UNIVERSITÉ DU QUÉBEC À MONTRÉAL

# EFFETS DE L'ASYMÉTRIE DANS LA DISTRIBUTION DES PARAMÈTRES DE DIFFICULTÉ AU SEIN D'UNE BANQUE D'ITEMS SUR L'ESTIMATION DES NIVEAUX D'HABILETÉ EN TESTING ADAPTATIF

MÉMOIRE
PRÉSENTÉ COMME
EXIGENCE PARTIELLE
DE LA MAITRÎSE EN ÉDUCATION

PAR
CHRISTIAN BOURASSA

OCTOBRE 2016

# REMERCIEMENTS

Je tiens d'abord à remercier monsieur Gilles Raîche, mon directeur de maîtrise à l'UQAM. C'est une sommité en mesure au Québec et tout spécialement dans mon domaine d'expertise, le testing adaptatif. C'est avec les yeux pleins d'espoir que je suis allé le rencontrer en février 2011 avec en main ce que je croyais être un test adaptatif qui pourrait piquer sa curiosité. C'est avec de la lecture plein les bras que je suis retourné chez moi, déçu de constater que le testing adaptatif était beaucoup plus complexe que je le croyais et de ne rien y comprendre, mais curieux néanmoins et avide de connaissance. C'est une chance et un honneur d'écrire sous son égide.

J'aimerais aussi remercier monsieur Sébastien Béland, étudiant au doctorat avec le même père scientifique, monsieur Raîche, quand j'ai débuté ma maîtrise. Très généreux de son savoir et vulgarisateur hors pair, il a grandement facilité le rapprochement entre moi et mon sujet de recherche!

Ensuite, j'aimerais remercier ma conjointe et mes quatre enfants. Leur patience et leur compréhension ont tamisé quelque peu ma culpabilité de consacrer autant de temps à mon projet.

Je dois aussi des remerciements à monsieur Patrick Charland et madame Julia Poyet, tous deux professeurs à l'UQAM, qui très tôt ont éveillé mon intérêt pour les études supérieures, montrant qu'elles étaient accessibles et nécessaires.

Enfin, je tiens à remercier mes parents, amis et membres de mon entourage qui, chacun à sa façon, m'ont encouragé dans ce projet.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

## RÉSUMÉ

Cette recherche s'intéresse à la constitution des banques d'items et à son impact sur le comportement de tests adaptatifs qu'elle alimente. Plus précisément, elle veut déterminer à quel degré une asymétrie dans la distribution des paramètres de difficulté des items peut avoir un effet sur le nombre d'items administrés et l'estimation du niveau d'habileté. À cette fin, puisque celle-ci, contrairement à une étude à partir de données réelles, permet d'avoir un contrôle strict sur les caractéristiques de la banque d'items, une stratégie de simulation de patrons de réponses à un test adaptatif est retenue. 14 banques d'items ont été générées, selon deux combinaisons possibles de la taille de la banque d'items et sept valeurs du coefficient d'asymétrie de la distribution du paramètre de difficulté des items de la banque. Pour chacune de ces 14 conditions, 5000 unités d'observation ont été produites, soient 1000 à chacun des cinq niveaux d'habileté réels à partir d'une modélisation logistique à trois paramètres issue de la théorie de la réponse à l'item.

Il apparait que l'asymétrie dans la distribution des paramètres de difficulté des items d'une banque a principalement des effets sur la longueur des tests adaptatifs pour des individus dérogeant de deux écarts types ou plus de la moyenne. Pour les individus dits extrêmes, avec une banque constituée de 1000 items, cette asymétrie peut même augmenter la longueur d'un test adaptatif de 450 items. Les effets sur le biais d'estimation et la précision sont généralement négligeables lorsque la banque n'est pas épuisée.



MOTS-CLÉS : testing adaptatif, théorie de la réponse à l'item, banque d'items, asymétrie, difficulté.

# INTRODUCTION

Lorsqu'un enseignant administre un test papier-crayon à ses élèves, il leur distribue une copie d'un même examen. Cet examen peut avoir été construit par l'enseignant lui-même ou par un collègue ou il peut avoir été trouvé en ligne ou être fourni par le Ministère de l'Éducation, etc. Quoi qu'il en soit, cet examen est habituellement destiné à un élève moyen parce que c'est celui qui se trouve en plus grand nombre dans les classes.

Pour évaluer l'élève moyen, il faut lui administrer des items à son niveau d'habileté afin qu'il en réussisse certains et qu'il en échoue d'autres. Si le test comprend trop de questions faciles, l'élève moyen les réussira et paraîtra comme un élève fort. Inversement, si le test comprend trop d'items difficiles, l'élève moyen en échouera plusieurs et paraîtra faible. Le test doit donc comprendre plusieurs items moyens et quelques items plus faciles et plus difficiles.

Bien que cet examen puisse bien circonscrire l'habileté des élèves dits moyens, qu'en est-il des élèves plus forts et plus faibles? Cet examen « moyen » peut-il dresser un portrait représentatif d'un élève « fort » ou « faible »? Surtout, cet examen « moyen » peut-il différencier deux individus similairement « forts » ou « faibles »? Un examen papier-crayon, aussi bien constitué qu'il puisse l'être, ne peut offrir une précision satisfaisante face à des individus de tous les niveaux d'habileté parce qu'il ne peut choisir les items qu'il croit mieux ajustés. L'outil qui le fait le mieux à l'heure actuelle se nomme le testing adaptatif, car il reproduit à travers différents procédés automatisés la démarche et le comportement d'un évaluateur avisé (Wainer, 1990, p. 10). Il s'agit d'une modalité d'administration de tests où le contenu de ceux-ci varie en fonction des réponses des élèves et ainsi en fonction de leur niveau d'habileté.

Cette recherche, bien qu'elle présente les assises théoriques et le fonctionnement du testing adaptatif, cherche surtout à étudier son application dans des conditions spécifiques. Elle veut connaître l'effet de la constitution d'une banque d'items sur le nombre d'items administrés et l'estimation du niveau d'habileté.

Dans cette perspective, la problématique de cette recherche débute par une mise en situation qui présente d'abord le testing adaptatif comme une façon comparable à d'autres d'évaluer un individu. Puis, l'angle de vue sur le problème de recherche se précise, passant des mécanismes de base en testing adaptatif aux enjeux dans l'étude des banques d'items. La recension des écrits circonscrit cet espace de recherche et met en relief les zones moins connues, moins étudiées. C'est alors que sont explicités le problème et les objectifs de la présente recherche.

Le cadre théorique de cette recherche vise à renseigner le lecteur sur trois domaines du champ de l'évaluation : la théorie classique des tests, la théorie de la réponse à l'item et le testing adaptatif. Suivant l'ordre chronologique dans lequel ces domaines ont été développés, la théorie classique des tests est d'abord présentée à travers quelques concepts qu'elle met de l'avant, soient ceux réutilisés par les autres domaines : le score réel, le score observé et l'erreur de mesure. Les forces et faiblesses de cette théorie sont brièvement exposées et c'est dans cet élan que la théorie de la réponse à l'item est introduite dans ses façons de remédier aux lacunes de la théorie classique des tests. Ensuite, les fondements et modèles de la théorie de la réponse à l'item sont présentés en profondeur parce qu'ils constituent l'armature du testing adaptatif. Enfin, le testing adaptatif est présenté avec tous les critères et règles qui le régissent. Évidemment, l'accent est mis sur les enjeux liés aux banques d'items parce que le problème de recherche, la question de recherche et les objectifs de recherche s'y rapportent tous.

Le chapitre dédié à la méthodologie présente le contexte dans lequel la recherche sera conduite. Pour la présente, des unités d'observation seront générées et celles-ci seront soumises à différents tests adaptatifs. La librairie catR, développée pour gérer l'administration de tests adaptatifs dans l'environnement R, est utilisée pour générer les sujets simulés, les banques d'items et pour conduire tous les tests adaptatifs de la recherche. La librairie fGarch est utilisée pour générer des distributions avec différentes valeurs du coefficient d'asymétrie. La génération de toutes ces données est abordée en prémisse de ce chapitre. Par la suite, les méthodes d'analyse des données sont abordées. Pour interpréter les effets de l'asymétrie au sein d'une banque d'items sur différentes valeurs liées à la passation de tests adaptatifs, différentes statistiques calculées sur les variables dépendantes retiennent l'attention. La recherche liste précisément les statistiques et variables retenues pour l'analyse.

Les résultats de cette recherche sont présentés sous la forme de constats sur les effets de l'asymétrie sur le biais d'estimation, la précision et la longueur des tests adaptatifs. Ces constats s'appuient sur les données issues des différentes analyses statistiques décrites dans le chapitre sur la méthodologie. Les données les plus éloquentes ont été retenues et sont présentées dans des tableaux à travers le chapitre.

Dans le chapitre sur la discussion et l'interprétation des résultats, les constats émis dans le chapitre précédent sont expliqués plus en profondeur. Les données et observations sur lesquelles ils reposent sont comparées à des données et observations d'autres recherches s'intéressant au testing adaptatif ou à la théorie de la réponse à l'item. De cette façon, les constats de la présente recherche sont renforcés, appuyés par les écrits antérieurs. Ensuite, les limites de cette recherche sont présentées. Enfin, le chapitre se termine par une ouverture sur des rapprochements possibles avec d'autres recherches et d'autres sujets connexes qui pourraient constituer des études intéressantes pour le domaine du testing adaptatif.

Pour terminer, la conclusion rappelle les objectifs spécifiques de la présente recherche et explique en quoi ils ont été atteints ou non. Après la conclusion suivent des annexes contenant le code utilisé pour générer les banques d'items et les passations de tests adaptatifs, pour conserver les données issues des analyses statistiques et pour produire certains graphiques.

# CHAPITRE I

# PROBLÉMATIQUE

## 1.1 Un scénario fictif

C'est à l'aide d'une tablette numérique que Marie doit passer un test visant à vérifier l'étendue de ses savoirs. Tous les étudiants de son groupe sont présents, chacun une tablette à la main, et après quelques consignes d'un surveillant le test commence. Marie a étudié toute la session durant et elle se considère prête à toute éventualité. Elle a l'habitude de terminer la première et d'obtenir des résultats au-dessus de la moyenne. À chaque test, elle s'exaspère de voir certains étudiants, toujours les mêmes, qui arrivent non préparés. Malheureusement, ils réussissent néanmoins à s'en tirer en trichant : copiant les réponses du voisin, s'échangeant des papiers ou faisant un usage très discret de leur téléphone cellulaire. Toutefois, le test d'aujourd'hui est différent, le surveillant l'a dit. Il s'agit d'un test adaptatif, soit un test qui s'adapte aux réponses de chaque individu. On lui explique que les questions qui lui seront posées seront différentes de celles posées aux autres, mais que toutes couvriront les mêmes contenus. Marie se permet un petit sourire en observant les réactions des étudiants mal préparés, toujours les mêmes, à ces annonces. Ces derniers ne pourront plus demander d'aide à personne. Ils devront maintenant se fier à leurs connaissances et leur chance et non au talent des autres. On lui explique aussi que le test devrait être aussi difficile pour l'un que pour l'autre. En effet, chacun se verra administrer des items d'une difficulté telle qu'il ne soit jamais trop facile ni trop difficile d'y répondre. Ces choix, c'est l'ordinateur qui les fait selon les réponses aux items déjà administrés. On lui explique enfin que le test sera plus court pour certains et plus long pour d'autres. Le test a différentes conditions d'arrêt, dont l'atteinte d'une certaine précision minimale.

Hormis quelques items très faciles et d'autres, apparemment impossibles, Marie a eu l'impression d'être constamment mise à l'épreuve pendant le test, comme si l'ordinateur savait comment la défier sans la décourager. En discutant avec d'autres étudiants, tous semblent s'être sentis de la sorte, même s'ils ont répondu à des items bien différents. Aussi, elle a été surprise de connaître son résultat tout de suite à la fin de son test. Encore plus impressionnant : l'ordinateur l'a considérée deux écarts types au-dessus de la moyenne avant même que tous les résultats aient été compilés! Marie est ébahie, elle ne savait pas que les innovations en évaluation en étaient rendues là! Pourtant, ces innovations constituent en fait un champ de recherche datant déjà de plus de 30 ans!

## 1.2 Historique et fonctionnement du testing adaptif

*« Standardized testing of college-level skills recently began its long-awaited shift into the realm of computers »* pouvait-on lire dans le Chicago Sun-Times en 1986, alors que certains collèges américains introduisaient les premiers tests adaptatifs à leur processus de classement des nouveaux étudiants. Comme l'accès à certains programmes était restreint – et c'est encore le cas – les collèges voulaient savoir qui étaient les meilleurs étudiants. Avec un nombre de places limité, il était impossible d'établir un seuil minimal, il importait de connaître qui étaient les meilleurs parmi les meilleurs. Pour ce faire, il leur fallait un outil précis sur l'ensemble des niveaux d'habileté, surtout auprès des étudiants très forts, ce qu'un test traditionnel « papier-crayon » peinait – et peine encore – à faire. Le testing adaptatif, profitant d'un support technologique en constante évolution, offrait l'outil idéal.

Un test adaptatif s'adapte aux réponses de l'individu de manière à ce que les meilleures questions lui soient toujours posées (Veldkamp et Matteucci, 2013, p. 58). Une bonne question n'est ni trop facile ni trop difficile pour un individu. Le choix de la question – de l'item – appartient à l'ordinateur qui estime après chaque réponse le

niveau d'habileté de l'individu et qui sélectionne dans une banque un item sur mesure. Le test adaptatif se termine lorsqu'une précision minimale donnée a été atteinte ou lorsqu'un nombre d'items maximal donné a été administré. Puisque ce test se fait à l'ordinateur, le nombre d'individus à qui le test peut être administré est restreint au nombre d'ordinateurs disponibles. En contrepartie les individus se font tous administrer un test différent. Alors, tous les calculs, aussi complexes soient-ils, sont effectués rapidement et le résultat est disponible dès le dernier item complété. La figure 1.1 présente le déroulement d'un test adaptatif.

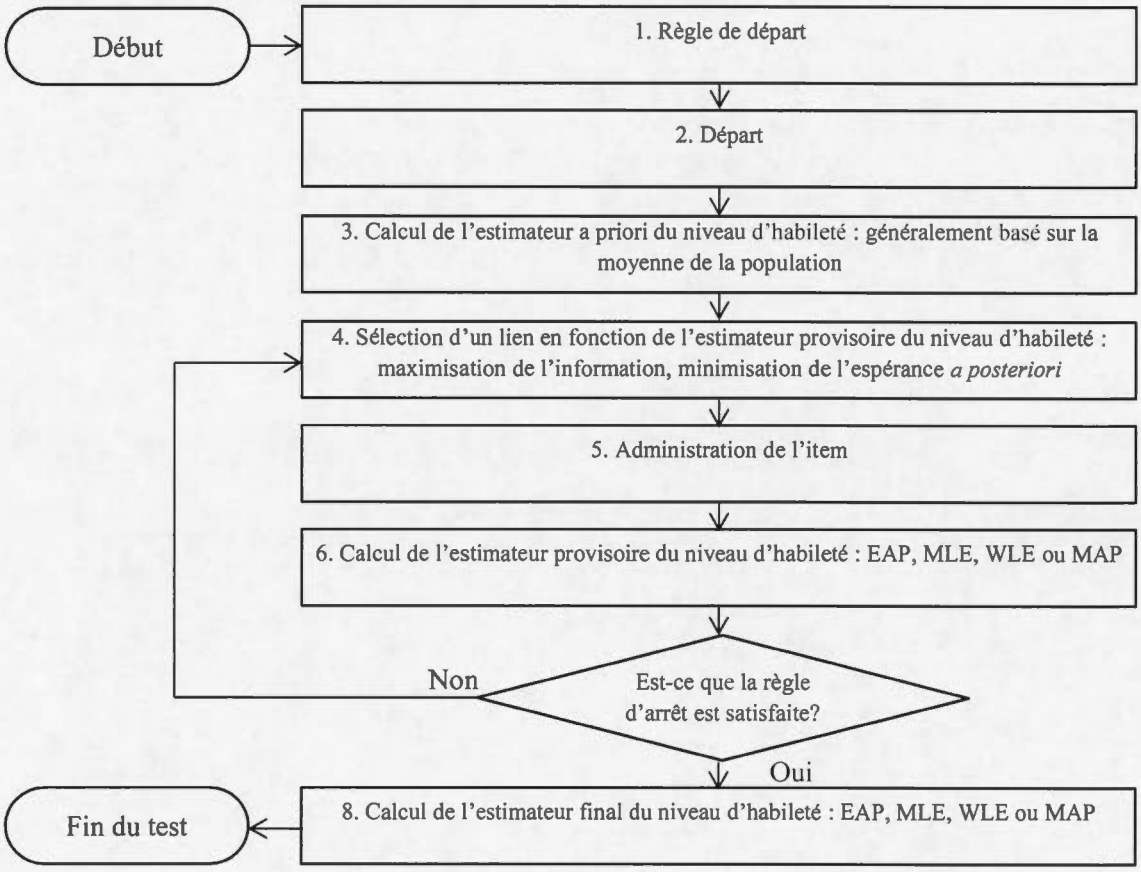


Figure 1.1 : Structure d'un test adaptatif basé sur la théorie de la réponse à item (tiré de Raîche, dans Bertrand et Blais, 2004, p. 325)

Lorsque des organisations[1] ont intégré le testing adaptatif à leurs pratiques évaluatives, ils ont investi énormément d'argent pour concevoir la plate-forme, monter des banques d'items, tester celles-ci et en assurer la maintenance. De nos jours, les étapes sont essentiellement les mêmes, quoique les ordinateurs ont été fortement réduits en taille et en prix. Ainsi, un établissement qui veut administrer des tests adaptatifs doit avoir la plate-forme pour ce faire et une banque d'items pour les nourrir. Bien qu'il soit plutôt simple de régler la question logicielle, alors qu'il existe des plates-formes de passation ou de simulation de tests adaptatifs gratuites (Concerto, catR, etc.), la question de la banque d'items est beaucoup plus épineuse. D'abord, tout test adaptatif doit reposer sur une banque d'items volumineuse afin qu'il y ait, pour chaque niveau d'habileté possible, un choix d'items suffisant sans quoi le test adaptatif ne peut administrer d'items à la juste mesure de l'individu (Reckase, 2010, p. 128). D'ailleurs, ces items, aussi nombreux soient-ils, doivent être vérifiés par des experts et testés auprès d'un échantillon représentatif de la population. Les données – des patrons de réponses dichotomiques, soient des vecteurs de 1 (succès) et 0 (échec) – ainsi recueillies sont ensuite utilisées pour calibrer la banque d'item, soit pour estimer les propriétés psychométriques des items de la banque. Enfin, il faut entretenir la banque d'items, c'est-à-dire régulièrement vérifier si certains items semblent dysfonctionnels et les supprimer, vérifier si des items sont surexposés ou sous-exposés, ajouter des items lorsqu'il en manque dans certaines plages de difficulté et vérifier les patrons de réponses des individus pour voir s'ils semblent appropriés ou non, pour ne nommer que ces quelques vérifications.

## 1.3 Recension des écrits

Des recherches se sont intéressées à ces vérifications, aux différents enjeux, défis, problèmes et innovations dans les banques d'items dédiées au testing adaptatif. En

[1] Les forces armées américaines, des Universités et une multitude de firmes conseil en ressources humaines en sont quelques exemples.

effet, Bjorner, Chang, Thissen et Reeve (2007) présentent très bien les étapes à suivre dans la construction d'une banque d'items, leur schéma est reproduit dans la figure 1.2.

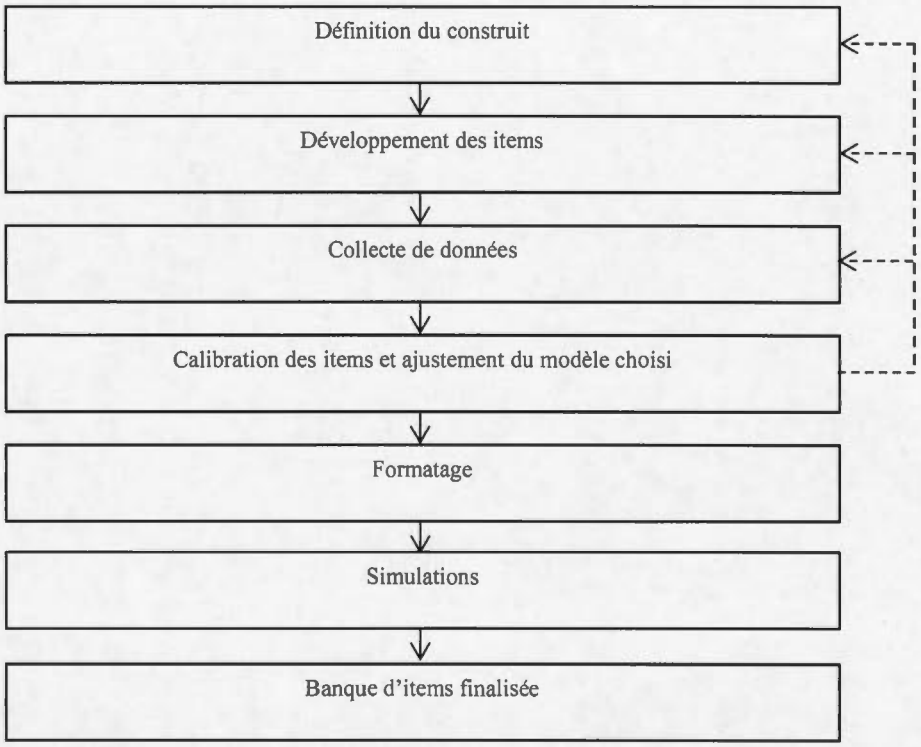


Figure 1.2 : Étapes dans la construction d'une banque d'items (d'après Bjorner, Chang, Thissen et Reeve, 2007, p. 101)

Aussi, Flaugher (1990) présente les caractéristiques désirables d'une banque d'items. Notamment, elle devrait contenir des items de haute qualité ajustés à différents niveaux d'habileté. Les items qu'elle contient devraient respecter les postulats inhérents au modèle choisi. Enfin, les versions papier-crayon et numériques de ces items devraient être comparables d'un point de vue opératoire et dans les résultats qu'ils admettent afin que toutes deux mesurent le même trait.

Reckase (2003) démontre comment l'utilisation de la méthode *bin and union* peut, justement, générer une banque d'items avec des caractéristiques enviables pour le testing adaptatif. Les items d'une banque sont distribués dans des paniers d'une largeur donnée selon leur paramètre de difficulté. Tous les items qui s'y trouvent sont reconnus équivalents. Le choix d'un item pour un individu se fait donc dans le *bin* – panier – où se situe le niveau d'habileté provisoire qui lui est reconnu. De cette façon, l'exposition des items est balancée et des items ajustés – mais pas nécessairement optimalement ajustés – sont administrés.

Bode, Lai, Cella et Heinemann (2003) exposent les analyses nécessaires à la validation d'une banque d'items composite créée à partir d'autres banques d'items à travers un exemple en médecine. Il faut d'abord conduire une analyse factorielle à partir des résultats d'individus de différents groupes afin de voir si les items administrés mesurent le même trait ou si plus d'un facteur semble expliquer les résultats observés. Ensuite, il faut conduire une analyse factorielle à même la banque afin de voir si les items qui la constituent mesurent aussi le même construit. Enfin, il faut vérifier si les items fonctionnent similairement à travers les groupes, en vérifiant si la hiérarchisation des items d'un groupe à l'autre, selon leurs paramètres, est comparable.

Mislevy, Rupp et Harring (2012) démontrent comment les statistiques $Q_3$ et $X^2$ peuvent aider à vérifier l'hypothèse d'indépendance locale au sein d'une banque d'items à réponses polytomiques. Zhang (2008) vérifie si l'hypothèse d'unidimensionnalité tient en présence d'items sensibles à d'autres dimensions. Sympson and Hetter (1985) proposent une méthode pour éviter la surutilisation des items dans laquelle une probabilité d'administration est calculée pour chaque item sélectionné pour un test adaptatif. Cette probabilité est notamment fonction du taux d'exposition actuel de l'item. Si un item n'est pas administré, un autre est sélectionné et le calcul recommence. Veldkamp, Verschoor et Eggen (2010) proposent des

méthodes pour contrôler simultanément la surexposition et la sous-exposition des items en testing adaptatif de sorte qu'il y ait un usage plus uniforme de l'ensemble des items de la banque.

En bref, une multitude de recherches s'intéressent spécifiquement aux banques d'items, mais d'autres s'intéressent plutôt à d'autres aspects du testing adaptatif : estimation des paramètres d'item et de personne, sélection d'item, règles de départ et de fin, etc.

## 1.4 Objectifs généraux de la recherche

La présente recherche s'intéresse à un aspect peu documenté des banques d'items dédiées au testing adaptatif. Elle veut découvrir comment le test réagit quand la banque d'items contient trop peu d'items ajustés au niveau de l'individu. Cette recherche compte trois objectifs généraux. D'abord, elle veut connaître l'impact d'une distribution asymétrique du paramètre de difficulté des items d'une banque sur différentes variables. Notamment, cette recherche veut savoir si pareille constitution d'une banque d'items peut affecter la précision et la longueur d'un test adaptatif qui y puise ses items. Aussi, elle veut voir quel effet cette asymétrie peut avoir sur le biais de l'estimateur du niveau d'habileté à différentes étapes d'un test adaptatif. Enfin, elle veut vérifier si les effets de l'asymétrie sont les mêmes avec des banques d'items de tailles différentes. Le concept même d'asymétrie n'a pas été beaucoup exploré en testing adaptatif; si quelques recherches utilisent le coefficient d'asymétrie d'un paramètre particulier, cette asymétrie est rarement au cœur de ces recherches.

## 1.5 Pertinences sociale et scientifique de la recherche

Pour le moment, cette recherche n'aura sans doute que de minces retombées dans les établissements d'enseignement du Québec puisque le testing adaptatif n'y est pas

encore une pratique très répandue. Toutefois, le testing adaptatif est bien établi ailleurs. En 2013, l'IACAT – une association internationale promouvant le testing adaptatif informatisé – listait 30 programmes largement utilisés dans le monde qui fonctionnent avec des banques d'items et des tests adaptatifs (http://www.iacat.org/content/operational-cat-programs). Le Québec, avec un peu de retard, suivra le pas éventuellement. D'ici là, cette recherche peut aider toute instance pratiquant le testing adaptatif dans l'entretien de ses banques d'items en proposant des valeurs acceptables du coefficient d'asymétrie dans la distribution des paramètres de difficulté des items.

Cependant, la contribution notable de cette recherche résidera dans ses apports au domaine scientifique que constitue le testing adaptatif. En effet, le testing adaptatif est encore jeune, chaque recherche qui lui est vouée ajoute à sa crédibilité et sa notoriété auprès des géants à qui sont consacrés des milliers de recherches depuis des lunes : motivation scolaire, difficultés d'apprentissage, etc. De plus, comme la construction des banques d'items constitue sans doute l'obstacle majeur, tant financièrement qu'opérationnellement, aux yeux des néophytes, cette recherche peut rendre, à sa façon, le testing adaptatif un brin moins complexe et, peut-être, un brin plus désirable!

