Richard Bertrand
Jean-Guy Blais

Modèles de mesure

L'apport de la théorie des réponses aux items

Presses de l'Université du Québec

Modèles de mesure

L'apport de la théorie des réponses aux items

PRESSES DE L'UNIVERSITÉ DU QUÉBEC
Le Delta I, 2875, boulevard Laurier, bureau 450
Sainte-Foy (Québec) G1V 2M2
Téléphone : (418) 657-4399 • Télécopieur : (418) 657-2096
Courriel : puq@puq.ca • Internet : www.puq.ca

Distribution :

CANADA et autres pays

DISTRIBUTION DE LIVRES UNIVERS S.E.N.C.
845, rue Marie-Victorin, Saint-Nicolas (Québec) G7A 3S8
Téléphone : (418) 831-7474 / 1-800-859-7474 • Télécopieur : (418) 831-4021

FRANCE

DISTRIBUTION DU NOUVEAU MONDE
30, rue Gay-Lussac, 75005 Paris, France
Téléphone : 33 1 43 54 49 02
Télécopieur : 33 1 43 54 39 15

SUISSE

SERVIDIS SA
5, rue des Chaudronniers, CH-1211 Genève 3, Suisse
Téléphone : 022 960 95 25
Télécopieur : 022 776 35 27

MODÈLES DE MESURE

L'apport de la théorie des réponses aux items

Richard Bertrand
Jean-Guy Blais

Avec la collaboration de
Gilles Raîche

2004

Presses de l'Université du Québec
Le Delta I, 2875, boul. Laurier, bur. 450
Sainte-Foy (Québec) Canada G1V 2M2

Catalogage avant publication de la Bibliothèque nationale du Canada

Bertrand, Richard, 1949-

Modèles de mesure : l'apport de la théorie des réponses aux items

Comprend des réf. bibliogr.

ISBN 2-7605-1103-0

1. Analyse d'items, Technique de l'. 2. Sciences sociales – Modèles mathématiques. 3. Sciences sociales – Méthodes statistiques. 4. Généralisabilité. 5. Tests et mesures en éducation. 6. Psychométrie. I. Blais, Jean-Guy, 1954- . II. Titre.

H61.25.B47 2004 300'.1'5118 C2004-940262-5

Nous reconnaissons l'aide financière du gouvernement du Canada par l'entremise du Programme d'aide au développement de l'industrie de l'édition (PADIÉ) pour nos activités d'édition.

Mise en pages : INFO 1000 MOTS INC

Illustrations : MARYLÈNE BERTRAND

Couverture : RICHARD HODGSON

1 2 3 4 5 6 7 8 9 PUQ 2004 9 8 7 6 5 4 3 2 **1**

Dépôt légal – 2e trimestre 2004
Bibliothèque nationale du Québec / Bibliothèque nationale du Canada
Imprimé au Canada

Table des matières

PARTIE 1

CONCEPTS ET MÉTHODES

Introduction

Dans l'histoire de l'humanité, la mesure et la quantification ont figuré parmi les préoccupations fondamentales de l'homme ; elles constituent des quêtes aussi anciennes que le monde civilisé lui-même. L'élaboration de stratégies, de techniques et d'instruments standardisés permettant d'effectuer des relevés, des prévisions et des comparaisons ne date en effet pas d'hier. Pensons au fameux nilomètre du temps des Pharaons qui servait à prévoir les crues du Nil et contribuait à la planification des cultures sur les berges du grand fleuve. En l'an 1115 avant J.-C., en Chine, la dynastie des Chan procédait à la sélection des futurs bureaucrates à la faveur d'un concours où les candidats étaient sélectionnés en fonction de leurs résultats à une batterie de tests standardisés. Citons aussi les tentatives de nombreux gouvernants à travers l'histoire qui ont voulu recenser les populations, compter et classer les gens, afin d'en établir un portrait utile, c'est-à-dire principalement pour s'assurer que tous payaient les impôts. Les civilisations du passé qui furent stables assez longtemps ont ainsi en commun d'avoir mis au point et utilisé des procédures mathématiques et des instruments de mesure qui contribuèrent au développement de l'astronomie, de la comptabilité, de l'architecture et de la gestion de l'État.

Ainsi, la mesure et la quantification se sont introduites lentement mais sûrement dans toutes les sphères de l'activité humaine. Les sciences sociales ne sont pas en reste. Mais alors que la mesure et la quantification l'ont nettement emporté dans les sciences de la nature, où elles constituent le paradigme dominant du rapport à la connaissance, elles rencontrent des terrains d'application dans les sciences sociales qui engendrent des difficultés mettant en relief une réussite assez relative. La nature même des objets d'étude respectifs illustre bien les difficultés rencontrées. Par exemple, quand avons-nous vu la dernière fois un électron refuser de participer à une expérience, tenter de dissimuler la vérité ou ne pas faire preuve de motivation ?

Les mesures de notre quotidien et une bonne partie de celles effectuées en sciences de la nature sont réalisées au moment où l'instrument de mesure est mis en application. Le plus souvent, dans les sciences sociales et en éducation, la mesure ne se passe pas lorsque nous utilisons l'instrument ; elle survient après, lorsque nous analysons les données issues de la rencontre de sujets, répondants ou individus avec des questions, des énoncés, des images, etc. En effet, les unités de mesure dans les sciences sociales ne possédant pas l'attribut de la stabilité dans le temps et l'espace, elles doivent être contextualisées pour prendre forme et avoir du sens. Par exemple, un test d'intelligence développé pour les enfants ne donnera pas une mesure très valide avec des adultes, alors qu'un mètre donnera toujours une mesure avec le même degré de validité et de fidélité, peu importent les circonstances de son utilisation. Si la mesure survient au moment où nous analysons les données, il est nécessaire de se doter de balises pour déterminer quelles seraient les stratégies d'analyse les plus appropriées. Il s'agit du propos central de ce livre : déterminer quelles sont les stratégies de modélisation de la mesure qui peuvent être utiles aux chercheurs des sciences sociales et, dans une perspective pragmatique, comment les utiliser avec les avantages et les limites qu'on leur reconnaît actuellement. Parmi les différentes perspectives de modélisation de la mesure en sciences sociales, nous nous attarderons aux modèles de la théorie classique des tests, de la théorie de la généralisabilité et, surtout, de la théorie des réponses aux items (TRI).

Plusieurs volumes dédiés aux modèles de la théorie des réponses aux items ont été produits à ce jour. Les textes d'Embretson et Reise (2000), Thissen et Wainer (2001), Bond et Fox (2001), van der Linden et Glas (2000), McDonald (1999), van der Linden et Hambleton (1997), Baker (1992), Hambleton *et al.* (1991), Hambleton et Swaminathan (1985), Hulin *et al.* (1983) ou Lord (1980) en témoignent. Notre objectif n'est pas simplement d'ajouter un autre titre à cette liste déjà longue, encore que le besoin soit manifestement plus pressant pour le public francophone. Nous avons surtout voulu offrir aux consommateurs de mesures, notamment à ceux œuvrant en psychologie et en éducation, les bases des modèles les plus utilisés. Notre intention est de faire de ce volume autant un guide d'apprentissage qu'un manuel de référence. Il nous a semblé important, voire crucial, d'employer un

langage épuré autant que possible du jargon technique, si naturel en ce domaine, de manière à élargir le plus possible le spectre des lecteurs potentiels. Bien que des efforts aient déjà été consentis (Warm, 1978 ; Baker, 1985 ; Hambleton *et al.*, 1991) pour rendre accessibles les modèles de la théorie des réponses aux items, nous sommes d'avis qu'il y a un urgent besoin de présenter ces modèles sous la forme d'un manuel scolaire de niveau universitaire. Nous entendons par là un texte donnant les origines des modèles de mesure, comparant ces modèles, montrant les conditions d'utilisation et les principales applications à l'aide d'exemples provenant de l'éducation et de la psychologie. Même si notre intention est de nous attarder surtout aux modèles de la théorie des réponses aux items, nous désirons également comparer ces modèles à ceux de la théorie classique et à ceux de la théorie de la généralisabilité. En effet, nous ne croyons pas qu'il faille toujours employer les modèles de la TRI. Il existe des situations (faibles ressources, plans complexes, etc.) où il s'avère nécessaire de considérer d'autres modèles, comme ceux présentés dans le cadre de la théorie classique ou de la théorie de la généralisabilité. L'accent mis sur les modèles de la TRI est cependant justifié du fait qu'il existe déjà en français des volumes très accessibles consacrés soit aux rudiments de la théorie classique (p. ex., Laveault et Grégoire, 2002), soit à ceux de la théorie de la généralisabilité (p. ex., Bain, 1996 ; Cardinet et Tourneur, 1985). Très peu de volumes présentent et analysent les avantages et les inconvénients des modèles de la théorie classique, de la théorie de la généralisabilité et de la théorie des réponses aux items. Sirotnik (1987, p. 41), il est vrai, souligne avec pertinence que la théorie classique et la théorie de la généralisabilité sont des théories visant la justesse[1] de la mesure et donc basées sur la réplication de la mesure plutôt que sur l'étalonnage comme l'est la TRI, qui vise plutôt la précision que la justesse. Nous voulons aller plus loin et discuter de la pertinence d'utiliser des modèles provenant de l'une ou l'autre de ces trois théories en fonction de la situation de mesure à l'étude.

Sans vouloir minimiser l'apport des modèles de la théorie classique ou de la théorie de la généralisabilité, notre volonté de nous concentrer sur les modèles de la TRI se justifie également par l'extraordinaire vivacité entourant, au cours des dernières années, soit le développement de nouveaux modèles, soit la mise au jour de nouvelles applications. Ces nouveaux développements et ces nouvelles applications sont malheureusement présentés la plupart du temps d'une façon très technique et, du coup, passent souvent inaperçus aux yeux des consommateurs de mesures. Pourtant, les conséquences de ces développements sont considérables et, parfois même, remettent en question ou redéfinissent purement et simplement des notions classiques presque centenaires (p. ex., fidélité, discrimination).

1. Traduction libre d'*accuracy*.

À l'aube du troisième millénaire, il sera de plus en plus important d'expliquer la TRI de façon à ce que, comme le souligne Goldstein (1994), ses concepts soient accessibles aux *outsiders*. Cet auteur conclut son exposé, qui discute notamment du mauvais usage des modèles de la TRI, en stipulant : « Je perçois la démystification de la théorie des réponses aux items comme un pas dans la bonne direction[2]. » D'autant mieux que de plus en plus de personnes non spécialistes de la mesure mais se devant d'utiliser ou d'interpréter des mesures (avocats, juges, administrateurs, enseignants) s'impliquent dans les décisions de mesure. Linn (1989) soutient que « […] le testing a souvent fait l'objet de controverses dans le public […] les futurs acteurs du débat seront probablement des juges, des législateurs et des organismes à caractère administratif[3] […] » et Goldstein (1994) reprend : « Ce sera néanmoins intéressant de constater les conséquences des demandes toujours plus nombreuses de personnes non spécialisées dans notre domaine pour une plus grande ouverture, une meilleure prise en compte de nos responsabilités et une plus grande capacité à expliquer les procédures les plus ésotériques de façon à ce qu'elles soient accessibles à un public non spécialisé[4] ». Plusieurs années auparavant, Nunally (1978, p. 318) stipulait déjà : « La théorie des courbes caractéristiques d'items (théorie des réponses aux items) est très difficile à comprendre pour plusieurs personnes à cause de son caractère hautement mathématique[5]. » Fred Lord lui-même, dans la préface à son volume (Lord, 1980) portant sur la théorie des réponses aux items, insiste : « Les critiques vont sentir le besoin de recommander un livre sur la TRI qui n'exige pas le niveau de compréhension mathématique requis ici. Un tel besoin est légitime ; ce genre de livre sera bientôt rédigé, par d'autres[6] […]. »

Doit-on aller des observations au modèle ou vice-versa ? Est-ce que ce sont les données qui doivent s'ajuster à un modèle mathématique préétabli (à la manière de Procruste !) comme c'est le cas pour un modèle de la TRI, le modèle de Rasch, ou plutôt le modèle mathématique qui doit s'ajuster aux données, qui doit être choisi en fonction des observations, comme c'est le cas avec l'approche empirique des autres tenants de la TRI ? Il s'agit bien sûr d'une question piège puisque les deux cas de figure sont légitimes. Cela dépend

2. Traduction libre de : « *I perceive the demystification of item response 'theory' as a step in the right direction.* »
3. Traduction libre de : « […] *testing has frequently been a subject of public controversy* […] *the actors in the debates are much more likely to be judges, legislators and administrative agencies* […] ».
4. Traduction libre de : « *It will be interesting, nevertheless, to see the results of the increasing demands from outside the profession for some more openness, accountability and explanation of some of our more arcane procedures in terms which outsiders are able to understand.* »
5. Traduction libre de : « *ICC Theory is highly mathematical and thus difficult for many persons to understand.* »
6. Traduction libre de : « *Reviewers will urge the need for a book on item response theory that does not require the mathematical understanding required here. There is such a need ; such books will be written soon, by other authors* […]. »

de notre point de vue sur la question. Il n'y a pas de réponse simple. Par exemple, en statistique, est-ce la médiane ou la moyenne arithmétique qui constitue le meilleur indicateur (modèle) de la tendance centrale d'une distribution de données ? Suivant l'échantillon de données en mains, la moyenne peut fausser l'information véhiculée par ces données, surtout en présence de valeurs aberrantes, alors que la médiane peut en donner un meilleur portrait. Par contre, la médiane n'est pas très malléable lorsqu'il s'agit d'effectuer des analyses statistiques inférentielles complexes. C'est pourquoi, en général, on devra postuler, avant d'utiliser un modèle mathématique, que certaines conditions d'application sont remplies. Par exemple, avec suffisamment d'observations, on peut présumer que l'impact des valeurs aberrantes sur le calcul de la moyenne sera négligeable et pourra, à toute fin utile, être ignoré. Dans le modèle de l'analyse de la variance, on suppose que chaque distribution suit la loi normale avec équivariance. Devant un échantillon qui ne se conforme pas à ce credo, devrait-on changer de modèle et trouver un modèle qui s'ajuste aux données (en utilisant par exemple un modèle non paramétrique, *distribution-free*) ou encore reconnaître que l'on a affaire à un échantillon rare et transformer cet échantillon (en éliminant des valeurs aberrantes, par exemple, ou en effectuant des transformations mathématiques) de manière à ce que les conditions d'application du modèle soient satisfaites ? Ne vaudrait-il pas mieux évaluer l'impact du manque de conformité au modèle lors de l'interprétation des résultats (comme dans les études de robustesse d'un modèle au manque de respect des conditions d'application du modèle) et prendre des décisions en conséquence ? La modélisation mathématique comporte des règles, mais aussi sa part de risque. Comme nous le verrons, si la recherche du modèle parfait est légitime, l'atteinte de ce modèle tant recherché relève souvent de l'utopie.

Le présent volume se divise en deux parties. La première partie traite des concepts et méthodes nécessaires à la compréhension des modèles de mesure présentés dans le cadre de cet ouvrage, soit ceux de la théorie classique, ceux de la théorie de la généralisabilité et ceux de la théorie des réponses aux items. C'est au chapitre 1 que sont définies les notions de modèle, de mesure et de modèle de mesure. S'agissant de la théorie classique, présentée au chapitre 2, il est question notamment des notions d'erreur de mesure, de score vrai, d'erreur-type de mesure, de méthodes d'estimation de la fidélité et d'analyse d'items. Après avoir présenté, au chapitre 3, la théorie de la généralisabilité comme une extension naturelle de la théorie classique, les concepts de facette, de phases d'observation, d'estimation, de mesure et d'optimisation sont définis. Y sont aussi distinguées les notions de décision relative et de décision absolue. Le chapitre 4 aborde les concepts de base des modèles de réponses aux items, soit la courbe caractéristique d'item, l'erreur-type de mesure, la fonction d'information, la courbe caractéristique de test et l'efficacité relative. Puis sont présentés les modèles logistiques à 1, 2 puis 3 paramètres. Enfin, sont abordés succinctement des modèles plus complexes, qu'ils soient

polytomiques, multidimensionnels ou non paramétriques. Les conditions d'application des modèles unidimensionnels, logistiques et paramétriques, notamment l'indépendance locale et l'unidimensionnalité, font l'objet du chapitre 5. Il y est également question de la propriété d'invariance et de l'adéquation du modèle aux données. C'est au chapitre 6 qu'est abordée la difficile question de l'estimation des paramètres d'items (a, b, c) et de sujets (θ). La modélisation non paramétrique est brièvement discutée au cours de ce chapitre. Le dernier chapitre de cette première partie constitue une rupture par rapport aux six chapitres précédents. Il y est question de la très importante notion de validité. On distingue les principales méthodes de validation, le concept de biais et la méthode de l'analyse factorielle.

La deuxième partie se confine aux applications des modèles, notamment celles touchant la théorie des réponses aux items. On y traite des méthodes de détection de biais d'items au chapitre 8 et du testing adaptatif au chapitre 9.

CHAPITRE 1

Qu'entend-on par modèle de mesure ?

Dans ce chapitre, nous présentons différentes définitions des concepts de **modèle** et de **mesure**. Dans un premier temps, nous examinerons attentivement le concept de modèle. Il s'agit d'un concept qui peut, à première vue, effrayer les novices, peut-être parce qu'il évoque quelque chose de trop technique, peut-être aussi parce qu'il est défini d'une façon équivoque, notamment dans le cadre de l'étude des théories de la mesure. Comme nous le verrons, ce concept ne comporte pas toujours l'auréole ésotérique qu'on lui prête souvent. Ainsi, nous présenterons plusieurs facettes du concept de modèle de façon à bien le situer. Dans un deuxième temps, nous aborderons le concept de mesure en mettant en perspective les différentes acceptions de ce concept qui ont cours en sciences de la nature et en sciences sociales, notamment en contrastant la définition proposée par Stevens (1951), qui met de l'avant l'idée de niveaux d'échelle de mesure et les caractéristiques de la mesure telle qu'elle

s'est toujours pratiquée en physique, par exemple. Finalement, nous présenterons une définition de modèle de mesure qui nous guidera dans les chapitres subséquents.

1.1. LE CONCEPT DE MODÈLE

Si la consultation d'un dictionnaire constitue une façon légitime d'obtenir une définition acceptable d'un terme, il est encore mieux d'en consulter plusieurs. Nous avons consulté le *Petit Larousse*, le *Grand Larousse encyclopédique*, le *Petit Robert* et le *Robert méthodique* à la rubrique ***modèle***. Même si plusieurs définitions sont proposées dans chacun de ces dictionnaires, nous en avons retenu deux qui semblaient tout particulièrement pertinentes dans le contexte[1] qui nous intéresse.

Définition 1

*Un modèle est ce qui sert d'**objet d'imitation** pour faire quelque chose.*

Définition 2

*Un modèle est une **représentation simplifiée** d'un phénomène pour mieux l'étudier.*

Bien que différentes, mais loin d'être contradictoires, ces deux définitions se complètent et vont permettre de circonscrire les particularités de cette notion plurielle.

Les lignes qui vont suivre présentent des exemples de l'utilisation de modèles en tant qu'objet d'imitation et soulignent, du même coup, l'importance des modèles dans la vie de tous les jours. L'étude de la grammaire illustre nos propos : qui de nous, en effet, a oublié que la conjugaison des verbes aimer et finir (ou encore, pour les férus de latin, la déclinaison de *rosa*) sert d'abord et avant tout de modèle ? En peinture, on n'a pas besoin d'épiloguer longtemps pour comprendre l'importance de Mona Lisa comme modèle pour la *Joconde*. Dans le monde de la mode, les grands couturiers ne peuvent pas se passer des modèles qui portent et exposent leurs créations. Enfin, des personnages historiques comme John F. Kennedy, Mère Teresa, Nelson Mandela, le Mahatma Gandhi, Michael Jordan ou Albert Einstein sont souvent cités, à tort ou à raison, comme modèles. On a fait de ces individus des personnes admirables, des personnes à imiter.

1. Il est intéressant de noter, pour les personnes particulièrement férues de mesure et d'évaluation, qu'au plan étymologique le terme modèle a une double origine. Ce mot, en effet, vient du mot italien *modello*, qui signifie modèle, et du latin *modulus*, qui signifie mesure.

Dans les cours de biologie, les modèles de squelettes permettent d'étudier l'ossature des vertébrés de façon plus confortable que les vrais squelettes. Le célèbre globe terrestre est un outil absolument incontournable dans un cours de géographie. Pour les voyageurs, le recours à un plan du réseau routier d'une ville inconnue réduit considérablement les risques de faire fausse route. En architecture, la maquette de certains bâtiments est si précise qu'elle nous semble plus réelle que l'édifice lui-même. Que penser également des reconstitutions présentées lors d'un procès en vue d'analyser les circonstances entourant un accident ! Pour les amateurs des prévisions météorologiques au petit écran, on ne peut passer sous silence la représentation d'un pays ou d'un continent et, en surimpression, le mouvement des dépressions et des anticyclones. Voilà quelques exemples de modèles pour lesquels on ne parlerait pas d'objet d'imitation, mais plutôt de représentation simplifiée d'un phénomène.

Tout ceci montre jusqu'à quel point nous sommes littéralement submergés de modèles dans notre quotidien, certains servant surtout comme objets d'imitation et d'autres comme représentations simplifiées d'un phénomène. Mais la notion de modèle ne se limite pas à ces exemples concrets. La vie de plusieurs chercheurs scientifiques est en effet remplie de modèles. En statistique, par exemple, un échantillon peut être vu comme un modèle de la population si, par exemple, on considère que c'est une espèce de portrait miniature de la population à l'étude. En statistique encore, il est impensable de ne pas mentionner les modèles omniprésents que sont la courbe normale et la droite des moindres carrés. En économie, même si la réalité est beaucoup plus volatile, plusieurs modèles ont été développés en vue de prédire ou prévoir (avec plus ou moins de succès, mais qu'importe !) le taux d'inflation ou de chômage[2]. Qui de nous, enfin, ne se souvient pas avoir entendu des spécialistes confronter le modèle de Newton ($F = ma$) au modèle d'Einstein ($E = mc^2$) ?

Ces exemples mettent en lumière certains invariants qui peuvent qualifier un modèle. Tout d'abord, un modèle se doit d'être précis : il doit représenter la réalité de la façon la plus fidèle possible. Le plan du réseau routier d'une ville et la maquette d'un architecte, à titre d'exemples, doivent sans nul doute être construits avec la plus grande précision, sinon ils seraient rapidement mis de côté. Un modèle doit en outre être parcimonieux et permettre la présentation ou la reconstitution de phénomènes à un coût beaucoup plus abordable que l'original pour être d'une quelconque utilité. Il faut ainsi éviter ce qu'on appelle la surmodélisation. On peut s'en rendre compte tout particulièrement dans le cas de la reconstitution d'un accident lors d'un procès ou encore dans l'utilisation d'une maquette d'architecte, mais c'est la même chose

2. Bertrand et Valiquette (1986, p. 286) ont même montré que l'école pouvait être considérée comme une cause du chômage si l'on interprétait un modèle de corrélation de façon trop aveugle.

pour les modèles de régression en statistique. De plus, un modèle fait parfois ressortir un côté esthétique, comme le modèle d'un peintre ou même un modèle mathématique. Enfin, certains modèles sont souvent associés à des éléments visuels ou graphiques comme la maquette d'un développement immobilier, le globe terrestre, le plan du réseau routier d'une grande ville ou la droite des moindres carrés.

Dans les derniers paragraphes, nous avons volontairement mélangé les modèles présents dans notre quotidien aux modèles proprement mathématiques employés par les scientifiques, l'idée étant de renforcer l'analogie entre ces deux facettes des modèles. La prochaine section présente deux exemples qui permettent de suivre les étapes menant à l'élaboration d'un modèle pour représenter un phénomène observé : le premier exemple est d'utilité courante, alors que le second sert en analyse statistique des données.

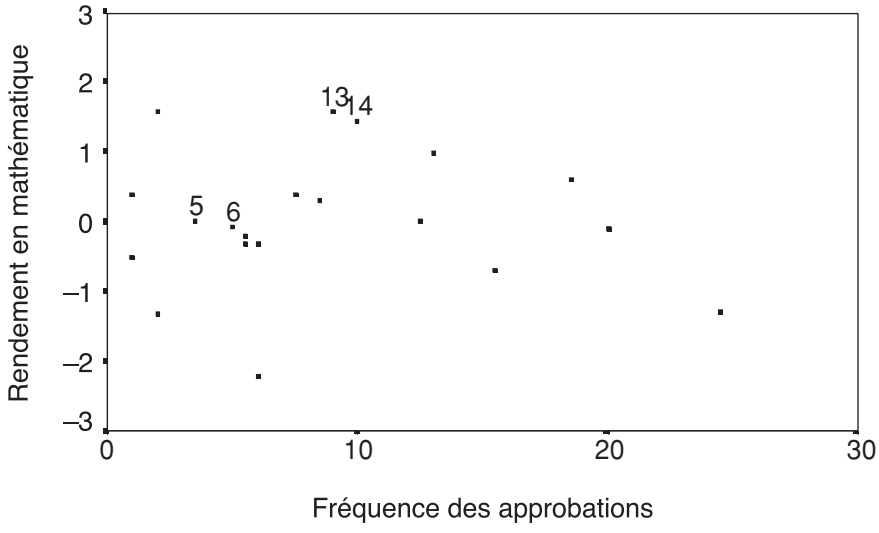
1.1.1. Genèse d'un modèle

Le calendrier est un système de division du temps en années, en mois et en jours, nous signale le *Petit Larousse* (1998). Mais comment ce système, ce modèle des jours de l'année, a-t-il vu le jour ? Les astronomes de Babylone ont tout d'abord observé que, dans un an, il y avait douze pleines lunes. Quoi de plus naturel alors que de diviser l'année en douze parties de 30 jours ! D'autant plus que ce modèle est compatible avec le très important modèle du cercle, lui-même plein de significations mystiques : il y aura 360 jours dans une année comme il y a 360 degrés dans un cercle. Au fil des ans, d'autres astronomes, plus malins, se rendent bien compte qu'une année comporte plus de 12 pleines lunes et donc plus de 360 jours. Ceux-ci calculent qu'il y a en fait 365 jours et 1/4 dans une année : c'est le modèle de l'année bissextile, le modèle julien (de Caius Julius César). Chaque cycle de quatre ans comportera donc trois années de 365 jours et une année de 366 jours. Pratiquement, cette mesure revient aujourd'hui à ajouter une journée au mois de février (le 29) pour les années qui sont des multiples de quatre comme 1984, 1988, 1992, 1996. Est-ce bien le modèle que nous connaissons aujourd'hui ? Eh bien non ! Des astronomes de la cour du pape Grégoire XIII montrèrent en effet, dès le xvi^{e} siècle, qu'il n'y avait pas 365 jours et 1/4 dans une année mais un peu moins… En fait, les calculs montrèrent qu'il y avait effectivement 365 jours, 5 heures, 46 minutes et 48 secondes dans une année : c'est le modèle grégorien. Il y aura donc, comme pour le calendrier julien, des années bissextiles (de 366 jours) à tous les quatre ans, sauf pour les années-centaines qui ne sont pas des multiples de 400. L'an 2000 par exemple est une année bissextile puisque 2000 est un multiple de 400, mais l'an 2100 ne sera pas une année bissextile : même si 2100 est bel et bien un multiple de quatre, c'est une année-centaine qui n'est pas un multiple de 400.

Le calendrier offre ainsi un parfait exemple d'un modèle que l'on utilise tous les jours et qui s'est raffiné avec le temps.

L'exemple suivant est tiré d'une étude internationale (Bertrand et Leclerc, 1984 ; Leclerc, Bertrand et Dufour, 1986) visant à relier des comportements d'enseignants de mathématique de deuxième secondaire au rendement scolaire de leurs élèves[3]. La figure 1.1 montre la relation obtenue entre la fréquence des approbations[4] créditées à un enseignant de mathématique et le rendement moyen de ses élèves. Chaque point représente un groupe particulier : 20 groupes-classes ont donc été observés en tout. L'abscisse de ce graphique représente la fréquence moyenne des approbations créditées à un professeur en une période normale de classe. L'ordonnée indique le rendement scolaire (en scores résiduels standardisés) d'une classe de mathématique. Quel est le modèle mathématique qui s'ajuste le mieux à cette relation ? Une observation sommaire des points de ce graphique montre qu'en général, plus un enseignant approuve un étudiant, en situation de grand groupe, plus le rendement en mathématique de son groupe tend à augmenter : une interprétation qui, du reste, paraît bien sensée. Par exemple, les enseignants des classes numérotées 5 et 6 approuvent en moyenne 4 à 5 fois par période de classe pendant que leurs élèves présentent un rendement moyen en mathématique.

FIGURE 1.1
Relation entre le nombre d'approbations par un enseignant de mathématique et le rendement scolaire moyen de son groupe

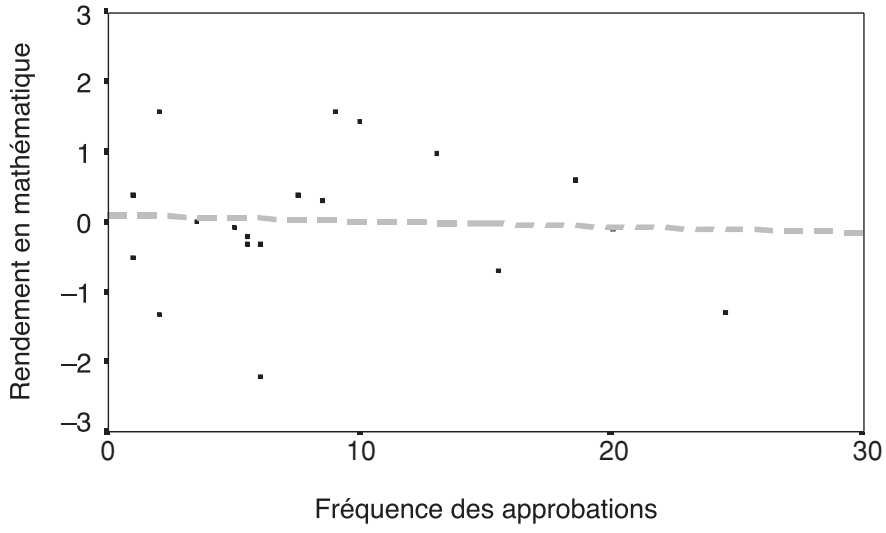


3. Intitulée « La classe et son environnement », cette étude se fonde sur les observations des comportements en classe par un certain nombre de juges ou d'observateurs formés à cette fin. Le rendement scolaire est standardisé.
4. Le comportement de l'enseignant que l'on nomme ici approbation peut se résumer ainsi : en contexte de grand groupe, le professeur approuve ce que viennent de dire les étudiants de son groupe.

Par ailleurs, les enseignants des classes numérotées 13 et 14, qui approuvent en moyenne 10 fois par période, voient leurs élèves obtenir un rendement plus élevé en mathématique.

Il paraît donc bien naturel d'avoir recours à un modèle linéaire pour représenter cette relation. La droite des moindres carrés[5] a été utilisée à cet effet, comme l'illustre la figure 1.2. On voit qu'elle ne rend compte que très grossièrement de la relation : la droite n'épouse (n'imite) pas très bien la forme générale suggérée par les observations. Le modèle linéaire semble assez imprécis et donc sujet à améliorations.

FIGURE 1.2
Relation entre le nombre d'approbations par un enseignant de mathématique et le rendement scolaire moyen de son groupe ; ajustement linéaire avec la droite des moindres carrés

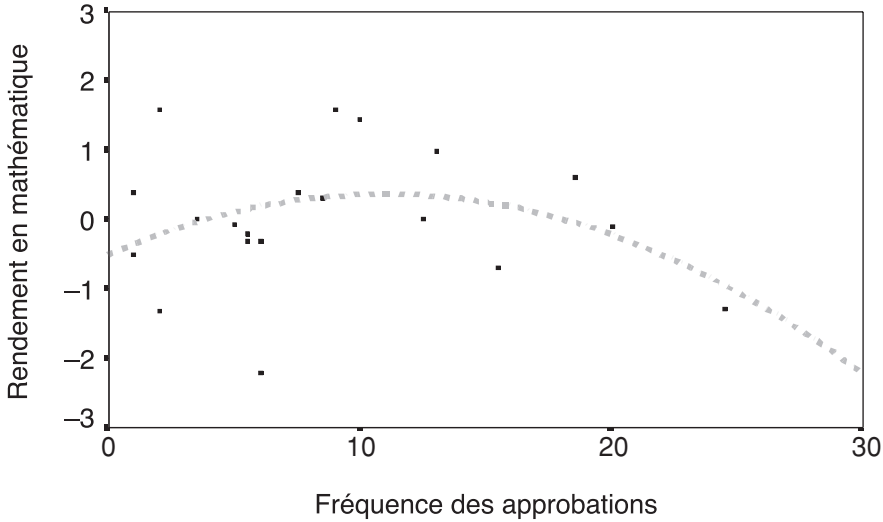


La figure 1.3 montre qu'une fonction quadratique[6] qui utilise une courbe au lieu d'une droite comme modèle épouse beaucoup mieux le nuage de points. C'est donc le modèle quadratique que nous retiendrons pour interpréter ces données : ainsi, plus un enseignant approuve ses étudiants, plus le rendement scolaire de ceux-ci augmente, jusqu'à un certain point de saturation

5. C'est la droite tracée à travers le nuage des 20 points de telle sorte que la distance moyenne entre chaque point et la droite soit réduite au minimum (Bertrand et Valiquette, 1986, p. 293).
6. Alors qu'un modèle linéaire (une droite) peut être décrit par une équation du type $Y = a + bX$, un modèle quadratique aura comme équation $Y = a + bX + cX^2$. Il reste maintenant à calculer le pourcentage de variance expliquée par les deux modèles.

où le rendement commence à diminuer si la fréquence des approbations continue d'augmenter (selon ce modèle, il semble qu'il y aurait des limites à approuver les comportements des élèves en classe). Sur le graphique, on peut voir que ce point de saturation se situe autour de dix approbations par période de classe. Cette interprétation nuance donc de façon fort intéressante celle proposée par le modèle linéaire et elle paraît tout aussi sensée !

