lorsque celles-ci sont multidimensionnelles. Il faut donc se fier aux avis des spécialistes du contenu en priorité en vue de dégager une dimension dominante.

En présence d'une dimension faiblement dominante, les paramètres des personnes estimés sont des combinaisons linéaires des valeurs correspondantes aux deux dimensions, mais la dimension dominante compte au 2/3. Les statistiques t permettent rarement d'identifier les items problématiques qui appartiennent à la seconde dimension en particulier lorsque la corrélation entre les deux dimensions est faible ou forte.

Lorsqu'une dimension est fortement dominante parmi les deux, au moins 75% de la valeur du paramètre sur celle-ci compte dans la constitution de la valeur estimée du paramètre des personnes par le modèle de Rasch unidimensionnel, et ce poids peut aller au-delà de 90%.

Cette étude a montré que la difficulté de l'item est conservée par le modèle de Rasch, quel que soit le degré de corrélation entre les deux dimensions, ou la dominance entre celles-ci dans le cadre du test bidimensionnel a structure simple.

Limites de l'étude

Deux limites de ce travail peuvent être soulignées : en premier lieu, les données ont été simulées pour nous permettre de travailler dans des conditions idéales, contrairement à la pratique. Deuxièmement, cette étude a été focalisée sur le cas des items à réponses dichotomiques, ce qui est une partie de la problématique sur la robustesse du modèle de Rasch. Les cas des items polytomiques à structure simple, ou à structure complexe (dans les deux cas d'items à réponses dichotomiques et polytomiques) dans laquelle un item peut saturer sur plusieurs dimensions n'ont pas été abordés et pourraient faire l'objet d'études supplémentaires.

Références

Akerman, T. A. (1991). The Use of Unidimensional Parameter Estimates of Multidimensional Items in Adaptive Testing. *Journal of educational measurement, 15*, 13-24.

Akerman, T. A. (1994). Creating a Test Information Profile for a Two-Dimensional Latent Space. *Applied Psychological Measurement, 18*, 257-275.

Bertrand, R., & Blais, J.-G. (2004). *Modèles de mesure. L'apport de la théorie des réponses aux items.* Québec: Presse de l'Université du Québec.

Besbeas, P. (2010). Estimation. (N. J. Salkind, Éd.) *Encyclopedia of Research Design*, pp. 420-423.

Blais, J.-G. (1987). *Effets de la violation du postulat d'unidimensionalité dans la théorie des réponses aux items.* Thèse de Doctorat non publiée. Faculté des sciences de l'éducation. Université de Montréal. 181 pages.

Blais, J.-G., & Laurier, M. D. (1995). The dimensionality of a placement test from several analytical perspectives. *Language Testing*, 12:72.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the Human sciences* (éd. second). New York: Routledge.

Brintari, E., & Golia, S. (2007). Unidimensionality in the rasch model: how to detect and interpret. *67*(3), 253-261.

Carmines, E. G., & Woods, J. A. (2005). Validity assessment. *Encyclopedia of social measurement, 3*, p. 933.

Chou, Y.-T., & Wang, W.-C. (2010). Checking Dimensionality in Item Response Models With Principal Component Analysis on Standardized Residuals. *Educational and Psychological Measurement, 70*(5), 717–731.

Christensen, K. B., & Kreiner, S. (2007). A Monte Carlo approach to unidimensionnality testing in polytomous Rasch models. *Applied psychological measurement, 31*(1), 20-30.

Christensen, K. B., Bjorner, J. B., Kreiner, S., & Petersen, J. H. (2002). Testing unidimensionnality in polytomous Rasch models. *Psychometrika, 67*(4), 563-574.

Chyi, H. I., & Lasorsa, D. L. (2002). An explorative study on the market relation between online and print newspapers. *The Journal of media economics, 15*(2), 91-106.

Chyi, I. H. (2005). Willingness to pay for online news: An empirical study on the viability of the subscription model. *Journal of media economics, 18*(2), 131-142.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory.* Harcourt: Harcourt College.

Cuesta, M., & Muniz, J. (1999). Robustness of item response logistic models to violations of the unidimensionnality assumption. *Psicothema, 11*, 175-182.

Cule, E., & De Iorio, M. (2012, Mai 3). *A semi-automatic method to guide the choice of ridge parameter in ridge regression.* Consulté le Janvier 4, 2013, sur http://arxiv.org/pdf/1205.0686v1.pdf

de Ayala, R. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.

Dimmick, J., & Eric, R. (1984). The theory of the niche: Quantifying competition among media industries. *Journal of communication*, 103-119.