# CHAPITRE II

# CONTEXTE THÉORIQUE

Le testing adaptatif, tel qu'utilisé dans la présente recherche, repose sur les modélisations issues de la théorie de la réponse à l'item. Les premiers modèles de cette théorie ont été développés par Lord (1952, 1953), Rasch (1960) et Birnbaum (1957, 1958a, 1958b, 1968). Ces modèles visaient à mieux circonscrire et mesurer un trait latent à l'aide de tests. C'est par la compréhension de la dynamique entre l'individu et les items qui lui sont administrés que ces chercheurs sont parvenus à estimer plus précisément le trait de l'individu qui a produit le résultat observé au test.

Borsboom, Mellenbergh et Van Heerden (2003) ont consacré tout un article au concept de trait latent et lui portent différents regards. Ils le présentent comme un concept théorique, formel, utilisé pour parler plus généralement d'une aptitude, compétence, attitude. Ils le présentent aussi comme un concept statistique, empirique, à la base de plusieurs modèles de mesure. Enfin, pour connecter ces deux conceptions du trait latent, les auteurs le définissent d'un point de vue philosophique, ontologique. Alors, par trait latent ou variable latente, on peut entendre intelligence, niveau d'habileté, connaissance ou compétence, mais il faut considérer aussi l'usage opérationnel qui en est fait.

Une variable latente est non observable et comme toute variable non observable ne peut être directement mesurable – « inherently unobservable, hence immeasurable » (Gibbs, 1975) – c'est par une mesure indirecte qu'il faut procéder. Plus précisément, il faut en estimer la valeur à partir des manifestations du trait – des réponses à des items – et espérer un maximum de précision. Ainsi, aucun test ne peut se vanter de pouvoir mesurer exactement un tel trait. La valeur de ce trait latent demeurera

toujours inconnue, mais les techniques d'estimation se raffineront et permettront des estimations toujours plus précises.

D'ailleurs, la théorie de la réponse à l'item constituait justement dans les années 60 un pas de géant en mesure, parce qu'elle offrait une meilleure précision que la théorie classique des tests, alors maîtresse dans le domaine depuis longtemps, en remédiant à certaines de ses limites (Hambleton, Swaminathan et Rogers, 1991, p. 1-31). Notamment, l'erreur de mesure pouvait être estimée à chaque niveau d'habileté parce qu'elle n'était plus liée à un groupe mais bien à une seule mesure. Quelques années et plusieurs recherches plus tard, des chercheurs ont développé une façon d'administrer des tests en s'arrimant aux principes de la théorie de la réponse à l'item. Ainsi, le testing adaptatif était né.

La structure du cadre théorique de cette recherche suit d'une certaine façon cette progression historique, allant de la théorie classique des tests au testing adaptatif. En effet, les concepts et fondements de la théorie classique des tests seront d'abord présentés, parce qu'ils aident à la compréhension de ces mêmes concepts et fondements repris et définis autrement dans la théorie de la réponse à l'item. Ils seront présentés à travers les problèmes de la théorie classique des tests en matière de précision. Par la suite, la théorie de la réponse à l'item sera introduite en démontrant comment elle remédie à certaines faiblesses de la théorie classique des tests et de quoi elle est constituée exactement. Puis, les modèles principaux de la théorie de la réponse à l'item seront exposés, ainsi que certains rouages nécessaires à la compréhension de la présente recherche. Il est important de noter que seuls les modèles à réponses dichotomiques – succès ($x = 1$) ou échec ($x = 0$) – seront présentés, car la présente recherche travaille avec des matrices dichotomiques. Enfin, le testing adaptatif sera présenté dans ses différents mécanismes et règles dont les différentes considérations en matière de banques d'items, soit l'intérêt de la présente recherche.

## 2.1 Problèmes de la théorie classique des tests

La théorie classique des tests a introduit trois concepts qui, par la suite, ont été récupérés et approfondis par d'autres théories : le score réel, le score observé et l'erreur de mesure :

$$X_j = T_j + E \tag{2.1}$$

où l'indice $j$ correspond à l'individu, $T_j$ correspond au score réel (aussi dénommé score vrai), soit un trait latent associé à cet individu, $X_j$ à son score observé et $E$ à l'erreur de mesure. Ce trait latent peut correspondre à une habileté précise, à une compétence ou à une mesure de l'intelligence en général. C'est un estimateur de cette valeur, de ce score réel, qui est recherché. Théoriquement, si un même test pouvait être administré à plusieurs reprises à un même individu $j$ sans qu'il découvre entre les administrations quels items il a réussis et manqués, alors la moyenne de ses résultats à ces tests constituerait son score réel $T_j$ (Bertrand et Blais, 2004, p. 39-40).

Le score observé $X_j$, est le résultat de l'individu à un test, soit la manifestation de son score réel à une épreuve. Ce score est souvent exprimé comme une proportion d'items réussis par rapport à un nombre total d'items administrés. Bien que le score réel d'un individu soit normalement invariable, le score observé n'est pas nécessairement stable. Certaines variables externes ou internes pourtant indépendantes du niveau d'habileté de l'individu peuvent avoir des impacts considérables sur sa performance à un test et produire de l'erreur dans l'estimation du score réel (Lord et Novick, 1968). Poursuivant l'exemple théorique amorcé en décrivant le score réel de l'individu $j$, le score observé $X_j$ constitue le score obtenu par cet individu $j$ à l'une des nombreuses administrations du test (Bertrand et Blais, 2004, p. 39).

L'erreur de mesure $E$ est ce qui différencie le score réel $T_j$ du score observé $X_j$ à chacune des administrations du test. Comme $E$ constitue l'écart entre $T_j$ et $X_j$, alors la moyenne des $E$ devrait correspondre à l'écart entre la moyenne des $T_j$ et la moyenne des $X_j$, mais comme la moyenne des $X_j$ donne $T_j$, alors la moyenne des $E$ est nulle. Il est généralement postulé que la distribution de $X_j$ suit une loi normale où les valeurs centrales sont plus fréquentes – autrement dit où l'individu $j$ réagit comme un individu de niveau $T$ – et où les valeurs extrêmes sont moins fréquentes – où l'individu $j$ réagit comme un individu plus fort ou plus faible que $T$. De ce fait, la distribution de l'erreur de mesure $E$ suit aussi, parallèlement, une loi normale. Quoi qu'il en soit, pour une simple mesure, $E$ constitue un écart et plus il est petit, plus la mesure est précise. Il est difficile d'identifier les causes exactes d'un tel écart parce que les possibilités sont nombreuses : validité du test, contexte de passation du test, patrons de réponse inappropriés des individus, etc. (Bertrand et Blais, 2004, p. 38-49).

Le test papier-crayon est généralement l'instrument de mesure de prédilection de la théorie classique des tests. Il s'agit d'un test constitué d'items à réponse construite ou choisie qui peut être administré à un individu ou à de grands groupes d'individus simultanément. C'est à partir de ces tests que sont estimés les scores réels et observés ainsi que les erreurs de mesure. Bien que l'utilisation de pareils tests facilite la correction et le calcul des résultats, celle-ci vient avec son lot d'inconvénients, surtout en ce qui concerne la précision des scores. En effet, un score précis devrait refléter ce dont un individu est réellement capable et pour ce faire, on devrait lui administrer un maximum d'items propres à son niveau d'habileté. Malheureusement, les tests papier-crayon sont conçus de façon à faire face à des individus de tous niveaux d'habileté et pour ce faire, ils sont constitués d'items de niveaux de difficulté très variés. Ainsi, les individus faibles se voient inévitablement administrer des items trop difficiles pour eux et les individus très forts, des items trop faciles pour eux. Alors, les calculs des scores ne reposent pas sur une quantité raisonnable d'items du bon niveau de

difficulté, surtout pour ces individus très forts ou très faibles. Ce type de test risque donc de ne contribuer que très peu à une estimation suffisamment précise du niveau d'habileté lorsque ce dernier est plutôt extrême (Mead et Drasgow, 1993, p. 450).

## 2.2 Avantages de la théorie de la réponse à l'item

La théorie de la réponse à l'item utilise différemment les concepts de score réel, de score observé et d'erreur de mesure. En effet, le score observé représente encore une proportion d'items réussis mais cette mesure est modulée par les paramètres des items administrés puis placée sur une échelle continue et affichée en score z. Cette mesure, notée $\hat{\theta}$, est une estimation du score réel $\theta$ et l'erreur de mesure correspond à l'écart entre $\hat{\theta}$ et $\theta$.

L'estimé du niveau d'habileté $\hat{\theta}$ est calculé selon les réussites et échecs d'un individu à des items aux paramètres connus et ce sont principalement ces paramètres qui déterminent la position de l'individu sur l'échelle, selon le modèle et l'estimateur utilisés, et non une proportion d'items réussis. Après tout, que signifie un score de 22/24 si tous les items administrés sont très faciles et que signifie un score de 6/22 si tous les items sont très difficiles? Un de ces scores est-il meilleur que l'autre? La théorie de la réponse à l'item a des réponses à ces questions pourtant insolubles selon la théorie classique des tests.

Ensuite, l'erreur de mesure correspond encore à l'écart entre le score réel et l'estimé qui en est fait. Toutefois, la théorie de la réponse à l'item raccorde l'erreur à un individu, à une mesure plutôt qu'aux résultats d'un groupe. Ainsi, il est possible qu'un lot d'items administrés à un individu indique une certaine erreur de mesure, mais qu'auprès d'un autre individu, ce même lot d'items en indique une différente. Ce faisant, il est possible de déterminer pour quels types d'individu un test, ou même

un item, est plus précis ou plutôt mieux ajusté. Selon la théorie de la réponse à l'item, la précision d'un test se mesure à l'aide de l'information que fournit chaque item selon l'individu et la racine carrée de la réciproque de l'information ainsi cumulée donne l'erreur de mesure, plus précisément son erreur type. Autrement dit, un même item ne fournit pas autant d'information pour deux individus de niveaux différents. Alors, il est possible d'identifier pour quelle étendue l'information calculée pour cet item est maximale ou, plus précisément, pour quels individus. Inversement, il est donc aussi possible de cibler quels items sont les plus informatifs pour un individu, de sorte qu'il soit même possible de lui créer un test sur mesure. La section 2.6 de ce chapitre présente et explique la fonction d'information (Baker, 2001, p. 106).

De plus, la théorie de la réponse à l'item, grâce à l'utilisation d'une même échelle de mesure pour les individus et les items (scores *z*), permet des analyses complexes que la théorie classique des tests ne permet pas. D'abord, comme les paramètres d'items et de personnes sont habituellement estimés à l'aide d'importants lots de données, alors les paramètres d'items sont considérés indépendants des individus grâce à qui ils ont été estimés. *De facto*, les niveaux d'habileté estimés sont indépendants des items sélectionnés pour l'estimation. Alors, une fois une banque d'items bien calibrée, ses items peuvent être utilisés dans n'importe quelle combinaison pour estimer l'habileté d'un individu (Hambleton et Swaminathan, 1985, p. 11), en autant qu'il provienne de la même population que ceux ayant fourni les données de départ. Cette vérification peut être effectuée à l'aide d'une analyse du fonctionnement différentiel desdits items, ce qui constitue d'ailleurs un autre avantage de la théorie de la réponse à l'item. En effet, l'analyse du fonctionnement différentiel d'items sert à détecter les items qui semblent léser certains groupes ou sous-groupes d'individus qui y répondent différemment des autres, indépendamment de leur niveau d'habileté

(Holland et Wainer, 1980, p. 4-5)[2]. Ensuite, elle facilite la détection de patrons de réponses inadéquats qui suggèrent des comportements particuliers d'individus : tricherie, négligence, sous-performance intentionnelle, réponses au hasard, etc. Plusieurs approches sont utilisées pour étudier le phénomène – l'utilisation d'indices de détection de patrons de réponses inappropriés, l'analyse des courbes de réponse des individus, etc. (Brassard, Béland et Raîche, 2011, p. 86) – et la théorie de la réponse à l'item permet l'utilisation d'indices paramétriques qui mesurent les écarts entre les scores observés et ceux attendus selon les paramètres et le modèle utilisé (Karabatsos, 2003, p. 279-282). Enfin, elle permet et facilite l'utilisation du testing adaptatif en assurant la mécanique de chaque rouage de cette plateforme complexe[3].

En somme, la théorie classique des tests comporte plusieurs avantages – simplicité d'utilisation, de compréhension et d'application –, mais elle ne peut rivaliser avec la théorie de la réponse à l'item en termes de précision et d'étendue d'applications. Toute situation d'évaluation qui, à son terme, veut mesurer précisément un niveau d'habileté et non seulement vérifier la réussite d'un pourcentage d'items aurait avantage à s'articuler autour des principes de la théorie de la réponse à l'item.

## 2.3 Fondements et modèles de la théorie de la réponse à l'item

La théorie de la réponse à l'item avance que la probabilité de réussite d'un individu à un item peut être calculée à partir du niveau d'habileté de l'individu et de certains paramètres de l'item. Différents modèles de régression logistique dépeignent cette interaction entre l'individu et l'item et tous considèrent la probabilité de réussite à un item fonction du niveau de l'habileté de l'individu à qui ledit item est administré.

---

[2] La section 2.9.10 est consacrée au fonctionnement différentiel d'items, un enjeu dans l'entretien des banques d'items vouées au testing adaptatif.

[3] La section 2.7 décrit le fonctionnement du testing adaptatif basé sur les modèles de la théorie de la réponse à l'item.

D'ailleurs, cette habileté doit être la seule variable indépendante en cause, autrement dit la seule variable qui explique un résultat à un item ou un test et il doit en être de même pour tous les items du test. Ces affirmations constituent les postulats de la théorie de la réponse à l'item, soient l'unidimensionnalité – un seul trait latent mis à l'épreuve – et l'indépendance locale – les résultats à des items n'ont aucune incidence sur les résultats à d'autres items, seule l'habileté est responsable (Hambleton et Swaminathan, 1985, p. 16-25). Ainsi, dans la théorie de la réponse à l'item, le calcul d'une probabilité de réussite se fait à l'aide du niveau d'habileté de l'individu et de caractéristiques propres à l'item, comme le montre l'équation 2.2 :

$$P(U_i = 1|\theta_j, a_i, b_i, c_i) = c_i + \frac{1-c_i}{1+e^{-Da_i(\theta_j - b_i)}} \quad (2.2)$$

Dans cette fonction, soit celle du modèle logistique à 3 paramètres (3PL), $\theta_j$ correspond au niveau d'habileté de l'individu $j$ et sa valeur varie entre $-\infty$ et $+\infty$ sur une échelle de scores $z$, où 0 le situe dans la moyenne et où chaque unité l'éloigne d'un écart-type de la moyenne vers des valeurs de plus en plus extrêmes. Ainsi, les valeurs de $\theta_j$ sont habituellement situées entre -3 et 3, là où se situe environ 99,93 % d'une population. Ensuite, la constante métrique $D$ de Haley est utilisée fréquemment dans les modèles logistiques pour permettre d'approximer une ogive normale. Cette constante est égale à 1,702 (Camilli, 1994, p. 294). Enfin, $a_i$, $b_i$ et $c_i$ sont des paramètres d'item qui, chacun à sa façon, agissent sur la probabilité de réussite d'un individu et donc modifient l'allure de cette courbe. Ces derniers paramètres seront expliqués plus en détail dans la section 2.5.

### 2.3.1 Courbe caractéristique d'item

La courbe associée à la fonction d'un modèle de la théorie de la réponse à l'item se nomme la courbe caractéristique d'item. La fonction précédente (équation 2.2)

permet justement d'en obtenir une. La courbe caractéristique d'item est une courbe non-linéaire dont tous les points se situent entre $f(x) = 0$ et $f(x) = 1$, car $f(x)$ correspond à une probabilité. Cette courbe permet notamment de visualiser les variations des probabilités de réussite à un item $i$ selon les différentes valeurs des paramètres de personne, ici $\theta_j$ et d'item, ici, en relation au modèle 3PL précédent, $a_i$, $b_i$ et $c_i$. Aussi, la courbe caractéristique d'item permet d'identifier le point d'inflexion, soit le $\theta_j$ où la tangente traverse la courbe ainsi que le $\theta_j$ où la probabilité de réussite est de 0,5 (Baker, 2001, p. 7-10). Un exemple est donné à la figure 2.1.

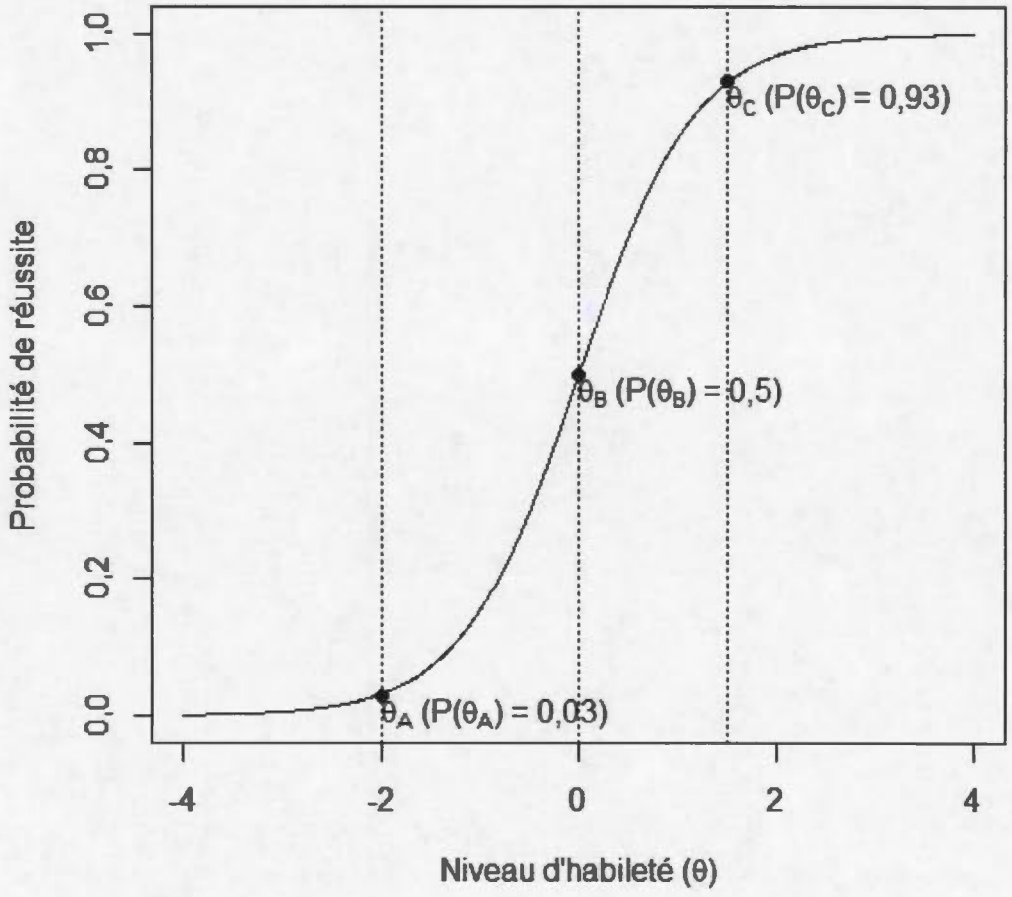


Figure 2.1 Courbe caractéristique d'un item modélisé selon le modèle logistique à un paramètre ($b = 0$).

La figure 2.1 montre bien que, naturellement, l'individu $A$ ($\theta_A$ = -2), très faible, a une moins grande probabilité de réussite ($P$ = 0,03) à l'item que l'individu moyen $B$ ($\theta_B$ = 0, $P$ = 0,5) ou que l'individu $C$ ($\theta_C$ = 1,5, $P$ = 0,93), plutôt fort.

Plusieurs modèles de réponses à l'item existent et autant de courbes caractéristiques d'item. Ceux-ci tiennent compte de différents paramètres de personne et d'item, mais comme le modèle logistique à 3 paramètres sera utilisé dans la présente recherche, il est préférable de maximiser les explications à son sujet et d'uniquement survoler les modèles à un et deux paramètres qu'il englobe.

### 2.3.2 Modèle 1PL

Dans la fonction de probabilité, l'utilisation du seul paramètre d'item $b$, soit le paramètre de difficulté, et de la constante de Haley ($D$ = 1,702) signifie que le modèle logistique à 1 seul paramètre d'item – abrégé 1PL – est utilisé et que seule la difficulté de l'item peut changer la forme de la fonction. Les paramètres $a_i$ et $c_i$ ne sont pas considérés dans ce modèle, comme si les valeurs 1 et 0 leur étaient respectivement attribuées, réduisant l'équation 2.2 à :

$$P(U_i = 1|\theta_j, b_i) = \frac{1}{1+e^{-D(\theta_j - b_i)}} \tag{2.3}$$

En modifiant le paramètre $b_i$ de la fonction de probabilité, la courbe se déplace sur l'abscisse, plus à gauche pour un item plus facile et plus à droite pour un item plus difficile, tout en demeurant parallèle à sa courbe d'origine, en conservant la même allure. Un exemple d'items modélisés selon le modèle logistique à un paramètre est présenté à la figure 2.2.

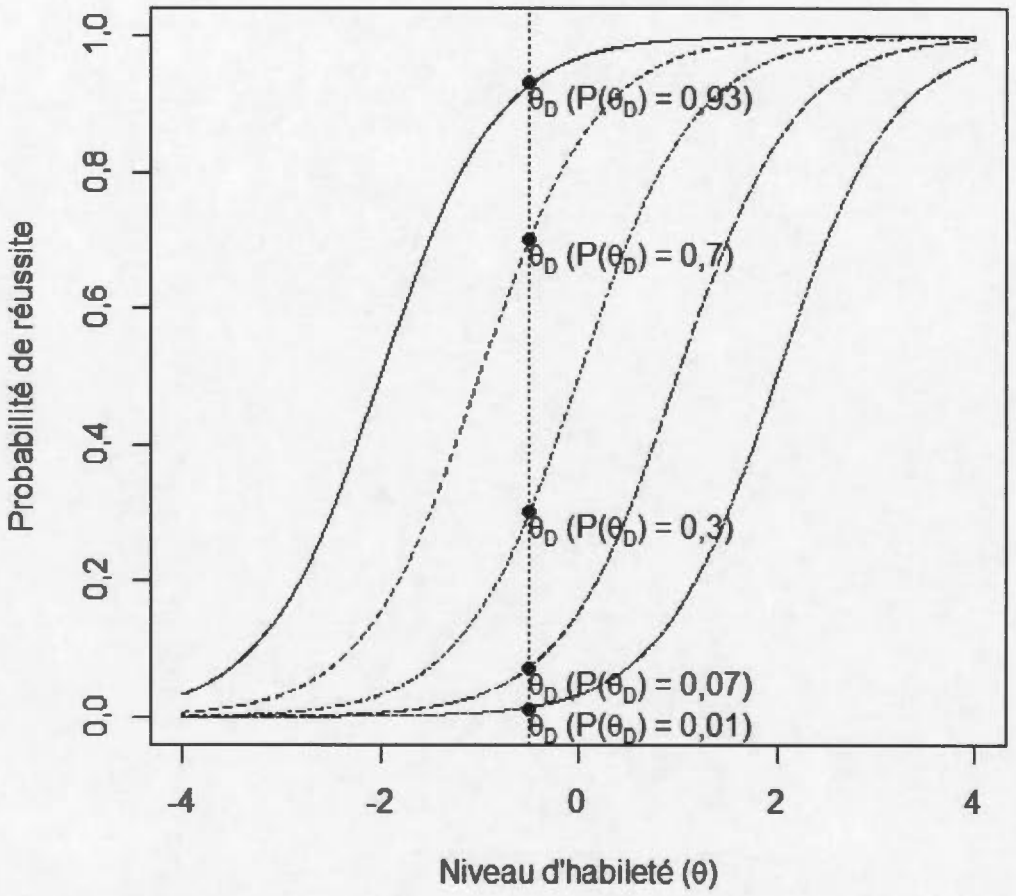


Figure 2.2 : Effets des variations du paramètre de difficulté sur la courbe caractéristique d'items suivant le modèle logistique à un paramètre ($b$ = [-2, -1, 0, 1, 2])

Par exemple, dans la figure 2.2, l'individu D ($\theta_D$ = -0,5) affiche différentes probabilités de réussite à l'item selon la valeur du paramètre de difficulté ($b$ = [-2, -1, 0, 1, 2]). Ce modèle est aussi nommé le modèle de Rasch. Rasch (1960) l'a proposé indépendamment des travaux effectués sur la théorie de la réponse à l'item par Lord. Ce modèle est largement utilisé vu sa simplicité d'utilisation et d'interprétation. De plus, la calibration d'une banque d'items modélisée selon le modèle de Rasch est une procédure beaucoup plus simple et nécessitant moins d'items que les modèles 2PL et 3PL plus complexes.

### 2.3.3 Modèle 2PL

Le modèle 1PL peut être restrictif, car certains items ne s'y ajustent pas et sont alors écartés du modèle. Des items présentant des courbes caractéristiques d'item non parallèles les unes aux autres, donc présentant des pentes différentes près de leur point d'inflexion, constituent un non-sens au sein de ce modèle, mais le modèle logistique à un paramètre peut être généralisé afin qu'il prenne en compte ces items discordants. Pour ce faire, il faut considérer un deuxième paramètre d'item dans le modèle logistique, soit le paramètre de discrimination $a_i$, habituellement fixé à 1 pour tous les items suivant le modèle 1PL. Ce paramètre montre à quel point un item fait réagir deux individus différemment, au-delà de la difficulté de l'item. Autrement dit, un item discriminant voit des individus de niveaux d'habileté différents présenter des probabilités de réussite plus divergentes qu'à la normale. D'ailleurs, l'emplacement du paramètre $a_i$ dans la fonction de probabilité du modèle 2PL (Birnbaum, 1968, p. 399-402) accentue la différence entre le niveau d'habileté de l'individu et la difficulté de l'item :

$$P(U_i = 1|\theta_j, a_i, b_i) = \frac{1}{1+e^{-Da_i(\theta_j - b_i)}} \tag{2.4}$$

Chaque item a son paramètre de discrimination propre et tout changement à la valeur de ce dernier se traduit par un changement de pente près du point d'inflexion de la courbe, là où la différenciation se fait le plus sentir. Un exemple d'items modélisés selon le modèle logistique à deux paramètres est présenté à la figure 2.3.