FIGURE 1.3
Relation entre la fréquence des approbations par un enseignant de mathématique et le rendement scolaire moyen de son groupe ; ajustement quadratique à l'aide d'une courbe (fonction quadratique)



1.1.2. Choix d'un modèle

On a déjà noté qu'il était souhaitable de rechercher, dans un modèle, certaines qualités d'ordre économique, esthétique ou graphique (visuel), en plus bien sûr de la précision, c'est-à-dire de sa capacité à reproduire la réalité de la façon la plus fidèle possible. Il faut ajouter que, pour un scientifique, il est également souhaitable de rechercher un modèle reconnu, éprouvé. Dans la recherche d'un modèle qui s'ajuste bien à la réalité, qui imite bien la réalité tout en étant une représentation simplifiée, il nous semble approprié d'adopter une attitude pragmatique fondée sur les observations qui vont suivre.

Premièrement, il n'est pas réaliste ni même souhaitable de retrouver toutes les qualités dans un même modèle : il faut s'habituer à faire des compromis. C'est parfois le côté esthétique qui prend le dessus, comme par exemple

dans le cas d'un modèle pour un grand couturier. Par contre, il ne fait pas de doute qu'une carte routière se doit d'abord et avant tout d'être précise. Dans le cas d'une maquette, l'architecte cherchera un moyen terme entre le côté esthétique et le côté précis de son œuvre.

Deuxièmement, il faut éviter, comme nous l'avons mentionné, de tomber dans ce que l'on pourrait appeler la surmodélisation : le modèle le plus sophistiqué n'est pas toujours celui qui nous sert le mieux. À titre d'exemple, combien d'entre vous étiez vraiment au fait des subtilités du calendrier grégorien, sans que cela ne vous empêche de consulter le calendrier quotidiennement ? D'un autre côté, que penser d'un modèle comme celui de la figure 1.4 pour représenter la relation entre la fréquence des approbations créditées à un enseignant de mathématique et le rendement scolaire de son groupe ? Le modèle observé est basé sur une méthode de lissage des moindres carrés pondérés (méthode *lowess*[7]). Il épouse presque parfaitement les observations, en tout cas mieux que le modèle quadratique et le modèle linéaire présentés plus haut : ce nouveau modèle est sans nul doute plus sophistiqué que les deux autres. Cependant, on y perd en simplicité (économie) et en esthétique : d'ailleurs, alors que le modèle linéaire et le modèle quadratique se prêtent très bien à une formulation algébrique (paramétrique), il n'en est pas de même pour ce modèle plus raffiné.

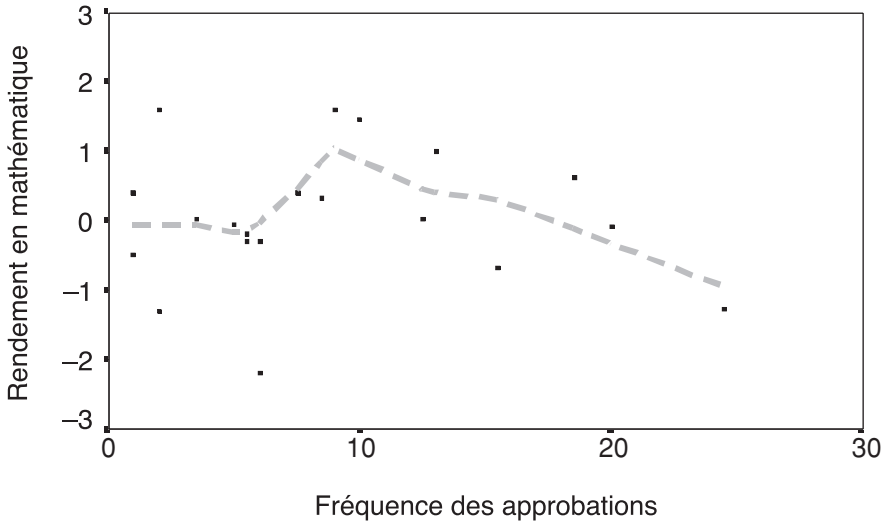
Troisièmement, même si un modèle semble convenir assez bien, n'oublions pas qu'il y aura toujours de meilleurs modèles : tout modèle est perfectible. À ce sujet, il n'y a qu'à questionner n'importe quel économiste ou n'importe quel météorologiste. L'idée est de rester ouvert à la possibilité de trouver quelque chose de mieux plutôt que de défendre son modèle à tout prix. « Tous les modèles sont faux, mais certains sont utiles », suggérait Box.

Quatrièmement, il est un peu illusoire de penser que tout modèle soit universellement reconnu. Les grands couturiers le savent bien : certains modèles de robes, de couvre-chefs ou d'habits ne conviennent pas à toutes les cultures et à toutes les époques de l'histoire. La mode évolue, on le sait, mais c'est aussi vrai dans des domaines scientifiques comme en météorologie, en économie ou encore en mesure et évaluation. Prenons un autre exemple : la question à choix multiple est un modèle d'item utilisé de façon presque routinière depuis plusieurs décennies en Amérique du Nord. Cependant, cela n'a pas toujours été le cas. En outre, est-on certain que ce modèle sera utilisé avec la même fréquence dans l'avenir (voir à ce sujet *The Times*[8]) ? Enfin ce modèle d'item est beaucoup moins courant en Europe et encore moins en Afrique. Dans le même domaine, on pourra noter qu'au XVII[e] siècle, à l'Université

7. Option d'ajustement des données présentée dans le logiciel SPSS.

Harvard, les candidats étaient admis uniquement sur la base d'une entrevue orale. Aux XVIII^e^ et XIX^e^ siècles, les choses changèrent : on vit l'apparition d'examens écrits et d'examens oraux devant des jurys composés de membres extérieurs à l'université. De nos jours, l'oral a presque disparu des pratiques d'évaluation des institutions d'enseignement. Ces exemples montrent que même s'il était préférable qu'un modèle possède une indépendance contextuelle, il appert qu'il n'en est rien et qu'il faut parfois accepter d'être beaucoup plus modeste.

FIGURE 1.4
Relation entre la fréquence des approbations créditées à un enseignant de mathématique et le rendement scolaire moyen de son groupe ; ajustement de type *lowess* à l'aide d'une courbe brisée



8. Dans sa livraison du 14 février 1993, ce quotidien de Trenton (New Jersey ayant pignon sur rue à quelques kilomètres seulement du siège social de l'Educational Testing Service (ETS), annonçait que le SAT (géré par ETS), « *the most famous (test) in America* », administré à 1,5 million d'élèves du secondaire pour estimer leur aptitude à entrer à l'université, changerait de format d'items, passant de questions à choix multiple à des questions à réponses construites.

1.2. LE CONCEPT DE MESURE

Au XX^e siècle, le concept de mesure a évolué principalement dans deux directions. La première et la plus ancienne est originaire de la définition de mesure proposée par Aristote et Euclide qui précise que la mesure est la **détermination du rapport entre deux quantités**. Ce rapport exprime une relation entre les grandeurs des quantités qui sont des propriétés empiriques appartenant au monde spatiotemporel. Par exemple, lorsque nous mesurons la longueur, nous effectuons des manipulations pour mesurer une propriété d'une chose, la longueur, qui existe empiriquement. Il s'agit de la définition classique du concept de mesure qui postule que tous les attributs que l'on peut mesurer sont par essence quantitatifs. Cette définition oblige la démonstration empirique de la présence d'une quantité. De plus, ce qui distingue la qualité de la quantité, c'est que cette dernière est constituée de parties que l'on peut additionner, plus précisément concaténer[9]. Ainsi la définition classique pourrait se résumer à ceci (voir Martin, 1999 ; Michell, 1999).

La **mesure** est l'estimation ou la découverte du **rapport** entre la **grandeur** d'un attribut quantitatif et **une unité** de cette même **grandeur**.

La définition d'Aristote et d'Euclide et la distinction qu'ils avaient établie entre quantitatif et qualitatif tinrent jusqu'au Moyen Âge. En fait, cette définition avait déjà au moins une rivale chez les Grecs. Platon et Pythagore soutenaient que le concept de nombre et le concept de grandeur étaient deux concepts différents. Alors que dans la définition classique, les nombres n'existent pas à l'extérieur des rapports entre grandeurs d'une même quantité, pour Pythagore les nombres ont une réalité à l'extérieur du monde observable ; ce sont des entités abstraites[10]. Celui-ci soutenait également que la réalité est fondamentalement quantitative. Il s'agit du même point de vue que l'on retrouve chez Galileo Galilei lorsqu'il parle de « compter ce qui peut être compté, mesurer ce qui peut-être mesuré et rendre mesurable ce qui ne l'est pas » et chez Lord Kelvin qui, de son côté, a dit : « Lorsque vous pouvez mesurer ce qui vous intéresse et l'exprimer avec des nombres, vous savez quelque chose à son sujet ; lorsque vous ne pouvez le mesurer et l'exprimer numériquement, ce que vous en savez est insatisfaisant et insuffisant. » Cette perspective prend d'ailleurs le nom de pythagoricisme : tout est mesurable et tout doit être mesuré. C'est ainsi que, grâce aux succès importants qu'obtenait la science quantitative, on en vint à soutenir, à partir du Moyen Âge, que la science ne pouvait être sans l'existence de la mesure.

---

9. Évidemment, le principe de l'addition était adéquat pour les quantités disponibles à l'époque, mais il n'est plus l'unique principe de combinaison pour toutes les quantités modélisées actuellement. Par exemple, est-ce qu'il y a addition des températures lorsqu'on mélange deux liquides ? Il y a plutôt équilibre, comme le prévoit la 2^e loi de la thermodynamique.
10. Il ne faut pas confondre les nombres avec les graphies 1, 2, 3, 4… Ces dernières sont des représentations symboliques commodes qui constituent un support visuel fonctionnel.

La deuxième direction que prend la mesure au XX^e siècle a obtenu une impulsion importante grâce aux travaux de S.S. **Stevens** (1951) qui a défini

> la **mesure** comme **l'assignation de nombres** à des objets ou des phénomènes selon des **règles**.

Sa définition se voulait une réponse au rapport produit par la British Association for the Advancement of Science, aussi appelé rapport Ferguson (1940), qui remettait en question l'existence même de la mesure en psychologie. Ce rapport basait ses observations sur la définition de mesure produite par un de ses membres, N.R. Campbell (1920). Pour ce physicien de formation,

> la **mesure** consiste en la **représentation** par des **nombres** des **propriétés** d'objets ou de phénomènes et des relations entre ces objets.

Évidemment, cette définition suppose aussi que les propriétés d'ordre et d'additivité des nombres se retrouvent dans la structure des objets empiriques. D'autres auteurs ont poussé plus à fond cette analogie, qui a donné lieu à un fort courant de recherche qui porte le nom de théorie représentationiste ou axiomatique de la mesure (voir Krantz *et al.*, 1971, 1989, 1990). On peut remarquer dans cette définition que les nombres deviennent externes au monde de la réalité ; ils constituent un système d'abstractions que l'on met en relation d'isomorphisme avec un système d'objets empiriques. La définition de Campbell, considérée alors comme la définition stricte de la mesure, remettait toutefois en question toutes les tentatives de mesure entreprises dans le domaine de la psychologie. Ainsi, Stevens reprit une partie de la définition de Campbell, mais pour la teinter d'opérationalisme. En effet, pour Stevens, ce qui est primordial pour la mesure consiste à établir une règle bien définie pour attribuer les nombres ; le problème central de la mesure devient donc le développement de ces règles opérationnelles. La définition de Stevens est considérée comme une définition souple de la mesure. En toute logique, Stevens proposa ensuite des niveaux de mesure qui correspondent à ce que nous désirons que les nombres représentent. Nous y reviendrons dans une section ultérieure.

Les débats sur l'existence de la mesure ont repris de la vigueur au cours des quinze dernières années précisément à cause des développements entourant la théorie des réponses aux items. En effet, nonobstant le fait que la modélisation avec la TRI facilite une perspective différente sur les tests et les items, les modèles possèdent des propriétés qui leur permettent de produire des estimations sur des échelles ayant des propriétés s'approchant de l'additivité. Cette dernière remarque est particulièrement fondée pour ce qui concerne le modèle de Rasch qui, selon Wright (1997), est le modèle de la TRI qui répond le mieux aux exigences de la mesure fondamentale.

Loin de nous l'idée de vouloir régler les problèmes des fondements de la mesure et de la quantification dans cet ouvrage ; celui-ci a un objectif plus pragmatique, qui est d'outiller l'utilisateur de modèles de mesure afin de

le rendre plus averti et plus conscient des avantages et désavantages des différentes avenues qui s'offrent pour la modélisation de la mesure. Nous pensons tout de même que le lecteur doit être conscient des différentes perspectives qui existent dans le déploiement de la mesure dans le secteur des sciences sociales.

Cette trop brève exploration historique et philosophique des fondements de la mesure pourrait laisser le lecteur en appétit. L'objectif de ce livre n'étant pas d'approfondir les concepts dans ces directions, nous pouvons lui recommander de consulter les textes de Martin (1999), Michell (1999) ou Berka (1983) pour y retrouver le développement des idées que nous venons d'esquisser dans cette introduction à la mesure.

Au-delà des définitions et des domaines d'application, nous proposons donc une façon de voir en quoi consiste essentiellement l'acte de mesurer et ses limites. Nous traiterons par la suite des quatre niveaux d'échelles de mesure de Stevens et nous tenterons, enfin, une définition de la notion de modèle de mesure.

1.2.1. Mesure des attributs physiques

En consultant le *Petit Larousse*, le *Grand Larousse encyclopédique*, le *Petit Robert* et le *Robert méthodique* à la rubrique **mesure**, nous retenons que celle-ci consiste en la

> détermination de la **valeur** de certaines **grandeurs** par **comparaison** avec une **grandeur constante** de même espèce.

Par identification, on définit également la mesure comme le résultat de ce processus de détermination d'une grandeur. Cette définition se prête particulièrement bien à des mesures physiques comme la taille ou le poids d'un objet ou d'un individu. À la lecture de manuels classiques de physique employés au niveau secondaire, tel celui de Benoît, Gauthier et Laberge (1962), il est intéressant de noter que l'on insiste sur trois composantes d'une mesure :

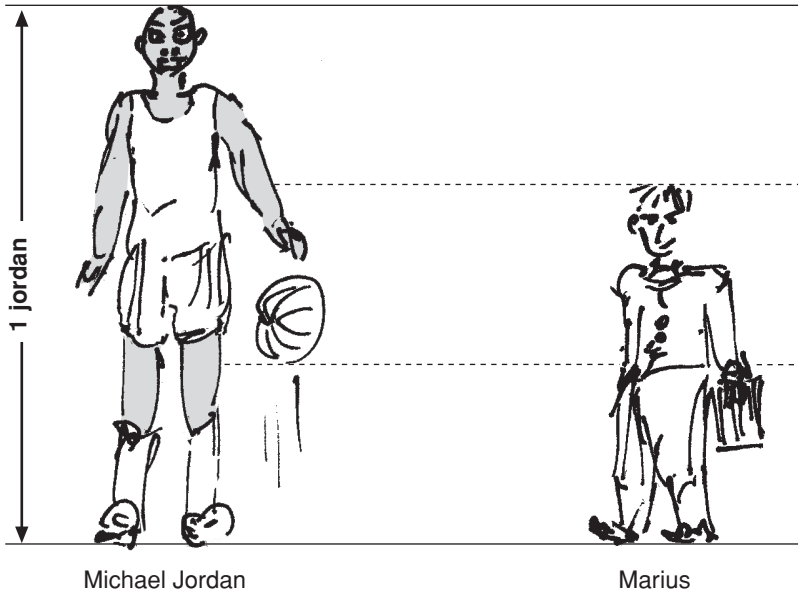
- le nombre lu sur l'instrument de mesure ;
- l'unité de mesure ;
- la précision de la mesure.

Ces trois composantes sont indissociables et suggèrent qu'un énoncé comme « Jean mesure 165 » n'a aucun sens. Même un énoncé plus complet du type « Jean mesure 165 cm » est difficilement acceptable puisque nous n'avons aucune idée de la précision de la mesure. À cet égard, un énoncé tel « Jean mesure 165 cm ± 2 cm » est déjà plus acceptable puisqu'il fait référence aux trois composantes d'une mesure : le nombre lu sur l'instrument, 165, l'unité de mesure, le centimètre (cm), et la précision de la mesure, ± 2 cm. Cette information nous indique en fait que la taille de Jean se situe, selon toute probabilité, entre 163 et 167 centimètres.

L'unité de mesure utilisée ici est arbitraire. Nous avons employé le centimètre, mais nous aurions pu tout aussi bien opter pour le mètre[11] (et dire que « Jean mesure 1,65 m ± 0,02 m ») ou encore le pouce ou le pied. Il est généralement facile de passer d'une unité de mesure à une autre en effectuant une transformation linéaire : par exemple, 1 pouce = 2,54 centimètres. Mais une unité de mesure de longueur n'est pas nécessairement toujours aussi classique. Pensons, par exemple, à la coudée ou à l'année-lumière. On peut même imaginer des unités de mesure de longueur plus concrètes comme le jordan[12]. La figure 1.5 montre que Marius mesure environ 2/3 jordan. On peut s'interroger cependant sur l'utilité d'une telle unité de mesure. C'est pourquoi nous employons plutôt des unités pour lesquelles nous avons adopté des conventions d'interprétation et d'utilisation pour rapporter la mesure de la longueur ou toute autre mesure. Ceci dit, plusieurs conventions peuvent se chevaucher, même pour une mesure aussi simple que la longueur. Pourquoi les Américains utilisent-ils encore le pouce alors que les Canadiens sont passés au système métrique ? Une recherche historique permettrait de découvrir les origines du système métrique et celui du système anglo-saxon et de trouver les raisons sociopoliticoéconomiques qui font adhérer un pays à un système d'unités de mesure plutôt qu'à un autre. Tout ça pour dire qu'il s'agit en partie de conventions qui permettent aux individus de mieux communiquer, de mieux échanger. Mais c'est tout de même le phénomène de la standardisation de la mesure qui la rend si attrayante et si universelle ; cette standardisation stimule sa diffusion. De plus, sans la symbolique nécessaire à la transmission des informations que convoie la mesure, il n'y aurait pas de langage commun et la mesure ne pourrait être transmise d'une génération à une autre. Mais il y a plus dans la mesure que de simples conventions, il y a aussi le désir de découvrir de vraies relations quantitatives chez les variables étudiées et de les traduire en régularités sous la forme d'unités de mesure. L'interaction entre les conventions et la démonstration de l'existence de ce qu'on veut mesurer forme d'ailleurs l'essence de l'histoire de la mesure depuis l'adoption généralisée du système métrique.

11. Le mètre (et ses dérivés comme le centimètre ou le kilomètre) est une unité de mesure de longueur très populaire, car il peut être défini très précisément. Traditionnellement, on définissait le mètre comme « un dix-millionième du quart du méridien terrestre ». Récemment, on a trouvé une définition plus moderne et beaucoup plus précise : un mètre est « la longueur parcourue par la lumière en exactement 1/299 792 458 de seconde » !
12. Michael Jordan est un ex-joueur de basketball américain très populaire, peut-être un des plus prolifiques de tous les temps. C'est un modèle pour plusieurs jeunes.

FIGURE 1.5
Si l'on utilise une unité de mesure comme le jordan, on voit que Marius mesure environ 2/3 jordan.



La précision d'une mesure physique dépend notamment de la graduation de l'instrument (au moment où l'instrument est construit) et de l'habileté de l'expérimentateur (au moment où l'instrument est utilisé). Un trait à tous les centimètres sur une règle ne permet pas la même précision qu'un trait à tous les millimètres. Par ailleurs, on sait bien que l'habileté à mesurer d'un observateur peut varier tout aussi bien d'une occasion d'observation à l'autre que d'un observateur à l'autre. C'est donc en répétant une expérience (comme la mesure de la longueur) que l'on pourra se rendre compte de l'ampleur de l'infidélité de la procédure de mesure. Notons que si la précision est une qualité de l'instrument qui renvoie à la graduation de celui-ci, la justesse renvoie à l'instabilité de la mesure.

1.2.2. Mesure des attributs psychologiques

Comme le soulignent justement Crocker et Algina (1986, p. 4), parmi d'autres, des mesures physiques comme la taille ou le poids d'un individu sont plus faciles à définir et à obtenir que des mesures psychologiques telles l'habileté à lire ou l'attitude envers les mathématiques. Ceci tient à plusieurs causes dont le fait que les variables psychologiques mentionnées ne sont observables que

très indirectement (par les réponses données aux items d'un test par exemple), qu'elles sont en général moins stables dans le temps et que les instruments servant à effectuer la collecte de ces mesures sont plus faillibles. Pourtant, en dépit de telles fluctuations, l'acte de mesurer s'inspire essentiellement des mêmes principes.

Nous avons préalablement vu la mesure comme la détermination de la valeur de certaines grandeurs par comparaison avec une grandeur constante de même espèce. Cette définition est, comme nous l'avons mentionné, plus adaptée aux mesures physiques qu'aux mesures psychologiques. Les définitions qui ont cours en sciences sociales prennent plutôt appui sur la proposition de Stevens (1951) qui parle de l'assignation de nombres à des sujets[13] selon certaines règles. Torgerson (1958) précise que cette assignation touche l'une ou l'autre des propriétés (anxiété sociale, attitude envers les mathématiques, etc.) des sujets plutôt que les sujets eux-mêmes. Blalock (1982) parle d'un procédé général par lequel des nombres sont assignés aux (propriétés des) sujets de façon à relier certaines opérations physiques (Marius est plus grand que Mario) aux opérations mathématiques (175 cm > 172 cm). Dans la même veine, de Gruijter et van der Kamp (1984) précisent qu'il s'agit d'une représentation d'un système de relations empiriques par un système de relations numériques : c'est donc l'assignation de nombres aux propriétés des sujets de telle sorte que les relations entre les propriétés des sujets (Marius est plus habile en mathématique que Mario) soient représentées par les relations entre les nombres (84 % > 75 %). Nous verrons à la section 1.3 comment cette définition de Stevens, amendée par Torgerson, permettra de définir ce que nous entendons par un modèle de mesure, toujours dans le contexte de la psychologie et de l'éducation.

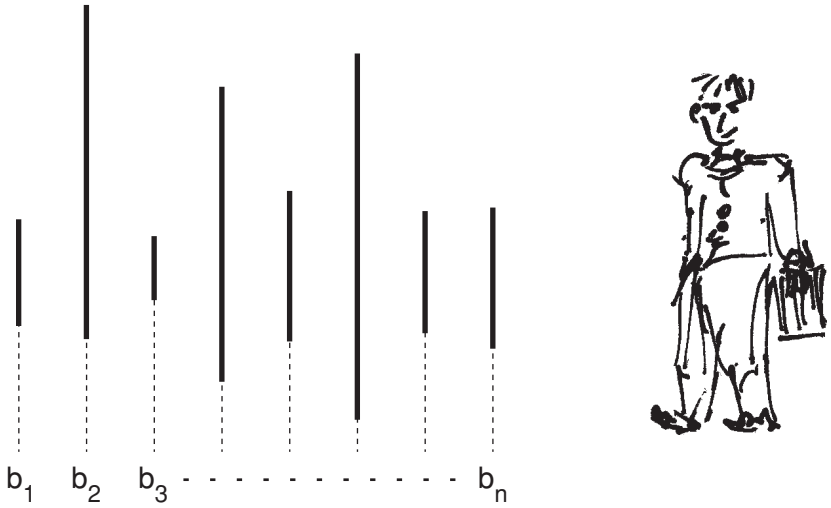
1.2.3. L'acte de mesurer

Au-delà des définitions, n'y a-t-il pas un ou des invariants dans la façon de mesurer quelque chose ou quelqu'un ?

Sirotnik (1987, p. 27) propose l'analogie suivante. Supposons que l'on ait, à notre disposition, une série de n bâtons de diverses longueurs, que l'on peut symboliser par b_1 , b_2 , b_3 … b_n. Pour mesurer la taille de Marius à l'aide de ces bâtons, il suffit de comparer sa taille à la longueur de chacun des bâtons tel qu'illustré à la figure 1.6. Si la taille de Marius dépasse la longueur de certains bâtons, elle est, en revanche, inférieure à la longueur de certains autres bâtons. La taille de Marius, selon cette procédure, est déterminée par la quantité de bâtons plus petits que lui.

13. Puisque que ce texte ne considère que la mesure des individus, nous parlons de sujets plutôt que d'objets.

FIGURE 1.6
Mesurer la taille de Marius revient à la comparer à la longueur de chacun des bâtons.



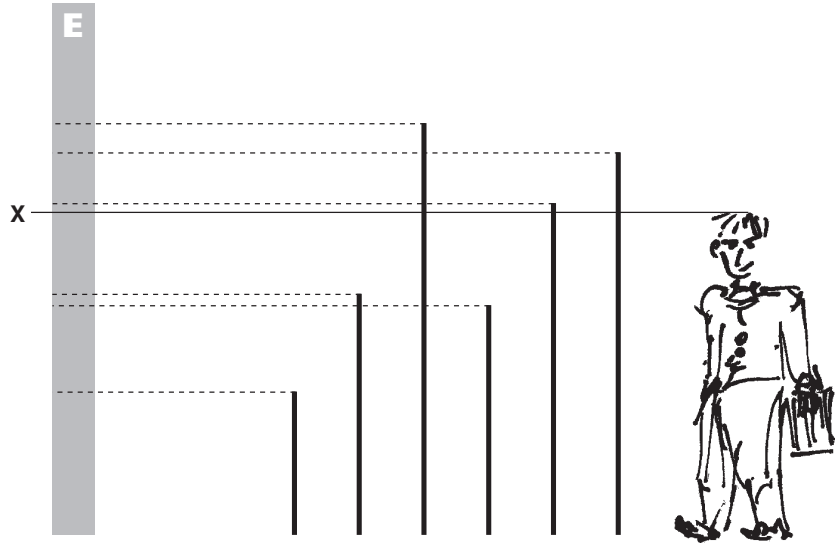
Cette façon de procéder revient de fait à construire un instrument de mesure de la façon suivante. Choisissons d'abord un bâton, plus long que le plus long des n bâtons déjà en notre possession, et représentons-le par la lettre **E** (pour échelle). Plaçons ensuite, sur le bâton **E**, des marques correspondant à la longueur de chacun des n autres bâtons, tel qu'illustré à la figure 1.7. Posons encore, sur le bâton **E**, une autre marque, que l'on symbolise par **X**, correspondant à la taille de Marius. Ainsi, la taille de Marius est égale au nombre de marques situées au-dessous de **X** : à la figure 1.7, puisqu'il y a trois marques au-dessous de **X** , la taille de Marius est égale à trois. Notons ici l'absence d'unité de mesure et d'estimation de la précision de la mesure[14] : ceci n'a pas vraiment d'importance pour le moment et nous y reviendrons plus loin. Précisons toutefois qu'il serait relativement facile de définir comme unité de mesure le plus petit bâton, pourvu que chacun des n – 1 autres bâtons ait une longueur qui corresponde à l'un ou l'autre des n – 1 premiers multiples de la longueur de ce plus petit bâton.

La construction de cet instrument n'est pas très différente en substance de la construction d'une règle commune de 30 centimètres. En effet, si cette règle est graduée en millimètres, cela revient à identifier 300 bâtons,

14. Plus spécifiquement, un bâton sera dit plus petit que Marius si celui-ci dépasse le bâton d'au moins $u/2$ où u est la plus petite unité de graduation de l'échelle.

dont la longueur varie entre 1 millimètre et 30 centimètres, chacun ayant un millimètre de longueur de différence. Il faut ensuite trouver un bâton **E** plus long que les autres et faire une marque qui corresponde à la longueur de chacun des 300 bâtons identifiés plus haut pour que la règle soit construite. L'avantage marqué de ce dernier instrument tient au fait qu'il a une unité de mesure (le millimètre) et offre la possibilité d'estimer la précision de la mesure.

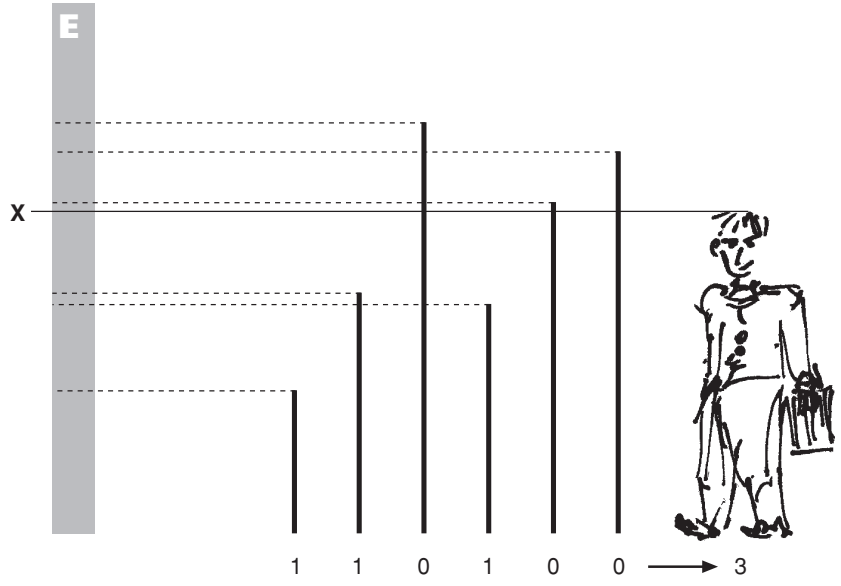
FIGURE 1.7
Le bâton noté E est un instrument constitué à partir des marques qui représentent la longueur de chacun des bâtons.



Quelle analogie peut-on faire maintenant entre la procédure décrite plus haut et la procédure de construction d'un instrument de mesure des attributs psychologiques d'un individu comme on en retrouve en éducation ?

Il s'agit tout d'abord de représenter la difficulté d'un item d'un test par la longueur d'un bâton et l'habileté d'un individu (p. ex., l'habileté à lire) par la taille d'une figurine comme celle de Marius à la figure 1.8. Ainsi, plus un item sera difficile, plus le bâton qui le représente sera long. De même, plus un individu aura une habileté élevée, plus la taille de la figurine qui le représente sera grande. Selon la procédure décrite plus haut et suivant notre analogie, mesurer l'habileté de l'individu revient à la comparer à la difficulté de chacun des items. Chaque fois que l'habileté de l'individu dépasse la difficulté d'un item, on lui attribue un point. En appliquant cette règle dans le cadre de la figure 1.8, on attribuera à Marius le score de 3.

FIGURE 1.8
Marius a un score de 3 puisqu'il dépasse 3 items (bâtons).



La figure 1.9 montre comment on peut générer le score de trois individus qui n'ont pas la même habileté. Marius et Mario obtiennent chacun un score de 3 alors que Marcel se voit attribuer un score de 5. À n'en pas douter, cette procédure présente certaines carences puisqu'elle ne parvient pas à différencier Marius et Mario, qui n'ont pourtant pas la même habileté : ils obtiennent un score identique de 3 même si leur habileté est manifestement distincte.

Afin d'améliorer cette procédure, il faut avoir à notre disposition des items de difficulté intermédiaire entre l'habileté de Marius et l'habileté de Mario, c'est-à-dire, toujours selon notre analogie, des bâtons plus courts que la taille de Marius et plus longs que la taille de Mario, tel qu'illustré à la figure 1.10. En ajoutant deux items dont la difficulté se situe entre l'habileté de Marius et l'habileté de Mario, les scores reflètent mieux l'habileté relative des trois individus.

Cet exercice basé sur notre analogie suggère plusieurs commentaires. Il permet de prendre conscience du fait que plus il y a d'items (de bâtons) en jeu, plus on a de chances de bien distinguer l'habileté des individus. En contrepartie, plus l'instrument comporte d'items, plus la procédure est coûteuse. Nous observons également que plus la difficulté des items est de l'ordre de grandeur de l'habileté des individus (c'est-à-dire plus la longueur des bâtons se rapproche de l'ordre de grandeur des tailles des individus), plus l'instrument

FIGURE 1.9
Marius, Mario et Marcel obtiennent des scores respectifs de 3, 3 et 5.