Dimmick, J., Chen, Y., & Li, Z. (2004). Competition between the internet and traditional news media: the gratification-opportunities niche dimensions. *The Journal of media economics, 17*(1), 19-33.

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

Erling, B. A. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*(1), 123-140.

Fons, J. (1986). The Robustness of Rasch estimates. *Applied Psychological Measurement*, 45-57.

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical Results related to the Robustness of the Rasch Model. *Applied Psychological Measurement*, 175-186.

Green, W. H. (2005). *Econometric Analysis.* Dehli: Pearson Education.

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

Haig, B. D. (2010). Models. *Encyclopedia of Rresearch Design*, pp. 827-831.

Hambleton, R. K., & Russel, W. J. (1993). Comparison of classical test theorty and item response theory and their application to test development. *Educational Measurement: Issues and practice*, pp. 38-47.

Hambleton, R. k., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* California: SAGA publications.

Hattie, J. (1985). Methodology Review: Assessing unidimensionality of tests and items. *Applied Psychology Measurement , 9*, 139-164.

Jasper, F. (2010). Applied Dimensionality and Test Structure Assessment With the START-M Mathematics Test. *The International Journal of Educational and Psychological Assessment, 6*(1), 104-125.

Kahraman, N., & Thompson, T. (2011). Relating uinidimensional IRT parameters to multidimensional response space: A review of two alternative projection IRT models for scoring subscales. *Journal of educational measurement, 48*, 146-164.

Kirisci, L., Hsu, T.-c., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement*, 146-162.

Larousse. (2012). *Le Petit Larousse Illustré.* Paris: Larousse.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2*(3), 266-283.

Linacre, J. M. (1999). Understanding Rasch Measurement: Estimation Methods for Rasch Measures. *Journal of Outcome measurements, 3*(4), 382-405.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*(2), 157-162.

Magno, C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. *The International Journal of Educational and Psychological Assessment, 1*, 1-11.

Nandakumar, R. (1994). assessing dimensionality of a set of item responses- Comparison of different approaches. *Journal of educational measurement, 81*(1), 17-35.

Raîche, G. (2005). Critical Eigenvalue Sizes in Standardized Residual Principal Components Analysis. *Rasch Measurement Transactions , 19*(1).

Reckase, M. D. (2009). Multidimensionnal item response theory. *Statistics for Social and Behavioral Sciences*. Michigan, United States: Spinger.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology, 8*(33).

Smith, E. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of applied measurement, 3*(2), 205-231.

Stout, W. (1987). A non parametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4), 589-617.

Teol, P.-C., Osman, M., & Ramayah, T. (2011). Testing the dimensionality of Consumer Ethnocentrism Scale (CETSCALE) among a young Malaysian consumer market segment. *African Journal of Business Management, 5*(7), 2805-2816.

van der Wurff, R. (2011). Are News Media Substitutes? Gratifications, Contents, and Uses. *Journal of media economics, 24*, 139-157.

Verhelst, N. (2001). Testing unidimensionality of the Rasch model . *Methods of Psychological Research Online, 6*(3), 231-271.

Wagner, R., Tchatcher, P. K., & Piasta, S. (2010). latent variable. *Encyclopedia of Research design*. (N. J. salkind, Éd.) Thousand Oaks, CA.

Walker, C. M., & Beretvas, N. S. (2003). Comparing multidimensional and unidimensional proficiency classifications: multidimensional IRT as a diagnostic aid. *Journal of educational measurement*, 255-275.

Wang, W.-H., Wilson, M., & Adams, R. (1996). Rasch models for multidimensionality between items and within items. *Objective measurement. Theory into practice*, 139-155.

Way, W., Ansley, T. N., & Fortsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two dimensionnal data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*, 239-252.

Wilson, M., Allen, D. D., & Li, J. C. (2006). Improving measurement in health education and healthj behavior research using item response modeling: Comparison with the classical test theory approach. *Health education research. Theory & practice*, i19-i32.

Wollenberg, A. L. (1982). A simple and effective method to test the dimensionnality axiom of the Rasch model. *Applied psychological measurement, 6*(1), 83-91.

Wright, B., & Mok, M. (2004). An overview of the family of Rasch measurement models. Dans E. V. Smith, & R. Smith, *Introduction to Rasch measurement* (pp. 1-24). Minesota: JAM press.

Zeller, R. A. (2005). Measurement error, issues and solutions. *Encyclopedia of social measurement, 2*, p. 666.