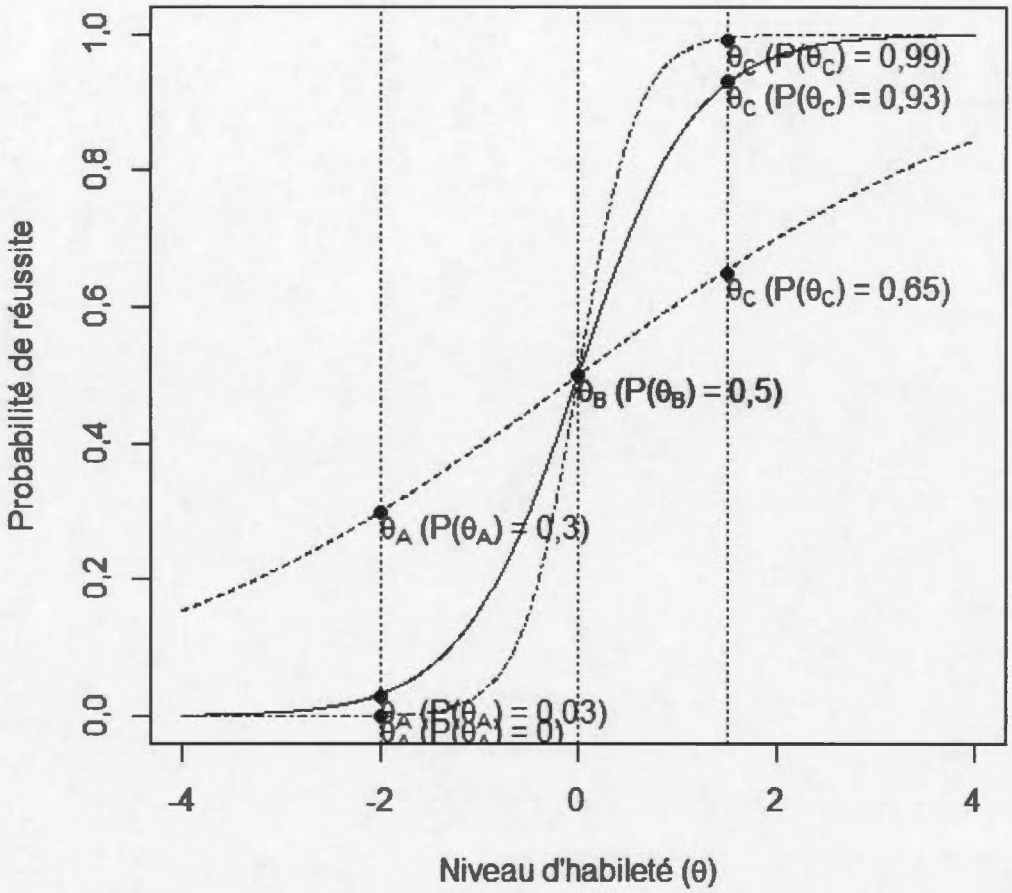


Figure 2.3 : Effets des variations du paramètre de discrimination sur la courbe caractéristique d'items modélisés selon le modèle logistique à deux paramètres ($b$ = 0, $a$ = [0,25, 1, 2])

Une pente abrupte dans la courbe caractéristique d'item traduit donc un fort pouvoir de discrimination de l'item en question, soit une plus grande différence dans les probabilités de réussite de deux individus de niveaux d'habileté différents. Par exemple, suivant la figure 2.3, des individus de niveaux d'habileté -2, 0 et 1,5 présentent des probabilités de réussite de 0,001, 0,5 et 0,99 à l'item dont la pente est très abrupte ($a$ = 2). Ces mêmes individus présentent des probabilités de réussite de 0,3, 0,5 et de 0,65 à un item tout aussi difficile ($b$ = 0), mais dont la pente est plus douce ($a$ = 0,25). La différence entre les probabilités de réussite de ces individus est

donc beaucoup plus importante lorsque la pente de la courbe caractéristique d'item est abrupte, soit lorsque le paramètre de discrimination $a_i$ d'un item est élevé. En ce sens, un item discriminant est un item qui peut différencier des individus de niveaux d'habileté différents en leur attribuant des probabilités de réussite différentes. Tuerlinckx et De Boeck (2005) proposent d'autres interprétations plus techniques à ce paramètre : le paramètre $a_i$ comme étant une quantité d'information à collecter avant qu'une réponse ne soit donnée audit item et le paramètre $a_i$ comme étant fonction d'une dépendance entre des parcelles d'information.

### 2.3.4 Modèle 3PL

Si la recherche démontre que toutes les courbes caractéristiques d'items d'un test ne sont pas toutes parallèles, elle démontre aussi que certaines courbes caractéristiques d'item ne débutent pas à $P(\theta) = 0$, même lorsque $\theta$ tend vers $-\infty$. Les items à choix multiples, par exemple, ont une propension à présenter pareilles courbes caractéristiques d'item de par leur conception. En effet, même un individu aussi faible qu'imaginable peut réussir un item à choix multiple en y répondant au hasard. Cette probabilité minimale de réussir l'item ne peut être bien représentée par les modèles logistiques à 1 ou 2 paramètres, mais elle peut être prise en compte dans un modèle logistique en généralisant le modèle 2PL. Un troisième paramètre d'item doit alors être considéré pour constituer le modèle logistique 3PL (Birnbaum, 1968, p. 404). L'ajout d'un paramètre de pseudo-chance $c_i$, qui correspond à l'asymptote inférieure de la courbe caractéristique d'item, est alors tout indiqué.

$$P(U_i = 1|\theta_j, a_i, b_i, c_i) = c_i + \frac{1-c_i}{1+e^{-Da_i(\theta_j - b_i)}} \tag{2.5}$$

La position du paramètre de pseudo-chance dans l'équation démontre bien que sa valeur est une probabilité en soi – la probabilité de réussite minimale – et qu'elle

s'ajoute à la probabilité de réussite qui intègre les deux autres paramètres d'item et l'habileté de l'individu. Tout changement à ce paramètre modifie donc la probabilité minimale de réussite – $P(\theta_{\infty})$ – comme le démontre la figure 2.4.

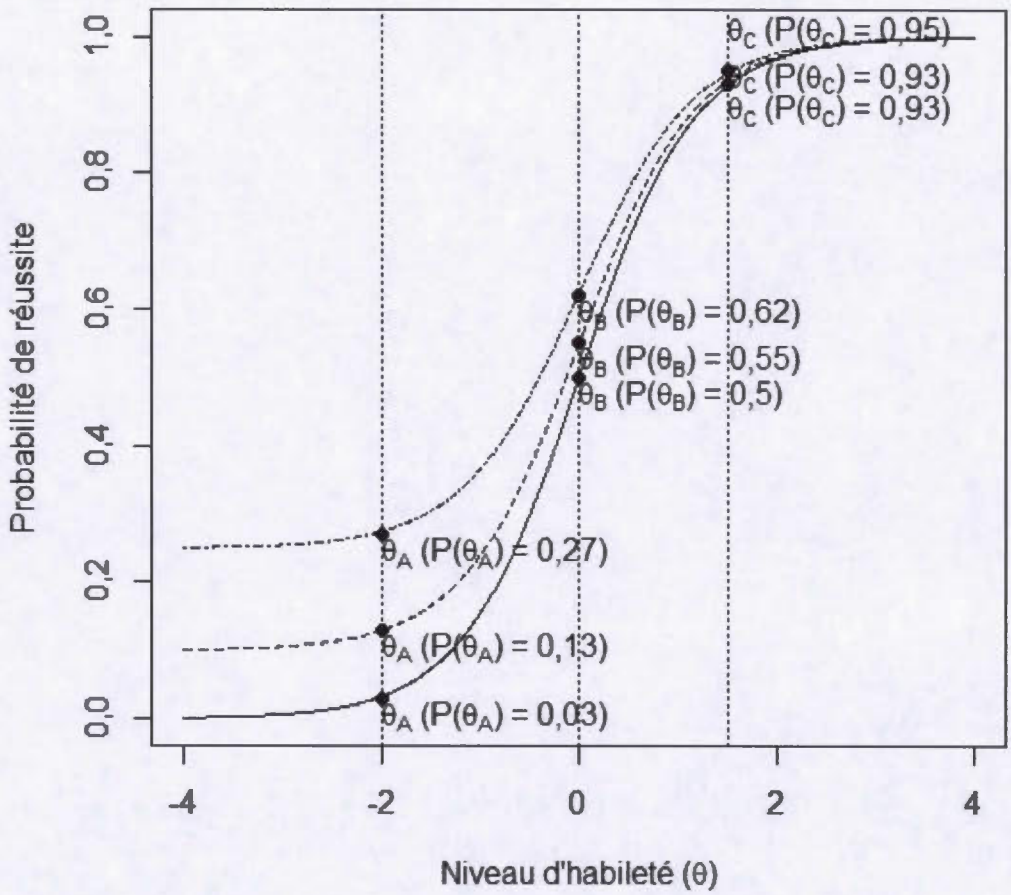


Figure 2.4 Effets des variations du paramètre de pseudo-chance sur la courbe caractéristique d'items selon le modèle logistique à trois paramètres ($b = 0$, $a = 1$, $c = [0, 0{,}1, 0{,}25]$)

Le paramètre de pseudo-chance n'a que très peu d'effet sur la probabilité de réussite d'un individu très fort. Après tout, un individu qui a déjà une forte probabilité de réussite à un item de par son niveau d'habileté (ex. $P(\theta) = 0{,}95$) aura-t-il une plus forte probabilité de réussite parce que l'item est à choix de réponse? Les individus

très forts ou pour lesquels un item est très facile n'ont normalement pas recours à la chance ou au hasard pour y répondre. Ce principe se traduit dans le modèle par une différence dans les probabilités de réussite plus importante chez les individus plus faibles parce qu'eux sont plus près de l'asymptote inférieure modifiée par $c$. Par exemple, la figure 2.4 montre bien que les probabilités de réussite à l'item pour l'individu $C$, plutôt fort ($\theta_C$ = 1,5), varient peu aux différentes valeurs de c (0, 0,1, 0,25), l'étendue de ces probabilités est de 0,02. Par contre, les probabilités de réussite de l'individu $A$, plutôt faible ($\theta_A$ = -2), varient beaucoup plus à ces mêmes valeurs de $c$, où l'étendue est de 0,24.

Il pourrait sembler approprié d'interpréter le paramètre de pseudo-chance comme une probabilité minimale de réussite établie selon un nombre de choix de réponses possibles. Cependant, ce paramètre est plus complexe qu'il ne parait, les données peuvent en attester autrement. En effet, sa valeur ne dépend pas directement du nombre d'options possibles, elle est plutôt estimée à partir des patrons de réponses d'un groupe. Ainsi, il se pourrait même qu'un item à réponse ouverte, donc sans choix de réponse, ait un paramètre de pseudo-chance élevé si l'estimation des paramètres d'item semble mieux ajustée ainsi, mais alors la signification dudit paramètre de pseudo-chance est encore plus difficile à interpréter. Toutefois, il est tout à fait possible que l'estimation des paramètres d'item donne à $c_i$ une valeur tout à fait représentative de sa constitution, par exemple une valeur de près de 0,2 pour un item à cinq choix de réponse (Reckase, 2009, p. 23-25).

## 2.4 Estimation des paramètres d'item

Il y a plusieurs façons d'estimer les paramètres d'item dans les différents modèles de régression logistique. Une des méthodes classiques, celle de l'estimation par maximum de vraisemblance conjointe (JML), consiste à procéder de façon itérative selon deux étapes alternées : d'ailleurs, certains lui préfèrent la dénomination

d'estimation par maximum de vraisemblance alternée. Il s'agit d'abord de fixer les valeurs des paramètres de personne ($\theta_J$) à l'aide des données observées. Généralement, on effectue cette première approximation en calculant un score $z$ associé au pourcentage de bonnes réponses à chacune des personnes. Dans une seconde étape, ayant en main une première approximation des paramètres de personnes, les paramètres d'item ($a_I$, $b_I$ et $c_I$) sont calculés par maximum de vraisemblance. Les valeurs des paramètres ainsi obtenus, il s'agit de revenir à l'étape 1 pour calculer les paramètres de personnes par maximum de vraisemblance. Les paramètres sont estimés ainsi les uns à partir des autres jusqu'à l'obtention d'une précision satisfaisante, lorsque les valeurs des paramètres d'item et de personne ne varient plus de façon significative. Hambleton et Swaminathan (1985, p. 125-149) ont bien documenté ce procédé. Baker et Kim (2004, p. 84-92) proposent une autre méthode, le maximum de vraisemblance marginal (MML), où seuls les paramètres d'items sont estimés en supposant une distribution donnée des niveaux d'habileté.

## 2.5 Estimation des paramètres de personne

Lorsque les paramètres d'items sont connus et qu'un niveau d'habileté doit être estimé à partir d'un patron de réponses à ces items, plusieurs méthodes peuvent être utilisées. Toutes ces méthodes s'intéressent à expliquer la probabilité d'observer un patron de réponses $u$. Un postulat de base des modèles à un, deux et trois paramètres est que, lorsque le niveau d'habileté du répondant est fixé, la probabilité d'une réponse à un item est indépendante de la probabilité de réponse à un autre item du même test. C'est ce qu'on nomme l'indépendance locale : locale, car cette indépendance n'est valide que lorsque le niveau d'habileté est fixé. Cette probabilité se traduit par l'équation ci-dessous, soit la fonction de vraisemblance.

$$L(u_1, u_2, \dots, u_n|\theta) = \prod_{i=1}^{n} P_i^{u_i}(\theta) \cdot Q_i^{1-u_i}(\theta) \qquad (2.6)$$

Dans cette équation, $P_i$ correspond à la probabilité de réussite à l'item $i$ et $Q_i$ correspond à $1 - P_i$, soit à la probabilité de non-réussite à cet item. Le logarithme naturel de la fonction de vraisemblance est souvent utilisé, car sous cette forme c'est la somme des composantes qui est utilisée plutôt que leur produit, ce qui simplifie les calculs :

$$\ln L(u_1, u_2, \dots, u_n|\theta) = \sum_{i=1}^{n}[u_i \ln P_i(\theta) + (1 - u_i) \ln Q_i(\theta))] \quad (2.7)$$

Pour trouver le $\theta$ qui maximise cette fonction, la procédure itérative de Newton-Raphson est habituellement utilisée. Cette dernière utilise le rapport de la dérivée première de cette fonction sur sa dérivée seconde pour corriger l'estimateur jusqu'à ce que ce rapport soit minime. Comme la dérivée première de la fonction de log-vraisemblance correspond à la pente de la tangente de cette fonction à un point donné, le point le plus élevé de cette courbe devrait afficher une pente nulle et, donc, un rapport nul ($(\ln L)' = 0$). Le $\theta$ à cet emplacement de la fonction constitue l'estimateur du maximum de vraisemblance.

La figure 2.5 présente les courbes de log-vraisemblance de deux individus ayant répondu à 3 items modélisés selon le modèle logistique à 3 paramètres ($a$ = [1,5, 1,2, 1,1], $b$ = [1, 0,4, 1,7], $c$ = [0,03, 0,1, 0,18]).

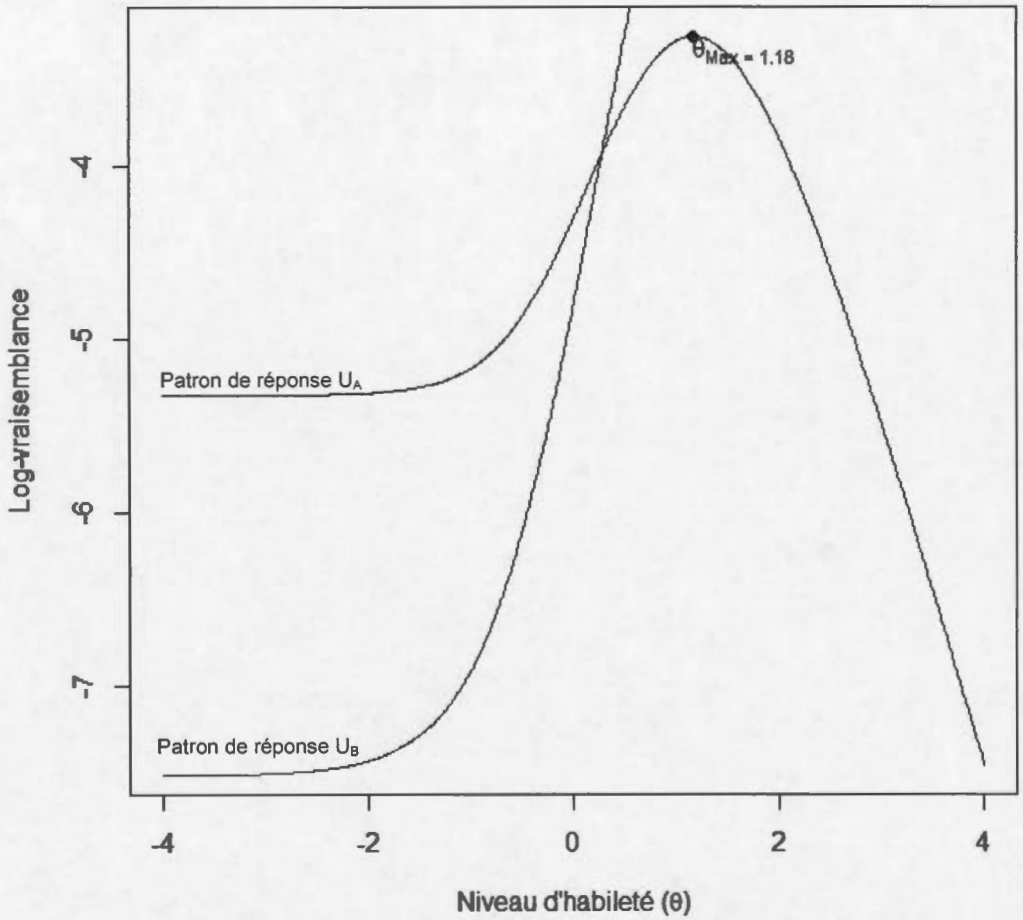


Figure 2.5 : Courbe de log-vraisemblance après 3 items pour les individus A et B ($\theta_A = \theta_B = 1,5$)

Dans la figure 2.5, l'estimateur du maximum de vraisemblance du patron de réponse [1, 0, 1] de l'individu A se trouve à $\hat{\theta}_A = 1,18$. La procédure de Newton-Raphson l'a ciblé en escaladant la fonction peu à peu, vérifiant à chaque itération la valeur de la pente jusqu'à ce que les variations soient pratiquement nulles. Malgré la faible quantité d'items, l'estimateur du maximum de vraisemblance de cette fonction ($\hat{\theta}_A = 1,18$) est plutôt près du niveau d'habileté réel de l'individu $A$ ($\theta_A = 1,5$). L'estimateur du maximum de vraisemblance du patron de réponse de l'individu B est, par contre, introuvable. En effet, en présence d'un patron de réponse dit parfait – ici

[1, 1, 1] – la courbe de log-vraisemblance tend vers l'infini et le mode de cette fonction, qui constitue l'estimateur du maximum de vraisemblance, est alors incalculable ($\theta_B = +\infty$). Il peut même arriver qu'une fonction affiche deux modes lorsque le nombre d'items administrés est trop faible et que les réponses à ces items défient les attentes (Samejima, 1973, p. 223-225). Toute une gamme d'adaptations est alors possible.

Malgré ses quelques limites, l'estimateur du maximum de vraisemblance est l'un des estimateurs les plus utilisés. Cependant, plusieurs autres estimateurs peuvent être utilisés. Ces derniers peuvent être moins biaisés, ainsi plus précis – surtout quand peu d'items ont été administrés – ou simplement mieux s'ajuster à une réalité. Par exemple, suivant une voie bayésienne, l'estimation du maximum *a posteriori* (MAP) consiste à estimer le mode d'une distribution *a posteriori* de $\theta$ après avoir modulé le la fonction de vraisemblance de $\theta$ par la fonction de densité *a priori* de $\theta$ (Swaminathan et Gifford, 1982, 1985 et 1986), soit dans une fonction :

$$\hat{\theta}_{MAP} = L(u_1, u_2, \ldots, u_n|\theta) \cdot g(\theta) \tag{2.8}$$

où $L(u_1,u_2,\ldots, u_n|\theta)$ correspond à la fonction de vraisemblance et $g(\theta)$ à la distribution *a priori* de $\theta$. Comme il considère la densité *a priori*, cet estimateur a tendance à être moins polarisé, moins extrême après quelques items, en revanche, il est plus biaisé que l'estimateur du maximum de vraisemblance parce qu'il tend $\theta$ vers des valeurs plus centrales (Bock, 1997, p. 25). Toujours sous la bannière bayésienne, la moyenne de cette même distribution peut aussi être utilisée comme estimé de $\theta$ et il s'appelle alors l'estimateur de l'espérance *a posteriori* (EAP) et s'exprime ainsi :

$$\hat{\theta}_{EAP} = \frac{\sum_{k=1}^{K} \theta_k \cdot L(u_1,u_2,\ldots,u_n|\theta_k) \cdot g(\theta)}{\sum_{k=1}^{K} L(u_1,u_2,\ldots,u_n|\theta) \cdot g(\theta)} \tag{2.9}$$

où k représente un des K points de quadrature sur l'étendue de $\theta$, $L(u_1,u_2,\ldots, u_n|\theta)$ correspond à la fonction de vraisemblance et $g(\theta)$ à la distribution *a priori* de $\theta$ (Chen, Hou et Dodd, 1998, p. 575-576). Il s'agit d'une approximation d'une intégrale (Wang, 1997, p. 6-7).

Plus récemment, Warm a proposé une variation du maximum de vraisemblance *a posteriori* où la fonction de densité *a priori* est remplacée par une pondération du niveau d'habileté : soit le maximum de vraisemblance pondéré (Warm, 1989). Cette pondération vise à atténuer le biais de premier ordre dans l'estimation de $\theta$. Pour ce faire, Warm a isolé le biais générique de l'expression mathématique d'un estimateur pondéré et a assemblé une fonction qui, en pondérant la fonction de vraisemblance, élimine asymptomatiquement ce biais. L'expression mathématique de cet estimateur ressemble en tous points à l'estimateur du maximum *a posteriori* (MAP) :

$$\hat{\theta}_{WL} = L(u_1, u_2, \ldots, u_n|\theta) \cdot f(\theta), \qquad (2.10)$$

mais $f(\theta)$ correspond maintenant à une fonction qui élimine le biais de premier ordre. Lorsque les modélisations logistiques à un ou deux paramètres sont utilisées, $f(\theta)$ est proportionnel à la racine carrée de la fonction d'information de Fischer (équation 2.11 de ce mémoire) (Magis, 2015; Magis et Raîche, 2012), soit l'*a priori* non informatif de Jeffrey (1946). Les valeurs du niveau d'habileté les plus informatives au sens de Fisher se voient alors appliquer un poids plus importants dans l'estimation. Le maximum de vraisemblance pondéré présente l'avantage de fournir un estimateur du niveau d'habileté beaucoup moins biaisé que les autres estimateurs lorsque le nombre d'items administrés est faible. Son erreur type se calcule comme celle de l'estimateur par maximum de vraisemblance. C'est pourquoi ce sera cet estimateur qui sera retenu plus loin à la section portant sur la méthodologie. La littérature abonde de nouvelles techniques d'estimation développées pour pallier les problèmes des estimateurs

actuellement utilisés. Considérant la complexité des calculs impliqués, nous ne présentons pas ici le détail des calculs.

## 2.6 Information et précision

Dans la théorie de la réponse à l'item, la précision de l'estimateur se mesure à l'aide de l'information, un concept introduit par Fisher en 1950. Pour distinguer cette quantité d'autres statistiques similaires, elle est plus précisément souvent dénommée, l'information au sens de Fischer. L'idée est que plus on détient d'information sur une mesure, plus cette mesure est précise. La quantité d'information que fournit chaque item dépend de son ajustement à l'individu. Utilisant le modèle logistique à 3 paramètres ainsi qu'une estimation par maximum de vraisemblance ou par maximum de vraisemblance pondérée, la fonction d'information d'item est calculée par l'équation suivante :

$$I_i(\theta) = D^2 a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left(\frac{P_i(\theta)-c_i}{(1-c_i)}\right)^2 \tag{2.11}$$

La fonction d'information présente l'avantage d'être cumulative et ainsi l'information totale à un test *I* correspond à la somme de l'information calculée à chacun des items. La racine carrée de l'inverse de l'information cumulée de ces items donne la précision de l'estimateur, soit son erreur type :

$$S_\theta = \sqrt{I(\theta)}^{-1} = \sqrt{\sum_{i=1}^{n} I_i(\theta)}^{-1} \tag{2.12}$$

grâce à laquelle un intervalle de confiance peut être dressé autour de θ. Il est à noter que l'erreur-type se calcule ainsi pour les estimateurs du maximum de vraisemblance et du maximum de vraisemblance pondéré mais que les autres estimateurs, bayésien par exemple, ont leurs méthodes propres.

## 2.7 Fonctionnement du testing adaptatif

Le testing adaptatif est une pratique millénaire d'abord employée pour sélectionner l'entourage de l'empereur en Chine antique. Ces tests mobilisaient alors un « maître » par individu pour la passation d'un test, et ce maître choisissait parmi différentes questions la plus appropriée à administrer jusqu'à ce que l'individu prouve sa valeur (Wainer, 1990, p. 2). Avec l'avènement des micro-ordinateurs, ce maître est maintenant émulé à travers différents procédés psychométriques. Ce faisant, un test adaptatif peut maintenant être administré à de nombreux groupes simultanément sans mobiliser autant de « maîtres » qu'il y a d'individus. Les seules limites à son utilisation sont d'ordre technologique.

Dans un test adaptatif, un individu se voit administré chaque item selon ses réponses aux items précédents. Pour ce faire, le test octroie à l'individu un niveau d'habileté initial, ce qui constitue la règle de départ. Ensuite, le test recherche un item à administrer à l'individu et l'algorithme de recherche constitue la règle de sélection du prochain item. Enfin, le test estime le niveau d'habileté de l'individu à partir des réponses collectées aux items déjà administrés et vérifie si le test doit prendre fin, ce qui constitue la règle de fin. Ces trois règles seront approfondies dans les sections suivantes.

### 2.7.1 Règle de départ d'un test adaptatif

Un test adaptatif développé autour des principes de la théorie de la réponse à l'item et administré par ordinateur est régi par trois règles distinctes. Tout d'abord, la règle de départ sert à déterminer sur quelle base le premier item du test est choisi. Wainer (1990, p. 109) avance qu'en absence de renseignements sur l'individu, lui supposer un niveau d'habileté moyen ($\theta = 0$) est tout à fait raisonnable, compte tenu que les

probabilités qu'il soit d'un niveau extrême sont faibles, d'autant plus qu'après quelques items le test se sera recadré près du niveau réel de l'individu. C'est d'ailleurs la pratique qui semble la plus répandue. D'autres chercheurs proposent de considérer d'autres renseignements sur l'individu – le groupe d'appartenance par exemple – qui permettraient de lui supposer un niveau d'habileté plus précis et de lui administrer dès le départ des items ajustés à son habileté. Enfin, Wainer et Lewis (1989, p. 11-12) proposent l'utilisation de mini-tests (*testlets*), soient de petits ensembles d'items administrés dès le départ les uns après les autres, mais les mêmes pour tous les répondants, afin de contrer quelques problèmes liés à l'utilisation d'algorithmes adaptatifs.