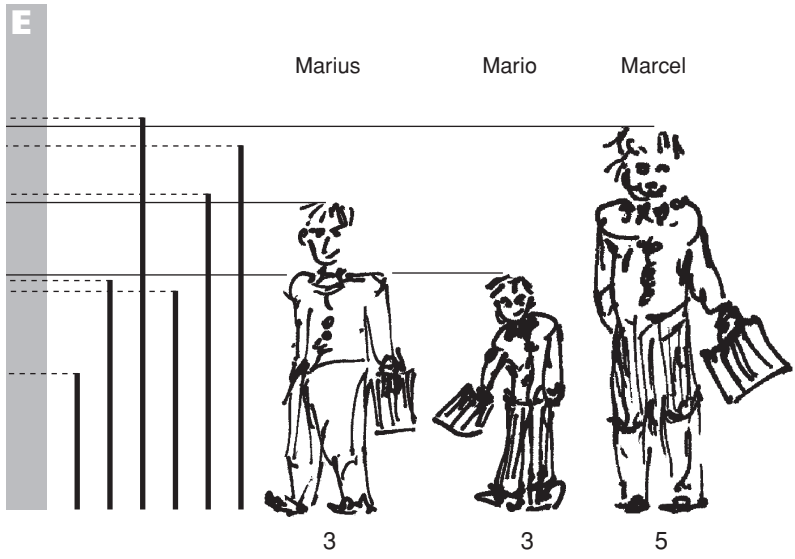
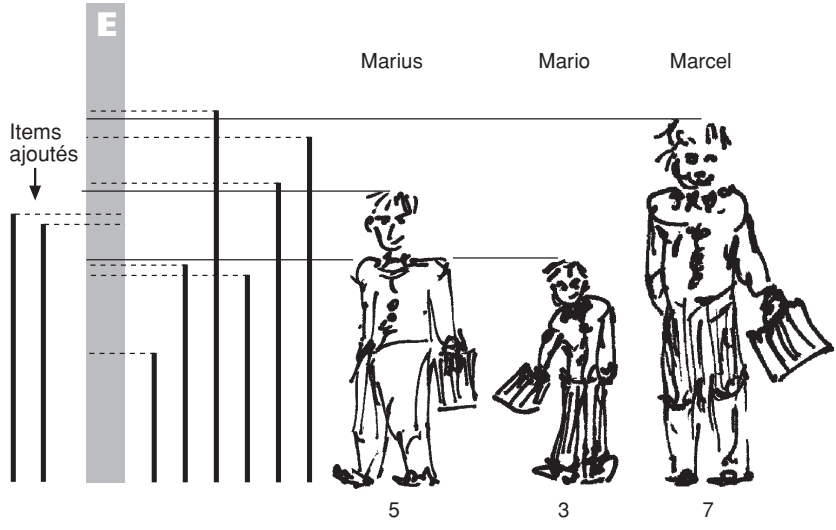


FIGURE 1.10
Ajout de deux items qui a pour effet que Marius, Mario et Marcel obtiennent des scores différents, soit respectivement 5, 3 et 7.



est précis, un résultat déjà noté par plusieurs auteurs dont Lord (1980, p. 114). Il faut aussi souligner les limites d'une telle analogie. Premièrement, les scores attribués aux individus ne prennent pas en compte la difficulté des items, en ce sens qu'un point est ajouté au score si un item est réussi, indépendamment de la difficulté de l'item. Il aurait été envisageable de pondérer différemment des items qui ne sont pas du même niveau de difficulté. Deuxièmement, lorsqu'une personne possède une taille plus grande qu'un bâton A, la personne est aussi certainement plus grande que tout bâton B plus petit que le bâton A. Or, ceci n'est pas le cas lorsqu'une personne passe un test : si un item A est plus difficile qu'un item B et que la personne réussit l'item A, elle ne réussit pas toujours l'item B. C'est pourquoi, comme nous le verrons, la popularité des modèles de mesure déterministes sera moins grande que celle des modèles probabilistes.

L'intérêt de cette analogie est bien sûr d'ordre théorique puisqu'elle suppose la connaissance initiale de la difficulté de chaque item et de l'habileté de chaque individu avant même de lui attribuer un score. De plus, il n'a pas vraiment été question des unités de mesure et de la précision de la mesure. Nous y reviendrons plus loin.

1.2.4. Niveaux d'échelles de mesure

Nous venons de voir que la procédure visant à mesurer une caractéristique pouvait s'appliquer aussi bien aux attributs physiques que psychologiques. Cependant les quantités, les scores ou, mieux, les mesures que l'on obtient à la suite de cette procédure ne peuvent pas toutes être traitées de la même façon. Si nous prenons la mesure de la taille de Marius et que nous obtenons 180 cm, nous pourrons dire que sa mesure est deux fois plus grande que celle de son petit frère Pierrot, dont la taille est de 90 cm. Mais peut-on raisonnablement affirmer que 40 °C est une température deux fois plus chaude que 20 °C ? Par contre, la différence (15 cm) entre la taille de Marius, 180 cm, et la taille de Marielle, 165 cm, peut vraiment être considérée identique à la différence entre les tailles de Jeannot, 105 cm et de Pierrot, 90 cm. Nous pouvons également considérer comme identiques la différence entre 30 °C et 40 °C et la différence entre 10 °C et 20 °C : dans les deux cas, il s'agit d'une différence de 10 °C.

Si Pauline réussit 30 des 60 items du test de géographie, peut-on affirmer qu'elle est deux fois plus compétente en géographie que Thierry qui a réussi 15 items du même test ? Peut-on conclure, en outre, que la différence de compétence entre Pauline et Thierry (15 items) est la même que la différence de compétence entre Pierre et Hélène qui, eux, ont réussi respectivement 55 et 40 items ? Il n'existe pas de réponse très claire à ces deux dernières questions. Cela dépend, dira-t-on. Si, maintenant, seul le rang des étudiants à ce test de géographie est considéré, peut-on dire que Pierre qui a obtenu le 2^{e} rang est deux fois plus compétent que Paul qui est 4^{e}, ou que la différence

de compétence entre Paul et Pierre est la même qu'entre Annie, 7e et Fleur, 9e ? Dans ce cas, on peut être plus catégorique et répondre non puisqu'il ne s'agit que de rangs.

Traditionnellement, on fait état de quatre types d'échelles de mesure suivant la nature des mesures et des opérations mathématiques que l'on peut effectuer avec ces mesures.

Au premier niveau se retrouve l'échelle dite **nominale** qui, de fait, n'est pas vraiment une échelle de mesure puisqu'elle se limite à une classification des personnes plutôt qu'à une véritable mesure de leurs caractéristiques. Le fait d'assigner le numéro 99 au dossard d'un joueur de football ne le rend pas de facto plus ou moins habile (11 fois plus habile ?) qu'un joueur qui porte le numéro 9. Ça ne ferait aucun sens d'additionner ou de multiplier ces nombres. Il ne s'agit que de codes, pas de mesures. On ne peut guère calculer que des fréquences et des proportions à l'aide de ces mesures.

L'échelle **ordinale** concerne les rangs. Elle est obtenue en assignant des rangs ou encore des nombres qui seront considérés comme tel. Par exemple, le fait qu'un enseignant donne des rangs à ses étudiants après une série d'examens constitue l'échafaudage d'une échelle ordinale. Les étudiants seront mis en ordre. On ne pourra cependant pas conclure que la différence entre l'étudiant qui a obtenu le premier rang et l'étudiant qui a obtenu le deuxième rang est la même que la différence entre les étudiants situés aux quatrième et cinquième rangs. Ce type d'échelle se prête bien au calcul de la médiane et de l'étendue interquartile.

L'échelle **d'intervalle** (ou à intervalles égaux) est basée sur le fait que des différences égales entre n'importe quelle paire de scores ont le même sens. L'échelle Celsius, par exemple, est une échelle d'intervalle puisqu'on peut considérer comme identiques la différence entre 25 °C et 40 °C et la différence entre 10 °C et 25 °C. Un écart de 15 °C a la même signification partout sur cette échelle. C'est pourquoi le calcul de la moyenne, de l'écart-type et de la corrélation de Pearson prennent ici tout leur sens.

L'échelle **proportionnelle** est une échelle d'intervalle particulière : elle contient une vraie valeur de zéro. Par exemple, le poids en kilogrammes ou encore la taille en centimètres donnent lieu à des échelles proportionnelles. Si un objet pèse 0 kg, on pourra dire qu'il n'a pas de poids. Ce n'est pas le cas pour la température, par exemple, puisque 0 °C ne signifie pas qu'il n'y a pas de température ! On pourra dire qu'un individu qui pèse 100 kg a un poids deux fois plus important qu'un autre individu pesant 50 kg.

De toute évidence, si l'on peut observer plusieurs exemples d'échelles proportionnelles en sciences physiques, il n'en est pas de même en psychologie ou en éducation. Même les véritables échelles d'intervalle ne sont pas légion. En effet, un écart de 15 items entre deux étudiants à un test n'a pas nécessairement la même signification en termes de compétences pour tous les couples d'étudiants séparés par le même écart. C'est pourquoi nous devons

souvent présumer, parfois même tacitement, que les scores à un examen ou à une échelle d'attitude se situent sur une échelle d'intervalle. En effet, il n'est pas aisé d'identifier une méthode éprouvée pour infirmer ou confirmer avec assurance que les scores à un examen se situent sur une échelle d'intervalle. Ceci dit, notre décision est lourde de conséquences. Ainsi, par exemple, le fait de calculer la moyenne et l'écart-type de la distribution des scores à un examen suppose que nous présumons être en présence d'une échelle à intervalles égaux ; autrement, ces statistiques n'ont plus qu'une signification symbolique.

1.3. MODÈLE DE MESURE

Nous avons dit qu'un modèle était un objet d'imitation, une représentation simplifiée d'un phénomène pour mieux l'étudier. Nous avons dit également que, selon la définition qui a cours en sciences sociales, la mesure consiste à assigner des nombres à des propriétés, des objets ou des sujets selon certaines règles. Mais avant d'assigner ces nombres, il faut tout d'abord construire un instrument permettant de recueillir des informations qui pourront se voir attribuer l'étiquette de mesures. Il faut donc distinguer deux étapes : la construction de l'instrument en tant que tel et l'attribution d'un nombre, d'une mesure à un individu à l'aide de cet instrument.

Nous dirons qu'un **modèle de mesure** consiste en un **plan** formé d'une série de **règles** à suivre (à imiter) afin

- de construire un ou des instruments de mesure possédant des propriétés comme la précision et la validité ;
- d'assigner à des individus à qui on a administré un instrument des nombres appelés scores, qui représentent au mieux l'habileté visée par l'instrument.

Selon cette conception, tout modèle de mesure aura donc une double fonction : construire des instruments de mesure et assigner un score à des individus. Il est important de le préciser, car les définitions de « modèle de mesure » ou d'un synonyme de cette expression peuvent varier d'un auteur à l'autre[15]. L'objectif ultime d'un modèle de mesure sera donc d'encadrer, à travers une série de règles à suivre, la construction d'un instrument de mesure d'un attribut donné (p. ex., habileté en mathématique, attitude envers l'école, etc.) et l'assignation à un individu d'un score qui reflète la quantité d'attribut possédée par cet individu. Nous pourrons nous fier à ce score d'autant mieux que les règles du modèle seront suivies. Comme nous le verrons, certains modèles comportent des règles simples à énoncer mais très difficilement vérifiables permettant d'encadrer la construction de l'instrument et l'assignation

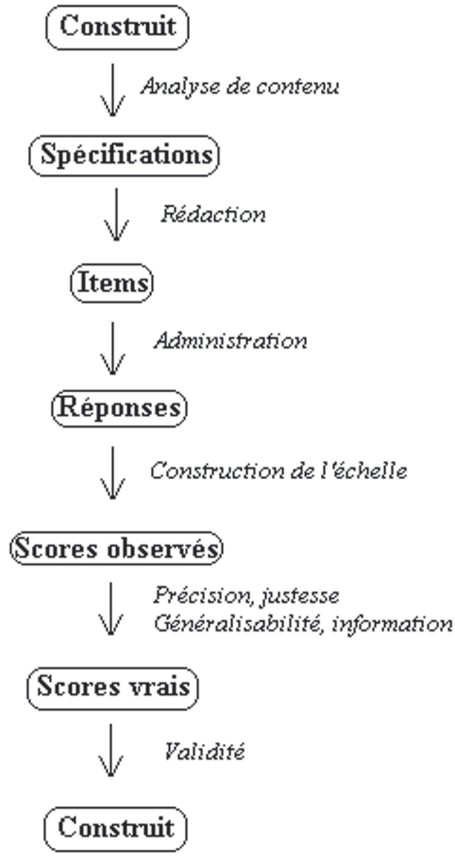
15. Au sens de Nunally (1978), nous nous intéressons donc exclusivement à l'étalonnage (*scaling*) des personnes.

des scores. D'autres modèles, par contre, comporteront des règles plus strictes qui permettront un meilleur encadrement, mais qu'il sera plus difficile de rencontrer.

Comment se fera le choix d'un modèle de mesure ? Sauf dans des cas particuliers (comme ceux exigeant l'analyse de plusieurs facettes, chapitre 3), il est très peu probable que le choix d'un modèle se fasse a priori. La modélisation des réponses aux items d'un instrument de mesure procède souvent par essais et erreurs. Un premier modèle est pressenti et on en vérifie les conditions d'application (les règles). Si elles sont respectées, le modèle peut être retenu, sinon il faut envisager un autre modèle. De même, on peut éliminer des items parce que les données qu'ils provoquent sont incompatibles avec le modèle. Et le processus peut se poursuivre de la sorte en prenant en considération, d'une façon plus ou moins explicite, les qualités d'un modèle (section 1.1) que sont la parcimonie et la précision.

Pour terminer ce chapitre, nous présentons à la figure 1.11 une adaptation de ce que Suen (1990, p. 6) a appelé le processus psychométrique et que nous appellerons le **processus de mesure** pour rappeler qu'il touche tout autant à la psychométrie (la mesure en psychologie) qu'à l'édumétrie (la mesure en éducation). Tout part d'un construit que nous voulons mesurer : l'habileté à produire une dissertation philosophique, l'habileté à résoudre des problèmes mathématiques, l'estime de soi, l'attitude envers l'école. Ce construit provient du chercheur, il est théorique et rien ne garantit son existence, sinon l'accumulation d'observations empiriques à son sujet. L'analyse des principales caractéristiques de ce construit permet de déboucher sur une définition opérationnelle et un tableau de spécifications. Ce tableau sert de guide à des spécialistes de contenu pour rédiger des items. Ces items servent de stimuli pour la production d'une réponse selon une procédure technique bien établie. Ils sont ainsi administrés à un échantillon de sujets et les réponses aux items sont collectées. C'est à partir d'ici que le modèle de mesure, tel que nous l'entendons, prend tout son sens. Il faut tout d'abord vérifier les conditions d'application du modèle. Puis il s'agit de construire une échelle (préférablement d'intervalle) en y situant les items et d'assigner un score aux individus à l'aide de cette échelle. Dans la mesure où l'instrument est suffisamment fidèle (précis, juste, généralisable, informatif), les scores observés représenteront les scores vrais. Enfin, l'étape de validité a comme objectif de vérifier si les inférences faites à partir de ces scores sont conformes ou si elles trahissent le construit préalablement défini.

FIGURE 1.11
Représentation schématique des différentes étapes d'un processus de mesure (adaptée de Suen, 1990, p. 6).



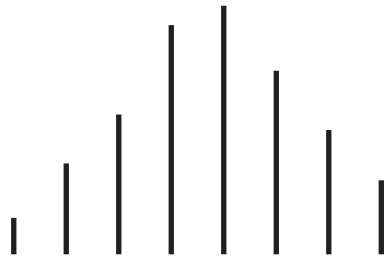
Questions d'approfondissement

1. Trouvez un modèle dont vous vous servez à tous les jours et qui correspond plutôt à notre définition 1 qu'à notre définition 2. Quel est le qualificatif qui sied le mieux à ce modèle : précis, économique, esthétique ou graphique ?
2. Trouvez un modèle dont vous vous servez à tous les jours et qui correspond plutôt à notre définition 2 qu'à notre définition 1. Quel est le qualificatif qui sied le mieux à ce modèle : précis, économique, esthétique ou graphique ?
3. Identifiez un modèle bien connu aujourd'hui et dont vous pouvez retracer la genèse.
4. Pourquoi peut-il être inapproprié de faire appel à la surmodélisation ?
5. Construisez un instrument de mesure primitif avec des blocs de bois de 1 à 10 cm. À l'aide de cet instrument, pouvez-vous mesurer la hauteur des livres qui vous entourent ? Si oui, dites pourquoi. Sinon, dites comment on pourrait confectionner un instrument de mesure qui puisse le faire avec les dix mêmes blocs de bois. Quelle sera votre unité de mesure ? Que pensez de la précision de votre instrument ? Faites une liste des objets que vous pouvez mesurer adéquatement avec votre instrument de mesure.
6. Donnez un exemple qui montre en quoi la corrélation de Pearson est un modèle beaucoup moins acceptable que le modèle de la corrélation partielle.

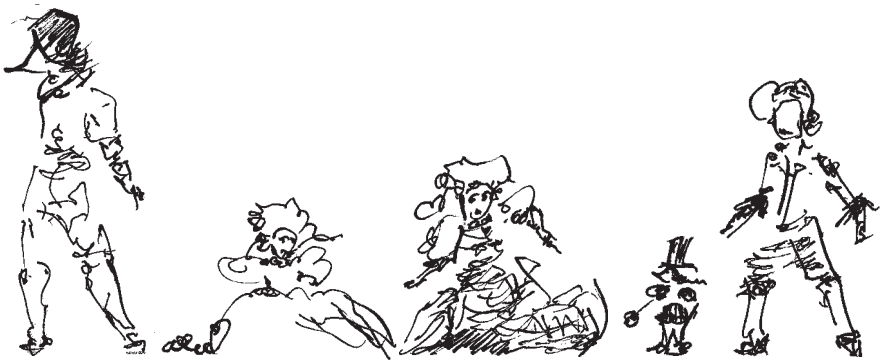
Exercices

1. Un test de huit items est administré à cinq individus. Dans la figure ci-dessous, la difficulté de chacun des items est représentée par la longueur d'un bâton. D'un autre côté, tel qu'illustré plus bas, l'habileté de chacun des individus est représentée par la hauteur de la figurine correspondante. Trouvez le score de chacun des individus. Est-il représentatif de leur habileté respective ? Quelle procédure pourrait-on suivre pour que le score des individus soit plus représentatif de leur habileté ?

   Voici les huit items :



   Voici les cinq individus :

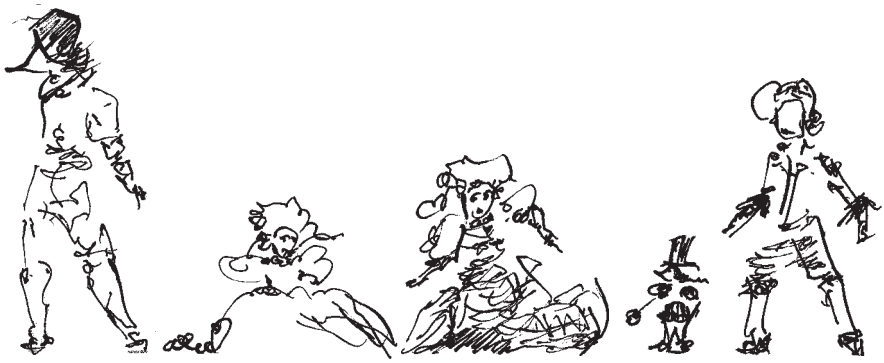


2. Trois enfants dans un parc décident de comparer leur taille avec, comme seul outil, un couteau. Ils veulent savoir qui est le plus grand, sans plus. Chacun leur tour, ils s'adossent à un arbre et l'un des enfants fait une marque sur l'arbre qui correspond le mieux à la taille de l'enfant adossé. Une fois les trois marques faites sur l'arbre, l'un des trois s'écrie : « Je suis le plus grand. » De quel type d'échelle de mesure est-il question ici ?

3. Trois jeunes veulent comparer leur taille avec, comme seul outil, une petite branche trouvée à l'orée d'un sous-bois. Afin de mesurer leur taille ils comptent, tour à tour, le nombre de longueurs de cette branche qui y correspond le mieux. Paul mesure 25 longueurs de branche, Pierre, 22 longueurs de branche et Jean, 27 longueurs de branche. De quel type d'échelle de mesure est-il question dans ce cas ?

Corrigé des exercices

1. Le score indiqué au bas de chaque figurine est représentatif de l'habileté de tous les individus, sauf pour les deux individus qui ont obtenu le score 8. Afin de mieux mesurer l'habileté de ces deux individus il faudrait ajouter des items plus difficiles (bâtons plus longs) que l'habileté de l'individu situé à l'extrême droite tout en étant plus faciles (bâtons plus courts) que l'habileté de l'individu situé à l'extrême gauche.



2. D'une échelle ordinale, puisque c'est seulement de l'ordre relatif des tailles qu'il est question.

3. D'une échelle proportionnelle, puisqu'il est possible de trouver la distance relative entre chaque couple de tailles et qu'il y a un vrai zéro.

CHAPITRE 2

Les modèles de mesure dans le cadre de la théorie classique

L'objectif de ce chapitre est de présenter les principales caractéristiques d'un modèle de la théorie classique que nous appellerons, pour simplifier, le modèle classique. La compréhension de plusieurs méthodes et concepts développés au cours des ans pour la modélisation de la mesure en éducation et en psychologie rend nécessaire une présentation au moins sommaire de ce modèle classique. Il n'est donc pas nécessaire de présenter tous les détails du modèle classique : plusieurs volumes (dont certains en français) ont déjà accompli cette tâche avec un certain succès. Nous renvoyons le lecteur intéressé à approfondir la théorie classique à Allen et Yen (1979), Crocker et Algina (1986), Gulliksen (1950), Laveault et Grégoire (2002), Lord et Novick (1968), Suen (1990) et Traub (1994).

Nous allons donc nous attarder tout particulièrement aux caractéristiques du modèle standard couramment employé en théorie classique. Ce modèle fait partie de ce que Lord et Novick (1968, p. 173) et Allen et Yen (1979, p. 239) appellent un modèle de score vrai avec des postulats faibles (*weak true-score model*) par opposition aux modèles de score vrai avec des postulats forts (*strong true-score models*) que sont le modèle binomial et le modèle de Poisson. Cependant, comme les applications en éducation et en psychologie du modèle binomial et du modèle de Poisson sont beaucoup plus clairsemées, nous ne les présentons pas, préférant renvoyer le lecteur à Lord et Novick (1968, chap. 21 à 24). Allen et Yen (1979, p. 242-253) donnent aussi une présentation sommaire de ces deux modèles. Puisque nous ne traiterons que d'un modèle de la théorie classique, nommément le modèle standard, nous l'appellerons tout simplement, à partir de maintenant, le modèle classique.

2.1. CARACTÉRISTIQUES DU MODÈLE CLASSIQUE

Zoé s'est montrée plus ou moins satisfaite d'avoir obtenu un score de 64 à l'examen unique du ministère de l'Éducation en mathématique. Jusqu'à quel point ce score reflète-t-il l'habileté réelle de Zoé en mathématique ? Zoé a pu être incommodée lors de son examen pour une raison ou pour une autre : maladie, grande fatigue, peine d'amour, etc. L'examen final préparé par les spécialistes du Ministère comportait peut-être, cette année-là, des questions particulièrement difficiles. La question de savoir si l'examen mesurait bel et bien l'habileté en mathématique et pas autre chose touche la validité ; nous l'aborderons dans un autre chapitre. Pour le moment, présumons que l'examen mesurait l'habileté en mathématique et demandons-nous plutôt si Zoé obtiendrait le même score si l'examen était passé dans des situations différentes. Après tout, si Zoé avait été plus en forme, peut-être aurait-elle obtenu un meilleur score que 64 ! Inversement, elle aurait pu tout aussi bien être moins en forme pour faire cet examen et obtenir un score inférieur à 64.

Le modèle de mesure dit classique permettra d'évaluer jusqu'à quel point un score comme 64 obtenu par Zoé à l'examen du Ministère reflète bien sa compétence en mathématique.

L'équation de base du modèle classique est donnée[1] par :

$$\mathbf{X} = \mathbf{V} + \mathbf{E} \qquad (2.1)$$

où **X** est le **score observé** d'un individu, **V** est le **score vrai** de cet individu et **E** est l'**erreur de mesure**.

1. Pour être plus précis, il faudrait dire que pour un test noté t administré à une répétition notée i et à un individu noté j alors $X_{ijt} = V_{jt} + E_{ijt}$. On verra plus loin pourquoi il est si important de tenir compte des indices i, j et t pour comprendre les notions de score vrai et d'erreur de mesure.

L'équation de base signifie que, selon le modèle classique, le score observé **X** à un test est constitué de deux composantes additives : **V** et **E**. Le score observé à un test est obtenu lors d'une administration particulière[2] de ce test. Chaque individu qui a subi ce test, à ce moment particulier, a donc un score observé. Ce score observé varie d'une répétition à l'autre du même test. Typiquement, le score observé peut être une fonction de la somme des items réussis d'un test lorsque ces items sont corrigés de façon dichotomique : 1 pour une bonne réponse, 0 pour une mauvaise réponse. Ainsi dans le cas de Zoé, si l'examen comportait 100 items à choix multiple et qu'elle en a réussi 64, son score observé est de 64, tout simplement. Nous ne connaissons et connaîtrons ni son score vrai ni son erreur de mesure.

Au contraire du score observé d'un individu, le score vrai et l'erreur de mesure d'un individu ne sont pas connus : ce sont des entités théoriques qui composent le score observé mais qui ne sont pas observables. En d'autres termes, ce sont ni plus ni moins que des abstractions conceptuelles (Lord, 1980, p. 5). En revanche, il est possible de proposer des définitions crédibles de ces deux concepts théoriques à partir de situations fictives.

Imaginons que l'examen ait pu être administré à Zoé un nombre très élevé de fois sans qu'elle se souvienne des réponses d'une fois à l'autre et qu'elle ait obtenu les scores présentés au tableau 2.1. C'est donc dire que Zoé aurait obtenu le score observé de 57, dix fois sur 100, le score observé de 58, 15 fois sur 100, et ainsi de suite. Quel serait le score (vrai) qui représenterait le mieux son habileté en mathématique telle que mesurée par l'examen du Ministère. À première vue, aucun des scores n'est un candidat plus acceptable qu'un autre. On est plutôt tenté de répondre que la moyenne des scores obtenus par Zoé, pondérée[3] par leur fréquence relative, constituerait un bon compromis. Il s'agit donc ici de $(57 \times 0{,}10) + (58 \times 0{,}15) + (59 \times 0{,}10) + (60 \times 0{,}10) + (62 \times 0{,}15) + (63 \times 0{,}20) + (64 \times 0{,}20) = 61$.

Ainsi, dans cette situation fictive, 61 pourrait être considéré comme le score vrai de Zoé tel que mesuré par l'examen de mathématique du Ministère.

Tableau 2.1
Scores de Zoé lors de plusieurs répétitions d'un même examen de mathématique

Score observé de Zoé (X)	**Fréquence relative**	**Score vrai de Zoé** (V)	**Erreur de mesure** (E)
57	0,10	61	– 4
58	0,15	61	– 3
59	0,10	61	– 2
60	0,10	61	– 1
62	0,15	61	+1
63	0,20	61	+2
64	0,20	61	+3

2. Chaque administration particulière d'un test sera appelée répétition.
3. La moyenne pondérée peut aussi être appelée espérance mathématique.

Nous pouvons donc définir de façon générale le score vrai d'un individu à un test donné comme la moyenne des scores observés obtenus lorsque le même test est administré à cet individu un très grand nombre de fois (un nombre de fois indéterminé !). La différence entre le score observé et le score vrai, obtenue à chaque répétition du test, est appelée l'erreur de mesure. De toute évidence, compte tenu des concepts définis à la section 1.2, l'erreur de mesure est associée au concept de justesse plutôt qu'au concept de précision.

Notons que :

- le score observé est une entité réelle, connue, variable d'une répétition à l'autre du test ;
- le score vrai est une entité non observable, inconnue, fixe d'une répétition à l'autre du test ;
- l'erreur de mesure est une entité non observable, inconnue, variable d'une répétition à l'autre du test ;
- l'erreur de mesure est aléatoire[4], en ce sens qu'elle est parfois positive, parfois négative et parfois nulle, sans toutefois que l'on puisse le prédire ;
- un score vrai est intimement lié à un individu particulier et à un test particulier : ainsi, le score vrai changera non seulement d'un individu à un autre, mais aussi d'un test à l'autre.

Cette façon de définir le score vrai d'un individu ne se limite pas à la mesure des variables en éducation ou en psychologie. À la section 1.2, nous avons vu que l'acte de mesurer des caractéristiques d'un individu dépassait la nature des variables considérées. Si on pouvait obtenir un nombre indéterminé de fois la mesure de la taille de Marius, alors la moyenne de ces mesures serait un score vrai, plus précisément la taille vraie de Marius, par exemple 165 cm. Il est important de souligner que l'expression **taille vraie** ne renvoie pas à une taille physiquement vraie, une taille qui pourrait être mesurée hors de tout doute. La taille vraie, au sens où nous l'entendons ici, est définie par convention comme une moyenne ou encore, dans le jargon de la statistique, une espérance mathématique.

La figure 2.1 représente la distribution des mesures de la taille de Marius. Tout écart entre la mesure de sa taille observée et sa taille vraie, pour une répétition donnée de la mesure, serait considérée comme une erreur de mesure. La figure 2.2 renvoie à la distribution des erreurs de mesure. Comme on le voit, certaines erreurs de mesure sont positives, d'autres négatives, la moyenne de ces erreurs de mesure étant nulle. Il faut encore noter que les distributions présentées aux figures 2.1 et 2.2 sont identiques, sauf pour la moyenne. La variance et la forme de ces deux distributions sont rigoureusement les mêmes.

---

4. L'erreur aléatoire affecte la fidélité, par opposition à l'erreur systématique (dont nous discuterons plus loin) ou biais, qui affecte la validité.

FIGURE 2.1
Distribution de fréquences des mesures de la taille de Marius

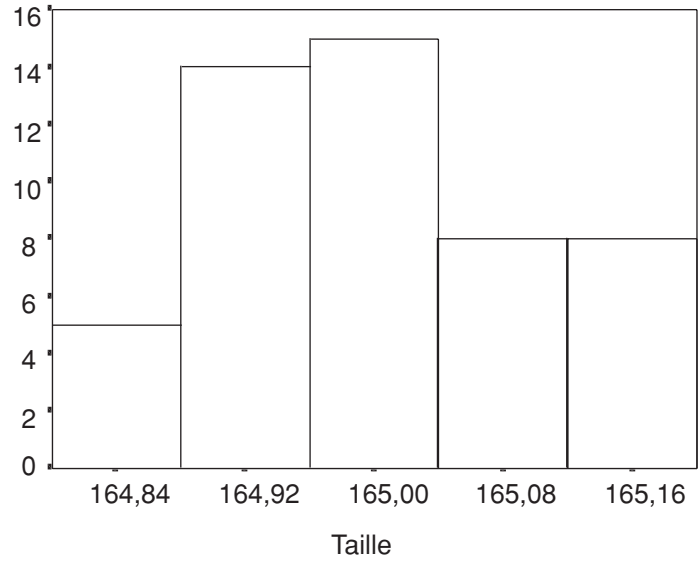
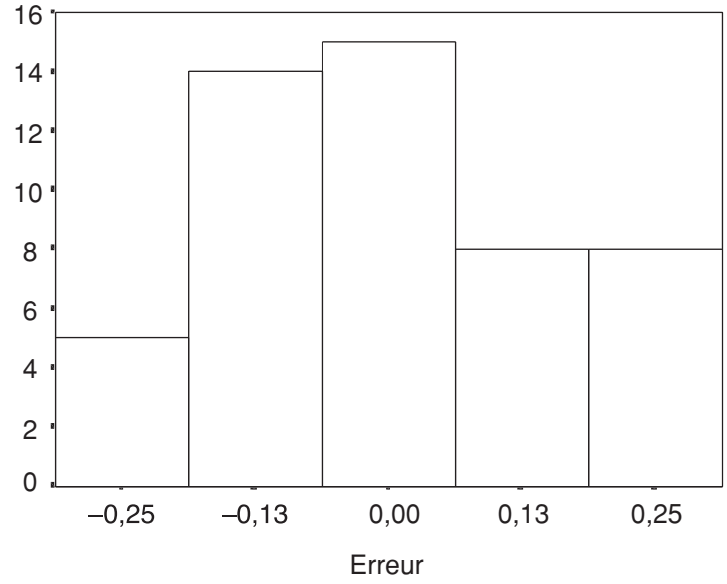


FIGURE 2.2
Distribution de fréquences des erreurs de mesure propres à la taille de Marius



L'exemple de la mesure de la taille d'un individu a eu l'avantage de permettre une représentation des concepts de score vrai et d'erreur de mesure dans un contexte relativement plausible, contrairement à l'exemple précédent, fondé sur d'hypothétiques répétitions du même examen à un individu. Voici un exemple supplémentaire, s'appuyant celui-là sur des attributs psychiques, qui permet de concrétiser encore un peu plus les notions de score vrai et d'erreur de mesure. Le tableau 2.2 donne les réponses (1 ou 0) d'une classe de 23 étudiants à un test de huit items[5]. Le total indique le score observé de chaque étudiant à ce test : ce qui revient à définir, dans le cadre du modèle classique, le score observé comme la somme des réponses aux items. Par exemple, l'étudiant 1 a réussi 5 items, il a donc un score observé $X_1 = 5$. L'étudiant 2, par ailleurs, a réussi 7 items et il a un score observé $X_2 = 7$. Enfin, l'étudiant 18 a réussi 4 items et il a un score observé $X_{18} = 4$. Mais qu'est-ce que ces scores observés disent à propos des scores vrais correspondants et des erreurs de mesure correspondantes ? Jusqu'ici, rien ! Mais regardons plutôt le tableau 2.3.