### 2.7.2 Règle de sélection du prochain item dans un test adaptatif

Une deuxième règle sert à déterminer le prochain item qui sera sélectionné selon le dernier estimé du niveau d'habileté de l'individu[4]. Ce choix peut être fait selon différents critères. D'abord, lorsque l'information maximale de Fisher (MFI) est utilisée comme critère de sélection, l'information que chaque item peut fournir à $\hat{\theta}$ est calculée et l'item non encore administré pour lequel l'information est maximale est choisi (Brown et Weiss, 1977, p. 8-9). Techniquement, il faut trouver l'item qui maximise l'équation 2.10 :

$$I(\theta)_{MFI} = argmax[D^2 a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \frac{P_i(\theta)-c_i}{(1-c_i)}] \qquad (2.13)$$

Dans les modèles logistiques qui considèrent le paramètre de discrimination $a_i$, l'utilisation de ce critère peut engendrer certains problèmes. En effet, comme le paramètre de discrimination est très déterminant dans le calcul de l'information, alors

[4] Différents estimateurs peuvent être utilisés pour estimer provisoirement l'habileté d'un individu. La section 2.5 liste quelques estimateurs fréquemment utilisés.

des items peu ajustés, mais fortement discriminants peuvent être sélectionnés et mener à des estimations provisoires de $\theta$ erronées, distantes de la valeur réelle de $\theta$ (Davey et Parshall, 1995, p. 3-4). C'est pour cette raison, mais aussi pour éviter la surexposition de ces items hautement discriminants que Hau et Chang (2001) suggèrent d'administrer des items peu discriminants en début de test, alors que $\theta$ n'est pas très bien circonscrit. Suivant la méthode $a$-stratifiée (A-STR), les items d'une banque doivent être regroupés selon leur paramètre de discrimination $a$. Les items d'un test sont d'abord puisés[5] dans le premier bassin contenant les items les moins discriminants. Ils sont ensuite puisés dans des bassins contenant des items avec des valeurs de $a$ de plus en plus élevées, pour terminer avec les items les plus discriminants, donc les plus déterminants alors que $\theta$ est beaucoup mieux défini. Suivant cette méthode, l'exposition des items est mieux encadrée et la mesure n'en est pas moins fiable. (Hau et Chang, 2001, p. 253-265)

Il est aussi possible d'utiliser un critère qui, au lieu de maximiser l'information, tente plutôt de diminuer la variance de l'estimateur après l'administration de l'item, soit le critère visant la minimisation de l'espérance de la variance *a posteriori* (MEPV). Pour ce faire, la procédure calcule l'espérance de la variance de l'estimateur $\hat{\theta}_{EAP}$ pour chaque item de la banque et choisit l'item qui la garde au minimum. Cette variance se calcule ainsi :

$$EPV_i = P_i(\hat{\theta}_j) \cdot Var(\theta|u_1, u_2, \dots, u_k, 1) + Q_i(\hat{\theta}_j) \cdot Var(\theta|u_1, u_2, \dots, u_k, 0) \qquad (2.14)$$

où k représente le nombre d'items administrés. La variance a posteriori est calculée à partir du patron des réponses aux $k$ items réellement administrés et des deux réponses possibles à l'item $i$, non encore administré, soit 1 (succès) ou 0 (échec). L'item qui

[5] Le choix de l'item dans le bassin se fait avec le critère de l'information maximale de Fisher.

minimise cette variance est celui qui est sélectionné en choisissant ce critère pour la sélection du prochain item (Magis, Raîche et Barrada, 2015).

Plusieurs autres critères peuvent être sélectionnés : l'approche ancillaire d'Urry, l'information par le maximum de vraisemblance pondéré, etc. Magis, Raîche et Barrada ont incorporé plusieurs de ces critères à leur librairie *catR* et les décrivent tous dans la documentation associée.

### 2.7.3 Règle de fin d'un test adaptatif

La règle d'arrêt détermine comment prend fin le test adaptatif. D'abord, il est possible de spécifier préalablement une certaine précision désirée et alors le test prend fin lorsque l'information cumulée des items administrés rencontre cette précision demandée. En temps normal, une bonne précision sera atteinte plus rapidement dans un test adaptatif parce que tous les items sont mieux ajustés au niveau d'habileté de l'individu. Ainsi, sur la base de ce critère d'arrêt, un test adaptatif devrait s'arrêter significativement plus rapidement qu'un test papier-crayon de longueur fixe. D'ailleurs, il est aussi possible de spécifier un nombre maximal d'items après quoi le test prend fin. Enfin, il est aussi possible d'arrêter le test lorsque les items administrés ne sont plus assez informatifs et n'améliorent plus la précision de l'estimateur du niveau d'habileté. Il est aussi possible, même plutôt commun, de combiner ces règles afin d'obtenir un estimateur précis sans gaspiller de temps. Raîche et Blais (2001) ont étudié les impacts de la variation de ces règles sur différentes statistiques liées à la passation d'un test adaptatif. Ils ont noté que l'administration d'environ 40 items est souvent suffisante pour obtenir une précision adéquate dans un test adaptatif.

Lorsque la règle d'arrêt est satisfaite, le dernier estimé du niveau d'habileté de l'individu est retenu[6] et l'information cumulée sert à calculer l'erreur de mesure de cet estimé. Le parcours de cet individu ne devrait ressembler en rien à celui d'un autre vu le caractère adaptatif de ce type de test. En effet, puisque les niveaux d'habileté de deux individus diffèrent, ce ne sont pas les mêmes items qui leur sont administrés alors ce n'est pas la même quantité d'information que cumule chaque passation. C'est d'ailleurs pourquoi chaque estimé a sa propre précision, sa propre erreur de mesure.

## 2.8 Avantages et limites du testing adaptatif

L'utilisation du testing adaptatif comporte un lot d'avantages intéressants que le testing classique ne peut égaler, en plus de ceux qu'il récupère de la théorie de la réponse à l'item qui en constitue les rouages. Notamment, il élimine les items trop faciles ou trop difficiles, inutiles dans l'estimation d'un niveau d'habileté. Ensuite, il permet d'établir un seuil d'information minimale à atteindre et d'en faire une règle d'arrêt du test. Ce faisant, il diminue le temps des passations, parce que moins d'items sont nécessaires – s'ils sont ajustés à l'individu – pour atteindre ce seuil. Il rend le plagiat beaucoup plus difficile parce que les individus ne se font pas tous administrer les mêmes items en même temps. De ce fait, davantage d'items sont exposés, mais pas aussi largement alors il devient plus difficile de préparer un individu à toute éventualité quand ces éventualités sont multiples. De plus, le testing adaptatif informatisé permet de recueillir d'autres types de traces : temps de réponse, utilisation d'outils, etc. Aussi, il fournit à l'individu une rétroaction immédiate, ce qui signifie que l'individu peut quitter en connaissant le score estimé pour lui par le système, sans aucun intermédiaire subjectif et sans aucun délai de correction. Enfin, il permet la comparaison de scores entre différentes versions d'un test alimentées par la

[6] Différents estimateurs peuvent être utilisés pour estimer officiellement l'habileté d'un individu. La section 2.5 liste quelques estimateurs fréquemment utilisés.

même banque. Dans la théorie classique des tests, c'est le nombre d'items réussis qui constitue le score, car tous reçoivent la même épreuve. Toutefois, dans un modèle qui permet l'administration d'items différents à des individus différents, cette information n'est plus pertinente parce qu'en temps normal tous les individus devraient réussir 50 % des items qui leur sont administrés. Alors, les scores estimés sont plutôt projetés sur une échelle qui permet ce genre de comparaisons.

Certes, le testing adaptatif comporte une foule d'avantages intéressants et surpasse dans maints domaines d'autres outils voués à l'évaluation. Cependant, il comporte son lot d'inconvénients. Martin (2003) dresse un bilan plutôt décevant du développement du testing adaptatif depuis ses balbutiements jusqu'en 2003, parce que sa mise en place est extrêmement complexe et que certains des avantages qui lui sont reconnus ne sont pas aussi scintillants qu'ils ne le paraissent. D'abord, les coûts sont exorbitants : équipements, comités d'experts pour le développement d'items, calibration des items avec des groupes-test, etc. sont autant de dépenses que peu d'organisations peuvent se permettre. Ensuite, la sécurité des items n'est pas autant assurée qu'on pourrait le croire. En effet, certains items de la banque sont souvent surexposés parce qu'ils fournissent beaucoup d'information sur un large spectre de $\theta$, Wainer (2000) avance même que 15 % à 20 % des items d'une banque constituent ensemble 50 % du lot d'items régulièrement administré. Alors, l'utilisation du testing adaptatif dans un contexte d'évaluation certificative ou à enjeux importants devient plus risquée. Malgré le portrait plutôt sombre qu'en dresse Martin en 2003, le testing adaptatif demeure très prometteur avec des coûts technologiques toujours à la baisse, une efficacité toujours à la hausse et de nouvelles techniques et approches pour prévenir, tempérer ou contrôler les difficultés rencontrées.

## 2.9 Banques d'items

La mise sur pied d'un test adaptatif, et surtout du système qui le supporte, est considérablement plus complexe et dispendieuse que la mise sur pied d'un test papier-crayon. Certes, l'achat et l'entretien du support technologique entraînent de nombreux coûts. Cependant, c'est l'élaboration et l'entretien de la banque d'items qui constituent le travail le plus ardu. En effet, le testing adaptatif demande une banque d'items volumineuse et bien calibrée afin d'être ajustée à des individus de tous niveaux d'habileté. Autrement dit, la banque d'items doit répondre aussi bien à un individu moyen qu'à un individu fort ou faible sans quoi les principaux avantages du testing adaptatif sont écartés. Le testing adaptatif, aussi précis et efficace qu'il puisse être, n'est valable que si la banque d'items est bien entretenue, régulièrement et avec assiduité. Sinon, elle génèrera des problèmes de tous genres qui, immanquablement, rendront stérile toute tentative de testing adaptatif à partir de cette banque (Bjorner, Kosinski et Ware, 2005, p. 107-110).

### 2.9.1 Dimensionnalité de la banque d'items

D'abord, la dimensionnalité de la banque d'items peut être problématique si l'administration de ses items à des individus fait émerger une structure latente multidimensionnelle et qu'aucune action n'est alors entreprise. Par structure multidimensionnelle, il faut entendre un construit formé de plus d'un trait latent, comme une série d'items en mathématiques formulés en situations problèmes qui ne mesurent pas qu'une habileté mathématique, mais aussi la compréhension de lecture des individus de par la formulation complexe de leurs intitulés. L'analyse factorielle, ou un procédé de la même famille, est utilisée pour détecter pareilles situations et pour mesurer l'influence de chaque trait latent sur les données observées. En conduisant régulièrement pareilles analyses sur les ensembles de données issus de passations de tests adaptatifs, les problèmes peuvent être détectés rapidement et les

décisions prises afin de respecter le postulat d'unidimensionnalité propre à la théorie de la réponse à l'item. Notamment, il pourrait être décidé de retirer les items apparemment problématiques de la banque ou d'en faire analyser le contenu par des experts (Anzaldua, 2002, p. 11).

### 2.9.2 Fonctionnement différentiel d'items

Aussi, certains items peuvent systématiquement pousser certains individus d'un même $\theta$, mais de groupes différents à répondre différemment de ce qui est normalement attendu d'eux. Ces items fonctionnent différemment des autres, car ils semblent mesurer d'autres caractéristiques que $\theta$; leur paramètre de difficulté semble modulé par une ou des sources de biais (Magis, De Boeck et Raîche, 2011, p. 31-32). Ce phénomène s'appelle le fonctionnement différentiel d'item et différentes méthodes existent pour en faire l'analyse. Ainsi, certains items peuvent voir les femmes réussir mieux que les hommes ou alors les Européens mieux que les Américains, au-delà des niveaux d'habileté des individus. La formulation de certains items peut être en cause, notamment par l'utilisation d'analogies (Schmitt, Holland et Dorans, 1992, p. 5-6). Les repères culturels auxquels font allusion ces items peuvent aussi être en cause (Huang, Church et Katigbak, 1997, p. 192-194). L'attitude des individus face à la technologie, leur aisance avec la technologie et leurs expériences avec la technologie peuvent avoir un impact sur leur performance à test reposant grandement sur la technologie (Sutton, 1993, p. 3). Enfin, des facteurs socioéconomiques peuvent aussi expliquer une part de ce fonctionnement différentiel, là où certains individus ont accès ou non à certaines ressources (Zwick, Donoghue et Grima, 1993, p. 233). Il est important de considérer ces items dysfonctionnels comme une menace à l'intégrité des données et de prendre les décisions qui s'imposent afin de conserver la banque d'items bien équilibrée et bien calibrée, surtout que le testing adaptatif a tendance à utiliser moins d'items, ce qui rend chacun d'eux plus important dans l'estimation du

niveau d'habileté (Zwick, Thayer et Wingersky, 1994, p. 121). Sinon, l'interprétation des scores en est faussée alors que certains individus sont évalués à la baisse.

### 2.9.3 Contrôle de l'exposition des items

Enfin, la surexposition des items de la banque est à considérer sans quoi elle facilite la tricherie, l'échange d'information. En effet, plus un item est administré de fois, plus son contenu est connu et peut être communiqué à d'autres. Par conséquent, il importe d'implanter des procédures qui contrôlent l'exposition des items, variant les items à administrer pour des individus semblables, au prix de légères pertes en information.

### 2.9.4 Constitution de la banque d'items et asymétrie

Des recherches se sont intéressées à des scénarios où les tests et les items qui les composent doivent répondre à un certain nombre de contraintes en plus de chercher à maximiser leur teneur en information. Ces contraintes peuvent être de différents ordres : des contenus spécifiques à couvrir, des items mutuellement exclusifs, des items interdépendants, une longueur prescrite, une limite de temps, etc. Si une banque d'items vouée à un test sous contraintes *(constrained test)* semble bien équilibrée, la complexité que constitue la recherche d'un item qui maximise l'information tout en satisfaisant des conditions particulières supplémentaires peut brouiller cet équilibre. En effet, il est possible qu'une banque satisfasse les demandes en information mais qu'elle soit incapable de satisfaire une, plusieurs, voire toutes les contraintes supplémentaires[7]. Dans pareille situation, la constitution de la banque serait en cause, mais pour des raisons différentes. Van der Linden (2000, p. 27-52) a mis de l'avant une approche dans l'assemblage de tests sous contraintes. Il propose l'utilisation de

[7] Voir les travaux de Timminga (1998, p. 280-291).

tests fictifs *(shadow tests)* dans la sélection du prochain item. Ces tests sont constitués de tous les items ajustés à $\hat{\theta}$ qui satisfont à toutes les contraintes du test. L'item disposant du plus d'information dans ce test fictif est sélectionné et administré à l'individu. Luecht (1998, p. 224-236), lui, propose un algorithme très complexe pour générer différentes versions d'un test à partir d'une banque d'items tout en intégrant un nombre colossal de contraintes de différents ordres. Cet algorithme s'appelle le *Normalized Weighted Absolute Deviation Heuristic* (NWADH) et il est basé, notamment, sur les travaux de Stocking et Swanson (1993, p. 277-292).

Toutefois, pour la présente recherche, la constitution de la banque d'items renvoie à la symétrie ou l'asymétrie dans la distribution des paramètres de difficulté de ses items. Lorsque mal développée, mal entretenue ou après avoir été dépouillée de ses items, une banque d'items peut se retrouver dans un piteux état. Elle pourrait ne plus être en mesure de bien répondre à des individus de tous niveaux parce que, par exemple, elle ne contient plus assez d'items d'un degré de difficulté donné. Un déséquilibre dans la distribution des paramètres d'items peut alors apparaître et se caractériser par différentes formes de distribution des paramètres d'items : celles-ci affichant alors des valeurs importantes de l'asymétrie ou de la kurtose. C'est généralement le paramètre de difficulté qui est le plus important à contrôler pour s'assurer que celui-ci couvre bien l'étendue des niveaux d'habileté à estimer alors l'asymétrie de sa distribution pourrait être importante à contrôler.

L'asymétrie au sein d'une distribution se caractérise par un manque de symétrie, soit par l'allongement de l'une de ses queues de distribution. Lorsque le coefficient d'asymétrie $a^3$ est positif, alors la queue de distribution de droite est allongée et lorsque le coefficient d'asymétrie $a^3$ est négatif alors la queue de distribution de gauche est allongée, comme le montre la figure 2.6.

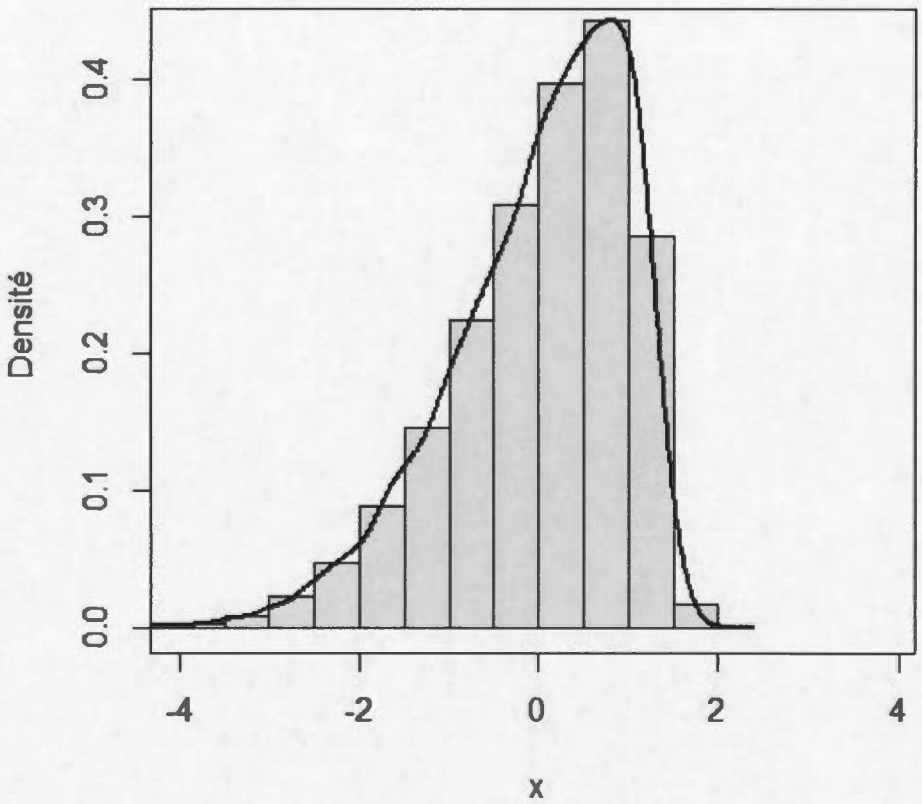


Figure 2.6 : Distribution asymétrique négative fictive

Une banque d'items dont la distribution des paramètres de difficulté est asymétrique n'a donc pas la même proportion d'items faciles, moyens et difficiles qu'une banque dite symétrique. Dans certaines situations, cette asymétrie peut être avantageuse. Par exemple, il pourrait être avantageux d'avoir davantage d'items difficiles dans une banque si des individus très forts doivent être départagés par un test adaptatif. Toutefois, cette asymétrie est rarement recherchée, un équilibre est souvent de mise.

## 2.10 Objectif spécifique

Considérant ce qui précède, il serait ainsi intéressant de vérifier en quoi la passation d'un test adaptatif est affectée par cette mauvaise constitution de la banque d'items

qui le nourrit. Plus précisément, et c'est l'objectif spécifique de cette recherche, nous désirons étudier les effets de l'asymétrie de la distribution du paramètre de difficulté des items d'une banque sur différentes variables liées à la passation d'un test adaptatif alimenté par cette banque : biais d'estimation, précision, durée du test, etc.

Cette recherche veut donc vérifier l'impact de différents degrés d'asymétrie dans les paramètres de difficulté des items sur le biais d'estimation du niveau d'habileté, l'erreur-type du niveau d'habileté et, enfin, la longueur des tests. Le prochain chapitre détaillera la méthodologie nécessaire pour atteindre cet objectif.

# CHAPITRE III

# MÉTHODOLOGIE

## 3.1 Simulation

14 banques d'items seront générées, selon deux combinaisons possibles de la taille de la banque d'items ($n_b$ = 200 ou 1000) et sept valeurs du coefficient d'asymétrie de la distribution de probabilité du paramètre de difficulté $b_i$ des items de la banque ($a^3$ = [-6, -3, -1, 0, 1, 3, 6]). Pour chacune de ces 14 conditions, 9000 unités d'observation seront produites, soient 1000 à chacun des neuf niveaux d'habileté réels suivants : $\theta$ = [-4, -3, -2, -1, 0, 1, 2, 3, 4]. Ces unités d'observation correspondront à des sujets simulés à qui seraient administrés différents tests adaptatifs. Ainsi, 126 000 patrons de réponses (9*1000*14) seront produits au hasard.

Le modèle logistique à 3 paramètres sera utilisé pour générer les données simulées. Préalablement, le paramètre de difficulté des items de chacune des banques d'items sera généré au hasard à l'aide de la fonction `rsstd()` de la librairie `fGarch` du logiciel R qui permet la génération de variables suivant une distribution student *t* normale asymétrique (Wuertz et Chalabi, 2013). Cette fonction reçoit cinq paramètres : un nombre d'observations *n* à générer, la moyenne *mu*, l'écart-type *sd*, un paramètre de forme *nu* ainsi qu'un coefficient d'asymétrie *xi*.

La fonction génère d'abord une distribution uniforme de *n* valeurs dans un intervalle autour de 0, de longueur 1 et délimité selon *xi* : plus à gauche si *xi* est fortement négatif et plus à droite si *xi* est fortement positif. Cet ensemble sert à en générer un autre où chaque valeur vaut $xi^1$ ou $xi^{-1}$, selon le signe de chaque valeur dans le vecteur

de départ. Le coefficient d'asymétrie dans ce vecteur correspond approximativement, à cette étape, au paramètre *xi*. Toutefois, les valeurs ne prennent que deux modalités et la moyenne de l'ensemble ne vaut pas zéro comme le prescrit une distribution student *t* normale asymétrique. Alors, des valeurs aléatoires tirées d'une distribution normale student t avec *nu* degrés de liberté sont générées dans un troisième vecteur. Ce vecteur est divisé par le précédent afin que l'asymétrie soit conservée. Cependant, la moyenne au sein de ce vecteur est encore loin de 0, alors les valeurs sont ramenées vers le centre tout en conservant un certain équilibre de sorte que l'asymétrie soit préservée au maximum. Par contre, l'asymétrie est souvent atténuée par les nombreuses modulations des valeurs générées. Alors, il faut choisir des valeurs de *xi* très extrêmes pour obtenir une très forte asymétrie.

Pour obtenir des distributions dont l'asymétrie affiche des valeurs de -6, -3, -1, 0, 1, 3 et 6, des simulations avec les paramètres de cette distribution ont été effectuées. Il s'est avéré que pour la banque de 1000 items les valeurs cibles de l'asymétrie correspondaient aux valeurs 2,05534, 4,20715, 9, 5, 9, 4.20715 et 2,05534 du paramètre *nu* et aux valeurs -5000, -100, -1.5, 1,5, 1,5, 100 et 5000 du paramètre *xi*. Pour la banque de 200 items, ces valeurs étaient respectivement de 2,0202, 4,1163325, 5,06, 5, 5,06, 4,1163325 et 2,0202 pour *nu* et de 5000, -2, 1, 1,5, -1, 2 et 5000 pour *xi*. Les paramètres de discrimination $a_i$ et de pseudo-chance $c_i$, pour leur part, seront distribués respectivement selon une loi log-normale (moyenne de 0 et écart-type de 0.1225) et une loi bêta (*a* égal à 1 et *b* égal à 12). Les valeurs de ces paramètres seront les mêmes pour toutes les banques de même taille. La loi log-normale tient compte du fait que le paramètre de discrimination est strictement positif (Baker et Kim, 2004, p. 186-187), tandis que la loi bêta tient compte du fait que le paramètre de pseudo-chance est borné par les valeurs 0 et 1 (Baker et Kim, 2004, p. 188-189).

Les patrons de réponses seront générés à l'aide de la fonction `simulateRespondents` de la librairie `catR` du logiciel R (Magis, Raîche et Barrada, 2015). Cette fonction permet de produire chaque patron de réponses selon un niveau d'habileté réel donné et les caractéristiques des items disponibles dans la banque d'items. Un estimé du niveau d'habileté réel est produit après chaque réponse afin de guider le choix du prochain item. Pour cette recherche, après l'administration de chaque item, le niveau d'habileté sera estimé par la méthode du maximum de vraisemblance pondéré (WLE) de Warm (1989) qui la présente comme étant moins biaisée que d'autres méthodes plus classiques.

Le prochain item sera sélectionné par la maximisation de la quantité d'information au sens de Fischer (MFI) de chaque item, quantité qui aura été préalablement calculée par l'une des fonctions appelées par `simulateRespondents`. Comme la génération des passations de tests adaptatifs s'annonce complexe et longue, il importe de sélectionner une méthode simple d'utilisation et rapide afin de ne pas allonger les traitements. De toute façon, cette méthode est largement utilisée, donc reconnue et appuyée par la communauté scientifique, et constitue *de facto* un bon choix indépendamment de la complexité des traitements.

Tous les tests adaptatifs prendront fin lorsque l'erreur-type sera au maximum de 0,4 ou lorsque tous les items de la banque auront été administrés. Toutes les estimations seront bornées à $|\hat{\theta}| \leq 4$ de façon à englober plus de 99% de la population. Par conséquent, les tests adaptatifs destinés à des unités d'observation dites extrêmes n'auront pas le loisir d'estimer $\theta$ librement, ayant des bornes qui ramèneront continuellement les estimés vers des valeurs plus centrales[8]. Ceci pourrait avoir quelques effets sur le biais ou sur l'erreur-type parce qu'il est plus difficile de

[8] Les tests adaptatifs estiment habituellement θ en le circonscrivant dans un intervalle qui, d'item en item, devient de plus en plus étroit, jusqu'à l'obtention d'une précision acceptable. Avec des bornes à [-4, 4], ces intervalles sont plus comprimés.

différencier, par exemple, un $\hat{\theta}$ réellement à -4 d'un autre $\hat{\theta}$ à -4 mais empêché d'aller au-delà.

## 3.2 Méthode d'analyse des données

Pour les deux tailles de la banque d'items, les neuf valeurs du niveau d'habileté réel ainsi que pour chacune des sept valeurs du coefficient d'asymétrie $a_3$ de la distribution de probabilité du paramètre de difficulté des items de la banque d'items des statistiques descriptives seront calculées. Plus spécifiquement, le minimum, le maximum, la médiane, la moyenne, l'écart type et le coefficient d'asymétrie du niveau d'habileté estimé, de l'erreur type estimée du niveau d'habileté ainsi que du nombre d'items administrés seront calculés. Aussi, le biais d'estimation du niveau d'habileté et son erreur-type, soit l'erreur-type empirique, seront calculés pour chaque regroupement. Certaines de ces statistiques seront présentées en préambule des résultats de cette recherche.

Pour permettre de juger du biais d'estimation, les moyennes des niveaux d'habileté estimés seront comparées aux valeurs des niveaux d'habileté réels. Les écarts types des niveaux d'habileté estimés seront comparés aux moyennes des erreurs types des niveaux d'habileté estimés : on pourra ainsi évaluer l'adéquation du calcul théorique de la précision du niveau d'habileté estimé. Le calcul de l'asymétrie des niveaux d'habileté estimés, des erreurs types estimées des niveaux d'habileté ainsi que des nombres d'items administrés permettront de vérifier si les niveaux d'habileté estimés suivent une loi normale et si les valeurs des erreurs types théoriques des niveaux d'habileté estimés ainsi que les nombres d'items administrés tendent vers des valeurs inférieures ou supérieures à leur moyenne respective. Tous les calculs seront effectués à l'aide du logiciel R.

# CHAPITRE IV

# RÉSULTATS

Dans ce chapitre, les résultats des analyses statistiques conduites seront présentés en trois temps : les effets de l'asymétrie au sein d'une banque d'items sur l'estimation, la précision et la durée d'un test adaptatif.

Dans ce chapitre, les résultats des analyses statistiques conduites seront présentés en trois temps : les effets de l'asymétrie au sein d'une banque d'items sur l'estimation, la précision et la durée d'un test adaptatif. Ainsi, les effets de l'asymétrie sur les valeurs estimées des niveaux d'habileté retournées par le test et sur le biais d'estimation seront d'abord présentés. Ensuite, les effets de l'asymétrie sur les erreurs-types de ces estimés ainsi que sur l'erreur-type empirique seront présentés. Enfin, les effets de l'asymétrie sur le nombre d'items devant être administrés à chaque passation seront présentés. Bien entendu, toutes ces données – estimés, erreurs-types, biais, nombres d'items – sont reliées entre elles alors même si elles sont présentées dans des sections dédiées, elles sont utilisées et référencées au besoin pour aider le lecteur à comprendre les effets de l'asymétrie dans une perspective à la fois séquentielle et globale. D'ailleurs, c'est dans cette optique que sont présentées ici quelques statistiques descriptives sur les banques d'items générées.

Tableau 4.1 : Statistiques descriptives sur le paramètre de difficulté *b* des différentes banques d'items générées pour la présente recherche

| $a^3$ | $a^3 = -6$ | $a^3 = -3$ | $a^3 = -1$ | $a^3 = 0$ | $a^3 = 1$ | $a^3 = 3$ | $a^3 = 6$ |
|---|---|---|---|---|---|---|---|
| | | | $N = 200$ | | | | |
| Moyenne | -0,03 | -0,07 | -0,12 | 0,04 | 0,12 | 0,07 | 0,03 |
| Écart-type | 0,34 | 1,23 | 1,04 | 0,93 | 1,04 | 1,23 | 0,34 |
| Médiane | 0,06 | 0,21 | -0,01 | -0,05 | 0,01 | -0,21 | -0,06 |
| Minimum | -2,72 | -7,61 | -5,42 | -2,21 | -3,24 | -1,84 | -0,14 |
| Maximum | 0,14 | 1,84 | 3,24 | 2,4 | 5,42 | 7,61 | 2,72 |
| | | | $N = 1000$ | | | | |
| Moyenne | 0,01 | 0,01 | -0,02 | -0,01 | 0,02 | 0,01 | 0,01 |
| Écart-type | 0,32 | 1,02 | 1,03 | 1,03 | 1,03 | 1,02 | 0,32 |
| Médiane | 0,1 | 0,29 | 0,16 | -0,04 | -0,16 | -0,29 | -0,1 |
| Minimum | -4,06 | -8,94 | -5,28 | -3,01 | -2,69 | -1,02 | -0,23 |
| Maximum | 0,23 | 1,02 | 2,69 | 3,81 | 5,28 | 8,94 | 4,06 |

Les résultats sont tous présentés sous la forme de constats qui sont expliqués et détaillés à l'aide de tableaux lorsque nécessaires. Puisque les données sont nombreuses, elles sont choisies et non toutes systématiquement affichées afin de ne pas alourdir la présente.

## 4.1 Effets de l'asymétrie sur l'estimation en testing adaptatif

Le biais d'estimation est l'écart moyen entre la valeur réelle et la valeur estimée dans un lot de données. Lorsque cet écart est considérable, l'estimateur est dit « biaisé », ce qui signifie qu'il pointe systématiquement à côté de la valeur réelle. En effet, si le

biais est positif, c'est que l'estimateur a tendance à pointer plus haut que la valeur réelle. Si le biais est négatif, c'est que l'estimateur a plutôt tendance à pointer plus bas que la valeur réelle. Un estimateur biaisé peut être vu comme un canon mal aligné qui tire systématiquement à droite ou à gauche de sa cible jusqu'à ce que l'angle soit changé. Or, en mesure et évaluation, lorsqu'un niveau d'habileté doit être estimé, c'est que le niveau d'habileté réel n'est pas connu. Il est alors difficile de vérifier si « le canon a le bon angle » et il est aussi difficile de distinguer cette erreur systématique – le biais – de l'erreur de mesure. Heureusement, la présente recherche fonctionne avec des sujets simulés à qui ont été assignés des niveaux d'habileté connus. Ainsi, le biais d'estimation calculé pour chaque administration d'un test adaptatif à un groupe d'individus d'un niveau d'habileté donné permet de vérifier si différents degrés d'asymétrie dans la distribution des paramètres de difficulté des items d'une banque peuvent amener un test adaptatif à pointer systématiquement plus haut ou plus bas que le niveau d'habileté réel, fixé à une même valeur pour tous les individus du groupe.

Le tableau 4.2 contient les valeurs du biais d'estimation dans différents contextes d'administration : la taille $N$ de la banque utilisée, le coefficient d'asymétrie $a^3$ dans la distribution des paramètres de difficulté d'item et le niveau d'habileté $\theta$ fixe et unique des individus à qui est administré le test adaptatif. Ce tableau a pour but de faciliter la compréhension des constats décrits ci-dessous.

Tableau 4.2 : Biais d'estimation du niveau d'habileté selon différentes tailles de banques d'items ($N$), différents coefficients d'asymétrie ($a^3$) et différents niveaux d'habileté réels ($\theta$)

| $a^3$ | $\theta = -4$ | $\theta = -3$ | $\theta = -2$ | $\theta = -1$ | $\theta = 0$ | $\theta = 1$ | $\theta = 2$ | $\theta = 3$ | $\theta = 4$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $N = 200$ | | | | | |
| $a^3 = -6$ | -0,37 | -0,08 | -0,11 | -0,09 | -0,01 | 0,06 | 0,14 | 0,13 | 0,22 |
| $a^3 = -3$ | -0,28 | -0,03 | -0,06 | -0,05 | -0,04 | 0,03 | 0,11 | 0,13 | 0,22 |
| $a^3 = -1$ | -0,24 | -0,08 | -0,05 | -0,03 | 0,00 | 0,04 | 0,06 | 0,09 | 0,22 |
| $a^3 = 0$ | -0,31 | -0,09 | -0,08 | -0,02 | -0,01 | 0,02 | 0,04 | 0,10 | 0,22 |
| $a^3 = 1$ | -0,29 | -0,06 | -0,07 | -0,05 | 0,00 | 0,05 | 0,07 | 0,08 | 0,21 |
| $a^3 = 3$ | -0,31 | -0,08 | -0,10 | -0,05 | 0,02 | 0,07 | 0,07 | 0,10 | 0,21 |
| $a^3 = 6$ | -0,39 | -0,07 | -0,12 | -0,09 | -0,01 | 0,05 | 0,10 | 0,12 | 0,23 |
| | | | | $N = 1000$ | | | | | |
| $a^3 = -6$ | -0,23 | -0,10 | -0,08 | -0,04 | -0,04 | 0,08 | 0,13 | 0,11 | 0,22 |
| $a^3 = -3$ | -0,19 | -0,05 | -0,03 | -0,02 | 0,00 | 0,03 | 0,08 | 0,12 | 0,22 |
| $a^3 = -1$ | -0,19 | -0,03 | -0,03 | -0,02 | -0,02 | 0,03 | 0,05 | 0,10 | 0,21 |
| $a^3 = 0$ | -0,23 | -0,07 | -0,04 | -0,03 | 0,00 | 0,02 | 0,04 | 0,05 | 0,21 |
| $a^3 = 1$ | -0,25 | -0,13 | -0,06 | -0,02 | 0,01 | 0,02 | 0,03 | 0,05 | 0,20 |
| $a^3 = 3$ | -0,27 | -0,13 | -0,10 | -0,04 | -0,01 | 0,01 | 0,03 | 0,05 | 0,18 |
| $a^3 = 6$ | -0,26 | -0,15 | -0,11 | -0,09 | 0,01 | 0,07 | 0,08 | 0,09 | 0,21 |

4.1.1 Constat 1 : L'asymétrie a peu d'effet sur le biais d'estimation auprès d'individus faibles, moyens et forts ($\theta = [-3, 3]$)

À première vue, l'asymétrie n'a que peu d'effet sur le biais d'estimation. En effet, comme le montre le tableau 4.2, le biais d'estimation absolu demeure faible – généralement moins de 0,10 – à tous les coefficients d'asymétrie, indépendamment de

la taille de la banque d'items, lorsque les niveaux d'habileté réels à estimer sont compris entre -3 et 3, soit pour 98,6% de la population.

Néanmoins, certaines passations ont mené à des estimations marginales, loin du niveau d'habileté réel. Par exemple, lorsque la banque à 200 items avec l'asymétrie ($a^3$) à -6 a été utilisée pour estimer $\theta$ lorsqu'il est à -3, le test adaptatif a produit quelques estimés épars. Notamment, le niveau d'habileté d'un individu a été estimé à -1,04 soit près de deux écart-types au-delà de son niveau réel. Toutefois, l'effet de masse vient tempérer ces estimations irrégulières. Le biais prend en considération cet écart, certes, mais il englobe aussi toutes les estimations mieux réussies.

4.1.2 Constat 2 : L'asymétrie a peu d'effet sur le biais d'estimation auprès d'individus dits « extrêmes » ($\theta = -4$ ou $\theta = 4$).

Pour les niveaux d'habileté dits « extrêmes » – -4 et 4 – les biais d'estimation sont légèrement supérieurs – 0,18 à 0,39 – mais néanmoins pas exagérément élevés et toujours inférieurs à l'erreur-type fixée comme règle d'arrêt, soit 0,4. Ces différences sont dues à la constitution même des banques d'items. Elles n'ont pas suffisamment d'items ajustés à des individus extrêmes. Par exemple, la banque à 200 items dont l'asymétrie dans la distribution des paramètres de difficulté est de 6 comprend très peu d'items faciles. La valeur la plus faible du paramètre de difficulté des items de la banque est de -0,14. Cette valeur maximale passe à -0,23 dans une banque à 1000 items. Ces items administrés à des individus moyens seraient très révélateurs. Cependant, pour des individus « extrêmes », ces items ajoutent peu à l'information que cumule le test adaptatif. Avec peu d'information cumulée, l'estimation finale de $\theta$ n'est pas très précise et l'intervalle de confiance est large. Alors, les valeurs estimées peuvent dévier largement des vraies valeurs.

## 4.2 Effets de l'asymétrie sur la précision en testing adaptatif

Pour cette recherche, la règle d'arrêt en vigueur stipule que les tests adaptatifs cessent lorsque l'erreur-type atteint 0,4 ou lorsque tous les items de la banque ont été administrés. Par conséquent, cette recherche ne peut pas vérifier la valeur de l'erreur-type lorsqu'inférieure à 0,4. Parallèlement, elle ne permet pas de comparer les erreur-types de deux estimés si toutes les deux atteignent la précision attendue. Toutefois, l'asymétrie au sein des paramètres de difficulté des items fait en sorte que plusieurs individus se voient administrés tous les items d'une banque lorsqu'elle contient 200 items. Alors, il est intéressant de voir où en est l'erreur-type, où en est la précision lorsque ladite banque d'items est épuisée.

Il est important de spécifier que la banque d'items à 1000 items n'a jamais été épuisée pour une passation. Ainsi, même lorsque des items nullement ajustés aux individus leur ont été administrés, l'information, aussi minime soit-elle, a continué de s'accumuler et la précision souhaitée a toujours été atteinte même s'il fallait 785 items pour ce faire. Par conséquent, les effets de l'asymétrie sur la précision de la mesure ne seront considérés que pour les tests conduits avec les banques à 200 items. Cependant, les banques à 200 items ont rarement été épuisées pour des individus moyens, dont $\theta$ est entre -2 et 2. Alors, les seules passations retenues pour étudier les effets de l'asymétrie sur la précision sont celles des individus dits extrêmes – dont $\theta$ est inférieur à -2 ou supérieur à 2 – conduites sur l'une des banques à 200 items qu'elles ont épuisées.

Le tableau 4.3 présente l'erreur-type moyenne lorsqu'un test adaptatif nécessite l'administration de tous les 200 items de la banque. Il affiche aussi le nombre de passation par niveau d'habileté (avec un maximum potentiel de 1000) qui ont épuisé la banque d'items et sur lequel est calculé l'erreur-type moyenne. Ce calcul est fait

pour les individus de niveau $\theta$ = (-4, -4, 3 et 4) et pour tous les coefficients d'asymétrie $a^3$.

Tableau 4.3 : Erreur-type moyenne après l'administration de tous les items de la banque lorsque celle-ci contient 200 items et nombre de passations $N$ sur lequel l'erreur-type moyenne est calculée

| | $\theta = -4$ | | $\theta = -3$ | | $\theta = 3$ | | $\theta = 4$ | |
|---|---|---|---|---|---|---|---|---|
| $a^3$ | Moy | N | Moy | N | Moy | N | Moy | N |
| $a^3 = -6$ | 0,72 | 971 | 0,55 | 607 | 0,45 | 230 | 0,52 | 931 |
| $a^3 = -3$ | 0,59 | 970 | 0,49 | 524 | 0,43 | 61 | 0,47 | 829 |
| $a^3 = -1$ | 0,55 | 942 | 0,47 | 280 | 0,44 | 62 | 0,47 | 835 |
| $a^3 = 0$ | 0,63 | 951 | 0,50 | 404 | 0,42 | 18 | 0,45 | 723 |
| $a^3 = 1$ | 0,64 | 962 | 0,51 | 471 | 0,42 | 21 | 0,43 | 689 |
| $a^3 = 3$ | 0,65 | 947 | 0,51 | 429 | 0,42 | 66 | 0,46 | 845 |
| $a^3 = 6$ | 0,77 | 974 | 0,57 | 688 | 0,44 | 125 | 0,50 | 893 |

4.2.1 Constat 3 : Les niveaux d'habileté d'individus extrêmement faibles ne sont presque jamais estimés avec précision après 200 items, indépendamment de l'asymétrie.

Plus de 95 % des individus extrêmement faibles ($\theta$ = -4) ont terminé leurs tests adaptatifs à 200 items sans que la précision désirée n'ait été atteinte. Lorsque la banque d'items est symétrique ($a^3$ = 0), elle n'arrive pas à nourrir le test adaptatif d'un individu aussi faible que $\theta$ = -4 : l'erreur-type moyenne est de 0,63. Lorsque le coefficient d'asymétrie de la banque est négatif, ce qui devrait avantager un individu aussi faible que $\theta$ = -4 par rapport à un individu moyen[9], cette dernière peine tout de

[9] Lorsque le coefficient d'asymétrie $a^3$ d'une variable est négatif, la queue de distribution de cette variable est plus allongée vers les valeurs négatives, ce qui devrait lui permettre de puiser davantage de valeurs négatives extrêmes au prix d'une plus petite densité dans cette zone.

même à fournir des items sur mesure et l'erreur-type moyenne est au minimum à 0,55, soit encore loin de la valeur souhaitée.

Les individus extrêmement forts ($\theta = 4$) sont sensiblement dans la même situation, mais les valeurs des erreurs-types moyennes sont plus près des valeurs acceptables, variant entre 0,43 et 0,52 et c'est plutôt 82 % des individus qui se voient administrer tous les items de la banque. Ainsi, les effets de l'asymétrie sur la précision sont plus marquants chez les individus faibles, possiblement parce qu'au sein du modèle logistique à 3 paramètres, le paramètre de pseudo-chance $c_i$ vient fausser la donne en haussant légèrement la probabilité de réussite des individus faibles, ce qui élève une partie de la courbe et change la règle de l'ajustement item/individu décrite en 2.3.4.

### 4.2.2 Constat 4 : L'asymétrie a peu d'effet sur l'erreur-type en testing adaptatif.

Aux valeurs retenues pour la présente recherche, le niveau d'habileté à estimer est beaucoup plus déterminant dans la valeur de l'erreur-type à la fin du test que le coefficient d'asymétrie. En imposant une condition d'arrêt au test adaptatif, l'erreur-type ne peut fluctuer jusqu'à sa précision maximale sans quoi tous les items seraient systématiquement administrés et on ne serait plus en présence d'un test adaptatif, mais bien d'un test classique. Alors, comme aucune passation ne peut générer une erreur-type bien inférieure à la valeur prescrite, ici 0,4, et que pour la majorité de la population le test prend fin parce que cette valeur est atteinte, il est plutôt difficile de comparer l'effet de l'asymétrie sur l'erreur-type. Le tableau 4.2 montre la valeur de l'erreur-type moyenne lorsque le test nécessite l'administration de tous les items, mais il ne le fait que pour des valeurs marginales de $\theta$. Lorsque $\theta$ est entre -2 et 2, le test prend fin avant l'administration de tous les items du test 99,99 % du temps et alors l'erreur-type est de 0,4, la valeur minimalement acceptable. En somme, le niveau d'habileté a plus d'impact que l'asymétrie sur la précision parce qu'aucune banque d'items n'est bien ajustée à des valeurs extrêmes de $\theta$.

## 4.3 Effets de l'asymétrie sur la longueur d'un test adaptatif

L'un des objectifs du testing adaptatif est de rendre plus efficace le processus d'évaluation notamment en diminuant la longueur d'un test en ne sélectionnant que les meilleurs items pour l'individu (Van der Linden et Glas, 2007, p. 802). Ces items sur mesure donnent à eux seuls plus d'information sur l'individu qu'un lot plus vaste comprenant des items faciles, moyens et difficiles, sans égard au niveau d'habileté de l'individu : les sections 2.2, 2.6 et 2.8 l'expliquent bien. Cependant, en présence d'une banque d'items déséquilibrée, il peut s'avérer difficile d'administrer les bons items à un individu si par malheur la banque d'items ne dispose pas d'items ajustés à son niveau d'habileté réel.

Les tableaux 4.4, 4.5, 4.6 et 4.7 présentent les longueurs moyennes, minimales et maximales de tests adaptatifs à différents degrés d'asymétrie pour différents niveaux d'habileté réels. On y remarque une grande variabilité de la longueur moyenne des tests, allant de 16,13 à 524,72. Ces tableaux aident à comprendre les constats 5 et 6 qui concernent les effets de l'asymétrie sur la longueur d'un test adaptatif.

### 4.3.1 Constat 5 : L'asymétrie a peu d'effet sur la longueur d'un test adaptatif pour les individus moyens ou qui dérogent au maximum d'un écart-type de la moyenne.

Tableau 4.4 : Longueur moyenne, minimale et maximale des tests adaptatifs avec une petite banque d'items ($N$ = 200) à différents coefficients d'asymétrie ($a^3$) pour des individus dérogeant au maximum d'un écart-type de la moyenne ($-1 \leq \theta \leq 1$)

| $a^3$ | $\theta = -1$ | | | $\theta = 0$ | | | $\theta = 1$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Moy. | Min. | Max. | Moy. | Min. | Max. | Moy. | Min. | Max. |
| $a^3 = -6$ | 28,82 | 19 | 59 | 20,28 | 19 | 29 | 27,00 | 19 | 61 |
| $a^3 = -3$ | 27,85 | 20 | 44 | 21,56 | 19 | 31 | 21,72 | 19 | 38 |
| $a^3 = -1$ | 24,08 | 20 | 38 | 20,96 | 20 | 26 | 24,21 | 20 | 36 |
| $a^3 = 0$ | 22,86 | 21 | 38 | 22,88 | 21 | 30 | 23,98 | 21 | 32 |
| $a^3 = 1$ | 24,88 | 20 | 41 | 21,22 | 20 | 26 | 23,16 | 20 | 38 |
| $a^3 = 3$ | 22,14 | 19 | 41 | 21,29 | 19 | 29 | 26,52 | 19 | 39 |
| $a^3 = 6$ | 29,16 | 19 | 73 | 20,14 | 19 | 31 | 26,78 | 19 | 59 |

De $\theta = -1$ à $\theta = 1$, la longueur moyenne des tests adaptatifs ne varie que très peu aux différents degrés d'asymétrie. Lorsqu'une petite banque d'items est utilisée ($N = 200$), et que l'asymétrie est pratiquement nulle ($a^3 = 0$), la longueur du test varie de 21 à 38 items et les longueurs moyennes sont de 22,86 ($\theta = -1$), 22,88 ($\theta = 0$) et 23,98 ($\theta = 1$), soient des valeurs très près les unes des autres. Lorsque l'asymétrie est fortement négative ($a^3 = -6$), la longueur varie de 19 à 61 items avec des longueurs moyennes de 28,82 ($\theta = -1$), 20,28 ($\theta = 0$), 27,00 ($\theta = 1$), des valeurs plus distantes. L'impact de l'asymétrie se fait moins sentir sur la moyenne, mais certes sur le maximum. Cependant, même si les nombres d'items administrés peuvent, dans certains cas, être augmentés considérablement, ces valeurs demeurent acceptables compte tenu qu'un test papier-crayon unique doit comprendre davantage d'items parce qu'il doit estimer précisément l'habileté d'individus de différents niveaux, et donc comprendre plusieurs items de différents degrés de difficulté. Il en va de même lorsque l'asymétrie est fortement positive ($a^3 = 6$), où on observe des longueurs entre 19 et 73 et des longueurs moyennes de 29,16 ($\theta = -1$), 20,14 ($\theta = 0$), 26,78 ($\theta = 1$), des valeurs dispersées, mais encore acceptables.

Tableau 4.5 : Longueur moyenne, minimale et maximale des tests adaptatifs avec une banque d'items volumineuse ($N$ = 1000) à différents coefficients d'asymétrie ($a^3$) pour des individus dérogeant au maximum d'un écart type de la moyenne ($-1 \leq \theta \leq 1$)

| $a^3$ | $\theta = -1$ | | | $\theta = 0$ | | | $\theta = 1$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Moy. | Min. | Max. | Moy. | Min. | Max. | Moy. | Min. | Max. |
| $a^3 = -6$ | 22,95 | 15 | 48 | 16,13 | 15 | 24 | 21,70 | 15 | 65 |
| $a^3 = -3$ | 20,17 | 17 | 28 | 17,56 | 16 | 25 | 18,25 | 16 | 32 |
| $a^3 = -1$ | 20,49 | 16 | 29 | 17,52 | 16 | 23 | 17,84 | 16 | 26 |
| $a^3 = 0$ | 19,00 | 17 | 25 | 18,63 | 17 | 24 | 18,64 | 17 | 25 |
| $a^3 = 1$ | 18,26 | 16 | 31 | 17,35 | 16 | 23 | 20,12 | 16 | 26 |
| $a^3 = 3$ | 18,63 | 16 | 35 | 17,37 | 16 | 23 | 19,84 | 16 | 27 |
| $a^3 = 6$ | 23,03 | 15 | 57 | 16,18 | 15 | 29 | 21,96 | 15 | 40 |

Avec des banques plus volumineuses ($N$ = 1000), le même phénomène est observé, les longueurs des tests adaptatifs varient peu à travers les différents degrés d'asymétrie, mais encore moins avec une plus grande sélection d'items. Ainsi, jusqu'à un écart-type de la moyenne, les longueurs des tests varient au plus entre 15 et 65 et les longueurs moyennes se situent toutes entre 16,13 et 23,03.

4.3.2 Constat 6 : L'asymétrie a beaucoup d'effet sur la longueur d'un test adaptatif pour les individus qui dérogent d'au moins deux écarts-types de la moyenne.

Les effets de l'asymétrie sur la longueur d'un test adaptatif sont beaucoup plus manifestes auprès d'individus à deux-écarts-types de la moyenne et plus. Comme les items bien ajustés sont plus rares, des items de plus en plus loin des individus doivent être utilisés. Bien que les effets de l'asymétrie sur la longueur d'un test adaptatif soient indépendants de la taille de la banque d'items, il est intéressant de les étudier séparément : d'une part avec une petite banque d'items et les limitations qu'elle met en place et d'autre part avec une banque d'items volumineuse.

Tableau 4.6 : Longueur moyenne, minimale et maximale des tests adaptatifs avec une petite banque d'items ($N$ = 200) à différents coefficients d'asymétrie ($a^3$) pour des individus dérogeant d'au moins deux écarts-types de la moyenne ($-2 \geq \theta \geq 2$)

| $a^3$ | $\theta = -4$ | | | $\theta = -3$ | | | $\theta = -2$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Moy. | Min. | Max. | Moy. | Min. | Max. | Moy. | Min. | Max. |
| $a^3 = -6$ | 197,79 | 53 | 200 | 167,77 | 21 | 200 | 64,92 | 22 | 200 |
| $a^3 = -3$ | 197,59 | 53 | 200 | 150,04 | 32 | 200 | 46,41 | 24 | 200 |
| $a^3 = -1$ | 194,96 | 55 | 200 | 110,11 | 28 | 200 | 35,98 | 22 | 118 |
| $a^3 = 0$ | 195,25 | 46 | 200 | 127,88 | 25 | 200 | 35,49 | 21 | 200 |
| $a^3 = 1$ | 196,40 | 37 | 200 | 138,30 | 31 | 200 | 41,99 | 22 | 200 |
| $a^3 = 3$ | 195,52 | 67 | 200 | 135,82 | 28 | 200 | 39,97 | 19 | 125 |
| $a^3 = 6$ | 198,49 | 85 | 200 | 174,42 | 41 | 200 | 68,60 | 22 | 200 |
| $a^3$ | $\theta = 2$ | | | $\theta = 3$ | | | $\theta = 4$ | | |
| | Moy. | Min. | Max. | Moy. | Min. | Max. | Moy. | Min. | Max. |
| $a^3 = -6$ | 56,76 | 21 | 156 | 135,66 | 31 | 200 | 197,71 | 71 | 200 |
| $a^3 = -3$ | 33,63 | 20 | 98 | 85,72 | 24 | 200 | 186,41 | 48 | 200 |
| $a^3 = -1$ | 35,54 | 22 | 76 | 85,74 | 31 | 200 | 187,05 | 47 | 200 |
| $a^3 = 0$ | 29,06 | 23 | 69 | 62,37 | 26 | 200 | 174,92 | 38 | 200 |
| $a^3 = 1$ | 32,16 | 20 | 68 | 66,36 | 28 | 200 | 173,81 | 41 | 200 |
| $a^3 = 3$ | 38,33 | 23 | 72 | 88,83 | 28 | 200 | 189,78 | 45 | 200 |
| $a^3 = 6$ | 54,27 | 20 | 141 | 129,20 | 30 | 200 | 195,33 | 100 | 200 |

Tout d'abord, lorsqu'une petite banque d'items est utilisée pour un test adaptatif, moins d'items d'un même niveau de difficulté sont disponibles alors le test adaptatif doit recourir à des items moins ajustés au niveau d'habileté du répondant. Donc, moins d'information est cumulée et davantage d'items sont nécessaires pour atteindre l'information – et parallèlement l'erreur-type– souhaitée. Lorsque la banque, en plus de contenir peu d'items, est fortement asymétrique, la longueur des tests s'en fait ressentir auprès des individus plus marginaux, très forts ou très faibles. Suivant le

tableau 4.6, la longueur moyenne des tests avec des petites banques fortement asymétriques pour des individus à deux ou trois écarts-types de la moyenne augmente minimalement de 27,70 items, ce qui constitue presque un deuxième test pour eux! Dans certains cas, la longueur moyenne va même jusqu'à doubler, passant par exemple de 62,37 à 135,66 items où $\theta = 3$ et $a^3 = 6$! Administrer en moyenne 63 items de plus qu'il ne le faut normalement n'est pas acceptable, surtout considérant que le testing adaptatif vise justement à écourter les tests!