On y présente les scores observés (X_{ij}), où l'indice i réfère à l'une ou l'autre des quatre répétitions[6] d'un test de résolution de problèmes mathématiques et l'indice j réfère à l'un ou l'autre des 23 étudiants. Considérons, aux fins de cette présentation, que le score vrai (V_j) de chacun des 23 étudiants est donné par la moyenne des scores observés à ces quatre[7] répétitions. Ainsi, l'étudiant 1 a obtenu un score de $X_{11} = 5$ à la première répétition, un score de $X_{21} = 7$ à la deuxième, un score de $X_{31} = 8$ à la troisième et un score de $X_{41} = 4$ à la quatrième répétition : son score vrai est donc $V_1 = (5 + 7 + 8 + 4) / 4 = 6$. Par ailleurs, à chacun des scores observés de chaque étudiant (X_{ij}) est associée une erreur de mesure (E_{ij}). Encore ici, les erreurs de mesure de chaque étudiant peuvent être positives ou négatives. Par exemple, l'erreur de mesure de l'étudiant 1 à la première répétition est $E_{11} = X_{11} - V_1 = 5 - 6 = -1$. L'erreur de mesure de l'étudiant 1 à la deuxième répétition est donnée par $E_{21} = X_{21} - V_1 = 7 - 6 = 1$. Cependant, comme dans le cas de la mesure de l'habileté mathématique de Zoé ou de la taille de Marius, la somme des erreurs de mesure

5. Ces données proviennent d'une recherche avec 759 étudiants de quatrième secondaire à qui on a administré quatre tests de huit items de résolution de problèmes mathématiques (Bertrand *et al.*, 1993). L'exemple décrit ici à des fins pédagogiques ne comporte qu'un nombre très limité d'individus et d'items ; il ne faudrait pas y voir une situation-type d'application du modèle classique. Comme nous le verrons plus loin, plus le nombre d'individus et le nombre d'items augmentent, plus les estimations des paramètres des individus et des items auront tendance à se stabiliser.
6. En réalité, il s'agit de quatre tests de huit items. Mais comme le contenu de chaque item d'un test en particulier correspond au contenu d'un item de chacun des trois autres tests, nous considérons, aux fins de cet exposé, qu'il s'agit de quatre répétitions d'un seul test.
7. Strictement parlant, pour calculer le score vrai, il aurait fallu obtenir un nombre indéterminé de répétitions et pas seulement les quatre répétitions présentées ici.

est nulle, pour chacun des étudiants. Par exemple, pour l'étudiant 1, la somme des quatre erreurs de mesure donne (– 1) + 1 + 2 + (–2) = 0. Le lecteur pourra vérifier que cette somme est nulle pour chacun des 23 étudiants.

Cet exemple fait ressortir une caractéristique importante de l'erreur de mesure : elle est particulière à chaque situation de testing et à chaque individu. Par ailleurs, on peut noter un effet compensatoire : l'étudiant 3, qui a obtenu un résultat faible à la première répétition, performe plutôt bien à la troisième répétition, son erreur de mesure étant 1,5.

TABLEAU 2.2
Réponses de 23 étudiants à un test de huit items de résolution de problèmes mathématiques

	Item								
Étudiants	1	2	3	4	5	6	7	8	Total
1	1	0	1	1	0	1	1	0	5
2	1	1	1	1	1	1	0	1	7
3	1	1	1	1	0	0	0	0	4
4	1	1	1	1	1	1	0	0	6
5	1	1	1	1	0	1	1	1	7
6	1	1	1	0	1	0	0	0	4
7	1	1	1	1	1	1	1	0	7
8	0	0	1	1	0	1	1	0	4
9	1	1	0	0	1	1	1	0	5
10	1	1	1	1	1	1	1	0	7
11	1	1	1	0	0	1	0	0	4
12	1	1	1	1	0	0	0	0	4
13	1	1	1	0	1	1	1	0	6
14	1	1	1	0	1	1	1	1	7
15	1	1	1	0	1	0	0	0	4
16	1	1	1	1	0	0	1	1	6
17	1	1	1	1	0	1	1	0	6
18	0	1	0	0	1	1	1	0	4
19	1	0	1	1	1	1	1	1	7
20	1	1	1	0	1	1	0	0	5
21	1	1	1	1	1	1	1	1	8
22	1	1	1	1	1	1	1	1	8
23	1	0	1	1	1	1	1	1	7

Les exemples présentés ici permettent de se faire une idée intuitive des concepts théoriques que sont le score vrai et l'erreur de mesure. Il est d'autant plus important de bien connaître l'existence de ces deux concepts théoriques que ce n'est pas vraiment le score observé qui devrait nous intéresser chez un individu mais bien son score vrai et ce, dans toute situation de testing. Écoutons Lord (1980, p. 5) : « V, et non X, est la quantité qui nous

intéresse. Lorsqu'un candidat à un poste sort d'une salle d'examen, c'est V et non X qui détermine sa capacité à bien performer ultérieurement. On ne peut observer V mais on peut faire des inférences utiles à son sujet[8] ».

TABLEAU 2.3
Scores observés (X_{ij}) et erreurs de mesure (E_{ij}) pour 23 étudiants à 4 répétitions d'un test de résolution de problèmes mathématiques

Étudiants	X_{1j}	E_{1j}	X_{2j}	E_{2j}	X_{3j}	E_{3j}	X_{4j}	E_{4j}	V_j
1	5	– 1,00	7	1,00	8	2,00	4	– 2,00	6,00
2	7	2,25	4	– 0,75	4	– 0,75	4	– 0,75	4,75
3	4	– 2,50	7	0,50	8	1,50	7	0,50	6,50
4	6	– 0,75	6	– 0,75	8	1,25	7	0,25	6,75
5	7	0,25	7	0,25	8	1,25	5	– 1,75	6,75
6	4	– 1,75	7	1,25	7	1,25	5	– 0,75	5,75
7	7	– 0,75	8	0,25	8	0,25	8	0,25	7,75
8	4	– 0,75	5	0,25	6	1,25	4	– 0,75	4,75
9	5	– 1,50	7	0,50	7	0,50	7	0,50	6,50
10	7	– 0,50	7	– 0,50	8	0,50	8	0,50	7,50
11	4	– 0,25	4	– 0,25	6	1,75	3	– 1,25	4,25
12	4	– 2,50	8	1,50	7	0,50	7	0,50	6,50
13	6	– 1,25	8	0,75	8	0,75	7	– 0,25	7,25
14	7	– 0,50	8	0,50	7	– 0,50	8	0,50	7,50
15	4	– 1,00	4	– 1,00	6	1,00	6	1,00	5,00
16	6	– 1,25	8	0,75	8	0,75	7	– 0,25	7,25
17	6	– 1,00	8	1,00	8	1,00	6	– 1,00	7,00
18	4	– 2,00	7	1,00	7	1,00	6	0,00	6,00
19	7	0,00	7	0,00	7	0,00	7	0,00	7,00
20	5	– 0,75	8	2,25	6	0,25	4	– 1,75	5,75
21	8	0,25	8	0,25	7	– 0,75	8	0,25	7,75
22	8	0,00	8	0,00	8	0,00	8	0,00	8,00
23	7	– 0,50	7	– 0,50	8	0,50	8	0,50	7,50

Mais il y a plus. Ces concepts théoriques de score vrai et d'erreur de mesure sont omniprésents quel que soit le modèle d'interprétation employé : qu'il provienne de la théorie classique, de la théorie de la généralisabilité ou encore de la théorie des réponses aux items. Même si, dans le cas de ces deux dernières théories, ces concepts théoriques sont souvent étouffés par une masse d'équations et de postulats.

8. « *T, not X is the quantity of real interest. When a job applicant leaves the room where he was tested, it is T, not X, that determines his capacity for future performance. We cannot observe T, but we can make useful inferences about it.* »

2.2. QUELQUES PROPRIÉTÉS DU MODÈLE CLASSIQUE

Cette section a pour objectif de présenter les principales propriétés du modèle classique, en tout cas celles qui sont susceptibles de nous servir dans ce volume. Nous ne fournirons pas toujours de preuves formelles de ces propriétés, mais plutôt des pistes de nature empirique qui permettront, nous l'espérons, de satisfaire l'intuition. Le lecteur intéressé à obtenir des preuves mathématiques pourra consulter Allen et Yen (1979), Lord et Novick (1968) ou Traub (1994). Nous avons également tenté d'en établir quelques-unes en annexe.

2.2.1. La moyenne des erreurs de mesure

La donnée de l'équation 2.1 et la définition du score vrai permettent d'en arriver à une série de propriétés souvent considérées à tort comme des postulats.

Par exemple, comme nous l'avons observé à la section précédente, la définition même du score vrai permet d'énoncer cette propriété :

- La moyenne des erreurs de mesure pour un individu à qui on a administré un test un très grand nombre de fois est nulle.

Il est facile de se convaincre empiriquement de la véracité de cette propriété en calculant, au tableau 2.3, la moyenne des quatre erreurs de mesure de chacun des 23 étudiants. On peut cependant vérifier cette propriété avec un peu plus de rigueur. En effet, supposons que l'on représente par n le nombre de fois qu'un test a été administré à un individu, ce nombre étant le plus grand possible. Selon l'équation 2.1, chaque erreur de mesure d'un individu pour une répétition donnée i est $E_i = X_i - V$ où X_i est le score observé à la répétition i, V est le score vrai de l'individu et i varie de 1 à n. Ainsi la moyenne de ces E_i est la moyenne des différences comme $X_i - V$; c'est donc aussi la différence entre la moyenne des X_i et la moyenne des V. Mais cette différence est nulle, parce que la moyenne des X_i est, par définition, le score vrai V et la moyenne des V est aussi égale à V (c'est comme additionner V n fois et diviser cette somme par n).

2.2.2. Relation entre les scores vrais et les erreurs de mesure

Comme le mentionnent Crocker et Algina (1986, p. 113), en prenant appui sur la propriété précédente, il est facile de se rendre compte de la véracité de la propriété suivante :

- La corrélation entre les erreurs de mesure et les scores vrais d'un ensemble (très grand) d'individus à qui on administre un test est nulle.

Cette propriété signifie qu'il n'y a pas de relation telle que par exemple « plus les individus sont habiles et plus leur erreur de mesure est faible » ou encore « plus les individus sont habiles et plus leur erreur de mesure est élevée ». Notons que cette propriété tient dans une situation idéale où rien ne peut

perturber la mesure. Elle ne tiendrait pas dans le cas où, par exemple, les étudiants les plus faibles réussissent à obtenir une copie de l'examen final avant qu'il ne soit administré. Ces étudiants faibles auraient alors chacun une erreur de mesure très élevée (positive !) et il s'en suivrait une corrélation négative entre le score vrai et l'erreur de mesure. Allen et Yen (1979, p. 58) mentionnent d'autres exemples où cette propriété ne pourrait tenir. Évidemment, comme en pratique on ne connaît ni le score vrai, ni l'erreur de mesure, il sera difficile de vérifier le tout concrètement si nous ignorons que des étudiants ont eu accès aux questions avant la passation du test. C'est pourquoi on essaie le plus possible, dans les tests à enjeux critiques[9], de standardiser les conditions de passation pour contrôler une partie de l'erreur de mesure.

2.2.3. Relation entre les erreurs de mesure associées à deux tests

Comme le montrent Crocker et Algina (1986, p. 111) :

- ♦ La corrélation est nulle entre les erreurs de mesure à un premier test et les erreurs de mesure à un second test pour une population d'individus à qui on a administré les deux tests.

Cette propriété signifie qu'il n'y a pas de lien linéaire entre les erreurs de mesure d'un groupe d'individus à un test et celles du même groupe d'individus à un autre test. On ne peut donc prédire directement les erreurs de mesure d'individus à un test à partir des erreurs des mêmes individus à un autre test. Tel que formulé par Lord (1980, p. 9) et décrit par Allen et Yen (1979, p. 58-59), dans plusieurs cas, cette propriété ne se vérifie pas. Pensons par exemple à la situation où deux examens uniques du ministère de l'Éducation doivent être administrés une journée de tempête de neige et que cette tempête ne touche que la moitié des élèves. Dans ce cas, les erreurs de mesure associées à ces deux tests risquent d'être plus élevées et négatives pour les élèves affectées par la tempête. Ainsi, la corrélation entre les erreurs de mesure du premier test et celles du second test risque d'être positive.

2.2.4. Le parallélisme entre deux formes de test

Même s'il s'agit, à proprement parler, plus d'une définition que d'une propriété, nous avons tenu à l'énoncer ici :

- ♦ Lorsque, à deux formes du même test (avec scores observés X et X') chaque individu a le même score vrai, V = V', et que la variance des erreurs de mesure est la même, $\sigma^2_E = \sigma^2_{E'}$, alors on dit que les deux formes sont (strictement) parallèles.

9. Traduction libre de *high stake tests.*

Tout comme les scores vrais et les erreurs de mesure, les tests parallèles sont des entités théoriques qui n'existent que pour permettre de définir de nouveaux concepts d'une théorie de la mesure ou d'établir des relations entre ces concepts. Par exemple, la corrélation entre deux formes parallèles $\rho_{XX'}$ est une des façons de définir le très important concept de **fidélité**. Mais, fondamentalement, ce concept est également théorique et tout aussi inobservable. Nous verrons plus loin comment le modèle classique permet d'estimer la fidélité d'un test, c'est-à-dire comment évaluer l'impact de l'erreur de mesure présente dans les scores observés.

Voici enfin deux autres propriétés relatives à la variance observée qui découlent des précédentes propriétés, en supposant que l'on ait affaire à un groupe d'individus très grand (à une population !).

- La variance des scores observés d'un test (aussi dite variance totale) est égale à la somme de la variance des scores vrais et de la variance des erreurs de mesure :

$$\sigma_X^2 = \sigma_V^2 + \sigma_E^2 \tag{2.2}$$

- La fidélité peut aussi être vue comme une proportion de variance vraie dans la variance totale (observée) :

$$\rho_{XX'} = \rho_{XV}^2 = \frac{\sigma_V^2}{\sigma_X^2} \tag{2.3}$$

Ces propriétés nous permettront de développer des approches pour cerner le score vrai ou, ce qui revient au même, d'estimer l'ampleur de l'erreur de mesure.

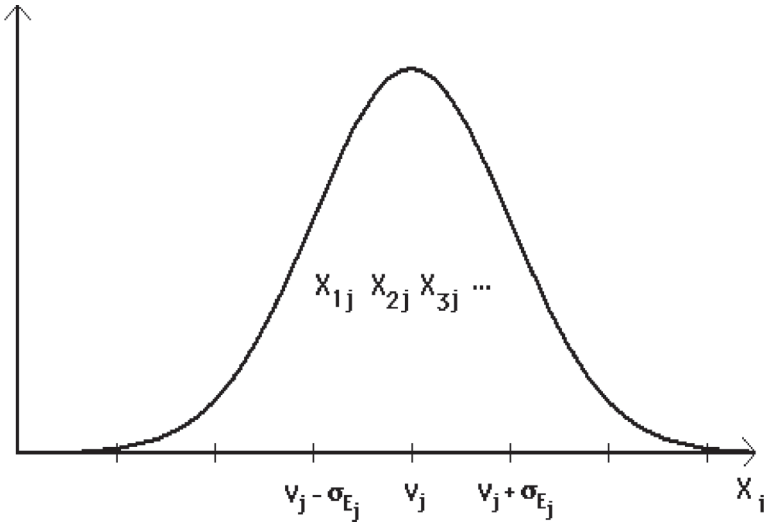
2.3. COMMENT APPRÉHENDER L'ERREUR DE MESURE

Mises à part les situations fictives, il n'est pas possible d'observer le score vrai et donc de quantifier la part d'erreur de mesure dans le score observé d'un individu qui vient de passer un test. Le score vrai et l'erreur de mesure sont des concepts théoriques et non observables. Pour jauger de la capacité du score observé à bien représenter le score vrai, donc pour quantifier l'erreur de mesure, il faudra procéder de manière indirecte en définissant un concept dont nous avons fait état à la section précédente : la fidélité.

Puisque, pour un individu donné j et une répétition i du test, le score vrai est fixe et que $X_{ij} = V_j + E_{ij}$, la variance des scores observés σ_{Xj}^2 est égale à la variance des erreurs de mesure σ_{Ej}^2. Ainsi, comme on le constate à la figure 2.3, s'il était possible d'administrer un test à l'individu noté j un nombre indéterminé de fois, l'écart-type de la distribution des scores observés X_{ij} (pour

i = 1, 2, 3,…), soit la valeur de $\sigma_{Xj} = \sigma_{Ej}$, permettrait de se faire une idée de la variabilité des mesures et d'en inférer les caractéristiques des erreurs de mesure.

FIGURE 2.3
Distribution des scores observés pour un individu de score vrai V_j



Comme on l'a vu, cette situation visant à administrer un test un nombre indéterminé de fois au même individu est purement hypothétique. Il n'est donc pas possible d'obtenir une distribution comme celle présentée à la figure 2.3 et de calculer la valeur de σ_{Ej}. Nous allons plutôt utiliser, en lieu et place, la distribution des observations disponibles, soit les scores observés en rapport avec le groupe d'individus qui a subi le même test que l'individu j. Dans ce cas, nous avions déjà indiqué (équation 2.2) que la variance (totale) des scores observés σ^2_X de ces individus était égale à la somme de la variance des scores vrais σ^2_V et de la variance des erreurs de mesure σ^2_E.

Comme le mentionne Traub (1994, p. 41) σ_E peut être interprétée, en quelque sorte, comme la moyenne (l'espérance mathématique) des σ_{Ej}, cette moyenne étant prise sur la population des individus. Nous appellerons σ_E l'erreur-type de mesure du groupe des individus et σ_{Ej}, l'erreur-type de mesure propre à l'individu j.

S'il n'est toujours pas possible d'obtenir la valeur de σ_E, ni d'ailleurs celles des σ_{Ej}, il sera par contre possible, comme nous le verrons plus loin, d'estimer la valeur de la **fidélité,** que nous avons déjà définie, à la section 2.2

comme la corrélation $\rho_{XX'}$ entre deux formes parallèles, X et X', d'un test ou encore, comme l'exprime l'équation 2.3, la proportion de variance vraie dans la variance observée.

Puisqu'il s'agit d'une proportion, la valeur inférieure que peut prendre la fidélité est de 0 et la valeur supérieure est de 1. Or, le concept de fidélité est intimement relié à l'erreur de mesure puisque, tel que démontré à l'annexe 2.1, la fidélité est égale à 1 si et seulement si l'erreur de mesure est nulle. Ainsi, une fidélité parfaite sera synonyme d'absence d'erreur de mesure : dans ce cas, le score observé d'un individu à un test représentera parfaitement son score vrai.

Il nous reste à trouver des façons d'estimer la fidélité. Pour y arriver, nous utiliserons tour à tour l'une des deux définitions de la fidélité : cette façon de faire permet de distinguer trois approches pour obtenir une estimation de la fidélité, chacune des approches étant basée sur une définition différente des mesures parallèles.

2.4. MÉTHODES D'ESTIMATION DE LA FIDÉLITÉ

2.4.1. La stabilité

Si un test est administré deux fois et que chaque répétition est considérée comme une mesure parallèle, l'estimation de la fidélité, que l'on nomme alors stabilité, consiste à calculer le coefficient de corrélation de Pearson entre les deux répétitions du même test.

S'il peut sembler très facile de calculer un simple coefficient de corrélation entre deux mesures, il faut néanmoins tenir compte de différents éléments contextuels avant de procéder à une expérimentation visant à estimer la stabilité. Premièrement, il faut s'assurer qu'il s'agit bien là d'une estimation de la fidélité utile pour le type de test à l'étude. Il ne serait pas très pertinent, par exemple, de planifier une expérimentation coûteuse qui viserait à estimer la stabilité d'un examen qui ne servirait qu'une seule fois. Deuxièmement, l'échantillon choisi pour cette expérimentation doit être représentatif de la population visée par le test. Troisièmement, les conditions de l'expérimentation (limite de temps, bruit, etc.) doivent refléter les conditions dans lesquelles le test sera habituellement administré. Quatrièmement, et surtout, l'intervalle de temps entre les deux répétitions du test doit prendre en considération des éléments comme la mémorisation des questions du test, la mortalité expérimentale des individus de l'échantillon, mais aussi le changement ou l'apprentissage chez les individus. Plus le laps de temps est court entre les répétitions, plus les individus risquent de mémoriser les questions. Par contre, plus le laps de temps est long, plus il risque de se présenter des problèmes associés à la mortalité expérimentale ou à l'apprentissage des individus. Notons enfin que les effets de mémorisation et d'apprentissage risquent d'être différents

d'un individu à l'autre et que ce sont justement ces effets d'interaction avec les individus qui vont affecter directement la valeur du coefficient de corrélation et, partant, l'estimation de la stabilité.

2.4.2. L'équivalence

Si deux formes d'un test considérées parallèles sont administrées au même groupe d'individus, l'estimation de la fidélité, que l'on nomme alors équivalence, consiste à calculer le coefficient de corrélation de Pearson entre les deux formes du test.

Lorsque plusieurs formes d'un test doivent être construites pour des questions liées à la sécurité par exemple, comme dans le cas des formes d'un test d'intelligence, il est nécessaire d'estimer l'équivalence de chacune des formes. Tout comme dans le cas de l'expérimentation visant à estimer la stabilité, l'échantillon d'individus choisi doit être représentatif de la population visée par les formes du test. De même, les conditions de l'expérimentation doivent être semblables à celles anticipées lors de l'administration régulière du test. Il faut en outre prévoir un laps de temps raisonnable entre l'administration de la première forme et l'administration de la seconde forme, de manière à limiter les effets de la fatigue. Souvenons-nous que ce sont les effets différentiels de la fatigue qui affecteront le coefficient de corrélation, donc la valeur de l'estimation de l'équivalence. Il est préférable de diviser l'échantillon d'individus en deux sous-échantillons équivalents : le premier sous-échantillon subira la première forme suivie de la seconde alors que le deuxième sous-échantillon subira la seconde forme suivie de la première.

2.4.3. La cohérence interne

S'il n'y a qu'une forme du test à l'étude et qu'il n'est pas nécessaire d'obtenir une estimation de la stabilité, il existe une proposition moins coûteuse pour l'estimation de la fidélité qui consiste à administrer le test une seule fois à un groupe d'individus. L'estimation de la fidélité obtenue de la sorte porte le nom de cohérence interne. Deux familles de méthodes ont été développées au cours du dernier siècle : les méthodes fondées sur la bissection et celles reposant sur l'analyse des covariances entre les parties d'un test (souvent les items).

*Méthodes fondées sur la bissection (***split-half***)*

Il s'agit de diviser le test en deux parties, considérées comme autant de mesures parallèles, et de calculer la corrélation entre ces deux parties : une sorte d'équivalence interne. Deux méthodes sont illustrées ici : la méthode de Spearman-Brown et la méthode de Rulon-Guttman.

Méthode de Spearman-Brown

Trois étapes sont nécessaires pour obtenir un estimé de cohérence interne à l'aide de la méthode de Spearman-Brown. Il faut d'abord diviser le test en deux moitiés (considérées parallèles), puis calculer la valeur du coefficient de corrélation entre les deux moitiés et, enfin, obtenir l'estimation de la fidélité (la cohérence interne) du test en appliquant la formule de Spearman-Brown à la valeur du coefficient de corrélation déjà calculée.

La division du test en deux moitiés ne paraît pas très complexe en soi, bien qu'il faille tenir compte au préalable du fait que ces moitiés doivent être considérées parallèles. En ce sens, cette partition du test ne peut s'effectuer au gré de l'utilisateur. Si on a proposé plusieurs façons de définir cette partition par le passé, certaines se sont montrées plus efficaces que d'autres : on peut penser, à titre d'exemples, à la partition impliquant les items pairs[10] du test d'un côté et les items impairs de l'autre, ou encore à celle visant à utiliser un échantillon aléatoire[11] de la moitié des items d'un côté et les autres items de l'autre côté, ou même à tenir compte du contenu des items pour les apparier avant de constituer les moitiés.

La valeur du coefficient de corrélation ainsi obtenue entre les deux moitiés du test constitue bien un estimé de fidélité (d'équivalence), mais seulement pour la moitié du test. Afin d'obtenir un estimé de fidélité pour le test entier, il est nécessaire de corriger cette valeur en utilisant la formule de Spearman-Brown donnée par[12] :

$$r_{XX'} = kr_{YY'} / [\ 1 + (k - 1)\ r_{YY'}]$$

Cette formule donne la façon de calculer le coefficient de fidélité estimatif d'un test X, noté ici $r_{XX'}$, lorsque celui-ci est k fois plus long qu'un test Y et ce, bien sûr, si l'on connaît $r_{YY'}$, le coefficient de fidélité de Y. Dans le cas qui nous concerne ici, comme Y est une moitié de X, alors k = 2 et le coefficient de cohérence interne par la **méthode de Spearman-Brown** est donné par

$$r_{XX'(S\text{-}B)} = 2r_{YY'} / (1 + r_{YY'})$$

où Y et Y' représentent les deux moitiés du test X.

Plus la valeur de $r_{XX'(S\text{-}B)}$ s'approche de 1, plus le test X est considéré fidèle (au sens de la cohérence interne).

---

10. Nous faisons allusion ici à l'ordre des items dans le test, étant entendu que les items faciles se trouvent le plus souvent au début du test et les items difficiles à la fin.
11. Si le nombre d'items est suffisant pour le justifier.
12. Nous noterons un estimé de fidélité par $r_{XX'}$.

Bien que cette méthode d'estimation de la fidélité semble attrayante, il ne faut pas oublier que la valeur de l'estimé de fidélité dépend, en partie du moins, de la façon dont les moitiés ont été constituées. Nous allons voir plus loin des méthodes d'estimation de la fidélité qui ne dépendent pas de la division du test en moitiés et qui en font des méthodes donnant des estimés plus stables.

Méthode de Rulon-Guttman

C'est Rulon (1939) qui a eu l'idée d'utiliser la différence entre les scores aux deux moitiés d'un test comme base pour proposer un estimé de la cohérence interne. Remarquant que la fidélité est une proportion de variance vraie dans la variance totale et s'appuyant sur l'équation 2.2, celui-ci a proposé le coefficient suivant :

$$r_{XX'(\text{Rulon})} = \frac{s_V^2}{s_X^2} = 1 - \frac{s_E^2}{s_X^2}$$

où E = Y – Y' représente l'erreur ou la différence entre les deux moitiés Y et Y' tandis que X = Y + Y' représente le test entier.

La valeur du coefficient de Rulon-Guttman est très près de la valeur du coefficient de cohérence interne obtenu par la formule de Spearman-Brown (voir le tableau 2.4). En fait, on peut montrer que si Y et Y' étaient des mesures parfaitement parallèles, ces deux valeurs seraient rigoureusement égales. L'annexe 2.3 donne une preuve mathématique de ce dernier énoncé.

Au fait, pourquoi l'appelle-t-on le coefficient de Rulon-Guttman ? Parce que, indépendamment des efforts de Rulon, Guttman (1945) a proposé un coefficient tout à fait similaire, mais en le formulant différemment de Rulon. Incidemment, cette formulation de Guttman n'est pas sans rappeler celle du coefficient KR-20 (Kuder et Richardson, 1937) ou encore celle du fameux coefficient alpha (Cronbach, 1951) que nous aborderons bientôt. Avec la même notation que précédemment, le coefficient de Guttman se formule comme suit :

$$r_{XX'(\text{Guttman})} = 2\left\{1 - \left(\frac{\left(s_Y^2 + s_{Y'}^2\right)}{s_X^2}\right)\right\}$$

L'annexe 2.2 démontre l'équivalence des formules de Rulon et de Guttman.

Précisons que lorsque les moitiés ne peuvent pas être considérées parallèles, le coefficient de Rulon-Guttman donne un estimé de la limite inférieure à la fidélité (Traub, 1994, p. 81).

Méthodes fondées sur les covariances (corrélations)

Les méthodes d'estimation de la fidélité fondées sur la bissection sont attrayantes de par leur simplicité, mais elles présentent toutes la même faiblesse : la valeur de l'estimé de fidélité dépend de la façon dont le test est divisé en moitiés. Les deux méthodes que nous allons maintenant présenter ne sont pas fragilisées par cette contrainte. Elles sont basées sur les corrélations ou les covariances entre les items. Nous traiterons tour à tour du coefficient alpha de Cronbach et du coefficient L_2 de Guttman.

Coefficient alpha de Cronbach

Au moment où Cronbach s'intéressait à différentes façons d'estimer la fidélité, soit au début des années 1950, tout était en place pour populariser le fameux coefficient alpha. En 1936, Kuder et Richardson avaient défini le KR-20, un cas particulier de ce qui allait être le coefficient alpha : le KR-20, en effet, permettait d'estimer la cohérence interne d'un test dont les items étaient corrigés de façon dichotomique en utilisant les mêmes ingrédients statistiques de base que le coefficient alpha. Hoyt, en 1941, avait défini une procédure, basée sur l'analyse de la variance, qui donnait un estimé de la cohérence interne identique à ce qu'allait donner le coefficient alpha. En 1945, Guttman définissait plusieurs coefficients qui visaient tous plus ou moins à fournir un estimé de la fidélité : un de ces coefficients, le L_3, était justement une forme de ce qui allait devenir le coefficient alpha.

Il y a plusieurs façons différentes de formuler le coefficient alpha. Nous avons choisi de présenter les formes les plus classiques.

Soit un test de n items dont on connaît la variance de chacun des items, s_i^2, la covariance entre les items, s_{ij} et la variance du test, s_X^2. Alors, le coefficient alpha est donné par l'une ou l'autre des deux formules suivantes :

$$\alpha = [n/(n-1)]\left[1-\left(\Sigma_i s_i^2 / s_X^2\right)\right] = [n/(n-1)]\left[\Sigma_{ij} s_{ij} / s_X^2\right]$$

C'est donc dire que la valeur prise par le coefficient alpha est d'autant plus élevée que les covariances s_{ij} entre les items sont elles-mêmes élevées. Notons, au passage, que la somme des covariances Σ_{ij}, est prise sur toutes les paires d'items i et j, où $i \neq j$. Remarquons que le fait de considérer les items comme les n parties du test n'est pas du tout restrictif. Ces n parties pourraient tout aussi bien être des regroupements d'items. À la limite, nous pourrions considérer n = 2 parties, notées Y et Y' ; le coefficient alpha reviendrait alors ni plus ni moins qu'au coefficient de Guttman, soit :

$$r_{XX'(\text{Guttman})} = 2\left\{1-\left(\frac{\left(s_Y^2 + s_{Y'}^2\right)}{s_X^2}\right)\right\}$$

Une troisième façon tout aussi importante de formuler le coefficient alpha est de considérer l'expression suivante où r_{iX} est la corrélation entre l'item i et le test X :

$$\alpha = [n/(n-1)]\left[1 - \left(\Sigma_i s_i^2\right)/\left(\Sigma_i s_i r_{iX}\right)^2\right]$$

Cette expression montre que la valeur de la cohérence interne est très intimement liée à la corrélation item-total r_{iX}, que nous appellerons plus loin l'indice de discrimination. Ainsi, plus les valeurs de l'indice de discrimination sont élevées, c'est-à-dire plus le degré d'association entre les items et le total est élevé, plus la valeur du coefficient alpha est élevée.

Il est nécessaire de préciser que le coefficient alpha est un estimé de la fidélité si les n parties (n items) du test peuvent être considérées parallèles. Autrement, et c'est la situation qui prévaut la plupart du temps, le coefficient alpha doit être considéré comme une limite inférieure de la fidélité du test.

Coefficient L_2 de Guttman

Pratiquement inconnu à cause de sa formulation rébarbative[13], le coefficient L_2 (Guttman, 1945) doit être, à notre avis, considéré avec beaucoup plus d'égard qu'il ne l'a été jusqu'ici.

$$L_2 = \left[\left(\Sigma_{ij} s_{ij}\right)/s_X^2\right] + \left\{[n/(n-1)]^{1/2}\left[\left(\Sigma_{ij} s_{ij}^2\right)\right]^{1/2}/s_X^2\right\}$$

À l'instar de Traub (1994, p. 89), nous recommandons l'utilisation de ce coefficient qui, comme le coefficient alpha, est une limite inférieure de la fidélité, mais qui possède toujours une valeur supérieure ou égale au coefficient alpha. On peut donc considérer qu'il donne un estimé de cohérence interne plus près de la réalité.

En d'autres termes, $\alpha \leq L_2 \leq \rho_{XX'}$.

Le tableau 2.4 présente une comparaison des valeurs prises par les quatre coefficients de cohérence interne décrits plus haut dans le cas des données du tableau 2.2. On peut y voir les valeurs de chacun des coefficients obtenus à partir des mêmes données. La première remarque que l'on peut faire concerne la très grande disparité entre les valeurs des coefficients obtenues par toutes ces méthodes, la valeur du coefficient L_2 offrant, dans ce cas-ci, une sorte de compromis.

---

13. La notation est la même que pour le calcul du coefficient alpha de Cronbach.

Tableau 2.4
Coefficients de cohérence interne obtenus par diverses méthodes à l'aide des données du tableau 2.2 (23 individus et 8 items)

Coefficients de cohérence interne			
$r_{XX'(Rulon)}$	$r_{XX'(S\text{-}B)}$	L_2	α
0,646	0,651	0,473	0,327

Il faut cependant remarquer que notre exemple ne comprend que huit items et 23 individus. En serait-il vraiment autrement avec un autre groupe ou avec un test comportant plus d'items et administré à un groupe d'individus beaucoup plus considérable ? Bien sûr, les valeurs des coefficients obtenues ici sont très sensibles puisqu'elles reposent sur les variances et les covariances entre les items. Or, il est entendu qu'avec de petits échantillons d'individus et d'items, il suffit de quelques valeurs plus ou moins aberrantes pour affecter les coefficients, parfois même de manière importante. Cette instabilité tend toutefois à s'estomper à mesure que la taille des échantillons s'accroît. Afin de concrétiser cette assertion, nous avons calculé les valeurs des mêmes coefficients à l'aide d'échantillons plus importants. Les résultats se trouvent aux tableaux 2.5 et 2.6. En considérant également les résultats du tableau 2.4, nous constatons ce qui suit :

- ♦ Plus la taille des échantillons (d'individus et d'items) augmente, plus les valeurs des coefficients estimant la cohérence interne augmentent.
- ♦ Plus la taille des échantillons augmente, moins il y a de variabilité entre les valeurs des quatre coefficients à l'étude.
- ♦ La variabilité entre les valeurs des deux coefficients basés sur la bissection est très faible dans tous les cas.

Tableau 2.5
Coefficients de cohérence interne obtenus par diverses méthodes à l'aide des données d'un échantillon de 100 individus et 14 items

Coefficients de cohérence interne			
$r_{XX'(Rulon)}$	$r_{XX'(S\text{-}B)}$	L_2	α
0,739	0,746	0,788	0,777

Tableau 2.6
Coefficients de cohérence interne obtenus par diverses méthodes à l'aide des données d'un échantillon de 1000 individus et 76 items

Coefficients de cohérence interne			
$r_{XX'(Rulon)}$	$r_{XX'(S\text{-}B)}$	L_2	α
0,929	0,929	0,920	0,918

2.5. ANALYSE D'ITEMS

Nous appelons **analyse d'items** la procédure à suivre pour examiner certaines caractéristiques métriques des items et du test. Il s'agit d'utiliser les différents indices et coefficients obtenus à partir des items et du test pour juger de la valeur des items et, en bout de ligne, de l'instrument. Lorsque le budget le permet, il est préférable d'effectuer cette analyse avant de prendre des décisions d'ordre administratif ou pédagogique à partir des résultats des individus au test. Il s'agit alors d'utiliser un groupe d'individus (un groupe cobaye) représentatif de la population initialement ciblée par le test, de lui administrer le test et de calculer les indices et coefficients voulus.

En pratique, cependant, il n'est pas toujours possible de mettre le test à l'essai avec un groupe cobaye. Il peut arriver que ce soit contre-indiqué de procéder avec un groupe cobaye pour des raisons déontologiques ou autres. Par exemple, si les individus du groupe cobaye savent que le test ne compte pas leur comportement risque d'être différent, situation susceptible d'affecter les réponses aux items. Ainsi, les indices et coefficients calculés à partir de ces réponses affectées ne constitueront pas un portrait fidèle de ce qui se serait produit en situation véritable de testing, biaisant de la sorte les résultats de l'analyse. De plus, si on emploie un groupe cobaye, il ne faut pas négliger les opérations visant à assurer la confidentialité des questions.

C'est pourquoi l'analyse d'items se réalise souvent *a posteriori*, une fois l'instrument administré en situation véritable de testing. Cette analyse peut mener à identifier voire à rejeter des items dits aberrants, à savoir ceux qui ne se comportent pas comme les autres items du test, ceux qui ne sont pas associés aux autres items du test ou, pire, ceux qui y sont associés négativement. En effet, souvenons-nous que dans une situation de mesure, un item vise en quelque sorte à être un portrait miniature du test et, par le fait même, doit être associé fortement et positivement aux autres items et au test.

2.5.1. Indices d'items

Avant d'effectuer une analyse d'items, il convient de présenter certains indices propres aux items qui serviront à déterminer leurs caractéristiques métriques.

Indice de difficulté

Le premier de ces indices est appelé l'indice de difficulté d'un item et est noté p_i. C'est la proportion d'individus d'un groupe donné qui réussissent l'item i. Donc,

> p_i = nombre d'individus réussissant l'item i / nombre d'individus du groupe.

L'indice p_i est en fait un indice de facilité puisque plus p_i augmente, plus l'item i est considéré facile. Cependant, suivant la tradition, nous conserverons l'expression consacrée d'indice de difficulté.

Indices de discrimination

La puissance de discrimination d'un item est son aptitude à faire la distinction entre les individus plus habiles et les individus moins habiles ou entre les individus ayant atteint un certain niveau d'habileté et ceux qui ne l'ont pas atteint. C'est bien souvent l'objectif premier d'un test que de pouvoir distinguer les individus forts des individus faibles ; c'est pourquoi le concept de discrimination et les indices associés seront d'une importance capitale lors de l'examen d'un item, voire d'un test. En fait, ce qu'on attend de tout item d'un test, c'est d'être en quelque sorte le portrait miniature du test. Idéalement, l'item devrait nous renseigner de la même façon que le fait le test lui-même. Il faut donc que l'item soit le plus associé possible au test. C'est pourquoi la corrélation de Pearson entre les scores à un item i et les scores au test X, r_{iX}, aussi appelée **corrélation item-total**, est souvent employée comme indice de discrimination. Lorsque l'item est corrigé de façon dichotomique (0 ou 1, par exemple), il est d'usage d'appeler r_{iX} la corrélation bisériale en point (Bertrand et Valiquette, 1986, p. 310).

Parce que le test X comprend notamment l'item i, la corrélation entre i et X est toujours biaisée à la hausse : c'est-à-dire qu'il est possible d'observer une corrélation positive et relativement forte entre i et X même si l'item i n'est pas bien relié aux autres items du test. C'est pourquoi on applique habituellement avec un test court une correction à cet indice de discrimination : **la corrélation item-total corrigée** est la corrélation entre les scores à l'item i et les scores à la variable X – i, c'est-à-dire le test X amputé de l'item i. Plus le test sera court, plus l'ampleur de la correction sera grande et plus il sera important d'effectuer cette correction. En pratique, il est plus sage de toujours utiliser la valeur corrigée.

Il est tout à fait possible et même fréquent qu'un item soit corrigé de façon dichotomique mais que cette dichotomisation soit en réalité bien artificielle[14], au sens où il serait légitime de considérer qu'il existe un continuum théorique entre 0 et 1. Autrement dit, il serait théoriquement possible d'obtenir un score à l'item de 0,25, 0,5 ou même 0,86. Si l'on peut supposer que ces scores théoriques possibles entre 0 et 1 suivent une distribution normale (Bertrand et Valiquette, 1986, p. 310), il est justifiable d'utiliser la **corrélation bisériale** entre l'item i et le test X, r'_{iX}, comme indice de discrimination. On peut montrer (Crocker et Algina, 1986, p. 318) que, si on note $1 - p_i = q_i$, alors :

$$r'_{iX} = r_{iX}\left[p_i q_i\right]^{1/2} / Y$$

14. Si on donnait le score 1 à toutes les personnes dont le père est vivant et le score 0 à toutes celles dont le père est mort, on aurait affaire à une variable dichotomique non artificielle.

où Y est l'ordonnée de la courbe normale centrée et réduite à l'endroit où l'abscisse z correspond à la surface p_i située sous la courbe et à la gauche de z (voir Hulin *et al.*, 1983, p. 238).

2.5.2. Un modèle de fidélité

Avant d'effectuer une analyse d'items, il nous a semblé utile de présenter un modèle de ce qui peut être considéré comme une fidélité parfaite pour les réponses aux items d'un test administré à un groupe d'individus donné. Ce modèle peut se présenter comme un tableau de réponses aux items (par exemple le tableau 2.2) qui donnerait des indices parfaits, des corrélations parfaites, des coefficients parfaits (c'est-à-dire égaux à 1) pour toutes les méthodes (connues) l'estimation de la fidélité. Par exemple, si la corrélation entre les deux moitiés d'un test est égale à 1, le coefficient d'estimation de la fidélité suivant l'approche de Spearman-Brown est égal à 2(1) / (1 + 1) = 1 ; il est donc parfait. Mais cela ne veut pas nécessairement dire que le coefficient alpha et le coefficient L_2 seront eux aussi parfaits.

Qu'est-ce à dire si toutes les corrélations entre les items sont parfaites, égales à 1 ? Dans ce cas, le patron de réponses doit être le même pour chacun des items, un peu comme au tableau 2.7. Ceci implique que chaque individu a soit un score parfait (8 / 8), soit un score nul (0 / 8), et que chaque item a le même indice de difficulté, soit ici 0,688 (11 / 16), et la même variance (de population), soit 0,215.

Examinons maintenant les indices de discrimination des items. Il va de soi que les corrélations bisériales en point sont toutes égales à 1, donc parfaites, puisque les patrons de réponses aux items sont les mêmes. En outre, les corrélations bisériales en point corrigées sont aussi égales à 1 puisque, même en enlevant l'item du total, le patron du test reste le même, tous les autres items ayant le même patron. Typiquement, au lieu d'avoir le patron original {8 0 8 8 […] 0 8}, le patron corrigé du test sera {7 0 7 7 […] 0 7} et la corrélation sera inchangée, soit égale à 1. Il est intéressant de noter que les corrélations bisériales seront dès lors plus grandes que 1. Si l'on se rapporte à la relation établie entre la corrélation bisériale et la corrélation bisériale en point on s'aperçoit qu'il s'agit d'un artifice[15] dû au fait que Y est toujours plus grand que $[p_i q_i]^{1/2}$ et que, dans notre cas, la corrélation bisériale en point est déjà égale à 1.

Qu'en est-il maintenant des coefficients de cohérence interne ? Tous les coefficients calculés à l'aide des méthodes fondées sur la bissection du test seront égaux à 1 puisque, peu importe comment les items sont divisés en

15. Nunally (1978, p. 136) donne de multiples raisons de ne pas utiliser la corrélation bisériale lors d'une analyse d'items : en voici peut-être une autre.

moitiés, le patron de chaque moitié Y et Y' sera toujours le même, soit {4 0 4 4 […] 0 4}. La corrélation entre les moitiés, mais aussi le coefficient de cohérence interne de Spearman-Brown, seront donc égaux à 1. Il en sera de même pour le coefficient de Rulon-Guttman, puisque la moitié Y étant égale à la moitié Y', E = Y – Y' = 0, donc $s_E^2 = 0$.

TABLEAU 2.7
Réponses de 16 individus à un test de huit items (données simulées)

	Item								
Étudiants	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	Total
1	1	1	1	1	1	1	1	1	8
2	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	8
4	1	1	1	1	1	1	1	1	8
5	1	1	1	1	1	1	1	1	8
6	1	1	1	1	1	1	1	1	8
7	1	1	1	1	1	1	1	1	8
8	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0
10	1	1	1	1	1	1	1	1	8
11	0	0	0	0	0	0	0	0	0
12	1	1	1	1	1	1	1	1	8
13	1	1	1	1	1	1	1	1	8
14	1	1	1	1	1	1	1	1	8
15	0	0	0	0	0	0	0	0	0
16	1	1	1	1	1	1	1	1	8

Reste à considérer le coefficient alpha et le coefficient L_2. Rappelons d'abord la formule du coefficient alpha (α) :

$$\alpha = [n/(n-1)]\left[1-\left(\Sigma_i s_i^2 / s_X^2\right)\right]$$

Il s'agit de calculer la variance du test, s_X^2, et la variance de chacun des items, s_i^2. Souvenons-nous que 11 individus ont un score parfait et cinq individus ont un score nul. Ainsi, la moyenne du test est égale à (11 × 8) / 16 = 5,5. Donc la variance du test est s_X^2 = {[11 × (8 – 5,5)^2] + [5 × (0 – 5,5)^2]} /16 = {68,75 + [151,25]} /16 = 220/16 = 55/4. Par ailleurs, la variance de chacun des items est égale à $s_i^2 = p_i q_i$ = 11/16 (5/16) = 55/256.

Or, comme chaque item a le même profil, il a aussi la même variance.

Donc, $\Sigma_i s_i^2 = 8(55/256) = 55/32$.

Et α = (8/7) [1– (55/32) / (55/4)] = (8/7) [1 – 4/32] = (8/7) [7/8] = 1.

La preuve que L_2 = 1 est redondante avec celle concernant le coefficient α, mais beaucoup plus technique : nous avons préféré la reproduire à l'annexe 2.4.

Ainsi, tous les coefficients d'estimation de la fidélité connus mènent à la même valeur, soit 1. C'est en ce sens que nous dirons que les scores présentés au tableau 2.7 constituent une sorte de modèle de fidélité parfaite, ce vers quoi devraient tendre les données pour être fidèles, dans le contexte de la théorie classique. La caractéristique principale de ce modèle est bien sûr de minimiser la variance entre les items. Que cette absence de variance entre les items mène à une fidélité parfaite n'est peut-être pas si étonnant. En effet, si les items sont vus comme des répétitions d'une mesure, puisque la note obtenue par chaque individu est la même pour les 8 items, on peut dire qu'il n'y a pas d'erreur de mesure dans ces données, un reflet, comme l'a déjà montré, d'une fidélité parfaite.

2.6. L'ERREUR-TYPE DE MESURE

Comme nous l'avons montré à la section 2.3, dans la perspective (théorique !) où on administre le même test à un individu, noté j, un nombre indéterminé de fois, l'écart-type de la distribution (figure 2.3) des scores observés de cet individu constitue ce qui est convenu d'appeler l'erreur-type de mesure propre à l'individu j, qu'on note σ_{Ej}. Cette entité est cependant théorique et non observable.

Nous savons également que la moyenne de ces erreurs-types de mesure constitue l'erreur-type de mesure du groupe des individus et se note σ_E. C'est cette erreur-type de mesure propre à un groupe d'individus que nous allons tenter d'estimer. Partant de l'équation 2.2, nous pouvons écrire

$$\sigma_E^2 = \sigma_X^2 - \sigma_V^2 = \sigma_X^2 - \sigma_X^2\left(\frac{\sigma_V^2}{\sigma_X^2}\right) = \sigma_X^2\left(1 - \frac{\sigma_V^2}{\sigma_X^2}\right) = \sigma_X^2\left(1 - \rho_{XV}^2\right)$$

En d'autres termes, si on utilisait le coefficient alpha comme estimé du coefficient de fidélité ρ_{XV}^2, il serait possible d'obtenir un estimé, noté s_E, de l'erreur-type de mesure propre à un groupe d'individus par :

$$s_E = s_X\,(1 - \alpha)^{1/2}.$$

Cette erreur-type peut ensuite servir à encadrer le score vrai en établissant un intervalle de confiance à 68 % autour d'un score observé[16] X_j. Pour un individu j par exemple, on sera certain à 68 % que son score vrai V_j se trouve dans l'intervalle $[X_j - s_E, X_j + s_E]$.

16. Traub (1994, p. 42) affirme qu'un intervalle de confiance à 68 % est suffisant autour d'un score individuel.

La plus grande difficulté que nous éprouvons face à cet estimé est qu'il vaut pour un groupe d'individus et non pour un individu en particulier. Or nous savons bien que, en théorie du moins, chaque individu j peut posséder une erreur-type de mesure σ_{Ej} distincte. Dans ce contexte, utiliser un estimé commun s_E à tous les individus du groupe est de nature à fausser la réalité. Pourquoi alors ne pas partitionner le groupe d'individus en sous-groupes en fonction de leur niveau d'habileté et obtenir un estimé pour chaque sous-groupe ? Ainsi, comme il se doit, tous les individus de même habileté auraient le même estimé d'erreur-type de mesure. Le problème se pose alors de définir ces sous-groupes d'individus de même habileté. Une façon facile de résoudre ce problème est d'utiliser le score observé au test comme un indicateur de l'habileté. Il s'agit ensuite de constituer un sous-groupe pour chaque score observé et d'obtenir une valeur de s_E pour chaque sous-groupe. Woodruff (1990) a montré qu'il s'agit là d'une procédure qui mène à des erreurs-types de mesure biaisées. C'est pourquoi il propose l'approche suivante.

Première étape : Diviser le test en deux parties (autant que possible) parallèles, par exemple en considérant les items pairs pour la partie 1 et les items impairs pour la partie 2.

Deuxième étape : Constituer les sous-groupes d'individus en se basant sur les scores totaux obtenus par les individus à la partie 1 du test.

Troisième étape : Effectuer une analyse sur les items de la partie 2 du test et ce pour chacun des sous-groupes d'individus formés à la deuxième étape afin d'obtenir, pour chaque sous-groupe d'individus, les valeurs de s_X et de α.

Quatrième étape : Calculer s_E pour chacun des sous-groupes à l'aide des valeurs de s_X et de α déjà calculées à la troisième étape ; on aura ainsi obtenu un estimé de l'erreur-type de mesure pour la moitié du test seulement.

Cinquième étape : Afin d'obtenir une erreur-type de mesure pour l'ensemble du test (Traub, 1994, p. 120), il faut calculer $\left(2s_E^2\right)^{1/2}$.

Cette procédure en cinq étapes peut sembler un peu fastidieuse, voire rébarbative, mais elle a le mérite de produire des erreurs-types de mesure non biaisées et le plus près possible des inaccessibles σ_{Ej}. Une des difficultés que nous avons rencontrées est liée à la taille de l'échantillon d'individus nécessaire pour constituer un nombre suffisant de sous-groupes avec, pour chacun, un nombre minimal d'individus. Traub (1994, p. 124) présente un exemple assez convaincant qui comprend 498 individus auxquels a été administré un test de 36 items.

Nous discutons maintenant d'un exemple consistant en un groupe de 1000 individus québécois de 13 ans à qui on a administré un test de mathématique de 76 items dans le cadre de l'enquête de l'IAEP2 (Lapointe, Mead et Askew, 1992). Nous avons défini la partie 1 du test comme l'ensemble des items impairs et la partie 2 du test comme celui des items pairs. Les sous-groupes d'individus ont donc été formés à partir du score total aux 38 items

impairs du test. Même si, théoriquement du moins, il était possible de constituer 39 sous-groupes, nous avons dû regrouper certains sous-groupes adjacents de taille insuffisante. Pour chacun de ces sous-groupes, on a effectué une analyse des 38 items pairs, qui a produit et les statistiques suivantes : s_X, α, s_E et $\left(2s_E^2\right)^{1/2}$. Notons que toutes ces statistiques s'obtiennent en analysant les scores totaux aux 38 items pairs tour à tour pour chacun des sous-groupes d'individus constitués à partir des scores aux items impairs. Voici, au tableau 2.8, les résultats que nous avons observés pour quelques-uns des sous-groupes.

TABLEAU 2.8
Calcul de l'erreur-type de mesure conditionnelle de Woodruff (1990) pour un test de 76 items administré à un groupe de 1000 individus

Sous-groupes (scores aux items impairs)	s_X	α	s_E	$\left(2s_E^2\right)^{1/2}$
16	3,6184	0,4613	2,6558	3,7558
20	3,2873	0,3230	2,7048	3,8252
23	3,1180	0,2884	2,6302	3,7197
26	3,9771	0,5825	2,5698	3,6342
27	3,6436	0,5358	2,4825	3,5107
32	2,7103	0,4143	2,0704	2,9333
33	2,6448	0,4011	2,0468	2,8946
36	2,5510	0,5500	1,7113	2,4201

La tendance soulignée par Traub (1994, p. 121) à produire des valeurs d'erreur-type de mesure plus petites pour les sous-groupes situés aux extrémités (les très faibles et les très forts) ne se vérifie que partiellement dans notre cas. Il faut dire, cependant, que nous avons été contraints de fusionner en un seul sous-groupe tous les sous-groupes de très petite taille formés des scores de 0 à 16.

Cette approche, bien qu'attrayante à plus d'un point de vue, exige cependant un nombre important d'individus et d'items : tant qu'à se plier à cette contrainte au niveau des échantillons d'individus et d'items, il faut se demander s'il ne vaut pas mieux utiliser l'un ou l'autre des modèles de la TRI qui permettent d'obtenir automatiquement une erreur-type de mesure propre à chaque individu. C'est une question que nous laissons ouverte pour le moment.

Exercices

1. Imaginons que l'on puisse répéter plusieurs fois l'administration d'un test de géographie à Pauline et que celle-ci ne se souvienne pas des questions d'une répétition à l'autre. Déterminez le score de Pauline qui représente le mieux son habileté en géographie telle que mesurée par ce test si la distribution des scores observés de Pauline aux diverses répétitions est la suivante :

Score	Fréquence relative
55	0,15
56	0,00
57	0,25
58	0,13
59	0,12
60	0,05
61	0,30

2. Trouvez une situation de testing de la vie courante où la propriété de corrélation nulle entre les scores vrais et les erreurs de mesure ne sera probablement pas respectée.

3. Générez deux échantillons de données dont le premier produira une fidélité plus grande mais une erreur-type de mesure de groupe plus petite que le deuxième.

4. Pourquoi les méthodes d'estimation de la fidélité basées sur une bissection sont-elles moins recommandables que les méthodes fondées sur les variances et les covariances ?

5. Trouvez une autre situation (d'autres données) que celle présentée au tableau 2.7 où toutes les valeurs des coefficients d'estimation de la fidélité seront égales à l'unité.

6. Trouvez un échantillon de données du type de celui présenté au tableau 2.7 qui produira des valeurs nulles pour tous les coefficients d'estimation de la fidélité.

7. Qu'arrive-t-il de la grandeur de l'intervalle de confiance à 68 % autour du score observé si s_x^2 est quadruplée ?

8. Que se passe-t-il si l'échantillon d'individus servant à l'estimation de la fidélité est très homogène ?

9. Un test de 76 items est administré à un échantillon de 1000 individus. Comment déterminer la cohérence interne de ce test si on connaît la cohérence interne de chacun des quatre sous-tests de 19 items de ce test, à savoir $\alpha_1 = 0{,}7597$, $\alpha_2 = 0{,}6917$, $\alpha_3 = 0{,}7600$, $\alpha_4 = 0{,}7425$?

10. Comment est-ce possible que, pour un test de mathématique de 25 items, la valeur de la cohérence interne soit de 0,79 en utilisant un échantillon d'étudiants de 16 ans mais seulement de 0,29 en utilisant un échantillon d'étudiants de 13 ans de même taille que l'échantillon précédent ?

11. Un test d'histoire est administré de manière répétée à deux personnes, Karine et Jean. Les distributions des scores observés de ces deux individus se trouvent au tableau suivant.

    a) Trouvez le score vrai de Karine et le score vrai de Jean.

    b) Selon vous, quel est le score vrai le plus fiable ?

Score observé de Karine à l'examen d'histoire (X)	**Fréquence relative**	**Score observé de Jean à l'examen d'histoire** (X)	**Fréquence relative**
76	0,10	76	0,00
78	0,20	78	0,00
80	0,15	80	0,50
82	0,25	82	0,50
83	0,10	83	0,00
85	0,20	85	0,00

12. Donnez un exemple de situation de mesure pour lequel la corrélation entre l'erreur de mesure et le score vrai risque d'être positive.

13. Est-ce possible que la moyenne des erreurs de mesure d'un individu soit positive ?

14. Pourquoi n'est-il pas possible d'observer l'erreur-type de mesure associée à un individu ?

15. Combien d'items faudrait-il ajouter à un test de 15 items pour que sa cohérence interne passe de 0,60 à 0,75 ?

16. Dans quel sens peut-on dire que la stabilité, l'équivalence et la cohérence interne sont des formes de fidélité ?

Corrigé des exercices

1. $V = (55 \times 0{,}15) + (57 \times 0{,}25) + \ldots + (61 \times 0{,}30) = 58{,}42$

3. Il s'agit de trouver deux échantillons d'individus A et B tels que l'échantillon A est plus homogène que l'échantillon B, soit $\sigma_{X_A} < \sigma_{X_B}$, mais dont $\alpha_A > \alpha_B$.

5. Tout tableau de données qui, comme le tableau 2.7, contiendra des individus qui auront obtenu le même score à tous les items.

7. Si s_X^2 est quadruplée, c'est donc que s_X est doublé et aussi s_E, puisque $s_E = s_X (1 - \alpha)^{1/2}$. La grandeur de l'intervalle de confiance à 68 % est elle aussi doublée.

9. Pourquoi ne pas prendre la moyenne des 4 valeurs du coefficient alpha puis utiliser la formule de Spearman-Brown ? La moyenne des 4 valeurs du coefficient alpha donne 0,7385 et est un bon estimé d'un sous-test de 19 items représentant bien les 76 items du test. En appliquant la formule de Spearman-Brown, l'estimé recherché devient

   $\alpha = 4\ (0{,}7385) / [1 + 3(0{,}7385)] = 0{,}9187$

10. a) $V_{Karine} = V_{Jean} = 81$

    b) Le critère doit être l'erreur-type de mesure. Or Jean possède une erreur-type de mesure plus faible que celle de Karine :
    $\sigma_{E_{Karine}} = \sigma_{X_{Karine}} = 2{,}90$ alors que $\sigma_{E_{Jean}} = \sigma_{X_{Jean}} = 1{,}01$

13. Non, la moyenne des erreurs de mesure est nulle à moins, bien sûr, qu'il ne s'agisse pas d'une erreur aléatoire. Ce type d'erreur non aléatoire, dite aussi systématique, surviendrait si, par exemple, on administrait un test d'intelligence informatisé à une personne d'un pays en voie de développement qui ne connaît pas bien le fonctionnement d'un ordinateur. Lorsque le test sera répété, la moyenne des erreurs de mesure a bien des chances d'être négative.

15. Quinze autres items puisque $0{,}75 = 2\ (0{,}6) / (1 + 0{,}6)$.

Annexe 2.1

Preuve de l'équivalence de $\rho^2_{XV} = 1$ et de $E_{ij} = 0$ pour tous les individus j à une seule répétition i du test

Si le symbole ⇔ tient la place de l'expression « équivaut à » alors :

$\rho^2_{XV} = 1 \Leftrightarrow \sigma^2_V = \sigma^2_X$, d'après la définition même de la fidélité.

De même, $\sigma^2_V = \sigma^2_X \Leftrightarrow \sigma^2_E = 0$, puisque $\sigma^2_X = \sigma^2_V + \sigma^2_E$.

De plus, $\sigma^2_E = 0 \Leftrightarrow \sigma^2_{E_j} = 0$ pour tous les individus j puisque σ^2_E peut être vue comme une moyenne des $\sigma^2_{E_j}$.

Puis $\sigma^2_{E_j} = 0 \Leftrightarrow \sum_i \left(E_{ij}\right)^2 = 0$ puisque la moyenne des E_{ij} est nulle.

Enfin, $\sum_i \left(E_{ij}\right)^2 = 0 \Leftrightarrow$ chaque $E_{ij} = 0$ pour chaque répétition i du test.

Annexe 2.2

Preuve de l'équivalence des formules de Rulon et de Guttman

Selon Rulon, $r_{XX'(\text{Rulon})} = \frac{s_V^2}{s_X^2} = 1 - \frac{s_E^2}{s_X^2}$ où E = Y – Y' et X = Y + Y'

Or, compte tenu de la formule de la variance d'une somme, soit $s_{A+B}^2 = s_A^2 + s_B^2 + 2r_{AB}s_A s_B$,

$$
\begin{aligned}
r_{XX'(\text{Rulon})} &= 1 - \frac{s_E^2}{s_X^2} = 1 - \frac{s_Y^2 + s_{Y'}^2 - 2r_{YY'}s_Y s_{Y'}}{s_Y^2 + s_{Y'}^2 + 2r_{YY'}s_Y s_{Y'}} \\
&= \frac{s_Y^2 + s_{Y'}^2 + 2r_{YY'}s_Y s_{Y'} - s_Y^2 - s_{Y'}^2 + 2r_{YY'}s_Y s_{Y'}}{s_Y^2 + s_{Y'}^2 + 2r_{YY'}s_Y s_{Y'}} \\
&= \frac{4r_{YY'}s_Y s_{Y'}}{s_Y^2 + s_{Y'}^2 + 2r_{YY'}s_Y s_{Y'}} \\
&= \frac{4r_{YY'}s_Y s_{Y'}}{s_X^2} \\
&= \frac{2\left[2r_{YY'}s_Y s_{Y'}\right]}{s_X^2} \\
&= \frac{2\left[s_X^2 - \left(s_Y^2 + s_{Y'}^2\right)\right]}{s_X^2} \\
&= 2\left[1 - \frac{s_Y^2 + s_{Y'}^2}{s_X^2}\right] \\
&= r_{XX'(\text{Guttman})}
\end{aligned}
$$

Annexe 2.3

Preuve de l'équivalence des indices de Rulon-Guttman et de Spearman-Brown si les moitiés Y et Y' sont parallèles

Partons de la formulation de Guttman soit $r_{XX'(Guttman)} = 2\left[1 - \frac{s_Y^2 + s_{Y'}^2}{s_X^2}\right]$

Si les deux moitiés Y et Y' sont parallèles alors $s_Y^2 = s_{Y'}^2$.

Donc,
$$\begin{aligned} r_{XX'(Guttman)} &= 2\left[1 - \frac{s_Y^2 + s_{Y'}^2}{s_X^2}\right] \\ &= 2\left[1 - \frac{2s_Y^2}{s_X^2}\right] \\ &= 2\left[1 - \frac{2s_Y^2}{s_Y^2 + s_{Y'}^2 + 2r_{YY'}s_Y s_{Y'}}\right] \\ &= 2\left[1 - \frac{2s_Y^2}{2s_Y^2 + 2r_{YY'}s_Y^2}\right] \\ &= 2\left[1 - \frac{1}{1 + r_{YY'}}\right] \text{ en simplifiant les } 2s_Y^2 \\ &= \frac{2r_{YY'}}{1 + r_{YY'}} \\ &= r_{XX'(S-B)}. \end{aligned}$$

Annexe 2.4

Preuve de la valeur de $L_2 = 1$ dans le cas des données parfaites du tableau 2.7

Remarquons que, pour tous les i et j, $r_{ij} = 1$ et $s_i = s_j$ alors $s_{ij} = r_{ij}s_is_j = s_i^2$.

De plus, il y a $n(n-1)$ combinaisons d'indices i et j différents dans la somme Σ_{ij}.

Enfin, il a déjà été montré que $s_X^2 = \frac{55}{4}$ et $s_i^2 = \frac{55}{256}$.

En conséquence,

$$L_2 = \frac{\sum_{ij} s_{ij}}{s_X^2} + \sqrt{\left(\frac{n}{n-1}\right)}\left[\frac{\sqrt{\sum_{ij} s_{ij}^2}}{s_X^2}\right]$$

$$= \frac{n(n-1)s_i^2}{s_X^2} + \frac{\sqrt{\left(\frac{n}{n-1}\right)}\sqrt{\left[n(n-1)s_i^2 s_i^2\right]}}{s_X^2}$$

$$= \frac{n(n-1)s_i^2}{s_X^2} + \frac{\sqrt{\left[n^2 s_i^2 s_i^2\right]}}{s_X^2}$$

$$= \frac{n(n-1)s_i^2}{s_X^2} + \frac{ns_i^2}{s_X^2}$$

$$= \frac{n^2 s_i^2}{s_X^2}$$

$$= \frac{64 \times \left(\frac{55}{256}\right)}{\frac{55}{4}}$$

$$= 1$$

CHAPITRE 3

Les modèles de mesure dans le cadre de la théorie de la généralisabilité

Au chapitre précédent, nous avons vu comment procéder à une analyse psychométrique d'un échantillon de données comme celles présentées au tableau 2.2 dans le contexte de la théorie classique. À l'aide de concepts comme la fidélité et l'erreur-type de mesure, nous avons pu, en quelque sorte, quantifier la part d'erreur de mesure présente dans les données collectées. Or, le modèle utilisé en théorie classique suppose que l'erreur de mesure est indifférenciée. Le modèle ne prévoit pas la différenciation des diverses sources d'erreur : ainsi, les erreurs dues aux correcteurs (sévérité, effet de halo, etc.) ne peuvent être différenciées des erreurs dues à l'individu (fatigue, plagiat, etc.) ni, d'ailleurs, des autres sources d'erreurs.

La théorie de la généralisabilité est une extension de la théorie classique qui permet justement de différencier les sources d'erreur de mesure, pourvu, bien sûr, qu'un plan ait été établi à l'avance pour mettre en évidence les sources d'erreur qui doivent être analysées. Les modèles de la généralisabilité permettront de considérer tous les aspects d'une situation de mesure : correcteurs, thèmes, items, moments, formes, etc. Il sera possible alors de quantifier toutes les sources d'erreur de mesure et donc de déterminer les sources les plus importantes afin, subséquemment, de les contrôler.

Ce sont les travaux de Cronbach (Cronbach *et al.*, 1963 ; Gleser, Cronbach et Rajaratnam, 1965 ; Cronbach *et al.*, 1972), puis ceux de Brennan (Brennan, 1979 ; Brennan, 1983 ; Crick et Brennan, 1982 ; Brennan, 2001) et de Cardinet (Cardinet et Tourneur, 1978, 1985 ; Cardinet, Tourneur et Allal, 1981) qui ont le plus marqué la théorie de la généralisabilité au cours des quatre dernières décennies. Les travaux de Smith (1978, 1980), Joe et Woodward (1976), Llabre (1980), Longford (1985) et Marcoulides (1986) ont aussi influencé le développement de cette théorie. C'est véritablement Lee Cronbach qui est considéré comme le père fondateur de cette théorie : trouvant trop flou le concept de fidélité, il voulait le remplacer par un concept qui tiendrait explicitement compte des différentes composantes d'une situation de mesure.