Pour les individus exceptionnels – à quatre écarts-types de la moyenne – les longueurs moyennes des tests à différents degrés d'asymétrie ne varient que très peu parce que la plupart des tests requièrent tous les items de la banque, ce qui garde les valeurs très près les unes des autres, soit toutes entre 173,81 et 198,49, un intervalle assez mince, mais avec des valeurs très élevées. En haussant la taille des banques, les tests ont certes plus de choix d'items, mais pour des individus si marginaux, à des forts degrés d'asymétrie, les tests se voient plutôt allongés, comme le montre le tableau 4.7.

Tableau 4.7 : Longueur moyenne, minimale et maximale des tests adaptatifs avec une banque d'items volumineuse ($N$ = 1000) à différents coefficients d'asymétrie ($a^3$) pour des individus dérogeant d'au moins deux écarts-types de la moyenne ($-2 \geq \theta \geq 2$)

| $a^3$ | $\theta = -4$ | | | $\theta = -3$ | | | $\theta = -2$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Moy. | Min. | Max. | Moy. | Min. | Max. | Moy. | Min. | Max. |
| $a^3 = -6$ | 413,17 | 58 | 592 | 110,74 | 31 | 368 | 39,91 | 19 | 103 |
| $a^3 = -3$ | 53,04 | 27 | 66 | 31,67 | 22 | 61 | 24,43 | 19 | 39 |
| $a^3 = -1$ | 44,18 | 24 | 59 | 29,30 | 22 | 49 | 23,16 | 19 | 36 |
| $a^3 = 0$ | 75,43 | 30 | 101 | 34,82 | 21 | 94 | 22,31 | 18 | 39 |
| $a^3 = 1$ | 135,62 | 26 | 181 | 50,81 | 19 | 179 | 26,19 | 16 | 49 |
| $a^3 = 3$ | 210,05 | 36 | 283 | 77,79 | 21 | 266 | 32,41 | 16 | 85 |
| $a^3 = 6$ | 524,72 | 68 | 785 | 138,04 | 37 | 419 | 54,29 | 17 | 152 |
| $a^3$ | $\theta = 2$ | | | $\theta = 3$ | | | $\theta = 4$ | | |
| | Moy. | Min. | Max. | Moy. | Min. | Max. | Moy. | Min. | Max. |
| $a^3 = -6$ | 49,81 | 15 | 136 | 121,12 | 34 | 265 | 241,95 | 54 | 283 |
| $a^3 = -3$ | 30,18 | 16 | 78 | 66,33 | 21 | 160 | 138,36 | 45 | 162 |
| $a^3 = -1$ | 25,19 | 16 | 58 | 44,66 | 20 | 106 | 91,21 | 29 | 107 |
| $a^3 = 0$ | 20,45 | 18 | 32 | 31,01 | 19 | 65 | 56,33 | 21 | 68 |
| $a^3 = 1$ | 22,75 | 17 | 31 | 27,93 | 21 | 43 | 38,63 | 26 | 45 |
| $a^3 = 3$ | 23,68 | 17 | 32 | 30,14 | 22 | 48 | 43,69 | 27 | 53 |
| $a^3 = 6$ | 36,61 | 18 | 96 | 82,96 | 29 | 227 | 193,06 | 49 | 229 |

Lorsque les banques d'items volumineuses ($N$ = 1000) sont utilisées auprès d'individus marginaux, les effets de l'asymétrie sont manifestes : les longueurs minimum diminuent parce que davantage d'items ajustés sont disponibles tandis que les longueurs moyennes sont plus courtes qu'avec de petites banques d'items pour la même raison et surtout, à un même $\theta$ la longueur moyenne des tests croit avec l'asymétrie. Aucun test n'utilise tous les items d'une banque avec les valeurs de $a^3$ et

$\theta$ retenues pour la présente recherche alors les longueurs moyennes sont libres de varier, contrairement aux tests avec de plus petites banques d'items ($N$ = 200), qui plafonnent plus rapidement (voir le tableau 4.6). Ainsi, il est plus facile de mesurer les effets de l'asymétrie avec de plus grandes banques. Par exemple, la longueur moyenne des tests pour des individus extrêmement faibles ($\theta$ = -4) passe de 75,43 ($a^3$ = 0) items à 524,72 items ($a^3$ = 6), soit 450,32 items de plus ou 7 fois sa longueur! Même dans le scénario le plus acceptable, où $\theta$ = 2 et $a^3$ passe de 0 à 6, la longueur moyenne diffère néanmoins de 16,16 items, une augmentation considérable.

## 4.4 Remarques sur les effets de l'asymétrie

En somme, l'asymétrie dans la distribution des paramètres de difficulté des items d'une banque a des effets considérables, mais principalement sur un seul élément, la longueur d'un test. En effet, un test adaptatif reposant sur une banque déséquilibrée peut être complété et afficher un biais d'estimation très faible, une erreur-type satisfaisante, mais il est possible que pour ce faire, le test ait eu recours à beaucoup plus d'items qu'il en faut normalement. Comme le testing adaptatif vise notamment à diminuer la longueur des tests, c'est contre-productif de se retrouver dans pareille situation. Dans certains cas, toutefois, lorsque le coefficient d'asymétrie est élevé et que le niveau d'habileté est très marginal, l'erreur-type est aussi affectée et dépasse grandement la valeur fixée pour terminer le test. Il s'agit de cas où il a fallu administrer tous les items de la banque.

Si une longueur maximale avait été imposée, 50 items par passation par exemple, moins d'information aurait été cumulée : alors les estimations auraient sans doute été davantage biaisées et les erreurs-types seraient dans plusieurs cas supérieures au minimum acceptable.

# CHAPITRE V

# INTERPRÉTATION ET DISCUSSION

Dans ce chapitre, les résultats rapportés précédemment seront expliqués plus en profondeur et leur valeur pour la présente recherche sera étudiée. Parallèlement, ces résultats seront vérifiés à la lumière d'autres recherches s'étant intéressées au testing adaptatif ou l'ayant utilisé pour une étude. Ainsi, chaque constat émis précédemment sera analysé et les résultats sur lesquels il est construit seront comparés à ceux de recherches analogues. À la lumière de ces comparaisons, les limites de la recherche seront ensuite présentées.

## 5.1 Effets de l'asymétrie sur le biais d'estimation en testing adaptatif

La présente recherche n'a pas pu établir de liens significatifs entre l'asymétrie dans la distribution des paramètres de difficulté d'une banque d'items et le biais d'estimation. Même en présence d'une forte asymétrie ($|a^3| = 6$), le biais est demeuré généralement plutôt faible, du moins toujours en deçà de l'erreur-type préalablement fixée. Pourtant, il était aisé de croire que l'administration répétée d'items mal ajustés, un peu trop faciles ou un peu trop difficiles, pourrait amener le test adaptatif à pointer systématiquement trop haut ou trop bas. D'ailleurs, c'est cette présomption qui a justifié l'utilisation de l'estimateur de vraisemblance pondéré (Warm, 1989), choisi justement parce qu'il est reconnu être moins biaisé que l'estimateur du maximum de vraisemblance, un classique.

Dans l'article où Warm présente ce nouvel estimateur, il le compare aux estimateurs du maximum de vraisemblance et bayésien à travers des tests Monte-Carlo. Les résultats sont manifestes : avec une même longueur de test, l'estimateur de

vraisemblance pondéré est nettement moins biaisé que les autres. L'auteur démontre à travers 3 tests de 10, 30 et 60 items que son estimateur est moins biaisé que l'estimateur du maximum de vraisemblance même après seulement 10 items :

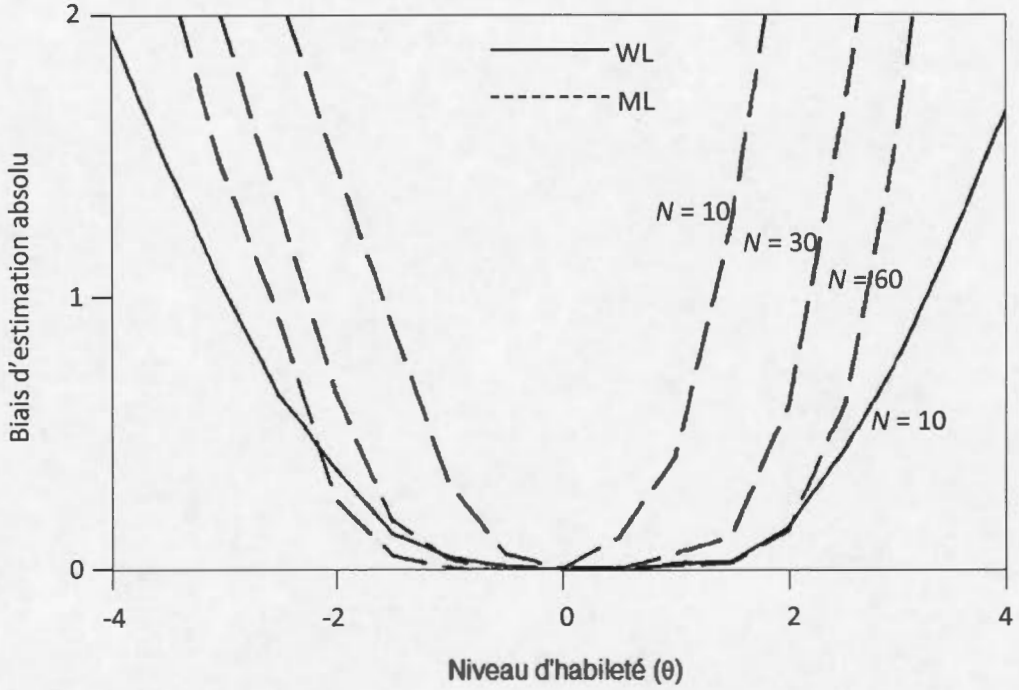


Figure 5.1 Biais d'estimation de trois estimateurs du maximum de vraisemblance à $N$ = 10, 30 et 60 et d'un estimateur de vraisemblance pondéré à $N$ = 10, tiré de Warm (1989, p. 434)

Cependant, l'estimateur de vraisemblance pondéré n'a pas été aussi étincelant dans la présente recherche. Quelques tests comparatifs ont été effectués *a posteriori* entre cet estimateur et l'estimateur du maximum de vraisemblance. Comme le montre le tableau 5.1, à $\theta$ = -4, l'estimateur de vraisemblance pondéré s'est avéré être davantage biaisé que l'estimateur du maximum de vraisemblance avec la plupart des banques d'items utilisées. Il est possible que des modifications au processus de production des patrons de réponses puissent altérer ce résultat : par exemple, en

générant des banques d'items variées, mais conservant les caractéristiques préalablement établies. Aussi, comme l'estimateur de vraisemblance pondéré (WLE) tend un peu plus vers le centre (Linacre, 2009, p. 1188-1189), il est possible qu'à des valeurs extrêmes de $\theta$, l'estimateur du maximum de vraisemblance (MLE) ait été plus stable.

Tableau 5.1 : Comparaison des biais d'estimation du niveau d'habileté entre l'estimateur de vraisemblance pondéré (WLE) et l'estimateur du maximum de vraisemblance (ML) à $\theta$ = -4 selon différentes tailles de banques d'items ($N$) et différents coefficients d'asymétrie ($a^3$)

| $a^3$ | $N = 200$ | | $N = 1000$ | |
|---|---|---|---|---|
| | WLE | ML | WLE | ML |
| $a^3 = -6$ | -0,37 | -0,25 | -0,23 | -0,20 |
| $a^3 = -3$ | -0,28 | -0,25 | -0,19 | -0,22 |
| $a^3 = -1$ | -0,24 | -0,18 | -0,19 | -0,21 |
| $a^3 = 0$ | -0,31 | -0,26 | -0,23 | -0,21 |
| $a^3 = 1$ | -0,29 | -0,23 | -0,25 | -0,21 |
| $a^3 = 3$ | -0,31 | -0,25 | -0,27 | -0,22 |
| $a^3 = 6$ | -0,39 | -0,20 | -0,26 | -0,20 |

En définitive, le biais est demeuré inférieur à l'erreur type dans pratiquement toutes les combinaisons de $N$, $\theta$ et $a^3$, mais, dans la présente recherche, il aurait pu l'être davantage avec un estimateur classique, notamment l'estimateur du maximum de vraisemblance.

## 5.2 Effets de l'asymétrie sur l'erreur-type en testing adaptatif

Dans la présente recherche, la règle de fin choisie uniformise la précision de la mesure de sorte que toute passation ne peut être plus précise que $S = 0,4$. Il est alors difficile de comparer les résultats ici obtenus à ceux de recherches similaires. Comme l'erreur-type est fixée à 0,4, c'est plutôt le nombre d'items nécessaires pour s'y rendre qui devient intéressant à analyser et à comparer. Néanmoins, l'erreur-type peut être différente de 0,4 à la fin d'un test adaptatif si ce test ne réussit pas à cumuler assez d'information après l'administration de tous les items de la banque.

Lorsque le niveau d'habileté réel est extrêmement faible ($\theta$ = -4), il est difficile de l'estimer à la fois rapidement et précisément, même qu'un choix s'impose! Si la précision est prioritaire, alors le test sera long parce que la banque ne contient pas suffisamment d'items ajustés à un niveau si faible. Davantage d'items devront être administrés afin de cumuler assez d'information. Si la rapidité est priorisée, alors la règle d'arrêt des tests sera modifiée en spécifiant un nombre d'items maximal à administrer – 50 items par exemple – et la précision en sera affectée, parce que peu d'information se cumule avec des items mal ajustés.

Raîche (2002) a étudié différentes façons de diminuer le biais d'estimation et l'erreur-type empirique en testing adaptatif lorsque l'estimateur de l'espérance a posteriori est utilisé. Dans cette recherche, il travaille avec des valeurs extrêmes du niveau d'habileté réel – jusqu'à $\theta$ = -9 – afin de voir comment se passe l'estimation dans pareil contexte avec des méthodes d'ajustement particulières. Toutefois, Raîche travaille avec des banques d'items optimales, qui sont toujours en mesure de fournir un item du niveau de difficulté prescrit, ce qui n'est pas le cas dans la présente recherche qui travaille avec des banques limitées et asymétriques. Néanmoins, l'auteur mentionne que même avec une banque d'items bien équilibrée l'erreur-type est très instable à des valeurs extrêmes de $\theta$. Comme les items à administrer sont bien

ajustés aux individus, l'erreur-type de l'estimateur se stabilise plutôt rapidement – même s'il est fortement biaisé à ces valeurs extrêmes (Wang et Vispoel, 1998, p.114).

Dans la présente recherche, mêmes les banques d'items volumineuses – $N = 1000$ – ne contiennent pas assez d'items pour administrer autant d'items parfaitement ajustés qu'il en faut lorsque $|\theta| > 2$. Alors, l'erreur-type en est affectée, ne variant que très peu lorsque les items ajustés viennent à manquer. L'effet est pire encore lorsque les petites banques d'items sont utilisées – $N = 200$ – parce que moins d'items sont adaptés au niveau d'habileté, moins d'items sont administrés au total et moins d'information est cumulée.

## 5.3 Effets de l'asymétrie sur la longueur d'un test adaptatif

Les banques asymétriques de la présente recherche contiennent plus d'items dont le niveau de difficulté est moyen – dont $b_i$ est près de 0 – que les banques symétriques. Leur distribution suit une loi normale asymétrique, soit une généralisation de la loi normale qui intègre deux paramètres de position et d'échelle pour contrôler la symétrie. Dans pareille distribution, la moyenne des valeurs donne 0, mais la densité n'est pas symétrique, car la queue est allongée à gauche ou à droite et une concentration de valeurs près de zéro vient contrebalancer le déséquilibre causé par cet étalement. Par exemple, les paramètres de difficulté de la banque à 1000 items au coefficient d'asymétrie $a^3 = -6$ forment une distribution normale asymétrique, comme le montre la figure 5.2.

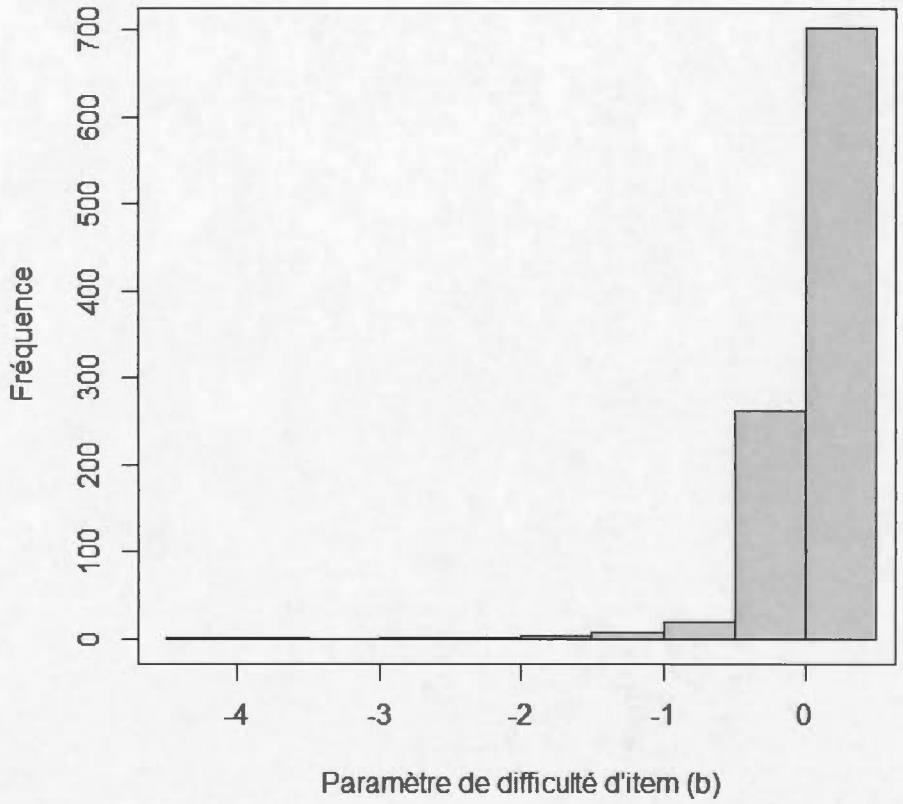


Figure 5.2 : Distribution du paramètre de difficulté *b* dans la banque à 1000 items au coefficient d'asymétrie $a^3 = -6$

Cette banque contient quelques items dont le paramètre de difficulté est entre -0,5 et 0, plusieurs items dont la difficulté est entre 0 et 0,5 mais rien au-delà de cette valeur. L'étendue des valeurs du paramètre *c* est beaucoup plus large sous zéro, avec des valeurs atteignant même -4,5, mais la fréquence de ces valeurs sous zéro est très faible. Cette banque d'items asymétrique dispose donc toujours d'un nombre considérable d'items moyens ($-1 \geq b \leq 1$). Lorsque ces items sont administrés à des individus moyens ($-1 \geq \theta \leq 1$), l'ajustement est optimal, l'information se cumule, la mesure se précise et l'asymétrie de la banque d'items n'y change rien, tant que des items ajustés peuvent être trouvés.

### 5.3.1 Effets de l'asymétrie sur la longueur d'un test pour des individus moyens

Gibbons, Weiss, Kupfer, Frank, Faglioni, Grochocinski, Bhaumik, Stover, Bock et Immekus (2008, p. 366) ont étudié comment l'utilisation du testing adaptatif peut faciliter le processus d'évaluation en santé mentale. Ils illustrent leurs constats à l'aide d'un exemple réel où le niveau d'habileté d'un individu a été estimé à $\hat{\theta} =$ 1,26 après l'administration de 26 items tirés de la banque MASS (Mood and Anxiety Spectrum Scales), soit avec l'atteinte d'une erreur type $S \leq 0{,}3$, un seuil de précision acceptable. Comme on peut le remarquer au tableau 5.2, en comparant cette passation aux passations de la présente recherche faites avec les individus les plus comparables – $\theta = 1$ – une certaine affinité est remarquée.

Tableau 5.2 : Longueurs moyennes des passations d'individus dont $\theta = 1$ avec les différentes banques d'items

| $a^3 = -6$ | $a^3 = -3$ | $a^3 = -1$ | $a^3 = 0$ | $a^3 = 1$ | $a^3 = 3$ | $a^3 = 6$ | Moyenne |
|---|---|---|---|---|---|---|---|
| | | | $N = 200$ | | | | |
| 27 | 21,72 | 24,21 | 23,98 | 23,16 | 26,52 | 26,78 | 24,77 |
| | | | $N = 1000$ | | | | |
| 21,7 | 18,25 | 17,84 | 18,64 | 20,12 | 19,84 | 21,96 | 19,76 |

Même les banques d'items les plus asymétriques de cette recherche ($|a^3| = 6$) produisent des passations qui sont environ aussi longues que celle nourrie par la banque MASS. De plus, même les plus petites banques d'items parmi les plus asymétriques ($N = 200$) génèrent des passations encore comparables. Certes, les contextes d'administration ne sont pas les mêmes. En effet, l'erreur type acceptable est de 0,4 dans la présente recherche et de 0,3 dans la recherche de Gibbons et *al.* Aussi, les banques d'items n'ont pas les mêmes tailles : 200 items ou 1000 items dans la présente recherche et 626 items dans la banque MASS. Les coefficients

d'asymétrie ne peuvent pas être comparés parce que cette information n'est pas disponible pour la banque MASS. Ensuite, cette banque est réelle alors que les banques d'items de la présente recherche sont fictives. Enfin, une passation choisie dans la recherche de Gibbons et *al.* pour la valeur de son estimé ($\hat{\theta} = 1{,}26$) est comparée à des passations générées dans la présente recherche à partir de différentes valeurs de $\theta$ mais aucune à $\theta = 1{,}26$, la plus près étant à $\theta = 1$. Malgré tout, ces différences ne sont pas assez importantes pour rendre les passations incomparables.

Le coefficient d'asymétrie de la banque d'items MASS n'est pas connu, mais même s'il était extrême ($|a^3| = 6$), la banque d'items serait néanmoins en mesure d'estimer plutôt rapidement un niveau d'habileté d'un individu moyen, c'est ce que cette comparaison avec les banques de cette recherche a mis en évidence.

### 5.3.2 Effets de l'asymétrie sur la longueur d'un test pour des individus marginaux

Lorsque le niveau d'habileté d'un individu est à deux écarts-types ou plus de la moyenne, les effets sont probants. En effet, tout test adaptatif nourri d'une banque d'items asymétrique nécessitera plus d'items qu'à la normale pour compenser la perte en information générée par l'administration d'items mal ajustés à ces individus. Dans la présente recherche, ces effets sont surtout visibles lorsque les banques d'items volumineuses sont utilisées. Comme les tests adaptatifs qu'elles nourrissent ne nécessitent jamais l'administration de tous leurs items, les différences d'un degré d'asymétrie à un autre ou d'un niveau d'habileté à un autre sont apparentes. Lorsque les petites banques d'items sont utilisées, les tests adaptatifs nécessitent l'administration de plusieurs items – voire tous – lorsqu'ils sont destinés à des individus au niveau d'habileté extrême ($|\theta| \geq 3$) alors, les différences, donc les effets, sont peu perceptibles. Le tableau 5.3 le démontre bien.

Tableau 5.3 : Longueurs moyennes des passations d'individus dont $|\theta| \geq 2$ avec les différentes banques d'items

| $\theta$ | $a^3 = -6$ | $a^3 = -3$ | $a^3 = -1$ | $a^3 = 0$ | $a^3 = 1$ | $a^3 = 3$ | $a^3 = 6$ | Moyenne |
|---|---|---|---|---|---|---|---|---|
| | | | | $N = 200$ | | | | |
| -4 | 179,79 | 179,59 | 194,96 | 195,25 | 196,4 | 195,52 | 198,49 | 191,43 |
| -3 | 167,77 | 150,04 | 110,11 | 127,88 | 138,3 | 135,82 | 174,42 | 143,48 |
| -2 | 64,92 | 46,41 | 35,98 | 35,49 | 41,99 | 39,97 | 68,6 | 47,62 |
| 2 | 56,76 | 33,63 | 35,54 | 29,06 | 32,16 | 38,33 | 54,27 | 39,96 |
| 3 | 135,66 | 85,72 | 85,74 | 62,37 | 66,36 | 88,83 | 129,2 | 93,41 |
| 4 | 197,71 | 186,41 | 187,05 | 174,92 | 173,81 | 189,78 | 195,33 | 186,43 |
| | | | | $N = 1000$ | | | | |
| -4 | 413,17 | 53,04 | 44,18 | 75,43 | 135,62 | 210,05 | 524,72 | 208,03 |
| -3 | 110,74 | 31,67 | 29,3 | 34,82 | 50,81 | 77,79 | 138,04 | 67,6 |
| -2 | 39,91 | 24,43 | 23,16 | 22,31 | 26,19 | 32,41 | 54,29 | 31,81 |
| 2 | 49,81 | 30,18 | 25,19 | 20,45 | 22,75 | 23,68 | 36,61 | 29,81 |
| 3 | 121,12 | 66,33 | 44,66 | 31,01 | 27,93 | 30,14 | 82,96 | 57,74 |
| 4 | 241,95 | 138,36 | 91,21 | 56,33 | 38,63 | 43,69 | 193,06 | 114,75 |

Cependant, le tableau 5.3 met surtout en évidence que l'asymétrie n'a pas tout à fait les mêmes effets sur des individus aux $\theta$ opposés. Considérant qu'une distribution asymétrique crée une concentration d'items près du centre d'un côté de zéro et une dispersion d'items sur un plus long intervalle de l'autre côté de zéro, le nombre d'items ajustés disponibles à $\theta = k$ ou $\theta = -k$ ne peut être le même. Par conséquent, l'information que détiennent ces items à $\theta = k$ ou $\theta = -k$ n'est pas la même. Par exemple, dans la banque de 1000 items où $a^3$ = -3, les tests adaptatifs administrés à des individus faibles à différents degrés ($\theta \leq$ -2) sont beaucoup plus courts que des tests adaptatifs prodigués à des individus forts ($\theta \geq 2$) avec la même banque. En effet, un test nécessite en moyenne 31,67 items pour bien circonscrire le niveau d'habileté

d'un individu très faible ($\theta$ = -3) alors qu'il en prend en moyenne plus que le double – 66,33 – pour estimer l'habileté d'un individu très fort ($\theta$ = 3). En observant la distribution des paramètres de difficulté des items de cette banque, cette situation s'explique d'elle-même. Le tableau 5.4 contient la distribution des paramètres de difficulté des items pour cette banque et pour toutes les autres banques de 1000 items.