Bien qu'il existe de bons textes présentant les bases de la théorie de la généralisabilité (Cardinet et Tourneur, 1985 ; Shavelson et Webb, 1991 ; Bain et Pini, 1996 ; Brennan, 2001), il nous a semblé justifié de présenter ces modèles dans le cadre de cet ouvrage, autant pour montrer leur distance par rapport aux modèles de la théorie des réponses aux items que parce qu'ils sont tout simplement incontournables pour analyser certaines situations de mesure. En 1983, Brennan affirmait même que la théorie de la généralisabilité constituait l'ensemble de modèles le mieux défini globalement. Car si les modèles de la TRI scrutent au microscope les tableaux à double entrée (individus $\times$ items), les modèles de la GEN ont une vision télescopique des multiples aspects des situations de mesure en prenant une distance par rapport à celles-ci, étendant l'analyse psychométrique à des tableaux comportant plus de deux entrées (correcteurs $\times$ individus $\times$ thèmes ; items $\times$ objectifs $\times$ moments $\times$ individus ; etc.).

Nous ne prétendons pas traiter des modèles de la généralisabilité de façon aussi exhaustive[1] que les ouvrages de référence cités plus haut. En revanche, nous aborderons les concepts de base et nous traiterons d'une procédure qui permettra aux utilisateurs de mener à bien une étude de généralisabilité en recouvrant, pour les calculs, à des logiciels comme ÉTUDGEN et EDUG.

1. Nous nous limiterons, en fait, aux situations de mesure comportant une seule composante de variance vraie.

3.1. LA GÉNÉRALISABILITÉ COMME EXTENSION DE LA THÉORIE CLASSIQUE

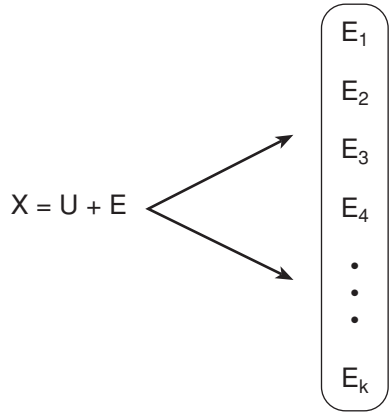
Avant de décrire l'équation de base de la généralisabilité qui prolonge l'équation de base de la théorie classique, il nous a semblé à propos de se référer à trois situations de mesure qui vont servir autant à présenter les concepts qu'à montrer l'incapacité du modèle classique à traiter ces situations « complexes » de façon efficace.

Situation A : Dans le contexte d'une production écrite en français, les deux thèmes produits par dix étudiants sont corrigés par trois correcteurs. Jusqu'à quel point peut-on se fier aux notes attribuées par les correcteurs pour décider qui réussit et qui échoue, si le seuil de réussite est fixé à 60 % ?

Situation B : Dans le contexte de l'observation en classe des comportements des enseignants, cinq enseignants sont observés par deux juges à cinq occasions chacun. Jusqu'à quel point peut-on se fier aux fréquences d'observation d'un comportement comme « le nombre de questions posées par l'enseignant en contexte de grand groupe » pour comparer les différentes pratiques pédagogiques des enseignants ?

Situation C : Dans le contexte d'un examen de géographie, vingt étudiants complètent une épreuve de six items, ceux-ci étant répartis en trois objectifs à raison de deux items par objectif. Jusqu'à quel point peut-on se fier aux résultats des étudiants pour décider quel étudiant est le meilleur en géographie ?

Comment le modèle classique pourrait-il approcher ces différentes situations de mesure ? Dans le cas de la situation A, on pourrait utiliser comme score observé la moyenne des deux thèmes et considérer les correcteurs comme l'unique source d'erreur ; mais alors l'erreur due aux thèmes ne pourrait être isolée. Dans le cas de la situation B, il faudrait que les deux juges fassent leur observation en même temps et qu'on agrège les résultats des cinq occasions pour chaque juge avant d'analyser les données, laissant en plan l'erreur due aux occasions d'observation. Dans le cas de la situation C, il serait possible d'obtenir un coefficient alpha pour chacun des trois objectifs et d'en faire la moyenne, mais comment quantifier alors des sources d'erreur comme l'interaction entre les étudiants et les objectifs (certains étudiants pouvant mieux réussir les items associés à un objectif donné qu'à un autre) ? Il faut se rendre à l'évidence : le modèle classique s'avère impuissant à englober toutes les caractéristiques de ces situations de mesure. Souvenons-nous que l'équation de base du modèle classique est $X = V + E$. Par opposition, l'équation de base en théorie de la généralisabilité est donnée par :



où U représente le score univers, le pendant du score vrai en théorie classique, et E est l'erreur de mesure, tout comme en théorie classique. Ici, cependant, **l'erreur** de mesure est **éclatée** en plusieurs sources d'erreur différentes E_1, E_2, … , E_k. C'est précisément l'apport original de cette théorie : identifier, quantifier puis contrôler les différentes sources d'erreur de mesure qui demeurent indifférenciées en théorie classique.

Rappelons que la théorie classique nous amenait à diviser la variance totale observée (σ^2_X) en deux parties, la variance vraie (σ^2_V) et la variance d'erreur (σ^2_E).

La théorie de la généralisabilité permettra de diviser la variance totale (σ^2_X) en k + 1 parties, soit la variance vraie ou univers[2] (σ^2_U) et la variance attribuable à chacune des k sources d'erreur $\left(\sigma^2_{E_1}, \sigma^2_{E_2}, \ldots, \sigma^2_{E_k}\right)$.

3.2. UNE IDÉE INFORMELLE DE LA GEN

Avant d'aborder la théorie de la généralisabilité de façon formelle, il nous semble important de fournir un support intuitif pour les différents concepts présentés subséquemment. Nous y parviendrons en prenant appui sur les trois situations de mesure dont il a été question à la section précédente. Il vaut la peine, tout d'abord, de décrire plus amplement les caractéristiques de ces trois situations.

Situation A : Dix étudiants ont rédigé deux productions écrites, chacune sur un thème donné (p. ex., le tabagisme dans la cour d'école). Trois enseignants ont corrigé chacune des deux copies des dix étudiants, comme on peut le voir au tableau 3.1. Nous considérons que ces dix étudiants constituent

2. Dans certains cas, plus rares, la variance univers peut elle-même être divisée en plusieurs « composantes » de variance. Tel qu'indiqué précédemment, nous ne traiterons pas ces cas plus complexes dans le présent ouvrage.

un échantillon aléatoire d'un univers (d'une population) d'étudiants de très grande taille[3], les deux thèmes ont été sélectionnés par un panel d'enseignants dans une banque de 12 thèmes et les trois correcteurs ont été choisis dans un groupe fixe de 58 correcteurs considérés compétents.

Tableau 3.1
Scores observés (X) en production écrite de dix étudiants (E) : trois correcteurs (C) ont corrigé les deux thèmes (T)

	E_1		E_2		E_3		E_4		E_5		E_6		E_7		E_8		E_9		E_{10}	
	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
C_1	63	59	69	71	63	62	71	75	71	70	58	61	64	62	73	71	71	72	71	70
C_2	53	60	57	70	60	61	55	65	56	64	53	60	55	65	70	75	58	66	65	73
C_3	72	77	73	71	51	51	81	80	83	84	75	71	77	76	81	83	80	83	77	75
X	**64,00**		**68,50**		**58,00**		**71,17**		**71,33**		**63,00**		**66,50**		**75,50**		**71,67**		**71,83**	

Nous voulons savoir jusqu'à quel point il est possible de se fier aux scores observés (X) des étudiants pour décider qui passe le seuil de 60 %, étant donné que ces scores observés sont en fait des moyennes des notes attribuées par les trois correcteurs à chacune des deux productions écrites des étudiants. En d'autres termes, nous voulons savoir jusqu'à quel point on peut généraliser des scores observés (X), lesquels sont des moyennes prises sur les six notes, aux scores univers (U), lesquels sont des moyennes (théoriques) provenant de l'univers des 58 correcteurs et de l'univers des 12 thèmes. Dit autrement, nous voulons savoir si les scores observés (X) sont près des scores univers (U), ce qui revient à se questionner sur l'importance des sources d'erreur (E_1, E_2, …, E_k) dans les scores observés.

L'objectif d'une étude de généralisabilité sera donc de déterminer les différentes sources d'erreur, de les quantifier pour connaître les plus influentes et, éventuellement, les contrôler. Quelles sont donc les sources d'erreur présentes dans cette situation de mesure ? Il y en a six en tout, lesquelles sont décrites ci-après.

L'effet **correcteurs** : Il s'agit de la différence de sévérité entre les correcteurs. On le voit en examinant les moyennes des notes attribuées par les trois correcteurs. Dans le cas du tableau 3.1, les moyennes des notes attribuées par les correcteurs C_1, C_2 et C_3 sont respectivement de 67,35, 62,05 et 75,05.

3. Strictement, il n'est pas possible de distinguer cet effet d'interaction triple de l'erreur expérimentale puisqu'il aurait fallu obtenir au moins deux notes par correcteur, par étudiant et par thème. Nous l'appellerons tout de même un effet d'interaction.

L'effet **thèmes** : Certains thèmes peuvent être plus difficiles à traiter que d'autres. Incidemment, au tableau 3.1, la moyenne des notes pour le thème T_1 est de 66,87 tandis qu'elle est de 69,43 dans le cas du thème T_2.

L'effet d'interaction **étudiants** × **correcteurs** : Certains correcteurs peuvent être plus sévères pour certains étudiants (p. ex., effet de halo). Par exemple, le correcteur C_3, le moins sévère des trois, a tout de même attribué les notes les plus faibles à l'étudiant E_3.

L'effet d'interaction **étudiants** × **thèmes** : Certains thèmes peuvent être plus difficiles à traiter par certains étudiants. À titre d'exemple, un thème en rapport avec la consommation de drogues peut inhiber un étudiant accroché au cannabis. Dans le tableau 3.1, cet effet d'interaction peut s'observer en comparant les moyennes des notes aux deux thèmes attribuées aux dix étudiants. Le thème T_2 est en général plus facile à traiter par tous les étudiants, sauf dans le cas de l'étudiant E_3.

L'effet d'interaction **correcteurs** × **thèmes** : Certains correcteurs ont pu éprouver des difficultés particulières à noter l'un ou l'autre des thèmes. Un correcteur récemment divorcé pourrait être enclin à noter plus sévèrement les copies portant sur un thème comme « le bonheur dans une famille unie ». Contrairement aux deux autres correcteurs, le correcteur C_2 a noté beaucoup plus sévèrement les copies du thème T_1 que les copies du thème T_2 .

L'effet d'interaction[4] **étudiants** × **correcteurs** × **thèmes** : Cet effet est présent lorsqu'un correcteur, par exemple, est porté à noter beaucoup plus sévèrement un thème particulier traité par un étudiant donné. L'étudiant E_4, qui obtient généralement de bonnes notes, a semblé traiter le thème T_1 de façon à déplaire au correcteur C_2.

Situation B : Cinq enseignants ont été observés par deux juges à cinq occasions différentes, chacun suivant le schéma présenté au tableau 3.2. Le protocole d'observation implique que les juges se concentrent sur les fréquences d'apparition de certaines pratiques des enseignants. Ici, par exemple, disons qu'il s'agit de la fréquence des questions posées par les enseignants en contexte de grand groupe pour une occasion d'observation donnée (p. ex., vendredi matin). L'objectif de l'étude est de relier les fréquences des pratiques des enseignants aux résultats moyens de leurs étudiants en mathématique. Les enseignants, tout comme les juges ou les occasions, constituent un échantillon aléatoire d'un univers de très grande taille. Il faut savoir que chaque couple de juges diffère d'un enseignant à l'autre. De même, les occasions varient d'un juge à l'autre et d'un enseignant à l'autre.

4. INF pour infini ou encore un nombre de très grande taille. C'est le cas, en pratique, lorsque N est beaucoup plus grand que n.

Tableau 3.2
Fréquences des observations de 5 enseignants (E) par deux juges (J) à 5 occasions (O) pour chaque juge

E_1										
J_{11}						J_{12}				
O_{111} 47	O_{112} 35	O_{113} 58	O_{114} 60	O_{115} 42		O_{121} 45	O_{122} 29	O_{123} 38	O_{124} 22	O_{125} 33
E_2										
J_{21}						J_{22}				
O_{211} 62	O_{212} 59	O_{213} 48	O_{214} 56	O_{215} 60		O_{221} 42	O_{222} 51	O_{223} 53	O_{224} 48	O_{225} 45
E_3										
J_{31}						J_{32}				
O_{311} 64	O_{312} 58	O_{313} 62	O_{314} 59	O_{315} 70		O_{321} 45	O_{322} 60	O_{323} 58	O_{324} 42	O_{325} 50
E_4										
J_{41}						J_{42}				
O_{411} 45	O_{412} 39	O_{413} 45	O_{414} 32	O_{415} 38		O_{421} 26	O_{422} 32	O_{423} 30	O_{424} 28	O_{425} 29
E_5										
J_{51}						J_{52}				
O_{511} 42	O_{512} 48	O_{513} 51	O_{514} 50	O_{515} 44		O_{521} 40	O_{522} 44	O_{523} 32	O_{524} 41	O_{525} 37

Pour chaque enseignant, le score observé s'obtient en prenant la moyenne des dix fréquences d'observation le concernant. Nous voulons savoir jusqu'à quel point on peut se fier aux scores observés des enseignants pour décider quel enseignant pose le plus de questions en contexte de grand groupe. En d'autres termes, nous voulons savoir jusqu'à quel point on peut généraliser des scores observés (X), lesquels sont des moyennes prises sur les dix fréquences d'observation, aux scores univers (U), lesquels sont des moyennes (théoriques) provenant des univers des juges et des occasions d'observation.

Seulement deux sources d'erreur sont présentes dans le cas de la situation B.

L'effet **juges** : Si, en moyenne, le premier juge observe plus souvent une pratique d'un enseignant donné (p. ex., le nombre de questions) que le deuxième juge, il y aura un effet dû aux juges. Dans le tableau 3.2, on voit que le premier juge observe systématiquement plus souvent cette pratique que le deuxième juge pour chacun des enseignants. Un tel effet peut se présenter, par exemple, si un juge est beaucoup mieux formé à observer une pratique chez un enseignant.

L'effet **occasions** : Si, en moyenne, une occasion d'observation donne lieu à des fréquences plus nombreuses que les autres, nous serons en présence d'un effet dû aux occasions. Ce pourrait être le cas, par exemple, si la première des cinq occasions d'observation pour chaque juge et chaque enseignant s'était toujours tenue le matin à 8 h 30.

Situation C : Un test de six items de géographie est administré à un échantillon de 20 étudiants dans le but de sélectionner les meilleurs étudiants. Trois objectifs sont visés par ce test, à raison de deux items par objectif tel que présenté au tableau 3.3. Les trois objectifs proviennent d'une banque de 25 objectifs de géographie alors que les items sont considérés échantillonnés d'un univers de très grande taille.

Les scores observés (X) sont les moyennes des notes aux six items. Les scores univers (U) sont les moyennes des notes qui seraient obtenues à tous les items de l'univers (infini) des items regroupés dans l'univers des 25 objectifs.

Nous voulons savoir jusqu'à quel point on peut se fier aux scores observés (X) des étudiants pour décider qui sont les meilleurs en géographie sur la base de ce test. En d'autres termes, nous voulons savoir jusqu'à quel point on peut généraliser des scores observés (X) aux scores univers (U).

TABLEAU 3.3
Notes de 20 étudiants (E) à un test de six items regroupés en trois objectifs

	Objectif 1		Objectif 2		Objectif 3	
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
E_1	0	1	1	1	0	1
E_2	0	0	1	0	1	1
E_3	0	1	1	0	0	0
E_4	0	0	1	0	1	0
E_5	0	1	0	1	1	0
E_6	0	1	1	1	1	1
E_7	0	0	1	1	0	0
E_8	1	0	1	0	0	1
E_9	0	1	1	1	1	0
E_{10}	1	0	0	1	0	0
E_{11}	0	0	1	1	1	0
E_{12}	0	1	1	0	1	1
E_{13}	0	1	0	0	1	1
E_{14}	0	1	1	0	0	1
E_{15}	1	1	1	0	1	1
E_{16}	0	1	1	1	1	1
E_{17}	0	1	1	0	1	1
E_{18}	0	0	1	0	0	1
E_{19}	0	0	1	1	1	1
E_{20}	1	0	0	1	1	1

Quatre sources d'erreurs sont associées à cette situation de mesure.

L'effet **items** : Toute variation systématique des notes d'un item à l'autre (c'est-à-dire la différence de difficulté des items) constitue une forme d'erreur.

L'effet **objectifs** : Si les items d'un objectif sont systématiquement plus difficiles que les items des autres objectifs, il s'ensuivra une erreur associée aux objectifs. Ce serait le cas, par exemple, pour un objectif qui n'a pas vraiment été suffisamment traité en classe.

L'effet d'interaction **étudiants** × **items** : Certains étudiants peuvent avoir de la difficulté à répondre à un ou à quelques items pour diverses raisons (fatigue, distraction, etc.), engendrant ainsi une autre forme d'erreur.

L'effet d'interaction **étudiants** × **objectifs** : Si un étudiant éprouve plus de difficulté avec un objectif particulier qu'avec les autres objectifs, une forme d'erreur due à l'interaction entre les étudiants et les objectifs est générée. Ce serait le cas, par exemple, si un étudiant était malade ou absent au moment où les concepts associés à cet objectif ont été présentés en classe.

La discussion de ces trois situations de mesure dénote une très grande diversité tant dans l'origine que dans la quantité des erreurs de mesure. Afin de poursuivre le développement formel des caractéristiques de la théorie de la généralisabilité, nous avons besoin de donner un nom aux différents concepts qui nous seront nécessaires, nous avons besoin d'un vocabulaire de base. C'est le but de la prochaine section.

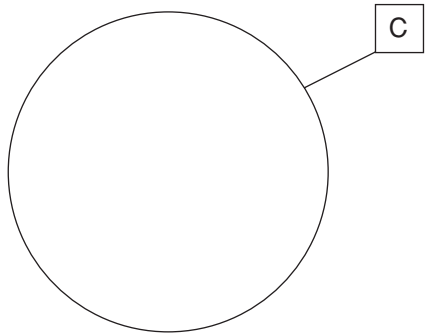
3.3. QUELQUES DÉFINITIONS

Une **facette** est une caractéristique de la situation de mesure, comme par exemple les correcteurs de la situation A, les juges de la situation B ou les objectifs de la situation C. Chaque facette est notée par une lettre : T pour thèmes, C pour correcteurs, E pour étudiants, J pour juges, etc.

Comme support visuel aux concepts qui seront présentés, Cronbach (1972) a proposé des **diagrammes** de type **Euler-Venn** où chaque facette est représentée par un cercle ou une ellipse. Nous croyons ce support visuel très utile, notamment, comme nous le verrons bientôt, pour identifier les sources d'erreur d'une situation de mesure. Ce support vient en quelque sorte compléter le tableau des données. En effet, s'il n'est pas toujours évident d'extirper toutes les sources d'erreur d'un tableau, nous verrons que l'exercice est nettement plus facile en employant un diagramme d'Euler-Venn.

Ainsi, la facette des correcteurs C se présenterait comme à la figure 3.1.

FIGURE 3.1
La facette des correcteurs (C) de la situation A se présente comme un cercle



Nous dirons qu'un **niveau** est une manifestation d'une facette. On distingue les niveaux **observés**, notés n et présentés dans le tableau des données, et les niveaux **admissibles**, notés N, qui sont ceux de l'univers.

Situation A : Trois facettes sont présentes dans cette situation : la facette C des correcteurs comprenant $n_C = 3$ niveaux observés, la facette T des thèmes avec $n_T = 2$ niveaux observés et la facette E des étudiants qui a $n_E = 10$ niveaux observés. Il y a $N_C = 58$ niveaux admissibles dans l'univers des correcteurs, $N_T = 12$ niveaux admissibles dans l'univers des thèmes et $N_E = INF$[5] niveaux admissibles dans l'univers des étudiants.

Situation B : Trois facettes sont présentes dans cette situation : la facette E des enseignants, qui comprend $n_E = 5$ niveaux observés, la facette J des juges, qui a $n_{J:E} = 2$ niveaux observés par niveau[6] de la facette E ; et la facette O des occasions, avec $n_{O:J:E} = 5$ niveaux observés par niveau de J et de E. De plus, $N_E = N_{J:E} = N_{O:J:E} = INF$.

Situation C : Encore ici, trois facettes sont présentes dans cette situation : la facette E des étudiants avec $n_E = 20$ niveaux observés, la facette O des objectifs comprenant $n_O = 3$ niveaux observés et la facette I des items avec $n_{I:O} = 2$ niveaux observés pour chaque niveau de la facette O. Notons que $N_O = 25$ et que $N_E = N_{I:O} = INF$.

5. Nous écrirons $n_{J:E} = 2$ pour signifier qu'il y a deux niveaux de J pour chacun des niveaux de E puisque la facette J est nichée dans la facette E, comme nous le verrons plus loin.
6. Bien que ce ne soit pas une règle générale, il faut tout de même préciser que, bien souvent, il n'y a qu'une seule facette de différenciation et qu'elle est constituée soit d'étudiants, d'enseignants ou, en tout cas, de personnes. Nous nous en tiendrons donc, dans cet ouvrage, à des situations constituées d'une seule facette de différenciation.

3.4. LES PHASES D'UNE ÉTUDE DE GÉNÉRALISABILITÉ

Suivant Cardinet et Tourneur (1985), quatre phases sont nécessaires afin de boucler une étude de généralisabilité et de répondre à des questions comme : « Jusqu'à quel point peut-on se fier aux scores observés (X) des étudiants pour décider qui passe le seuil de 60 % ? »

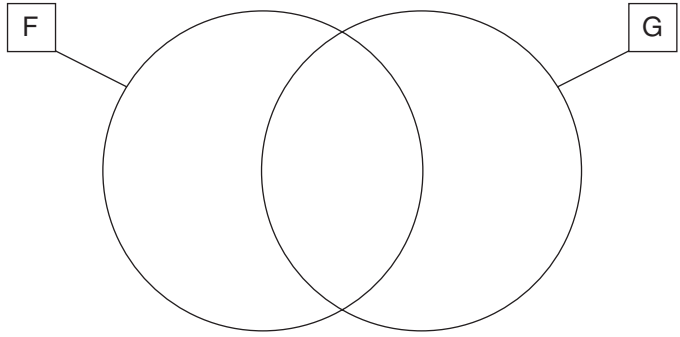
La première de ces phases se nomme l'observation, la deuxième, l'estimation, la troisième, la mesure et la quatrième, l'optimisation.

3.4.1. Phase d'observation

Il s'agit de déterminer, à l'aide du tableau des données (comme le tableau 3.1) et du diagramme d'Euler-Venn reflétant la situation de mesure à l'étude, le nombre de facettes, le nombre de niveaux observés pour chaque facette et, surtout, la relation qu'entretient chaque couple de facettes.

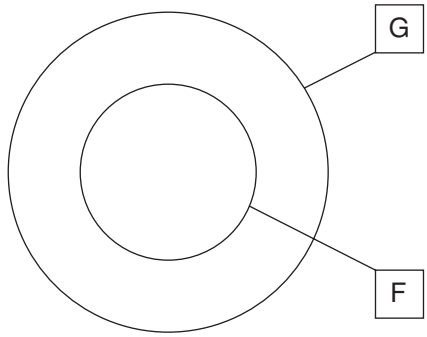
Deux facettes F et G sont dites **croisées** si chacun des niveaux observés de F est combiné à chacun des niveaux de G : on écrit alors F $\times$ G. Le croisement de deux facettes peut se représenter visuellement par deux diagrammes d'Euler-Venn en intersection.

FIGURE 3.2
Deux cercles F et G dont l'intersection représente le croisement des deux facettes, soit F $\times$ G.



Une facette F est dite **nichée** dans une facette G si une partie seulement des niveaux de F est associée à l'un ou l'autre des niveaux de G. On écrit alors F:G. Cette relation de nichage se représente graphiquement par une facette, F, incluse dans une autre, G. La facette G est dite alors nichante.

FIGURE 3.3
Si le cercle F est inclus dans le cercle G, la facette F est dite nichée dans la facette G et on écrit F:G.



Alors que la relation de croisement est commutative, c'est-à-dire $F \times G = G \times F$, il n'en est pas du tout de même pour la relation de nichage, car F:G et G:F reflètent deux situations fort distinctes.

Pour bien distinguer la relation de croisement et la relation de nichage, pensons à des correcteurs qui corrigent des copies d'élèves. Si tous les correcteurs corrigent toutes les copies d'élèves, la facette C des correcteurs est croisée avec la facette E des élèves. Le tableau des données pourrait ressembler à celui-ci.

TABLEAU 3.4
Les correcteurs (C) sont croisés avec les élèves (E)

	C_1	C_2	C_3
E_1	1	1	0
E_2	0	1	0
E_3	1	0	1
E_4	0	1	0
E_5	1	1	1

Par contre, si on assigne un échantillon différent d'élèves à chaque correcteur alors E sera nichée dans C, E:C, et on aura un dispositif comme celui qui est représenté au tableau 3.5.

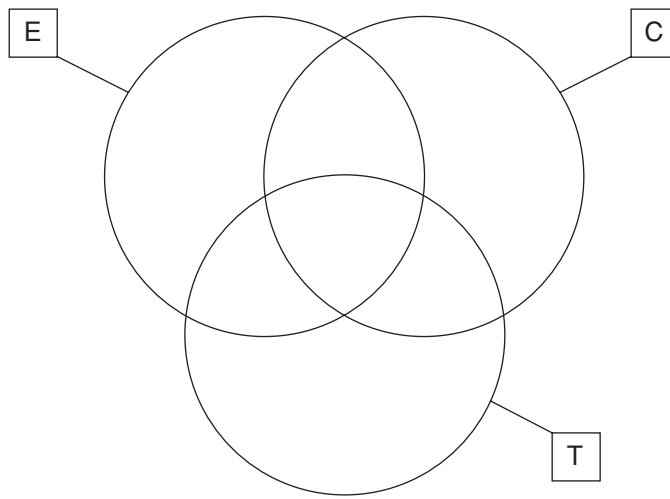
TABLEAU 3.5
Les élèves (E) sont nichés sous les correcteurs (C)

C_1					C_2					C_3				
E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	E_{13}	E_{14}	E_{15}
1	0	1	1	0	1	0	0	1	0	1	1	0	0	1

Afin d'approfondir les relations de nichage et de croisement, nous allons constituer le diagramme d'Euler-Venn pour chacune des trois situations de mesure discutées préalablement. Cela suppose bien sûr que soient connues, au préalable, les relations entre les facettes prises deux à deux.

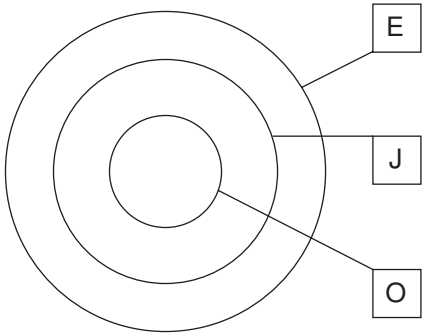
Situation A : Tous les correcteurs corrigent tous les étudiants, donc E × C. De plus, tous les thèmes sont abordés par tous les élèves, c'est-à-dire E × T. Enfin, tous les correcteurs corrigent tous les thèmes, donc C × T. Le plan ou devis d'observation de cette situation se lit E × C × T. La représentation visuelle à l'aide de diagrammes se présente comme à la figure 3.4.

FIGURE 3.4
Représentation du croisement des trois facettes de la situation A (E × C × T)



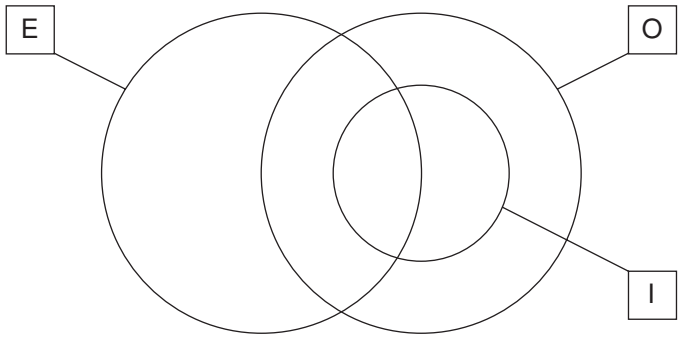
Situation B : Comme un échantillon distinct de deux juges est assigné à chacun des cinq enseignants, nous dirons que la facette J est nichée sous la facette E, ou J:E. De plus, comme un échantillon distinct de cinq occasions est assigné à chacun des juges, nous dirons que O est nichée sous J, ou O:J. Globalement le devis d'observation se lira O:J:E et la représentation visuelle aura l'allure de trois cercles emboîtés.

FIGURE 3.5
Représentation de la situation B où la facette O est nichée sous la facette J, elle-même nichée sous la facette E, O:J:E .



Situation C : Comme chacun des étudiants répond à chacun des items et touche à chacun des objectifs, les facettes E et I sont croisées, tout comme d'ailleurs les facettes E et O : ainsi, E × I et E × O. Or puisque les items diffèrent d'un objectif à l'autre, la facette I est nichée dans la facette O, soit I:O. Le devis global se lit E × (I:O) et la représentation visuelle est donnée par le graphique suivant.

FIGURE 3.6
Représentation de la situation C, où la facette I est croisée avec la facette E et nichée sous la facette O, soit E × (I:O).



3.4.2. Phase d'estimation

Cette phase de l'étude vise à identifier les univers associés à chacune des facettes. En d'autres termes, il s'agit de déterminer les niveaux admissibles relatifs aux facettes.

Si, pour une facette donnée F, tous les niveaux admissibles sont observés, donc si $N_F = n_F$, nous dirons que la facette est **fixe**. Ce serait le cas, par exemple, si, dans la situation C, il n'y avait que trois objectifs de géographie au programme. Comme les trois objectifs sont observés, la facette O serait considérée fixe.

Si, pour une facette F, $N_F > n_F$, nous dirons que la facette F est **aléatoire**. Par convention, nous dirons que la facette F est aléatoire **infinie** si $N_F > 100n_F$, c'est-à-dire si la taille de l'univers de la facette F est largement supérieure aux niveaux observés. Dans les autres cas, F sera considérée aléatoire **finie**.

Dans le cadre de la situation A, l'échantillon de 10 étudiants est considéré tiré d'un univers d'étudiants de très grande taille : nous dirons que la facette des étudiants est aléatoire infinie. Par contre, la facette T est aléatoire finie, puisque le nombre de thèmes observés est $n_T = 2$ alors que l'univers en contient $N_T = 12$. De même, la facette C des correcteurs est aléatoire finie du fait que $n_C = 3$ alors que $N_C = 58$.

Dans le cas de la situation B, autant la facette E des enseignants que la facette J des juges ou encore la facette O des occasions sont considérées aléatoires infinies.

En ce qui concerne la situation C, la facette E des étudiants est aléatoire infinie puisque l'échantillon de $n_E = 20$ étudiants est réputé provenir d'un univers de très grande taille. La facette I des items est aussi aléatoire infinie, l'univers des items étant en pratique très grand. Par contre, la facette O des objectifs est aléatoire finie, puisque les $n_O = 3$ objectifs observés de cette facette ont été choisis dans un univers fini de $N_O = 25$ objectifs.

3.4.3. Phase de mesure

Une fois toutes les facettes identifiées et nommées, une fois les niveaux observés et les niveaux univers précisés, il s'agit de déterminer quelles sont les facettes de **différenciation**, c'est-à-dire celles qui constitueront l'objet de la mesure et qui doivent être différenciées, et les facettes d'**instrumentation**, qui constituent l'instrument de mesure ou encore les conditions de la mesure. Afin de distinguer ces deux types de facettes, il faudra bien connaître les objectifs poursuivis dans le cadre d'une situation de mesure donnée.

Situation A : L'objectif est ici de différencier les étudiants qui passent de ceux qui ne passent pas le seuil de réussite de 60 %. C'est donc la facette E des étudiants[7] qui constitue la facette de différenciation. Les deux autres facettes, C et T, sont donc des facettes d'instrumentation. Les correcteurs tout autant que les thèmes font partie de l'instrumentation qui nous permettra de distinguer les étudiants.

Situation B : Puisque l'objectif est de différencier les enseignants quant au nombre de questions qu'ils posent, la facette E des enseignants est la facette de différenciation. Les juges, J, et les occasions, O, constituent autant de moyens d'obtenir la fréquence des questions posées par les enseignants : ce sont des facettes d'instrumentation.

Situation C : Comme nous nous intéressons au choix des meilleurs étudiants en géographie, la facette de différenciation est encore constituée des étudiants, E. Les instruments qui vont servir à établir les notes des étudiants sont les objectifs, soit la facette O et les items, la facette I : ce sont donc les deux facettes d'instrumentation.

3.4.4. Phase d'optimisation

La phase de mesure permettra d'obtenir un coefficient de généralisabilité, le pendant du coefficient de fidélité, qui indiquera jusqu'à quel point on peut différencier les niveaux observés de la facette de différenciation, le score de chaque niveau observé (aussi appelé score observé) étant obtenu en prenant la moyenne des valeurs (notes, fréquences, etc.) des niveaux observés des facettes d'instrumentation. Ce coefficient prendra, tout comme un coefficient de fidélité classique, des valeurs comprises entre 0 et 1. Si la valeur du coefficient n'est pas satisfaisante, la théorie prévoit une autre phase, dite d'optimisation, élaborant sur les conditions qui permettent d'améliorer cette valeur.

Compte tenu de la définition du coefficient de généralisabilité, quatre approches d'optimisation seront considérées :

1. augmenter la taille des niveaux observés des facettes d'instrumentation ;
2. diminuer la taille des univers des facettes d'instrumentation ;
3. effectuer une analyse de facettes ;
4. utiliser un coefficient critérié.

Le traitement de chacune de ces approches suppose la maîtrise de plusieurs concepts et l'habileté à interpréter une sortie informatisée d'un logiciel comme ÉTUDGEN et EDUG.

7. Même si nous savons que c'est une décision absolue qui nous intéresse dans ce cas !

3.5. LE COEFFICIENT DE GÉNÉRALISABILITÉ

3.5.1. Quelques définitions

Aussi appelé coefficient d'assurance par Cardinet et Tourneur (1985), le coefficient de généralisabilité se présente comme un coefficient de fidélité, à savoir une proportion de variance vraie, appelée **variance de différenciation**, dans la variance totale.

Rappelons que le coefficient de fidélité classique pouvait s'écrire, de façon générale

$$\rho^2_{XV} = \frac{\sigma^2_V}{\sigma^2_X} = \frac{\sigma^2_V}{\sigma^2_V + \sigma^2_E}$$

Dans le cadre de la théorie de la généralisabilité, nous écrirons

$$\rho^2_P = \frac{\sigma^2_P}{\sigma^2_P + \sigma^2_I}$$

où P renvoie à la facette de différenciation, soit celle des personnes, que ce soit des étudiants, des enseignants ou des individus de façon générale. Ainsi, la composante σ^2_P concerne la variance entre les personnes, appelée **variance de différenciation**. La **variance d'instrumentation** σ^2_I est beaucoup plus complexe et comporte en général plusieurs composantes. Une fois exprimée la variance d'instrumentation en fonction de ses différentes composantes et une fois connues les sources d'erreur les plus importantes, il sera possible de préciser les moyens de contrôler l'erreur associée à cette situation de mesure.

La composition de la variance d'erreur dépend de l'agencement des facettes entre elles, donc du devis d'observation, de la définition des univers obtenue lors de la phase d'estimation, du statut de chacune des facettes déterminé lors de la phase de mesure (à savoir s'il s'agit d'une facette de différenciation ou d'une facette d'instrumentation) et enfin du type de décision concerné par la situation de mesure à l'étude.

3.5.2. Décision relative, décision absolue

On distingue deux types de décision : la décision relative et la décision absolue. Si la situation de mesure prévoit une comparaison entre les niveaux observés de la facette de différenciation (donc les personnes), afin de connaître, par exemple, les meilleurs, nous avons affaire à une **décision relative**. Si, par contre,

nous voulons comparer les scores observés des niveaux de la facette de différenciation à un seuil quelconque, pour décider, par exemple, qui réussit et qui échoue, nous dirons qu'il s'agit d'une **décision absolue**.

Dans le cas de la situation A, nous sommes intéressés à une décision absolue puisqu'il s'agit de déterminer qui passe et qui ne passe pas le seuil de 60 %. Pour la situation B, nous voulons comparer la fréquence des questions des enseignants ; c'est donc une décision relative qui nous intéressera. Enfin, dans le cas de la situation C, il s'agit de trouver les meilleurs étudiants en géographie : encore ici, c'est à une décision relative que nous sommes renvoyés.

Si la situation de mesure appelle une décision relative par rapport à la facette des personnes P, la variance d'instrumentation σ^2_I s'écrira $\sigma^2_{\delta_P}$ et se nommera **variance d'erreur relative**. Si l'intérêt se porte sur une décision absolue, la variance d'instrumentation σ^2_I s'écrira $\sigma^2_{\Delta_P}$ et se nommera **variance d'erreur absolue**.

La composition de la variance d'erreur absolue ne sera pas la même que celle de la variance d'erreur relative. Ainsi, le coefficient de généralisabilité variera également en fonction du type de décision à prendre. C'est pourquoi, il est impératif de déterminer, avant tout, si l'intérêt de l'étude porte sur une décision relative ou une décision absolue.

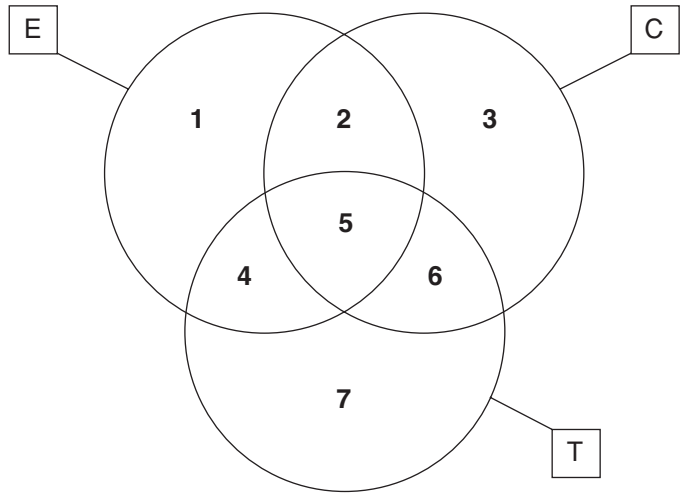
Il est d'usage de représenter les sources d'erreur à l'aide du diagramme d'Euler-Venn associé à la situation de mesure concernée. Nous allons utiliser la situation A pour distinguer les sources d'erreur associées à une décision relative des sources d'erreur associées à une décision absolue[8].

À la section 3.2 nous avons vu que la situation A comportait six sources de variance d'erreur (représentées ici par les régions 2 à 7 du diagramme de la figure 3.7) et une source de variance vraie (la région 1). Nous allons montrer que toutes les sources de variance d'erreur de cette situation font partie de la variance d'erreur absolue, mais que seules les sources de variance d'erreur en interaction avec la facette de différenciation E (régions 2, 4 et 5) font partie de la variance d'erreur relative.

La région 2 du diagramme concerne l'interaction entre les étudiants et les correcteurs : un correcteur peut être plus sévère envers un étudiant en particulier. Est-ce que ce type d'erreur affecte la variance d'erreur relative, donc le classement des étudiants ? Oui, car, au tableau 3.1, si le correcteur C_1 avait détesté l'élève E_5 au point de lui attribuer de très mauvaises notes, par exemple 51 et 50 au lieu de 71 et 70, le score observé de cet étudiant serait moins élevé que celui de l'étudiant E_4.

8. La facette de différenciation étant notée E, il est d'usage de noter σ^2_E la variance de différenciation.

FIGURE 3.7
Représentation des sources de variation relatives à la situation A : une composante de variance de différenciation (région 1) et six composantes de variance d'erreur



Est-ce que ce type d'erreur affecte la variance d'erreur absolue, donc la décision à savoir qui réussit et qui échoue ? Oui, de nouveau, car, toujours au tableau 3.1, si le correcteur le moins sévère globalement, c'est-à-dire le correcteur C_3, avait été aussi généreux envers l'étudiant E_3 qu'envers les autres étudiants, l'étudiant E_3 aurait pu passer le seuil de 60 %.

La région 3 du diagramme représente l'effet des correcteurs, certains étant plus sévères que les autres. Est-ce que ce type d'erreur affecte la variance d'erreur relative, donc le classement des étudiants ? Non, car pour observer un effet dû aux correcteurs, il faut qu'un correcteur soit systématiquement plus sévère (ou moins sévère) envers tous les étudiants, ce qui signifie ajouter ou enlever la même valeur à tous les scores observés. Au tableau 3.1, si on enlevait 5 points au score observé de tous les étudiants, cette opération ne changerait pas le classement des 10 étudiants.

Est-ce que ce type d'erreur affecte la variance d'erreur absolue, donc la décision à savoir qui réussit et qui échoue ? Oui, car le fait d'être systématiquement moins sévère envers tous les étudiants affecte tous les scores observés. Par exemple, au tableau 3.1, si le correcteur C_2 était moins sévère envers tous et ajoutait 10 points à la note qu'il a donnée à chacun des thèmes de chacun des étudiants, le score observé (X) de l'étudiant E_3 serait alors supérieur à 60.

On pourrait reprendre des arguments similaires pour les autres effets associés à cette situation de mesure et se rendre compte que la variance d'erreur relative n'est sensible qu'aux effets d'interaction avec les étudiants alors que la variance d'erreur absolue est sensible aux six effets qui correspondent à des sources d'erreur de cette situation. La figure 3.7 montre la région 1 qui correspond à la variance de différenciation[9] σ^2_E, soit la variance due aux élèves, les régions 2, 4 et 5 qui correspondent aux composantes de la variance d'erreur relative $\sigma^2_{\delta_E}$ et les six régions 2 à 7 qui correspondent aux composantes de la variance d'erreur absolue $\sigma^2_{\Delta_E}$.

Il en résulte que la variance d'erreur relative sera toujours plus petite ou égale à la variance d'erreur absolue, $\sigma^2_{\delta_E} \leq \sigma^2_{\Delta_E}$. Il sera donc toujours plus difficile de prendre une décision absolue qu'une décision relative ou en d'autres termes

$$\rho^2_{\Delta_E} \leq \rho^2_{\delta_E}$$

En se guidant sur le diagramme d'Euler-Venn (figure 3.7), on peut affirmer que les composantes de la variance d'erreur absolue comprennent toutes les régions du diagramme sauf celle associée à la variance de différenciation (la région 1 dans le cas de la situation A), c'est-à-dire les régions 2 à 7. Les composantes de la variance d'erreur relative ne comprendront que les régions associées à la variance d'erreur absolue se situant dans le cercle de la facette de différenciation (dans notre cas, la facette E), soit les régions 2, 4 et 5.

Cette façon de distinguer les deux types de décision et les deux formes de variance d'erreur qui leurs sont associées amène également une définition de deux types de coefficients de généralisabilité. Le coefficient relatif sera donné par :

$$\rho^2_{\delta_E} = \frac{\sigma^2_E}{\sigma^2_E + \sigma^2_{\delta_E}}$$

Tandis que le coefficient absolu sera :

$$\rho^2_{\Delta_E} = \frac{\sigma^2_E}{\sigma^2_E + \sigma^2_{\Delta_E}}$$

9. La preuve se trouve à l'annexe 3.2.

3.6. QUATRE APPROCHES D'OPTIMISATION

Nous connaissons les composantes de la variance d'erreur relative et les composantes de la variance d'erreur absolue. Mais ce ne sera pas encore suffisant pour justifier les approches d'optimisation visant à améliorer la généralisabilité. Il faut, en plus, connaître les formules qui définissent ces variances. Nous présentons maintenant ces formules, dans le cadre de la situation A, afin de justifier les approches d'optimisation décrites à la fin de la section 3.4. Les formules associées aux situations de mesure B et C se trouvent à l'annexe 3.1. Il ne sera pas nécessaire de discuter des formules associées à ces deux autres situations puisque, pour obtenir les valeurs des composantes de variance, on peut employer des logiciels conviviaux comme ÉTUDGEN et EDUG.

Nous avons dit que les composantes de la variance d'erreur relative étaient au nombre de trois, représentées par les régions 2, 4 et 5 sur le diagramme. La région 2 correspond à la variance d'interaction entre les étudiants et les correcteurs, que nous noterons σ^2_{EC}. La région 4 correspond à la variance d'interaction entre les étudiants et les thèmes, que nous noterons σ^2_{ET}. La région 5 correspond à la variance d'interaction entre les étudiants, les correcteurs et les thèmes : nous la noterons σ^2_{ECT}.

Il peut être montré (Cardinet et Tourneur, 1985) que

$$\sigma^2_{\delta_E} = \left[\frac{1}{n_C}\right]\left[\frac{N_C - n_C}{N_C - 1}\right]\sigma^2_{EC} + \left[\frac{1}{n_T}\right]\left[\frac{N_T - n_T}{N_T - 1}\right]\sigma^2_{ET} + \left[\frac{1}{n_C}\right]\left[\frac{N_C - n_C}{N_C - 1}\right]\left[\frac{1}{n_T}\right]\left[\frac{N_T - n_T}{N_T - 1}\right]\sigma^2_{ECT}$$

$$\sigma^2_{\Delta_E} = \left[\frac{1}{n_C}\right]\left[\frac{N_C - n_C}{N_C - 1}\right]\sigma^2_{EC} + \left[\frac{1}{n_T}\right]\left[\frac{N_T - n_T}{N_T - 1}\right]\sigma^2_{ET} + \left[\frac{1}{n_C}\right]\left[\frac{N_C - n_C}{N_C - 1}\right]\left[\frac{1}{n_T}\right]\left[\frac{N_T - n_T}{N_T - 1}\right]\sigma^2_{ECT} + \left[\frac{1}{n_C}\right]\left[\frac{N_C - n_C}{N_C - 1}\right]\sigma^2_{C} + \left[\frac{1}{n_T}\right]\left[\frac{N_T - n_T}{N_T - 1}\right]\sigma^2_{T} + \left[\frac{1}{n_C}\right]\left[\frac{N_C - n_C}{N_C - 1}\right]\left[\frac{1}{n_T}\right]\left[\frac{N_T - n_T}{N_T - 1}\right]\sigma^2_{CT}$$

Tableau 3.6
Statistiques relatives à l'étude de généralisabilité de la situation A (tableau 3.1)

Sources	Différenciation	Erreur relative	Erreur absolue	Pourcentage
E	18,753			
C			11,354	52,118
T			0,091	0,418
CT			1,242	5,701
EC		8,501	8,501	39,022
ET		0,000	0,000	0,000
ECT		0,597	0,597	2,740
Totaux	18,753	9,098	21,785	
Erreur-type		3,016	4,668	
ρ^2		0,673	0,463	

Le haut du tableau 3.6 présente les sources de variation puis la valeur de la composante de variance de différenciation σ^2_E, les valeurs des composantes de variance d'erreur relative et des composantes de variance d'erreur absolue, ainsi que le pourcentage relatif de variance associé à chaque composante. La partie du bas donne le total des valeurs des composantes de variance, les erreurs-type et les coefficients de généralisabilité relatif et absolu. À l'aide de l'erreur-type, dans le cas d'une décision absolue par exemple, il est possible de construire un intervalle de confiance autour des scores observés.

C'est à l'aide des valeurs inscrites dans ce tableau et des formules présentant la décomposition de la variance d'erreur en ses principales composantes que nous pourrons étudier les quatre approches d'optimisation présentées à la fin de la section 3.4 et que nous rappelons ici :

1. augmenter la taille des niveaux observés des facettes d'instrumentation ;
2. diminuer la taille des univers des facettes d'instrumentation ;
3. effectuer une analyse de facettes ;
4. utiliser un coefficient critérié.

Pour chacune des trois premières approches d'optimisation, nous proposons la procédure suivante pour améliorer la généralisabilité :

a) déterminer le type de décision : relative ou absolue ;
b) identifier les composantes de variance qui génèrent le plus d'erreur ;
c) augmenter n, diminuer N ou effectuer une analyse de facettes.

Voyons comment procéder dans le cas de la situation A.

Nous savons que cette situation implique une décision absolue et que nous devrons travailler sur les composantes de $\sigma^2_{\Delta_E}$ pour améliorer la généralisabilité. La dernière colonne du tableau 3.6 montre que les composantes de variance qui génèrent le plus d'erreur sont σ^2_C et σ^2_{EC} avec, respectivement, plus de 52 % et plus de 39 % de variance absolue. Les trois premières approches d'optimisation devront donc se concentrer sur des façons de

diminuer l'une ou l'autre de ces deux composantes. Afin d'y parvenir, nous allons prendre appui sur la formulation précédemment donnée de $\sigma^2_{\Delta_E}$ en fonction des composantes de variance.

La première approche implique l'augmentation du nombre de niveaux observés. En reprenant la formulation de $\sigma^2_{\Delta_E}$, il faudra donc augmenter le nombre de correcteurs n_C pour faire diminuer la variance d'erreur. En effet, n_C est au dénominateur de la partie de la formule qui comprend les composantes σ^2_C et σ^2_{EC}. Le fait d'augmenter n_C aura sûrement l'effet de diminuer $\sigma^2_{\Delta_E}$ si, bien sûr, les autres valeurs ne changent pas. En fait, on peut montrer[10] que le coefficient de généralisabilité absolu passe de 0,463 à 0,803 en augmentant le nombre de correcteurs de $n_C = 3$ à $n_C = 12$.

La deuxième approche implique la restriction du nombre de niveaux admissibles. Après tout, le fait de réduire l'univers de généralisation restreint l'ambition de l'étude et devrait donc, en toute logique, faciliter la généralisabilité. À la limite, le fait de fixer une facette d'instrumentation réduit son univers aux niveaux observés et anéantit, par le fait même, l'erreur liée à cette facette. En effet, par exemple, en revenant à la situation A et donc à la formulation de $\sigma^2_{\Delta_E}$, si la facette C est fixée, alors $N_C = n_C$, ce qui annule tous les termes de la variance d'erreur (relative ou absolue) affectés du terme $(N_C - n_C)$.

Si l'on restreint le nombre de niveaux admissibles d'une facette, ce sont les termes comme $(N_C - n_C / N_C - 1)$ ou $(N_T - n_T / N_T - 1)$ qui sont affectés à la baisse, atténuant ainsi l'effet des composantes de variance affectées par ces termes. Ainsi, toujours dans le cadre de la situation A, si le nombre de thèmes admissibles passe de $N_T = 12$ à $N_T = 6$, le terme $(N_T - n_T / N_T - 1)$ passe de $12 - 2 / 12 - 1 = 0{,}909$ à $6 - 2 / 6 - 1 = 0{,}800$.

La troisième approche implique d'effectuer une analyse de facettes, le pendant de l'analyse d'items en théorie classique. Rappelons que la variance d'erreur (relative ou absolue) est composée de termes comme $1/n_C\ (N_C - n_C / N_C - 1)\ \sigma^2_C$. Nous avons vu que le fait d'élever n_C ou encore d'abaisser N_C permettait de diminuer ces termes, donc la variance d'erreur. L'analyse de facettes affectera directement les composantes de variance comme σ^2_C. Une telle composante est le reflet de la différence de sévérité moyenne entre les correcteurs : plus les correcteurs différeront entre eux au niveau de la sévérité moyenne, plus cette composante sera élevée. Autrement dit, plus l'entente entre les correcteurs sera grande, moins cette composante sera élevée. L'analyse de facettes vise à identifier, pour une facette d'instrumentation donnée, le ou les niveaux qui auraient le plus d'impact sur la composante de variance,

10. Dans certaines circonstances, lorsque le seuil S est très près de la moyenne M, il est possible que la valeur du coefficient critérié phi-lambda soit inférieure à la valeur initiale du coefficient de généralisabilité absolu. Il vaut mieux alors conserver, comme référence, la valeur du coefficient de généralisabilité absolu que celle du coefficient critérié.

donc sur la variance d'erreur. Dans la situation A, effectuer une analyse de facettes implique, comme pour les deux autres approches d'optimisation, d'identifier d'abord les composantes qui génèrent le plus d'erreur. Comme nous l'avons déjà remarqué, ce sont les composantes σ^2_C et σ^2_{EC} qui génèrent le plus d'erreur. Or ces composantes impliquent toutes les deux la facette d'instrumentation C. C'est donc sur cette facette que s'exercera l'analyse. Si plusieurs procédures ont été proposées par le passé pour effectuer une analyse de facettes (Cardinet et Tourneur, 1985), nous proposons la procédure visant à éliminer les niveaux observés, un à un, puis à recalculer, à chaque fois, le coefficient de généralisabilité. Incidemment, cette méthode n'est pas sans rappeler l'analyse d'items classique.

En utilisant ce stratagème pour la facette C de la situation A, donc en éliminant, tour à tour, un des trois correcteurs et en recalculant le coefficient absolu, on obtient le tableau 3.7.

TABLEAU 3.7
Résultat de l'analyse de la facette C dans le cas de la situation A

Correcteur éliminé	Coefficient absolu
Aucun (situation initiale)	0,463
C_1	0,161
C_2	0,505
C_3	0,583

C'est donc en éliminant le correcteur C_3 qu'on maximise le coefficient de généralisabilité absolu. Ce résultat est conforme à ce que nous avons observé au tableau 3.1 : c'est bien le correcteur C_3 dont la moyenne de sévérité s'éloigne le plus des deux autres. Tout comme en analyse d'item classique, d'autres aspects de l'étude doivent être pris en compte avant d'éliminer un niveau d'une facette d'instrumentation, ici un correcteur. Mais si on en vient à la conclusion que nous désirons nous départir d'un correcteur, l'analyse de facettes montre que c'est au correcteur C_3 qu'il faut d'abord penser si nous sommes toujours désireux de prendre une décision absolue.

Bien sûr, s'il y avait plusieurs correcteurs, nous pourrions itérer cette procédure et rejeter encore un ou deux correcteurs qui ne s'entendent pas avec les autres. Nous pourrions aussi effectuer une analyse de facettes sur plus d'une facette d'un devis d'observation. Mais, comme pour l'analyse d'item classique, nous suggérons de n'enlever qu'un niveau d'une facette à la fois.

Encore deux remarques en rapport avec cette approche d'optimisation. Premièrement, elle nous semble relativement moins onéreuse et plus réaliste que les deux autres, à tout le moins dans le contexte de la situation A. En effet, si l'ajout de plusieurs correcteurs mène à une augmentation de la généralisabilité, les coûts qu'entraîne cette démarche risquent d'être disproportionnés. En outre, le fait de diminuer le nombre de niveaux admissibles,

donc de généraliser à un univers moins grand, risque de ne pas toujours être réaliste. Deuxièmement et en contrepartie, cette approche n'est pas toujours souhaitable ou même possible pour une facette d'instrumentation nichée. En effet, dans le cas de la situation B par exemple, enlever une occasion implique qu'il faille éliminer une occasion pour chaque juge et chaque enseignant, une perte d'information très importante.

Le recours à la quatrième approche d'optimisation, l'emploi d'un coefficient critérié, n'est valable que si la décision visée par l'étude est absolue comme dans le cas de la situation A. Incidemment, dans ce cas, $\rho^2_{\Delta_E} = 0{,}463$, signifiant que nous pouvons nous fier à 46,3 % aux données observées pour prendre une décision absolue à savoir qui réussit et qui échoue. Mais le coefficient de généralisabilité absolu ne fait aucun renvoi explicite au seuil de réussite S, que ce soit 50 %, 60 %, 70 % ou 75 %. En fait, comme nous allons le voir, ce coefficient suppose implicitement que ce seuil est la moyenne des scores observés X par les étudiants, soit le pire cas possible : voici pourquoi.

C'est autour de la moyenne que se trouvent la plupart des étudiants : plus on s'éloigne de la moyenne, moins il y a d'étudiants. En conséquence, plus le seuil S sera éloigné de la moyenne M, moins il y aura d'étudiants à cet endroit, donc moins on risquera de se tromper en prenant une décision absolue autour du seuil.

C'est à partir de cet argument que l'on peut justifier un coefficient, appelé critérié, qui prend en compte le seuil de réussite. Rappelons que, dans le cas de la situation A, le coefficient absolu est donné par

$$\rho^2_{\Delta_E} = \frac{\sigma^2_E}{\sigma^2_E + \sigma^2_{\Delta_E}}$$

Le coefficient critérié que nous proposons est le coefficient phi-lambda d'abord discuté par Brennan et Kane (1977). La formule du coefficient critérié est donnée par Brennan (2001) :

$$\rho^2_{\Delta_{E(S)}} = \frac{\sigma^2_E + (M-S)^2 - K}{\sigma^2_E + \sigma^2_{\Delta_E} + (M-S)^2 - K}$$

où

$$K = \frac{\sigma^2_E + \sigma^2_{\delta_E}}{n_E} + \left(\sigma^2_{\Delta_E} - \sigma^2_{\delta_E}\right)$$

On voit bien, d'après la formule du coefficient critérié, que plus le seuil S est éloigné de la moyenne M, plus la valeur du coefficient critérié est élevée.

Le tableau 3.8 donne les valeurs du coefficient critérié en fonction de plusieurs seuils dans le cas de la situation A. Rappelons, que, dans ce cas, M = 68,15 et la valeur du coefficient de généralisabilité absolu est de 0,463.

TABLEAU 3.8
Calcul du coefficient critérié (phi-lambda) en fonction de plusieurs seuils de réussite (Situation A ; M = 68,15)

Seuil	Coefficient critérié (phi-lambda)
50	0,939
60	0,762
70	0,235[11]
75	0,697

Pour bien montrer l'importance de considérer le seuil de réussite dans la valeur du coefficient de généralisabilité, remarquons à quel rythme la valeur du coefficient critérié augmente à mesure que le seuil s'éloigne de la moyenne. Souvenons-nous aussi que pour obtenir une valeur du coefficient de généralisabilité absolu de 0,8, donc dans le cas où l'on ne prenait pas en compte le seuil de réussite, il fallait $n_C = 12$ correcteurs. Par contre, en tenant compte du seuil voulu S = 60 dans le calcul du coefficient de généralisabilité, comme le fait le coefficient critérié, seulement $n_C = 3$ correcteurs suffisent pour obtenir une valeur tout à fait acceptable du coefficient de généralisabilité de 0,762. Notons enfin que pour obtenir une telle valeur (0,76) du coefficient de généralisabilité absolu, sans prendre en compte le seuil, cela prendrait $n_C = 10$ correcteurs.

3.7. LIMITES DE LA THÉORIE

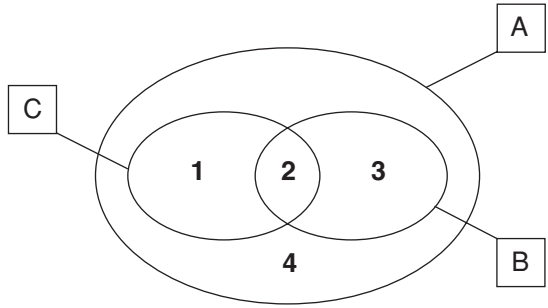
Webb (dans MacArthur, 1987, p. 199) signale deux limites importantes à l'utilisation de cette théorie. Tout d'abord la théorie, telle que présentée ici, ne permet pas l'emploi de devis non équilibrés : chaque cellule d'un tableau soumis à une étude de généralisabilité doit avoir le même nombre d'observations. Par exemple, dans la situation C, il ne pourrait y avoir 2 items pour l'objectif 1 et 4 items pour l'objectif 2, etc. : le nombre de niveaux d'une facette nichée est le même pour chaque niveau de la facette nichante. Ceci

11. Brennan (2001) propose un modèle de généralisabilité multivarié qui prend en compte les devis non équilibrés.

limite considérablement les possibilités d'utilisation de la théorie. La deuxième limite mentionnée par Webb concerne le nombre très appréciable d'observations à recueillir si l'on veut que les erreurs-type des composantes de variance demeurent raisonnablement petites. Il s'agit d'une limite déjà traitée par Smith (1978, 1980) et qui a peut-être sonné le glas à certaines utilisations de plans complexes contenant trop peu d'observations. Ceci dit, et malgré les limites propres à cette théorie, il faut bien avouer qu'il n'existe que très peu de choix pour traiter des plans à plus de deux facettes comme les situations A, B et C. En pratique, les chercheurs continueront certainement d'utiliser les modèles de la généralisabilité : nous espérons toutefois qu'ils sauront limiter leurs ambitions et traiter des plans relativement simples comme ceux présentés ici, même si cela peut sembler paradoxal, cette théorie étant justement développée pour traiter des plans très complexes.

Exercices

1. Imaginez une situation de mesure qui corresponde au diagramme d'Euler-Venn suivant. Précisez la phase d'observation, la phase d'estimation et la phase de mesure.



2. Décrivez une situation de mesure à 4 facettes comportant au moins une relation de nichage.

3. En utilisant les valeurs du tableau 3.6, dites quelle serait la valeur du coefficient de généralisabilité absolu si le nombre de thèmes était $n_T = 6$ plutôt que $n_T = 2$.

4. Construisez l'intervalle de confiance à 68 % autour du score observé de l'étudiant 1 de la situation A dans le cas où l'intérêt de l'étude est de prendre une décision absolue.

5. Prenant en compte les statistiques du tableau suivant, trouvez la valeur du coefficient de généralisabilité relatif si le nombre d'items par objectif passe de $n_{I:O} = 2$ à $n_{I:O} = 5$.

Sources	Différenciation	Erreur relative	Erreur absolue	Pourcentage
E	0,01184			
O			0,00007	0,2
I:O			0,00599	13,9
EO		0,0	0,0	0,0
EI:O		0,03706	0,03706	85,9
Totaux	0,01184	0,03706	0,04313	
Erreur-type		0,1925	0,2077	
ρ^2		0,242	0,215	

6. Trouvez la valeur du coefficient critérié de la situation A où S = 55.

7. Voici les valeurs des composantes de variance et les statistiques associées dans le cas de la situation B. Combien faudrait-il observer de juges pour que le coefficient de généralisabilité absolu soit de 0,8 ?

Sources	Différenciation	Erreur relative	Erreur absolue	Pourcentage
E	50,203			
J:E		27,469	27,469	87,5
O:J:E		3,935	3,935	12,5
Totaux	50,203	31,404	31,404	
Erreur-type		5,604	5,604	
ρ^2		0,615	0,615	

Corrigé des exercices

1. Afin d'évaluer la pertinence de la candidature de chacun des 12 pays participants (facette A), trois juges du Comité olympique (facette B) font des visites à chacun des pays à deux moments différents (facette C). Puisque les trois juges font les visites en même temps, la facette B est croisée avec la facette C. Puisque le trio de juges est différent d'un pays à l'autre, la facette B est nichée sous la facette A. Enfin, puisque les moments sont nécessairement distincts d'un pays à l'autre, la facette C est aussi nichée sous la facette A.

3. Le coefficient absolu ne change guère, passant de 0,463 à 0,478. En effet, les thèmes affectent peu les scores. Pour arriver à la valeur de 0,478, il suffit de diviser les composantes affectées par le terme $1/n_T$ non pas par 2, mais par 6, ce qui revient à diviser par 3 les composantes associées aux sources de variation T, CT, ET et ECT, déjà calculées dans le tableau 3.6. Ainsi,

$$\sigma^2_{\Delta_E} = 11{,}354 + (0{,}091)/3 + (1{,}242)/3 + 8{,}501 + (0{,}000)/3 + (0{,}597)/3$$
$$= 20{,}498$$

Et le coefficient de généralisabilité passe de sa valeur initiale de 0,463 à 0,478, soit (18,753) / (18,753 + 20,498).

5. La valeur du coefficient de généralisabilité relatif passerait à 0,444 en suivant la procédure utilisée dans le cas de l'exercice 3.

7. Il faut revenir aux formules originales du coefficient absolu et à la variance d'erreur absolue :

$$\sigma^2_{\Delta_E} = \frac{1}{n_{J:E}}\sigma^2_{J:E} + \left(\frac{1}{n_{J:E}}\right)\left(\frac{1}{n_{O:J:E}}\right)\sigma^2_{O:J:E}$$

$$\rho^2_{\Delta_E} = \frac{\sigma^2_E}{\sigma^2_E + \sigma^2_{\Delta_E}}$$

Puisque $\sigma^2_E = 50{,}203$, le fait de vouloir $\sigma^2_{\Delta_E} = 0{,}8$ équivaut à demander que $\rho^2_{\Delta_E} = \dfrac{50{,}203}{\left(50{,}203 + \sigma^2_{\Delta_E}\right)} = 0{,}8$

C'est-à-dire que $\sigma^2_{\Delta_E} = 12{,}5616$.

Or, d'après le tableau présenté à l'exercice 7, et comme $n_{J:E} = 2$ et $n_{O:J:E} = 5$,

$$\frac{1}{2}\sigma^2_{J:E} = 27,469 \text{ et } \left(\frac{1}{2}\right)\left(\frac{1}{5}\right)\sigma^2_{O:J:E} = 3,935$$

D'où $\sigma^2_{J:E} = 54,938$ et $\sigma^2_{O:J:E} = 39,35$.

Il faut trouver le nombre de juges $n_{J:E}$ tel que

$$\sigma^2_{\Delta_E} = \frac{1}{n_{J:E}}\sigma^2_{J:E} + \left(\frac{1}{n_{J:E}}\right)\left(\frac{1}{n_{O:J:E}}\right)\sigma^2_{O:J:E}$$

$$= \frac{1}{n_{J:E}}\left[54,938 + \left(\frac{1}{5}\right)39,35\right] = 12,5616$$

En isolant $n_{J:E}$ dans cette formule, on obtient $n_{J:E} = 5$.

Annexe 3.1

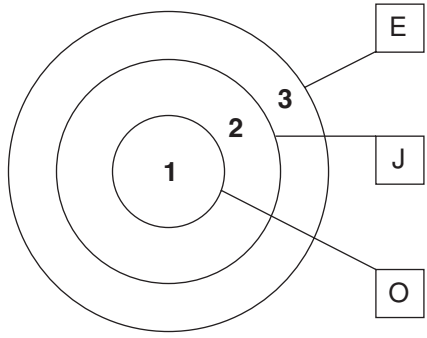
Décomposition de la variance d'erreur relative et de la variance d'erreur absolue pour les situations B et C

Situation B

Seulement trois composantes de variance sont présentes dans le cas de la situation B, comme on peut le constater au diagramme suivant. La région 1 concerne la variance due aux occasions, la région 2 concerne les juges et la région 3, les enseignants. C'est la région 3 qui se rapporte à la composante de variance de différenciation ; les deux autres régions font partie de la variance d'erreur. Mais puisque ces deux composantes d'erreur sont en interaction avec la facette de différenciation (E), la variance d'erreur relative est égale à la variance d'erreur absolue *dans ce cas*. Ainsi,

$$\sigma^2_{\Delta_E} = \sigma^2_{\delta_E} = (1/n_J)\,\sigma^2_{J:E} + (1/n_J)(1/n_O)\sigma^2_{O:J:E}$$

Puisque les facettes J et O sont aléatoires infinies, les termes ($N_J - n_J$ / $N_J - 1$) et ($N_O - n_O$ / $N_O - 1$) sont égaux à 1. De plus, il faut noter que, comme il y a deux juges par enseignant et cinq occasions par juge, $n_J = 2$ et $n_O = 5$.