Tableau 5.4 : Distribution des paramètres de difficulté des items des banques à 1000 items à tous les degrés d'asymétrie

| $b$ | $a^3 = -6$ | $a^3 = -3$ | $a^3 = -1$ | $a^3 = 0$ | $a^3 = 1$ | $a^3 = 3$ | $a^3 = 6$ |
|---|---|---|---|---|---|---|---|
| $b < -8$ | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| $-8 \leq b < -7$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $-7 \leq b < -6$ | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| $-6 \leq b < -5$ | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| $-5 \leq b < -4$ | 1 | 1 | 3 | 0 | 0 | 0 | 0 |
| $-4 \leq b < -3$ | 1 | 9 | 6 | 1 | 0 | 0 | 0 |
| $-3 \leq b < -2$ | 3 | 25 | 38 | 32 | 4 | 0 | 0 |
| $-2 \leq b < -1$ | 12 | 73 | 111 | 135 | 125 | 17 | 0 |
| $-1 \leq b < 0$ | 281 | 251 | 269 | 350 | 443 | 619 | 702 |
| $0 \leq b < 1$ | 702 | 619 | 443 | 324 | 269 | 251 | 281 |
| $1 \leq b < 2$ | 0 | 17 | 125 | 131 | 111 | 73 | 12 |
| $2 \leq b < 3$ | 0 | 0 | 4 | 25 | 38 | 25 | 3 |
| $3 \leq b < 4$ | 0 | 0 | 0 | 2 | 6 | 9 | 1 |
| $4 \leq b < 5$ | 0 | 0 | 0 | 0 | 3 | 1 | 1 |
| $5 \leq b < 6$ | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| $6 \leq b < 7$ | 0 | 0 | 0 | 0 | 0 | 3 | 0 |
| $7 \leq b < 8$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $b \geq 8$ | 0 | 0 | 0 | 0 | 0 | 2 | 0 |

Dans cette banque, il n'y a aucun item à un écart-type du niveau d'habileté d'un individu très fort ($\theta = 3$), soit $2 \leq b < 4$, et il faut aller à un écart-type additionnel, soit $1 \leq b < 5$ pour y trouver 17 items relativement ajustés à cet individu. Pour un individu très faible ($\theta = -3$), 34 items se trouvent à un maximum d'un écart-type de $\theta$, soit $-4 \leq b < -2$ et en allant à un écart-type additionnel, cette quantité d'items s'élève à 108! En manque d'items à $\hat{\theta} = 3$, le test adaptatif ira puiser des items loin de $\hat{\theta}$ plus rapidement. Ces items, moins ajustés, cumuleront beaucoup moins d'information, ce qui nécessitera plus d'items et allongera le test. En somme, même si l'asymétrie crée une concentration d'items moyens, ceux-ci ne fournissent pas autant d'information que des items plus près de $\theta$ mais en plus petit nombre peuvent donner.

### 5.3.3 Effets de l'asymétrie et de l'optimalité d'une banque sur la longueur d'un test

Dans sa recherche sur les différentes façons de diminuer le biais d'estimation et l'erreur-type empirique en testing adaptatif, Raîche (2002) a généré plusieurs passations de tests adaptatifs par des individus simulés de différents niveaux, notamment lorsque $\theta$ = [-9, -6, -3], des valeurs extrêmes. Toutefois, des banques optimales ont été utilisées alors chaque item est toujours parfaitement ajusté à l'individu et fournit un maximum d'information, ce qui rend les tests adaptatifs très courts comme le montre le tableau 5.5.

Tableau 5.5 : Longueurs de tests adaptatifs lorsque $\theta$ est extrême, tiré de Raîche (2002, p. 17-19)[10]

| $\theta$ | *N (où $S \leq 0,4$)* | *N (où $S \leq 0,3$)* |
|---|---|---|
| -9 | 22 | 29 |
| -6 | 18 | 25 |
| -3 | 13 | 20 |

Dans la présente recherche, même les items les plus ajustés aux individus ne le sont pas autant que ceux dans la recherche de Raîche, c'est pourquoi à $\theta$, $a^3$ et $S$ sensiblement égaux[11], les tests ne sont pas aussi longs : 13 items avec les banques d'items de Raîche et 34,82 items en moyenne dans la présente recherche. Néanmoins, en comparant la longueur d'un test adaptatif idéal à la longueur de tests adaptatifs asymétriques pour ce même genre d'individu, les différences sont béantes : 50,81 items lorsque l'asymétrie est relativement faible ($a^3$ = 1), 77,79 items lorsque l'asymétrie est considérable ($a^3$ = 3) et 138,04 items lorsque l'asymétrie est extrême ($a^3$ = 6), soit plus de 10 fois le nombre d'items nécessaires avec une banque d'items optimale.

Warm (1989) a lui aussi travaillé avec des banques optimales pour tester et comparer son estimateur de vraisemblance pondéré aux estimateurs du maximum de vraisemblance et bayésien. Les longueurs de tests qu'il observe pour des niveaux individus de niveau d'habileté $\theta$ = -3 sont semblables à celles de Raîche, donc plutôt différentes de celles de la présente recherche qui travaille avec des banques loin d'être optimales. La figure 5.3 présente les longueurs moyennes de tests observés à différents niveaux d'habileté dans la recherche de Warm.

---

[10] Raîche a testé différents correctifs à l'estimation de $\theta$ afin de voir leurs effets sur le biais. L'utilisation d'un estimateur adaptatif *a priori* lors des estimations provisoires de $\theta$ jointe à l'ajustement de l'intégrale lors de l'estimation finale semble donner les meilleurs résultats. Ce sont les longueurs des tests adaptatifs où ces correctifs ont été appliqués qui sont ici rapportées.

[11] $\theta$ = -3, $a^3$ = 0 et $S$ = 0,4

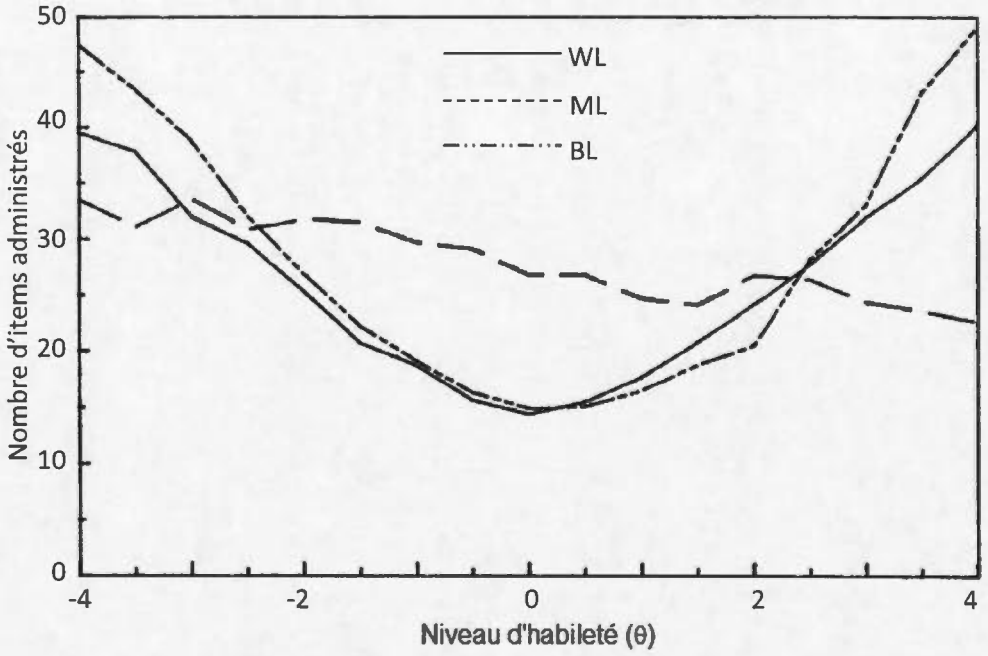


Figure 5.3 : Longueurs moyennes de tests adaptatifs optimaux utilisant différents estimateurs, pour différents niveaux d'habileté, tiré de Warm (1989, p. 439)

À $\theta$ = -3, les estimateurs de vraisemblance pondérée (WL), du maximum de vraisemblance (ML) et bayésien (BM), nécessitent tous entre 30 et 40 items pour atteindre la précision demandée. D'ailleurs, Warm a fixé à 0,22 cette précision minimalement acceptable pour les tests adaptatifs de sa recherche, ce qui est beaucoup plus précis que celle stipulée dans la règle de fin des tests adaptatifs de la présente – $S$ = 0,4. En effet, cette différence de 0,18 dans l'erreur-type maximale équivaut à une différence de 30,86 au regard de la quantité d'information :

$$S = \sqrt{\sum I_i(\theta)}^{-1}$$

$$\sqrt{\sum I_i(\theta)} = S^{-1}$$

$$\sum I_i(\theta) = (S^{-1})^2 = S^{-2}$$

$$\sum I_i(\theta) = 0{,}18^{-2}$$

$$I(\theta) = 30{,}86$$

Dans la présente recherche, lorsque $\theta$ = -3 et que la banque de 1000 items considérablement asymétrique ($a^3$ = 3) est utilisée, un test adaptatif nécessite en moyenne 77,79 items (voir les tableaux 4.6 et 5.3). Supposons qu'un test adaptatif est administré à un individu de ce niveau et que les 78 items les mieux ajustés à son niveau lui sont administrés. Les derniers items de ce test ne fourniront pas beaucoup d'information parce que les quelques items ajustés à une valeur aussi extrême de $\theta_j$ auront déjà été administrés. Des items beaucoup moins ajustés devront donc être sélectionnés. Pour bien comprendre cet exemple, le tableau 5.6 présente les quantités d'information fournies par les 5 items les mieux ajustés à $\theta_j$ = -3 et les 5 items les moins ajustés à $\theta_j$ = -3, parmi les 78 items les mieux ajustés à travers toute la banque.

Tableau 5.6 : Quantité d'information fournie par les 5 items les plus ajustés et les 5 items les moins ajustés à un individu extrêmement faible ($\theta_j$ = -3) parmi les 78 items les mieux ajustés à ce dernier en utilisant la banque de 1000 items fortement asymétrique ($a^3$ = 3)

| $I_i$ en ordre décroissant | | | | |
|---|---|---|---|---|
| 0,09 | 0,09 | 0,09 | 0,08 | 0,08 |
| | | (68 items dont $0,08 \leq I_i \leq 0,04$) | | |
| 0,04 | 0,04 | 0,04 | 0,04 | 0,04 |

Si, au lieu d'atteindre $S$ = 0,4 le test devait atteindre $S$ = 0,22 afin d'être aussi précis que les tests de Warm, il faudrait alors que les items supplémentaires cumulent ensemble 30,86 en information. En supposant que tous les items suivant le 78$^e$ donnent autant d'information que ce dernier item administré ($I_{78e}$ = 0,04), il faudrait au minimum 772 items supplémentaires pour atteindre la précision demandée par Warm, soit dix fois la quantité originale d'items! Un test adaptatif avec des banques

mieux équilibrées serait certainement allongé par un pareil changement dans l'erreur-type maximale, mais l'asymétrie vient ici en amplifier les effets parce que les items ajustés à des individus dits « extrêmes » sont plus rares, tel que le rapporte le constat 6 : l'asymétrie a beaucoup d'effet sur la longueur d'un test adaptatif pour les individus qui dérogent d'au moins deux écarts-types de la moyenne.

## 5.4 Limites et biais de la recherche

Bien qu'originale et pertinente, la présente recherche a ses limites. D'abord, l'utilisation d'unités d'observation simulées plutôt que des données provenant de répondants réels vient biaiser, mais légèrement, la recherche dans son ensemble. Plus spécifiquement, pour diverses raisons, les résultats d'évaluation peuvent ne pas être dus uniquement au niveau d'habileté du répondant et ainsi ne pas être issus d'une simple échelle de mesure unidimensionnelle. En effet, les humains ressentent des émotions et leurs comportements peuvent s'en faire ressentir et s'atténuer, s'amplifier ou changer. Par exemple, un individu peut répondre différemment à un test s'il est fatigué, non parce que son niveau d'habileté en est atteint, mais bien parce que dans ces conditions il est plus disposé à commettre des erreurs d'inattention. Le stress, la colère, le désespoir sont autant d'émotions qui peuvent jouer sur la performance d'un individu. Par ailleurs, certains tests de la présente recherche se sont avérés très longs, jusqu'à 785 items, et après tant d'items un répondant réel ne serait sûrement pas dans sa condition initiale et son comportement à même ce test en serait certainement affecté. Aussi, les humains peuvent choisir de se sous-classer dans le cadre d'un test en répondant au hasard ou en choisissant des mauvaises réponses. Inversement, ils peuvent aussi tricher et ainsi sembler plus forts qu'ils ne le sont réellement.

De multiples recherches ont modélisé ces différents comportements. La recherche de Raîche, Magis, Blais et Brochu (2012) propose une extension aux modèles 3PL et 4PL en ajoutant quatre paramètres de personne dont la fluctuation personnelle qui

dépeint la stabilité d'un individu dans sa performance à un test, la propension à la pseudo chance personnelle ainsi qu'un paramètre d'inattention personnelle. La recherche de Brassard, Béland et Raîche (2011), citée précédemment, présente différents indicateurs destinés à détecter des patrons de réponses inappropriés, sous-tendant des comportements incongrus. Cependant, toute cette instabilité propre à l'humain n'a pas été reproduite dans les données simulées afin de ne pas dénaturer la présente recherche. En effet, l'intérêt ici est de rendre compte des effets de l'asymétrie – et non de comportements inappropriés – sur différentes variables liées à des passations de tests adaptatifs.

Ensuite, seulement 14 banques d'items ont été générées pour la recherche, soit une banque pour chaque combinaison de $N$ (200, 1000) et de $a^3$ (-6, -3, -1, 0, 1, 3, 6). Il aurait été préférable de générer une banque d'items par passation afin de s'assurer que les effets de l'asymétrie ne soient pas dépendants des quelques banques d'items utilisées. Ainsi, pour les 1000 individus d'un niveau d'habileté donné, à un degré d'asymétrie donné et avec un volume donné, 1000 banques d'items auraient dû être générées. Les passations des tests adaptatifs auraient pu être comparées les unes aux autres comme c'est le cas dans la recherche, même si différentes banques les avaient alimentées. De cette façon, les résultats auraient été plus généralisables. Or, des contraintes d'ordre technologique ont obligé certains sacrifices. Les traitements nécessaires à la génération des banques d'items, des unités d'observation et surtout des passations de tests adaptatifs ont pris beaucoup de temps et nécessité excessivement de mémoire et d'espace disque. Générer 1000 fois plus de banques d'items n'était pas envisageable avec la technologie disponible, malheureusement.

Aussi, comme des banques d'items fictives ont été utilisées, elles n'ont pas été calibrées à partir de données issues d'une première administration. Alors, les valeurs des paramètres des items n'ont pas été déterminées à travers un processus d'estimation itératif comme l'estimation par maximum de vraisemblance conjointe

(Hambleton et Swaminathan, 1985, p. 125-149) ou marginale (Baker et Kim, 2004, p. 84-92). Plutôt, les valeurs des paramètres d'item ont été attribuées séquentiellement, les $a_I$, puis les $b_I$ et enfin les $c_I$. Ce faisant, les paramètres des items ne sont pas ajustés les uns aux autres. Il en résulte des items dont la combinaison des paramètres peut être douteuse.

Cependant, même si le lecteur doit en être informé, ce manque de consistance dans les paramètres d'item ne devrait pas affecter les résultats rapportés. En effet, ces items douteux font partie d'une banque, quelle qu'elle soit, et si un tel item est administré dans le cadre d'un test adaptatif c'est parce que le processus de sélection d'item aura déterminé que c'est l'item disponible qui maximise la fonction d'information au point où en est rendu le test, tout simplement.

Enfin, la génération de données asymétriques avec le package *fGarch* a produit des valeurs plus qu'invraisemblables. En effet, certaines combinaisons de *nu* – paramètre de forme – et de *xi* – paramètre d'asymétrie – ont produit des valeurs de $b$ à près de 9 écarts-types de la moyenne ($b$ = 8,94)! Pourtant, lorsque les paramètres d'item et de personne sont estimés à partir de données réelles, les valeurs de $b$ et de $\theta$ sont habituellement bornées dans un intervalle plus serré, [-4, 4] par exemple. Au-delà de ces valeurs, les valeurs sont improbables et de plus en plus aberrantes. Théoriquement, une banque d'items réelle ne devrait donc jamais contenir d'items dont le paramètre de difficulté va au-delà de cet intervalle. De plus, tout estimé du niveau d'habileté produit avec cette même banque ne devrait jamais dépasser cet intervalle. Cependant, la fonction du package *fGarch* qui génère des distributions asymétriques n'accepte pas de bornes ou d'intervalles. Alors, certaines banques d'items se retrouvent avec quelques items dont le niveau d'habileté est très extrême au lieu d'avoir plusieurs items « limites », près des bornes spécifiées. Retirer ces items ou en changer la valeur modifierait les coefficients d'asymétrie prescrits pour la présente recherche; alors ces items extrêmes ont été conservés et le déséquilibre

qu'ils causent peut affecter la longueur de certains tests adaptatifs. Il aurait peut-être été plus judicieux de ramener ces valeurs aux bornes les plus près [-4, 4] et de vérifier plus en profondeur l'impact de ces changements sur les coefficients d'asymétrie et sur le déroulement des tests en général.

# CHAPITRE VI

# CONCLUSION

La présente recherche avait comme objectifs spécifiques de circonscrire les effets de l'asymétrie dans la distribution des paramètres de difficulté au sein d'une banque d'items sur différentes variables liées à l'administration de tests adaptatifs.

## 6.1 Retour sur la méthodologie

Pour ce faire, des banques de 200 et de 1000 items à différents degrés d'asymétrie[12] ont été générées en suivant le modèle logistique à trois paramètres (3PL). Aussi, 1000 unités d'observation ont été générées à différentes valeurs de $\theta$[13]. Enfin, des tests adaptatifs ont été simulés pour chaque unité d'observation en utilisant chacune des 14 banques d'items. Les informations sur les passations de ces tests adaptatifs ont été conservées dans une structure multidimensionnelle au sein d'un environnement *R* et ce sont ces données qui ont été analysées afin d'en tirer des résultats.

## 6.2 Sommaire des résultats

D'une part, les résultats des différentes analyses démontrent que les effets de l'asymétrie sur le biais d'estimation et sur l'erreur-type sont négligeables. Cependant, les effets de l'asymétrie sur la longueur des tests sont plutôt considérables, surtout lorsque les niveaux d'habileté des individus sont extrêmes. Pour ceux-ci, les tests adaptatifs sont déjà longs parce que peu d'items sont ajustés à eux. Néanmoins, une forte asymétrie ($|a^3| > 3$) rend ces tests démesurément longs, plus de 800 items dans

[12] Les valeurs retenues de $a^3$ retenues pour la recherche sont [-6, -3, -1, 0, 1, 3, 6].
[13] Les valeurs retenues de $\theta$ retenues pour la recherche sont [-4, -3, -2, -1, 0, 1, 2, 3, 4].

certaines circonstances. Cette recherche met donc en évidence la nécessité de prescrire un nombre maximal d'items à administrer en plus de toute autre règle de fin, du moins lorsque les paramètres de difficulté forment une distribution asymétrique, d'autant plus que vérifier cette asymétrie est une opération fort simple. Alors, le premier objectif de cette recherche peut être considéré atteint.

D'autre part, il est plutôt difficile de déterminer des valeurs raisonnables du coefficient d'asymétrie, car de nombreuses autres variables entrent en ligne de compte : la taille de la banque d'items, le niveau d'habileté de l'individu, les paramètres de discrimination des items de la banque, etc. Néanmoins, dans toutes les situations abordées et étudiées, il apparait qu'un coefficient d'asymétrie $a^3$ situé entre -1 et 1 n'est jamais significativement dommageable sur le biais d'estimation, sur la précision ou sur la longueur d'un test. Toutefois, ce constat n'est pas assez soutenu pour en faire une valeur de référence. Plusieurs facteurs sont à considérer et l'étude de ces facteurs dépasse le cadre de cette recherche centrée sur les effets de l'asymétrie. Ce constat ouvre la porte à d'autres recherches visant à suggérer différentes valeurs acceptables de $a^3$ dans différents contextes.

## 6.3 Ouverture, pistes de recherche

En somme, les résultats de la présente recherche démontrent que l'asymétrie peut avoir des effets indésirables sur certaines variables liées à l'administration de tests adaptatifs. Il serait intéressant de pousser l'exercice plus loin en développant des fonctions qui détectent les cas où cette asymétrie est problématique. Ces fonctions pourraient être ajoutées à une procédure qui fait des vérifications diverses sur des banques d'items vouées au testing adaptatif. Dans cette optique, les travaux de Reckase (2007) sur les banques d'items *p*-optimales sont d'un grand intérêt.

Selon Reckase, une banque d'items optimale est toujours en mesure de trouver un item parfaitement ajusté à un individu, quel que soit son niveau d'habileté. Au sein du modèle de Rasch, ce concept de banque optimale se traduit en une banque toujours en mesure de sélectionner et d'administrer un item dont la difficulté correspond à l'estimé provisoire du niveau d'habileté ($b_i = \hat{\theta}_j$). De cette façon, la fonction d'information est toujours maximisée. Cependant, une banque d'items optimale doit pouvoir faire de même à toutes les valeurs de $\hat{\theta}$. En supposant que 20 items parfaitement ajustés suffisent pour un test, la banque devrait contenir 20 items pour chaque point de quadrature dans le spectre de $\theta$. Par exemple, l'intervalle de $\theta$ [-4, 4] divisé en tranches de 0,01 compte 801 points différents, 801 valeurs différentes de $\theta$. Une banque optimale couvrant cet intervalle devrait minimalement contenir 20 x 801 items, soient 16 020 items distribués selon une loi uniforme très précise, une tâche de conception colossale! De plus, comme les paramètres de ces items sont estimés à partir de données réelles et non fixés arbitrairement, il faut réunir assez d'items pour que 16 020 d'entre eux suivent la distribution prescrite, tout en supposant que les individus qui génèrent lesdites données réelles aient accepté de se voir administrer beaucoup plus d'items que ces quelques 16 020!

En fait, bien que l'idée d'une banque d'items optimale soit intéressante, elle relève de l'utopie. Cependant, considérant que la différence en information entre un item optimal et un autre presque optimal est minime, Reckase avance qu'il est possible de considérer plusieurs items comme optimaux s'ils peuvent fournir une proportion $p$ de l'information que donnerait l'item optimal. Par exemple, si $p$ = 0,9, alors tous les items qui peuvent fournir 0,9 x 0,25[14] = 0,225 en information sont dits $p$-optimaux et sont alors éligibles à être choisis comme prochain item pour $\hat{\theta}$. Grâce à cette $p$-optimalité, Reckase a pu diviser les échelles de $\theta$ et de $b$ en segments de largeur fixe

[14] Dans le modèle de Rasch, l'information se calcule en multipliant les probabilités de réussite et d'échec de $\theta$ à un item $i$, et l'information est à son maximum lorsque $b_i = \theta$, soit où P = 0,5 et Q = 0,5.

proportionnelle à $p$ qu'il appelle *bins*. Tous les items d'un *bin* sont dits $p$-optimaux pour les valeurs de $\theta$ à même ce *bin*. Il suffit qu'il y ait suffisamment d'items dans chaque *bin* pour que la banque soit dite $p$-optimale.

Le concept de $p$-optimalité de Reckase sert à mettre en lumière certaines caractéristiques d'une banque d'items vouée au testing adaptatif. Notamment, cette $p$-optimalité permet de déterminer si, à une valeur donnée de $p$, une banque d'items est disposée à estimer adéquatement toute valeur de $\theta$. Pour ce faire, la recherche de Reckase s'intéresse à la distribution des paramètres de difficulté des items d'une banque et, en ce sens, elle rejoint les intérêts de la présente recherche. Il serait intéressant de combiner la vérification de la $p$-optimalité de Reckase au sein d'une banque d'items à une vérification de l'asymétrie dans la distribution des paramètres de difficulté des items de cette même banque. Les croisements des résultats de ces vérifications pourraient donner des informations importantes sur ladite banque d'items, voire des indications sur les correctifs à y apporter advenant que les vérifications la considèrent inadéquate au testing adaptatif. Toutefois, pour ce faire, il faudrait adapter le concept de $p$-optimalité à des modèles plus complexes comme celui de la présente recherche, soit au modèle 3PL. Pour le modèle 3PL, l'information optimale devient alors difficile à mesurer parce qu'elle est fonction de $a$, un paramètre d'item sur une échelle continue, ce qui fait d'elle une variable continue, sans maximum fixe. Comment calculer la $p$-optimalité si l'optimalité ne peut être définie? De plus, les tests adaptatifs n'ont pas toujours un nombre prédéterminé d'items à administrer. Sans information optimale et sans un nombre d'items à administrer, la méthodologie de Reckase est instable et a besoin d'une méthodologie complémentaire pour retrouver un peu d'équilibre. Un croisement avec la présente recherche est donc tout indiqué, d'autant plus que la recherche de Reckase pourrait apporter beaucoup à la présente.

La fonction issue de ce croisement vérifierait si une banque d'items fortement asymétrique est utilisable néanmoins à l'aide du concept de *p*-optimalité de Reckase. Pour ce faire, il faudrait donner en argument à la fonction une valeur de $p$ ainsi que l'erreur-type maximale $S$ et elle retournerait un intervalle de la distribution de $\theta$ identifié *p*-optimal. Comme le modèle 3PL ne permet pas qu'une optimalité « maximale » soit calculée, l'optimalité du modèle de Rasch – $I_i(\theta) = 0{,}25$ – serait utilisée même s'il est possible dans les modèles 2PL et 3PL que l'information soit supérieure à cette valeur grâce au paramètre de discrimination $a$. En supposant que chaque item donne 0,25, il est possible de calculer l'information totale dont un test doit disposer pour atteindre $S$ en passant par les mêmes formules qu'en 5.1.4 :

$$S = \sqrt{\sum I_i(\theta)}^{-1}$$
$$\sqrt{\sum I_i(\theta)} = S^{-1}$$
$$\sum I_i(\theta) = (S^{-1})^2 = S^{-2}$$

En connaissant l'information totale, il est possible de calculer le nombre d'items nécessaire en supposant que chaque item fournisse 0,25 en information :

$$N = \frac{S^{-2}}{0{,}25}$$

Enfin, comme la *p*-optimalité est plus flexible que l'optimalité pure, une proportion $p$ du nombre d'items $N$ peut tenir lieu du nombre d'items $N$.

$$N = p \cdot \frac{S^{-2}}{0{,}25}$$
$$0{,}9 \cdot \frac{0{,}4^{-2}}{0{,}25} = 22{,}5$$

La *p*-optimalité d'un *bin* sera atteinte s'il contient au moins 22,5 items. La procédure pour segmenter l'échelle de $\theta$ en *bins* à partir de l'erreur-type $S$ prédéfinie demeure à déterminer. L'utilisation du modèle 3PL vient compliquer l'opération parce que les courbes d'information des items ne sont pas symétriques vu l'utilisation du paramètre de pseudo-chance $c$. Autrement dit, l'information varie selon qu'un item est réussi ou non : alors, l'information n'est pas prévisible contrairement au modèle de Rasch. Il devient alors difficile d'effectuer un découpage uniforme.

L'intervalle retourné par la fonction définirait pour quels $\theta$ une banque d'items est disposée à générer des tests adaptatifs. Cette information pourrait aider à la conception et à l'entretien d'une banque d'items. Toutefois, cette information ne serait d'aucune utilité une fois un test adaptatif entamé.

Le concept de *p*-optimalité de Reckase est vraiment très intéressant, même si son adaptation au modèle 3PL est plutôt complexe. D'ailleurs, la valeur de $p$ a beaucoup moins d'importance que le paramètre d'item $a$ qui est mis au carré avant de moduler l'information calculée. Une adaptation pourrait être pensée afin de donner davantage de poids à $p$. Néanmoins, un travail de rapprochement entre les travaux de Reckase et la présente recherche est indiqué.

D'autres avenues, qui seraient corolaires ou complémentaires à la présente recherche, pourraient être explorées. Notamment, l'étude des effets de la kurtose dans la distribution des paramètres de difficulté sur différentes variables liées à l'administration de tests adaptatifs, la surexposition d'items dans des banques où la distribution des paramètres de difficulté est asymétrique – comme dans la présente recherche – et l'étude des effets de l'asymétrie en testing adaptatif multidimensionnel sont autant de sujets de recherche intéressants liés à celui de la présente recherche.

# ANNEXE A

## FONCTIONS APPELÉES POUR LA GÉNÉRATION DES BANQUES D'ITEMS ET CODE POUR INITIER LA GÉNÉRATION

```
##############################################################################################
#                                                                                            #
# Fonction pour generer des banques d'items aleatoires selon 3PL et selon les                #
# parametres nu et xi (vs a3)                                                                #
#                                                                                            #
##############################################################################################
genBanq <- function(n, mean=0, sd=1, nu, xi, a3) {
  if(length(nu) != length(xi)) stop("nu and xi must be of the same length")
  if(length(nu) != length(a3)) stop("nu and a3 must be of the same length")
  bank <- NULL
  for (i in 1:length(nu)) {
    b       <- rsstd(n=n, mean=mean, sd=sd, nu=nu[i], xi=xi[i])
    temp    <- genDichoMatrix(items=n, cbControl=NULL, model="3PL",
                              aPrior=c("lnorm", 0, 0.1225), bPrior=c("norm", 0, 1),
                              cPrior=c("beta", 1, 12), seed=1)
    temp$b  <- b
    temp$a3 <- rep(a3[i], n)
    temp$n  <- rep(n, n)
    bank    <- rbind(bank, temp)
  }
  return(bank)
}


##############################################################################################
#                                                                                            #
# Code pour initier la generation des banques d'items                                        #
#                                                                                            #
##############################################################################################
Require(catR)
Require(fGarch)
set.seed(1)
tpar200 <- data.frame(
  nu = c(2.0202, 4.1163325, 5.06, 5, 5.06, 4.1163325, 2.0202),
  xi = c(-5000, -2, 1, 0, -1, 2, 5000),
  a3 = c(-6, -3, -1, 0, 1, 3, 6)
)
set.seed(1)
tpar1000 <- data.frame(
  nu = c(2.05534, 4.20715, 9, 5, 9, 4.20715, 2.05534),
  xi = c(-5000, -100, -1.5, 0, 1.5, 100, 5000),
  a3 = c(-6, -3, -1, 0, 1, 3, 6)
)
set.seed(1)
bank_200  <- genBanq(n=200, mean=0, sd=1, nu=tpar200$nu,
                     xi=tpar200$xi, a3=tpar200$a3)
set.seed(1)
bank_1000 <- genBanq(n=1000, mean=0, sd=1, nu=tpar1000$nu,
                     xi=tpar1000$xi, a3=tpar1000$a3)
```

# ANNEXE B

# FONCTIONS APPELÉES POUR LA GÉNÉRATION DES PASSATIONS DE TESTS ADAPTATIFS ET CODE POUR INITIER LA GÉNÉRATION DES STATISTIQUES

```
##############################################################################################
#                                                                                            #
# Fonction pour generer les statistiques, selon theta a partir d'un objet de la              #
# classe catResult                                                                           #
#                                                                                            #
##############################################################################################
stats_res <- function(x) {
 if (!class(x) == "catResult") stop("x is not from the class catResult.")
 x       <- x$final.values.df
 x$biais <- x$true.theta - x$estimated.theta
 tmean   <- tapply(x$estimated.theta, x$true.theta, mean, na.rm=TRUE)
 tsd     <- tapply(x$estimated.theta, x$true.theta, sd,   na.rm=TRUE)
 trange  <- tapply(x$estimated.theta, x$true.theta, range)
 tmed    <- tapply(x$estimated.theta, x$true.theta, median, na.rm=TRUE)

 bmean   <- tapply(x$biais, x$true.theta, mean, na.rm=TRUE)
 bsd     <- tapply(x$biais, x$true.theta, sd,   na.rm=TRUE)
 brange  <- tapply(x$biais, x$true.theta, range)
 bmed    <- tapply(x$biais, x$true.theta, median, na.rm=TRUE)

 smean   <- tapply(x$final.SE, x$true.theta, mean, na.rm=TRUE)
 ssd     <- tapply(x$final.SE, x$true.theta, sd,   na.rm=TRUE)
 srange  <- tapply(x$final.SE, x$true.theta, range)
 smed    <- tapply(x$final.SE, x$true.theta, median, na.rm=TRUE)

 nmean   <- tapply(x$total.items.administrated, x$true.theta, mean, na.rm=TRUE)
 nsd     <- tapply(x$total.items.administrated, x$true.theta, sd,   na.rm=TRUE)
 nrange  <- tapply(x$total.items.administrated, x$true.theta, range)
 nmed    <- tapply(x$total.items.administrated, x$true.theta, median, na.rm=TRUE)

 res <- data.frame(theta=unique(x$true.theta), mean=tmean, median=tmed, sd=tsd,
                   max=trange[[1]], min=trange[[2]],
                   meanb=bmean, medianb=bmed, sdb=bsd,
                   maxb=brange[[1]], minb=brange[[2]],
                   means=smean, medians=smed, sds=ssd,
                   maxs=srange[[1]], mins=srange[[2]],
                   meann=nmean, mediann=nmed, sdn=nsd,
                   maxn=nrange[[1]], minn=nrange[[2]])
 return(res)
}
```

```
##############################################################################################
#                                                                                            #
# Fonction pour generer les statistiques, selon theta a partir de fichiers .Rdata            #
# prealablement sauvegardes. Chacun des fichiers .Rdata a ete cree par la fonction           #
# genSimulation. L'objet dans lequel les resultats sont sauvegardes est toujours             #
# nomme res.                                                                                 #
#                                                                                            #
##############################################################################################
stats_res_from_file <- function(nitems){
  inputFiles <- as.vector(dir()[grep(".Rdata", dir())])
  if (is.null(inputFiles)) stop("No results files satisfy the criteria.")
  index      <- eval(parse(text=paste("grep(as.character(", nitems, "),
                    inputFiles)", sep="")   ))
  if (length(index)==0) stop("No results files satisfy the criteria.")
  inputFiles <- as.vector(inputFiles[index])
  x          <- NULL
  for( i in inputFiles){
   load(i)
   a3    <- strsplit(i, ".Rdata")[[1]][1]
   a3    <- strsplit(a3, "a3")[[1]][2]
   temp1 <- res
   temp2 <- data.frame(stats_res(temp1), a3=a3)
   x     <- rbind(x, temp2)
  }
  rownames(x) <- 1:dim(x)[1]
  return(x)
 }


##############################################################################################
#                                                                                            #
# Fonction pour les simulations des patrons de reponses. Creation d'un objet de              #
# classe catResult                                                                           #
#                                                                                            #
##############################################################################################
genSimulation <- function(theta, bank, start=list(nrItems=1, theta=0),
                          test=list(method="WL", itemSelect="MFI"),
                          stop=list(rule="precision", thr=0.4),
                          final=list(method="WL"), file=NULL){
  for (i in 1:length(theta)) {
    res <- simulateRespondents(theta, bank, maxItems = dim(bank)[1],
                               start = start, test = test,
                               stop = stop, final = final)
  }
  if(!is.null(file)) save(res, file=file)
  return(res)
}
```

```
#############################################################################################
#                                                                                           #
# Code pour initier la generation des statistisques pour les banques de 1000 items          #
#                                                                                           #
#############################################################################################
nSimulation <- 1000
theta       <- rep(c(-4,-3,-2,-1,0,1,2,3,4), nSimulation)
skew        <- c(-6,-3,-1,0,1,3,6)
skewLabels  <- c("M6", "M3", "M1", "P0", "P1", "P3", "P6")
bank        <- bank_1000
x           <- NULL
set.seed(1)
for (a3 in skew) {
 print(a3)
 resFile <- paste("items", dim(bank)[1]/length(unique(bank$a3)), "a3",
                  skewLabels[which(skew==a3)] , ".Rdata", sep="")
 temp1   <- genSimulation(theta=theta, bank=bank[which(bank$a3==a3), -c(5,6)],
                          file=resFile)
 temp2   <- data.frame(stats_res(temp1), a3=a3)
 x       <- rbind(x, temp2)
 }
rownames(x) <- 1:dim(x)[1]
x


#############################################################################################
#                                                                                           #
# Code pour initier la generation des statistisques pour les banques de 200 items           #
#                                                                                           #
#############################################################################################
nSimulation <- 1000
theta       <- rep(c(-4,-3,-2,-1,0,1,2,3,4), nSimulation)
skew        <- c(-6,-3,-1,0,1,3,6)
skewLabels  <- c("M6", "M3", "M1", "P0", "P1", "P3", "P6")
bank        <- bank_200
x           <- NULL
set.seed(1)
for (a3 in skew) {
  print(a3)
  resFile <- paste("items", dim(bank)[1]/length(unique(bank$a3)), "a3",
                   skewLabels[which(skew==a3)] , ".Rdata", sep="")
  temp1   <- genSimulation(theta=theta, bank=bank[which(bank$a3==a3), -c(5,6)],
                           file=resFile)
  temp2   <- data.frame(stats_res(temp1), a3=a3)
  x       <- rbind(x, temp2)
}
rownames(x) <- 1:dim(x)[1]
x


#############################################################################################
#                                                                                           #
# Code a utiliser lorsque les donnees ont ete prealablement produites avec                  #
# genSimulation                                                                             #
#                                                                                           #
#############################################################################################
stats_res_from_file(200)
stats_res_from_file(1000)
```

# ANNEXE C

## CODE POUR LA GÉNÉRATION DES GRAPHIQUES

```
##############################################################################################
#                                                                                            #
# Figure 2.1                                                                                 #
#                                                                                            #
##############################################################################################
item  <- cbind(1,0,0,1)
x     <- seq(-4,4,by=0.01)
y     <- Pi(x,item,D=1.7)$Pi
plot(x,y,type="l",xlab=bquote(paste("Niveau d'habilete (",theta,")")),
     ylab="Probabilite de reussite",main=titre)

# z est un vecteur qui contient les valeurs de theta ou des points dont desires
z  <- c(-2,0,1.5)
# Pour placer les valeurs de theta dans un vecteur et arrondir a deux decimales
yy <- round(Pi(z,cbind(1,0,0,1),D=1.7)$Pi,2)

for(i in 1:length(z)) {
 points(z[i],yy[i],pch=21,bg="black")
 text(x=z[i], y=yy[i],
      labels=bquote(theta[.(LETTERS[i])] * " (P = " * .(yy[i]) * ")"), adj=c(0,1))
 abline(,,,z[i],col = "black",lty = 3)
}


##############################################################################################
#                                                                                            #
# Figure 2.2                                                                                 #
#                                                                                            #
##############################################################################################
item <- matrix(c(1,1,1,1,1,-2,-1,0,1,2,0,0,0,0,0,1,1,1,1,1),5,4)
x    <- seq(-4,4,by=0.01)
y    <- Pi(x,item,D=1.7)$Pi
z    <- -0.5
plot(x,y,type="l",xlab=bquote(paste("Niveau d'habilete (",theta,")")),
     ylab="Probabilite de reussite",col="white")
for(i in 1:length(item[,1])) {
  y    <- Pi(x,matrix(item[i,],1,4),D=1.7)$Pi
  yy[i]   <- round(Pi(z,matrix(item[i,],1,4),D=1.7)$Pi,2)
  lines(x,y,col="black")
  points(z,yy[i],pch=21,bg="black")
  text(x=z, y=yy[i], labels=bquote(theta[D] * " (P = " * .(yy[i]) * ")"), adj=c(0,1))
}
abline(,,,z,col="black",lty=3)
```

```
#############################################################################################
#                                                                                           #
# Figure 2.3                                                                                #
#                                                                                           #
#############################################################################################
item  <- matrix(c(1,0.25,2,0,0,0,0,0,0,1,1,1),3,4)
x     <- seq(-4,4,by=0.01)
y     <- Pi(x,matrix(item[1,],1,4),D=1.7)$Pi
z     <- c(-2,0,1.5)
plot(x,y,type="l",xlab=bquote(paste("Niveau d'habilete (",theta,")")),
     ylab="Probabilite de reussite",col="white")
for(j in 1:length(item[,1])){
  y    <- Pi(x,matrix(matrix(item[j,],1,4),1,4),D=1.7)$Pi
  lines(x,y,col="black")
  for(i in 1:length(z)){
    yy[i] <- round(Pi(z[i],matrix(item[j,],1,4),D=1.7)$Pi,2)
    points(z[i],yy[i],pch=21,bg="black")
    text(x=z[i],y=yy[i],
         labels=bquote(theta[.(LETTERS[i])] * " (P = " * .(yy[i]) * ")"), adj=c(0,1))
  }
}
abline(,,,z,col="black",lty=3)


#############################################################################################
#                                                                                           #
# Figure 2.4                                                                                #
#                                                                                           #
#############################################################################################
item  <- matrix(c(1,1,1,0,0,0,0,0.1,0.25,1,1,1),3,4)
x     <- seq(-4,4,by=0.01)
y     <- Pi(x,matrix(item[1,],1,4),D=1.7)$Pi
z     <- c(-2,0,1.5)
c_adj <-c(0,1.25,0,0.25,0,-0.75)
plot(x,y,type="l",xlab=bquote(paste("Niveau d'habilete (",theta,")")),
     ylab="Probabilite de reussite",col="white")
for(j in 1:length(item[,1])){
  y    <- Pi(x,matrix(matrix(item[j,],1,4),1,4),D=1.7)$Pi
  lines(x,y,col="black")
  for(i in 1:length(z)){
    yy[i] <- round(Pi(z[i],matrix(item[j,],1,4),D=1.7)$Pi,2)
    points(z[i],yy[i],pch=21,bg="black")
  }
  for(k in 1:2){
    text(x=z[k],y=yy[k],
         labels=bquote(theta[.(LETTERS[k])] * " (P = " * .(yy[k]) * ")"), adj=c(0,1))
  }
  text(x=z[3],y=yy[3],
       labels=bquote(theta[.(LETTERS[3])] * " (P = " * .(yy[3]) * ")"),
                     adj=c_adj[((j*2)-1):(j*2)])
}
abline(,,,z,col="black",lty=3)
```

```
##############################################################################################
#                                                                                            #
# Figure 2.5                                                                                 #
#                                                                                            #
##############################################################################################
item <- matrix(c(1.5,1.2,1.1,1,0.4,1.7,0.03,0.1,0.18,1,1,1),3,4)
z    <- 1.5
u    <- c(1,0,1)
u2   <- c(1,1,1)
x    <- seq(-4,4,by=0.01)
y2   <- seq(-4,4,by=0.01)
for(i in 1:length(x)){
  y[i]  <- sum(u*log(Pi(x[i],item,D=1.7)$Pi) + (1-u)*log(1-Pi(x[i],item,D=1.7)$Pi))
  y2[i] <- sum(u2*log(Pi(x[i],item,D=1.7)$Pi) + (1-u2)*log(1-Pi(x[i],item,D=1.7)$Pi))
}
plot(x,y,type="l",xlab=bquote(paste("Niveau d'habilete (",theta,")")),
     ylab="Log-vraisemblance")
lines(x,y2,col="black")
points(1.18,-3.26,pch=21,bg="black")
text(x=1.18,y=-3.26,labels=bquote(theta["Max = 1.18"]), adj=c(0,1))


##############################################################################################
#                                                                                            #
# Figure 2.6                                                                                 #
#                                                                                            #
##############################################################################################
x  <- rsnorm(10000, mean = 0, sd = 1, xi = -2.5)
sk <- skewness(x)
hist(x, xlim=c(-4,4), breaks=11,
     probability=T,
     col='light grey', xlab='x', ylab='Densité', axes=T,
     main= paste('Asymétrie négative : ', round(sk,2)))
lines(density(x), col='black', lwd=2)


##############################################################################################
#                                                                                            #
# Figure 5.2                                                                                 #
#                                                                                            #
##############################################################################################
hist(bank_200[which(bank_200$a3==6),2],   breaks=10, col="grey",
                                          xlab="Parametre de difficulte d'item (b)",
                                          ylab = "Frequence",
                                          main="200 items")
hist(bank_1000[which(bank_1000$a3==6),2], breaks=10, col="grey",
                                          xlab="Parametre de difficulte d'item (b)",
                                          ylab = "Frequence",
                                          main="1000 items")



##############################################################################################
#                                                                                            #
# Pour effacer les figures de la memoire                                                     #
#                                                                                            #
##############################################################################################
dev.off(dev.list()["RStudioGD"])
```

# RÉFÉRENCES


Anzaldua, R. M. (2002). *Item banks : what, where, why and how*. Paper presented at the 25th Annual meeting of the Southwest educational research association. Austin, Texas: Southwest educational research association.

Baker, F. B. (2001). *The basics of item response theory* (2nd edition). College Park, Maryland: ERIC clearinghouse on assessment and evaluation.

Baker, F. B. et Kim, S. H. (Eds). (2004). *Item response theory: parameter estimation techniques*. Boca Raton, Florida: CRC press.

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

Birnbaum, A. (1957). *Efficient design and use of tests of a mental ability for various decision-making problems*. Series Report No 58-16. Project No 7755-23. San Antonio, Texas: USAF School of aerospace medecine.

Birnbaum, A. (1958a). *On the estimation of mental ability*. Series Report No 15. Project No 7755-23. San Antonio, Texas: USAF School of aerospace medecine.

Birnbaum, A. (1958b). *Further considerations of efficiency in tests of a mental ability*. Technical Report No 17. Project No 7755-23. San Antonio, Texas: USAF School of aerospace medecine.

Birnbaum, A. (1968). Some latent train models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, 395-479.

Bjorner, J. B., Chang, C. H., Thissen, D. et Reeve, B. B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of life research*, *16*(1), 95-108.

Bjorner, J. B.. Kosinski, M. et Ware Jr, J. E. (2005). Computerized adaptive testing and item banking. Dans P. Fayers et R. Hays (Eds), *Assessing quality of life in clinic trials*. Oxford: Oxford University Press.

Bock, R. D. (1997). A brief history of item response theory. *Educational measurement : Issues and practice, 14*(4), 21-33.

Bode, R. K., Lai, J. S., Cella, D. et Heinemann, A. W. (2003). Issues in the development of an item bank. *Archives of physical medicine and rehabilitation, 84*(Supplément 2), S52-S60.

Borsboom, D., Mellenbergh, G. J. et Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological review, 110*(2), 203-219.

Brassard, P., Béland, S. et Raîche, G. (2011). *Identification des patrons de réponses inappropriés à un test à partir des strategies qui sous-tendent les comportements des répondants*. Dans G. Raîche, K. Paquette-Côté et D. Magis (Eds), Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation: Volume 1 – La mesure. Québec, Québec: Presses de l'Université du Québec.

Brown, J. M. et Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries*. Research Report 77-6. Minneapolis, Minnesota: University of Minnesota – Psychometric methods Program.

Camilli, G. (1994). Teacher's corner: origin of the scaling constant d=1.7 in item response theory. *Journal of educational and behavorial statistics, 19*(3), 293- 295.

Chen, S.-K., Hou, L. et Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and psychological measurement, 58*(4), 569-595.

Davey, T. et Parshall, C. (1995). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, California: American Educational Research Association.

Flaugher, R. (1990). Item pools. dans H. Wainer (Ed.), *Computer adaptive testing: a primer*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Faglioni, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D. et Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatr Serv, 59*(4), 361-368.

Gibbs, J. P. (1975). *Crime, punishment and deterrence*. New-York, New-York: Elsevier North-Holland.

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

Hambleton, R. K., Swaminathan, H. J. et Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, California: Sage Publications.

Hau, K.-T. et Chang, H.-H. (2001). Item selection in computerized adaptive testing : Should more discriminating items be used first?. *Journal of educational measurement, 39*(3), 249-266.

Holland, P. W. et Wainer, H. (1993). *Differential item functioning*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Huang, C. D., Church, A. T. et Katigbak, M. S. (1997). Identifying cultural differences in items and traits. *Journal of cross-cultural psychology, 28*(2), 192-218.

IACAT - International Association for Computerized Adaptive Testing. (2013). *Operational CAT programs*. Récupéré le 2 avril 2016 de http://www.iacat.org/content/operational-cat-programs

Jeffreys, H. (1943). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, 186*, 453-461.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied measurement in education, 16*(4), 277- 298.

Linacre, J.M. (2009). The Efficacy of Warm's Weighted Mean Likelihood Estimate (WLE) Correction to Maximum Likelihood Estimate (MLE) bias. *Rasch Measurement Transactions*, *23*(1), 1188-1189.

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

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and psychological measurement, 13*(4), 517-549.

Lord, F. M. et Novick, M. R. (1968). *Statistical theories of mental test scores*. Charlotte, North Carolina: Information age publishing.

Luecht, R. M. (1998). Test assembly using optimization heuristics. *Applied psychological measurement, 22*(3), 224-236.

Magis, D. (2015). A note on weighted likelihood and Jeffrey's modal estimation of proficiency levels in polytomous response models. *Psychometrika, 80*(1), 200-204.

Magis, D., De Boeck, P. et Raîche, G. (2011). Comparaison empirique des méthodes classiques de détection du fonctionnement différentiel d'items en psychométrie. Dans G. Raîche, K. Paquette-Côté et D. Magis (Eds), *Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation : Volume 1 – La mesure*. Québec, Québec: Presses de l'Université du Québec.

Magis, D. et Raîche, G. (2012). On the relationships between Jeffreys modal and weighted likelihood estimation of ability under logistic IRT models. *Psychometrika, 77*(1), 163-169.

Magis, D., Raîche, G. et Barrada, J. R. (2015). *catR 2.0 - Application destinée à la simulation de tests adaptatifs*. Montréal, Québec : Université du Québec à Montréal.

Martin, R. (2003). Le testing adaptatif par ordinateur dans la mesure en éducation : potentialités et limites. *Psychologie et psychométrie, 24*(2-3), 89-116.

Mead, A. D. et Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: a meta-analysis. *Psychological bulletin, 114*(3), 449-458.

Mislevy, J. L., Rupp, A. A. et Harring, J. R. (2012). Detecting local item dependence in polytomous adaptive data. *Journal of educational measurement, 49*(2), 127-147.

Raîche, G. et Blais, J.-G. (2001). Étude de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt. *Mesure et évaluation en éducation, 24*(2-3), 23-39.

Raîche, G. et Blais, J.-G. (2002). *Practical considerations about expected a posteriori estimation in adaptive testing : Adaptive a priori, adaptive correction*

*for bias, and adaptive integration interval*. Paper presented at the biennial meeting of the International Objective Measurement Workshop. New Orleans, Louisiana : International Objective Measurement Workshop.

Raîche, G., Magis, D., Blais, J.-G. et Brochu, P. (2012). Taking atypical response patterns into account: a multidimensional measurement model from item response theory. Dans M. Simon, K. Ercykan et M. Rousseau (Eds), *Improving large-scale assessment in education. Theory, issues and practice* (p. 238-259). New York, New York: Routledge.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhague, Danemark: Danish Institute for Educational Research.

Reckase, M. D. (2003). *Item pool design for computerized adaptive tests*. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, Illinois: National Council on Measurement in Education.

Reckase, M. D. (2007). The design of p-optimal item pools for computerized adaptive tests. Dans D. J. Weiss (Ed.), *Proceedings of the 2007 Graduate Management Admission Council Conference on computerized adaptive testing*. Minneapolis, Minnesota: Graduate Management Admission Council.

Reckase, M. D. (2009). *Multidimensional item response theory*. New-York, New York: Springer.

Reckase, M. D. (2010). Designing item pools to optimize the functioning of a computerized adaptive test. *Psychological test and assessment modeling, 52*(2), 127-141.

Samejima, F. (1973). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38*(2), 221-233.

Schmitt, A. P., Holland, P. W. et Dorans, N. J. (1992). *Evaluating hypotheses about differential item functioning*. Paper presented at the Educational Testing Service/AFHRL Conference. Princeton, New Jersey: Educational Testing Service.

Stocking, M. L. et Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied psychological measurement, 17*(3), 277-292.

Sutton, R. E. (1993). *Equity issues in high stakes computerized testing.* Paper presented at the 1993 annual meeting of the American Educational Research Association. Atlanta, Geogia: American Educational Research Association.

Swaminathan, H. et Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of educational statistics, 7*(3), 175-192.

Swaminathan, H. et Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50*(3), 349-364.

Swaminathan, H. et Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*(4), 589-601.

Sympson, J. B. et Hetter, R. D. (1985). *Controlling item-exposure rates in computerized adaptive testing.* Paper presented at the 27th annual meeting of the Military Testing Association. San Diego, California: Navy Personnel Research and Development Center.

Thissen, D. (1990). Reliability and measurement. Dans H. Wainer (Ed.), *Computer adaptive testing: a primer.* Mahwah, New Jersey: Lawrence Erlbaum associates.

Timminga, E. (1998). Solving infeasibility problems. *Applied psychological measurement, 22*(3), 280-291.

Tuerlinckx, F. et De Boeck, P. (2005). Two interpretations of the discrimination parameter. *Psychometrika, 70*(4), 629-650.

Van der Linden, W. J. (2000). Chapter 2. Constrained adaptive testing with shadow tests. Dans W. J. Van der Linden et C. A. W. Glas (Eds), *Computerized adaptive testing: Theory and practice.* Dordrecht: Kluwer.

Van der Linden, W. J. et Glas, C. A. W. (2007). Statistical aspects of adaptive testing. Dans C. R. Rao et S. Sinharay (Eds), *Handbook of statistics 26.* Amsterdam: Elsevier.

Veldkamp, B. P. et Matteucci, M. (2013). Bayesian computerized adaptive testing. *Ensaio, 21*(78), 57-81.

Veldkamp, B. P., Verschoor, A. J. et Eggen, T. J. (2010). A multiple objective test assembly approach for exposure control problems in computerized adaptive testing. *Psicológica, 31*(2), 335-355.

Wainer, H. (1990). *Computer adaptive testing: a primer*. Mahwah, New Jersey: Lawrence Erlbaum associates.

Wainer, H. (2000). CATs: Whither and whence. *Psicologica, 21*(1-2), 121-133.

Wainer, H. et Lewis, C. (1989). *Toward a psychometrics for testlets*. Princeton, New Jersey: Educational testing service.

Wang, T. (1997). *Essentially unbiased EAP estimates in computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association. Chicago, Illinois: American Educational Research Association.

Wang, T. et Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of educational measurement, 35*(2), 109-135.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427-450.

Wuertz, D. et Chalabi, Y. (2013). *fGarch 3010.82 – Rmetrics – Autoregressive conditional heteroskedastic modelling*. Zurich, Switzerland: Rmetrics publishing.

Zhang, B. (2008). Application of unidimensional item response models to tests with items sensitive to secondary dimensions. *The journal of experimental education, 77*(2), 147-166.

Zwick, R., Donoghue, J. R. et Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of educational measurement, 30*(3), 233-251.

Zwick, R., Thayer, R. et Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied psychological measurement, 18*(121), 121-140.