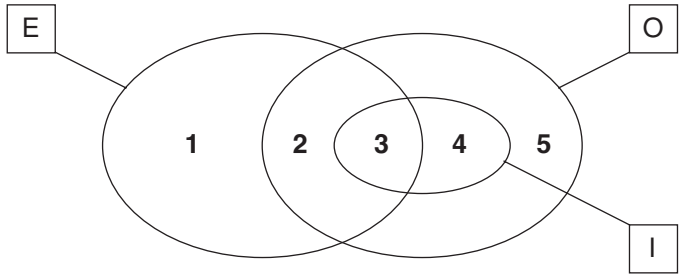
Situation C

Cinq composantes de variance sont présentes dans le cas de la situation C, comme on peut le constater au diagramme suivant. La région 1 concerne la variance de différenciation due aux étudiants. La région 2 et la région 3 se rapportent aux variances d'erreur relative et les régions 2 à 5 sont des composantes de variance d'erreur absolue. Ainsi,

$$\sigma^2_{\delta_E} = (1/n_O)(N_O - n_O/N_O - 1)\sigma^2_{EO} + (1/n_O)(1/n_I)\sigma^2_{EI:O}$$

$$\sigma^2_{\Delta_E} = (1/n_O)(N_O - n_O/N_O - 1)\left[\sigma^2_O + \sigma^2_{EO}\right] + (1/n_O)(1/n_I)\left[\sigma^2_{I:O} + \sigma^2_{EI:O}\right]$$

Puisque la facette I est aléatoire infinie le terme $(N_I - n_I / N_I - 1) = 1$. Il faut aussi noter que $n_I = 2$, et $n_O = 3$.



Annexe 3.2

Effet de l'augmentation du nombre de niveaux observés de $n_C = 3$ à $n_C = 12$ sur la valeur du coefficient de généralisabilité absolu dans le cas de la situation A

D'après le tableau 3.6, le coefficient de généralisabilité absolu $\rho^2_{\Delta_E}$ vaut 0,463 alors que la variance d'erreur absolue $\sigma^2_{\Delta_E}$ vaut 21,785 lorsque $n_C = 3$.

Rappelons la formule de la variance d'erreur absolue

$$\sigma^2_{\Delta_E} = \left[\frac{1}{n_C}\right]\left[\frac{N_C - n_C}{N_C - 1}\right]\sigma^2_{EC} + \left[\frac{1}{n_T}\right]\left[\frac{N_T - n_T}{N_T - 1}\right]\sigma^2_{ET}$$
$$+ \left[\frac{1}{n_C}\right]\left[\frac{N_C - n_C}{N_C - 1}\right]\left[\frac{1}{n_T}\right]\left[\frac{N_T - n_T}{N_T - 1}\right]\sigma^2_{ECT}$$
$$+ \left[\frac{1}{n_C}\right]\left[\frac{N_C - n_C}{N_C - 1}\right]\sigma^2_{C} + \left[\frac{1}{n_T}\right]\left[\frac{N_T - n_T}{N_T - 1}\right]\sigma^2_{T}$$
$$+ \left[\frac{1}{n_C}\right]\left[\frac{N_C - n_C}{N_C - 1}\right]\left[\frac{1}{n_T}\right]\left[\frac{N_T - n_T}{N_T - 1}\right]\sigma^2_{CT}$$

Essayons de voir quelle est la valeur de la variance d'erreur absolue lorsque $n_C = 12$. En d'autres termes, quel est l'impact du remplacement de $n_C = 3$ par $n_C = 12$ dans $1/n_C\ (N_C - n_C / N_C - 1)$? D'après cette formule, quatre composantes seraient affectées par cette expression, donc par un changement de n_C, le nombre de niveaux observés de la facette C : σ^2_C, σ^2_{EC}, σ^2_{CT}, σ^2_{ECT}.

Souvenons-nous que $N_C = 58$. Ainsi, lorsque $n_C = 3$,

$1/n_C\ (N_C - n_C / N_C - 1) = 1/3\ (58 - 3 / 58 - 1) = 0{,}322$

Alors que si $n_C = 12$,

$1/n_C\ (N_C - n_C / N_C - 1) = 1/12\ (58 - 12 / 58 - 1) = 0{,}067$.

Ainsi, multiplier les valeurs des 4 composantes ciblées du tableau 3.6 par 0,067/0,322 = 0,208 revient à remplacer $n_C = 3$ par $n_C = 12$ dans la formule de $\sigma^2_{\Delta_E}$.

Ce qui donne $0{,}208 \times (11{,}354 + 1{,}242 + 8{,}501 + 0{,}597) = 4{,}512$.

La valeur de la composante d'erreur absolue devient donc 4,512+ 0,091 = 4,603 plutôt que 21,785.

Et le coefficient de généralisabilité absolu passe de sa valeur initiale de 0,463 à 0,803, soit 18,753 / (18,753 + 4,603).

CHAPITRE 4

Concepts et modèles de base en théorie des réponses aux items

C'est au cours de ce chapitre, étape charnière de ce bouquin, que nous présenterons les bases de la théorie des réponses aux items (TRI). Ce sera, à notre sens, le premier texte substantiel en français concernant les fondements théoriques des modèles de réponses aux items. Déjà plusieurs documents pédagogiques ont été proposés en anglais sur le sujet : pensons, à titre d'exemples, aux textes de Warm (1978), Wright et Stone (1979), Lord (1980), Hulin, Drasgow et Parsons (1983), Baker (1985), Hambleton et Swaminathan (1985), Hambleton, Swaminathan et Rogers (1991), Embretson et Reise (2000) ou encore Thissen et Wainer (2001). Tous ces textes ont en commun de présenter à un public de chercheurs et d'étudiants universitaires les bases théoriques et les principales applications de la TRI, avec une certaine rigueur et un souci pédagogique évident. Il est, dès lors, loisible de se demander la place que prendra ou devrait prendre un autre manuel sur le sujet, hormis le fait que ce nouveau

texte est écrit dans la langue de Molière. Nous n'avons d'autre réponse à apporter que l'expérience des auteurs qui ont eu l'occasion, d'une part, d'enseigner les modèles de réponses aux items à quelques générations d'étudiants universitaires et, d'autre part, d'utiliser ces modèles dans différents contextes au Canada, aux États-Unis, en Europe et en Afrique. C'est cette double expérience qui, en nourrissant chacun des deux auteurs, a permis, croyons-nous, une présentation originale des concepts, procédures et applications propres à la TRI. Ajoutons également que peu de documents résolument pédagogiques ont vu le jour depuis dix ans en cette matière, les derniers volumes étant, selon nous, de facture beaucoup trop technique. Il nous a semblé approprié de discuter de façon pédagogique des nouveaux développements ou applications de la TRI survenus au cours de la dernière décennie.

Le présent chapitre traitera de l'historique de la modélisation logistique propre à la théorie des réponses aux items : comment en est-on arrivé au modèle logistique ? Quel est le lien avec l'ogive normale ? Quels sont les liens avec les modèles connus de Guttman et de Lazarsfeld ? Qu'entend-on par courbe caractéristique d'item ? Quelles sont les particularités propres aux modèles à un, deux ou trois paramètres ? Il sera également question de la courbe caractéristique de test et de son lien avec le score vrai. Nous traiterons, en outre, des fonctions d'information d'item et de test ainsi que de l'erreur-type de mesure qui peut être calculée pour chaque niveau d'habileté. Nous aborderons ensuite les principales conditions d'application de la TRI, notamment l'unidimensionalité, l'indépendance locale et l'ajustement des données au modèle : nous discuterons des façons de vérifier ces conditions, des conséquences du non-respect de l'une ou l'autre de ces conditions, ainsi que des procédés employés pour vérifier la propriété d'invariance. Enfin, nous mettrons en évidence les comparaisons avec les modèles déjà présentés de la théorie classique et de la théorie de la généralisabilité.

4.1. UNE NOUVELLE THÉORIE DE LA MESURE : POURQUOI ?

Reprenons ici l'exemple de Zoé (déjà abordé au chapitre 2), qui a obtenu un score de 64 % à l'examen de mathématique du ministère de l'Éducation. Nous nous sommes demandé jusqu'à quel point ce score reflétait son habileté en mathématique. Mais nous aurions pu également poser la question : d'où vient ce score de 64 % ? Comment a-t-il été calculé ? En théorie classique, de même d'ailleurs qu'en théorie de la généralisabilité, il est d'usage d'additionner ou de prendre la moyenne des résultats obtenus aux items, en considérant 0 pour une mauvaise réponse et 1 pour une bonne réponse. Ce score classique X est donc donné par la formule suivante :

$$X = \sum_{i=1}^{n} U_i$$

où $U_i = 1$ pour une bonne réponse à l'item i et $U_i = 0$ pour une mauvaise réponse à l'item i et où n est le nombre d'items considéré.

N'y a-t-il pas un meilleur indicateur de l'habileté de Zoé que ce score classique ? Ce score ne dépend en effet que du nombre d'items réussis. Si l'examen est très difficile, peu d'items seront réussis par Zoé et son score classique sera faible, auquel cas nous serons tentés d'affirmer que Zoé possède une bien piètre habileté en mathématique. Par contre, si l'examen est très facile, Zoé pourra réussir plusieurs items et ainsi obtenir un score élevé, auquel cas nous pourrions soutenir que Zoé possède une très bonne habileté en mathématique. De toute évidence, il serait préférable qu'un effet compensateur vienne corriger cette situation, par exemple en reconnaissant à Zoé plus d'habileté en mathématique pour la réussite d'items difficiles que pour la réussite d'items faciles. Ainsi, il est opportun de considérer que la contribution de chaque item à l'estimation de l'habileté de Zoé soit pondérée selon un critère qui tienne compte de certaines caractéristiques fixes de l'item. Comme nous allons le voir, c'est ce que propose la théorie des réponses aux items.

Si on s'attarde maintenant aux indices classiques propres aux items, il est aisé de constater que la valeur de l'indice de difficulté p_i, défini comme le nombre d'individus qui réussissent l'item i, dépend tout autant de la force moyenne du groupe à qui on administre l'item que de la difficulté de l'item i en tant que telle. De plus, ces deux effets sont confondus, dans le sens qu'il n'est pas vraiment possible de départager l'effet dû à la force moyenne du groupe de l'effet dû à la difficulté intrinsèque de l'item i. Il en est de même de l'indice de discrimination classique défini comme une forme d'association linéaire entre l'item et le total au test X, le cas le plus simple étant, comme nous l'avons constaté au chapitre 2, la corrélation de Pearson entre l'item et le total au test symbolisée par r_{iX}. Si le groupe d'individus auquel est administré le test est homogène en habileté, la variance du test sera faible, donnant ainsi peu de chances à la corrélation r_{iX} d'être élevée avec, en conséquence, le risque de conclure que l'item ne discrimine pas vraiment. Par contre, si le groupe d'individus testés est hétérogène, la corrélation r_{iX} pourrait être plus élevée et donc aussi l'estimation de l'ampleur de la discrimination de l'item i. De même, tel que nous l'avons souligné à la section 2.4.3, puisque le coefficient alpha de Cronbach, de loin l'indice de fidélité le plus employé, est fonction de la corrélation r_{iX}, il sera lui aussi affecté par les mêmes caractéristiques distributionnelles que l'indice de discrimination. Nous verrons que les paramètres d'items propres à la TRI que sont l'indice de difficulté, l'indice de discrimination et l'indice de pseudo-chance sont dits invariants par rapport à la distribution du groupe d'individus utilisé pour obtenir les estimés de ces paramètres. Cette propriété d'invariance est au cœur même de la justification de l'utilisation des modèles de la théorie des réponses aux items.

Les modèles de mesure proposés dans le contexte de la TRI sembleront plus complexes que les modèles présentés précédemment. Si la propriété d'invariance peut à elle seule être une justification suffisante pour choisir un modèle de la TRI plutôt qu'un modèle classique, il n'en demeure pas moins que les développements théoriques propres à cette théorie peuvent prendre une allure rébarbative en vertu de la complexité inhérente aux modèles. Cette apparente complexité doit cependant être modulée par les deux observations suivantes. Premièrement, les modèles de la théorie classique et de la théorie de la généralisabilité, s'ils semblent comporter moins d'éléments de complexité intrinsèque, souffrent cependant d'un manque de conformité à la réalité. Soyons honnêtes : quand peut-on réellement observer deux tests parallèles ? En outre, jusqu'à quel point peut-on estimer avec une précision suffisante un coefficient de généralisabilité d'un plan d'observation à cinq ou six facettes croisées ? De plus, si les conditions d'application associées à l'utilisation des modèles propres à ces deux théories sont moins explicites, il n'en demeure pas moins qu'elles devraient être vérifiées au même titre que les conditions associées à l'emploi des modèles de la TRI. Or, comme nous le savons, il s'agit là d'une entreprise périlleuse, justement à cause de ce manque de réalisme ! Deuxièmement, les concepts et procédures émanant de la TRI, si complexes soient-ils, sont la plupart du temps appuyés par un support visuel heuristique. Ainsi, le tracé d'une courbe caractéristique d'item permet-il de représenter jusqu'à quel point et où, sur l'échelle d'habileté, un item discrimine le mieux, alors qu'un énoncé classique comme $r_{iX} = 0{,}246$ exige, pour plusieurs, un effort mental important avant d'avoir une idée intuitive assez juste de la discrimination de l'item i.

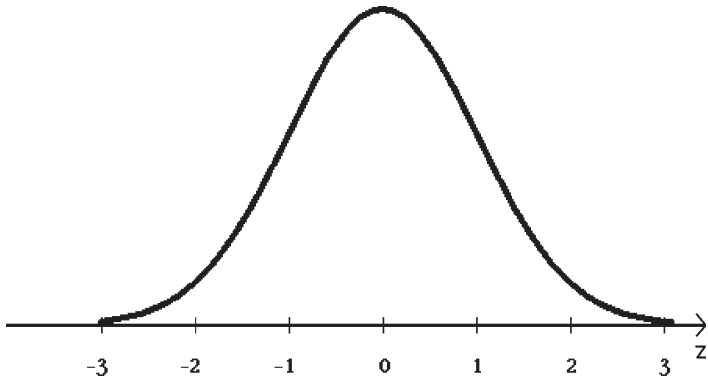
4.2. ORIGINE DE LA COURBE CARACTÉRISTIQUE D'ITEM

Comme pour toute théorie de la mesure, la théorie des réponses aux items vise, d'une part, l'estimation la plus pertinente et la plus précise possible de l'habileté des individus à partir de leurs réponses aux items et, d'autre part, l'évaluation des qualités psychométriques des items. Afin d'atteindre ce double objectif, la TRI prend appui sur un modèle mathématique que l'on peut représenter par une fonction reliant une variable latente, l'habileté de l'individu, à la probabilité de réussir un item : cette fonction, qui est à la base même de la théorie, est appelée **courbe caractéristique d'item**. Nous ne traiterons que des modèles dits unidimensionnels, à savoir ceux qui s'appuient sur un seul trait, une seule habileté : c'est pourquoi nous ne parlerons dans ce chapitre que de l'habileté des individus. Nous montrerons plus loin comment il est possible de formuler des modèles multidimensionnels.

4.2.1. CCI et courbe normale

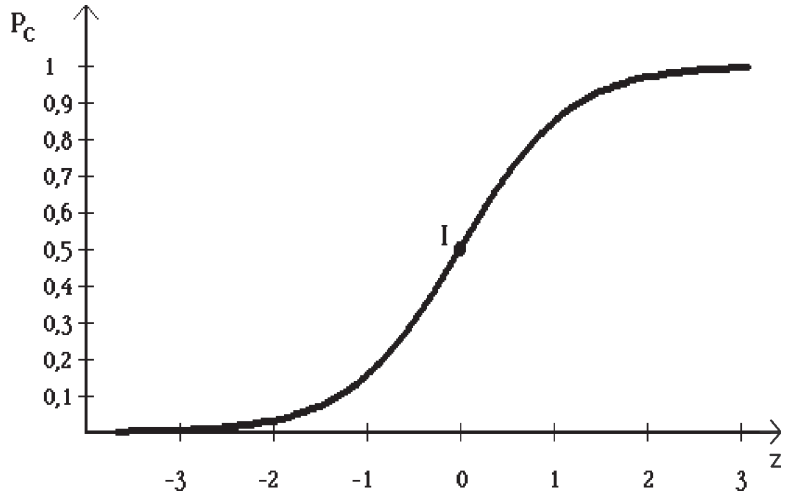
Une bonne façon d'appréhender ce concept de courbe caractéristique d'item, que l'on symbolisera dorénavant par **CCI**, est de le rapprocher d'un concept connu comme la courbe normale centrée et réduite. La figure 4.1 présente une telle courbe. L'échelle des scores *z* qui est ainsi produite est intéressante dans la mesure où elle permet de comparer l'habileté des individus entre eux : un score de z = 1 indique une plus grande habileté qu'un score de z = 0, par exemple. D'autre part, compte tenu des propriétés de la courbe normale, on sait que 50 % des scores sont inférieurs à z = 0 et 50 % des scores sont supérieurs à z = 0. On se souviendra également qu'un individu obtenant un score de z = 1 dépasse le score de 84 % des individus alors qu'un individu dont le score est z = –1 est dépassé par 84 % des individus.

FIGURE 4.1
Courbe normale centrée (moyenne = 0) et réduite (écart-type = 1)



Afin de mieux représenter ces informations, on peut produire une courbe comme celle de la figure 4.2. Il s'agit d'une ogive normale construite à partir des fréquences cumulées plutôt que des fréquences relatives comme dans la représentation classique de la loi de probabilité normale. Ce sont les proportions cumulées P_c qui sont reproduites ici. On remarquera tout particulièrement le point I, situé au centre de cette courbe et de coordonnées (0, 0,5). Il représente le fait qu'il y a 50 % des individus dont le score se situe au-dessus de z = 0 et 50 % des individus dont le score se situe au-dessous de z = 0. Le point I s'appelle **point d'inflexion,** car c'est à partir de ce point que la courbe passe du concave au convexe. On peut voir également, à partir de l'ogive normale, qu'à z = 1 correspond la proportion 0,84 signifiant que 84 % des individus ont une habileté inférieure à z = 1.

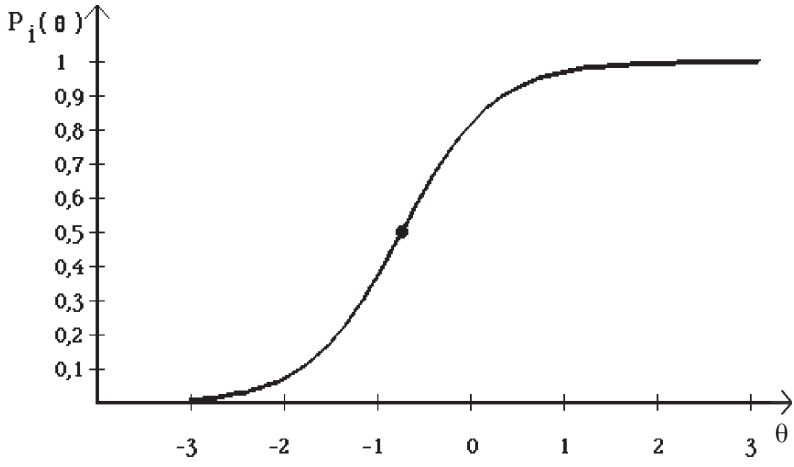
FIGURE 4.2
Ogive normale centrée (moyenne = 0) et réduite (écart-type = 1) : les proportions cumulées P_c et le point d'inflexion I



La forme de l'ogive normale (une forme en « S ») est, à peu de chose près, la forme typique que prend une courbe caractéristique d'item, comme nous allons l'observer prochainement. La figure 4.3 présente la CCI émanant d'un test de mathématique administré à un échantillon de 1000 élèves québécois francophones de 13 ans lors de l'enquête internationale de l'IAEP2[1]. L'échelle familière des scores z a été remplacée par une échelle similaire dite des scores θ, où θ est l'habileté mesurée par ce test de mathématique qui comprend 76 items. Les proportions cumulées P_c ont été remplacées par $P_i(\theta)$, qui donne une indication de la proportion des élèves d'habileté θ qui ont réussi l'item i. On peut définir plus formellement $P_i(\theta)$ comme la probabilité de réussir l'item i pour des élèves d'habileté θ. Par exemple, il est facile de se rendre compte qu'environ 80 % des élèves d'habileté moyenne (θ = 0) ont réussi cet item : c'est donc un item plutôt facile. Le point d'inflexion n'est pas exactement au même endroit que dans le cas de l'ogive normale (nous verrons bientôt pourquoi), mais il indique toujours le point où la courbe passe du concave au convexe. Incidemment, l'endroit où se situe le point d'inflexion donne une idée de l'endroit où l'item discrimine le mieux : ici, le fait qu'environ 40 % des élèves d'habileté θ = –1 mais plus de 80 % des élèves d'habileté θ = 0 aient réussi cet item témoigne de la discrimination sensible de l'item dans l'intervalle d'habileté [–1, 0].

1. Voir Lapointe *et al.* (1992).

FIGURE 4.3
Courbe caractéristique d'item d'un test de mathématique pour des élèves québécois francophones de 13 ans (modèle à deux paramètres)



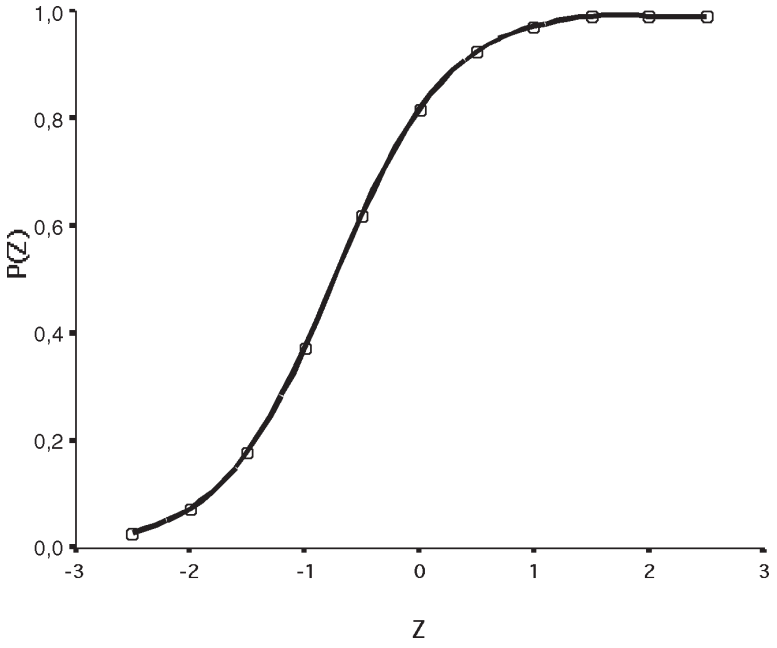
La façon dont cette courbe a été produite peut sembler mystérieuse pour les lecteurs non encore initiés aux multiples secrets inhérents à la théorie des réponses aux items. Les lignes qui suivent montrent qu'il existe un moyen de démystifier la construction de cette courbe. Notons, tout d'abord, qu'en abscisse se trouve l'habileté des élèves calculée en scores centrés et réduits (en scores z) et en ordonnée, la probabilité (proportion) de réussite de l'item. Nous allons tenter de reproduire une telle courbe sans avoir à estimer, comme cela a été fait pour obtenir la CCI de la figure 4.3, les valeurs des paramètres d'items ou des paramètres d'habileté des élèves. La figure 4.4 donne les proportions de réussite de cet item, P(z), pour certaines valeurs de z, à savoir ces valeurs standardisées[2] bien connues prises sur le score total du test de mathématique (76 items) : ces scores classiques serviront en quelque sorte d'estimé d'habileté des élèves. Nous avons, par la suite, regroupé ces scores classiques autour des valeurs z = –3, z = –2, z = –1, z = 0, z = 1, z = 2 et z = 3 et calculé la proportion d'élèves ayant réussi cet item autour de chacune de ces valeurs. Enfin, nous avons relié les points trouvés à l'aide d'une méthode d'interpolation (*spline*).

La figure 4.4 montre qu'environ 40 % des élèves ayant obtenu un score z = –1 ont réussi l'item. En outre, environ 80 % des élèves ayant obtenu un score z = 0 ont réussi l'item. Il s'agit bien de nombres qui sont suffisamment près de ceux obtenus à la figure 4.3 à l'aide d'une méthode d'estimation

2. Notons que la distribution de ces scores standardisés est, ici, approximativement normale.

beaucoup plus sophistiquée, utilisée de façon presque routinière en théorie des réponses aux items, et dont nous parlerons au chapitre 6. D'ailleurs, l'allure générale de la courbe représentée à la figure 4.4 n'est pas sans rappeler celle de la figure 4.3, bien que ces deux courbes aient été obtenues de façon fort différente.

FIGURE 4.4
Proportions de réussite d'un item calculées à partir des scores totaux standardisés d'un test de mathématique (les points ont été reliés en employant la méthode d'interpolation *spline* accessible sous SPSS 10)

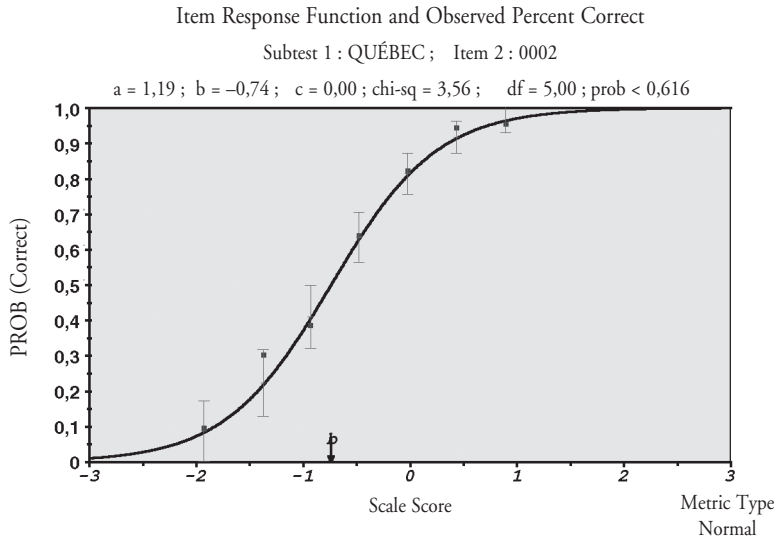


4.2.2. CCI et régression

Une autre façon de conceptualiser une courbe caractéristique d'item est de la comparer au tracé d'une courbe de régression. La figure 4.5 présente la courbe caractéristique de l'item de la figure 4.3 telle que produite par le logiciel BILOG-3 (Mislevy et Bock, 1990). Les points que l'on peut voir sur ce graphique constituent les proportions observées de réussite de l'item pour les élèves dont l'habileté correspond à l'abscisse (*scale score*) du point en question. La CCI elle-même est représentée par une ligne foncée. Les traits verticaux sur la CCI indiquent les intervalles de confiance à 95 % autour de la CCI. On note une certaine distance entre le point (l'observation) et la CCI (le modèle).

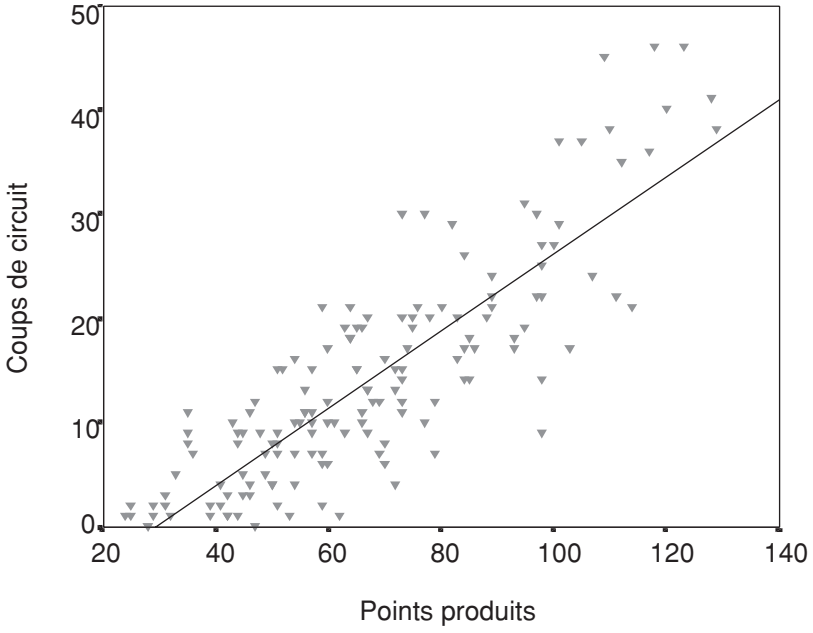
En réalité, la CCI tente d'épouser le nuage de points. Le tracé de cette CCI n'est pas sans rappeler celui d'une courbe de régression selon la méthode connue des moindres carrés.

FIGURE 4.5
Courbe caractéristique d'un item du test de mathématique pour les élèves québécois francophones de 13 ans



Voici maintenant un exemple qui illustre bien la production d'une droite des moindres carrés dans le contexte de la régression linéaire simple. La figure 4.6 montre la relation entre le nombre de coups de circuits (CC) et le nombre de points produits (PP) pour les 150 meilleurs joueurs de baseball majeur en 1993. Une régression linéaire effectuée à partir de ce nuage de points mène à une droite des moindres carrés comme celle de la figure 4.6 : certains points se situent près de la droite (ils sont bien prédits par le modèle) alors que d'autres en sont éloignés (ils sont moins bien prédits par le modèle). Il appert qu'en général la droite des moindres carrés se marie bien à la forme de ce nuage de points, indiquant par là que la droite des moindres carrés est un modèle approprié pour représenter la relation entre le nombre de coups de circuits (CC) et le nombre de points produits (PP) pour les 150 meilleurs joueurs de baseball majeur en 1993 (on a d'ailleurs calculé que la corrélation de Pearson dans ce cas vaut 0,85). Mais la réalité n'est pas toujours aussi simple. La section 1.1 a montré qu'il est parfois nécessaire d'utiliser un modèle quadratique pour rendre compte d'une relation entre des variables que l'on peut observer tous les jours comme la fréquence des approbations qu'un enseignant promulgue à ses élèves et leur rendement en mathématique.

FIGURE 4.6
Droite des moindres carrés obtenue à partir du nuage de points décrivant la relation entre le nombre de points produits (PP) et le nombre de coups de circuits (CC) des 150 meilleurs joueurs du baseball majeur en 1993



Alors que la droite des moindres carrés de la figure 4.6 est une manifestation de la régression linéaire simple et que la courbe quadratique de la figure 1.3 est obtenue à partir d'une régression polynomiale, la courbe caractéristique d'item de la figure 4.5 (qui est en fait la même que la CCI de la figure 4.3) a aussi été produite par une méthode de régression, dite logistique. Même si cette méthode de régression est un peu plus complexe, l'idée demeure essentiellement la même : ajuster une courbe à un nuage de points. Dans le cas qui nous intéresse, ce nuage de points est constitué des proportions de réussite des élèves qui se situent à tel ou tel niveau d'habileté. Alors que l'on utilise souvent la méthode des moindres carrés pour estimer[3] des courbes (une droite est une courbe) dans le cas de la régression linéaire (figure 4.6) et de la

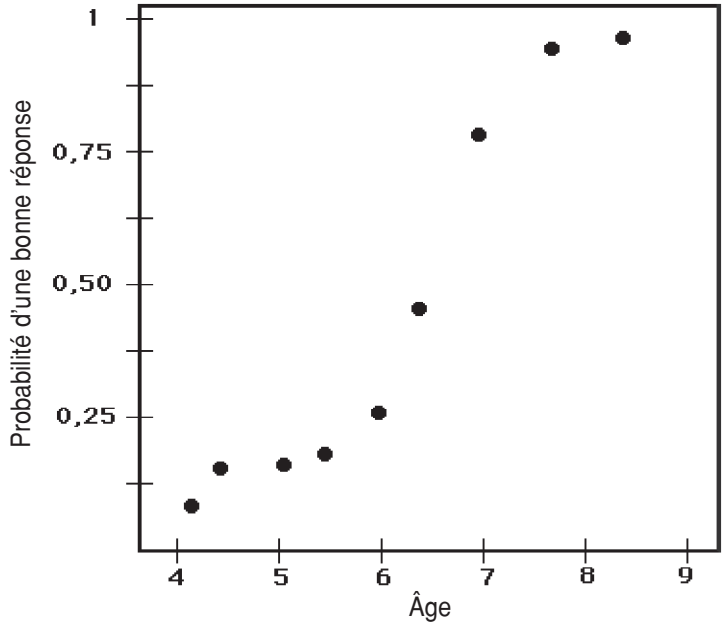
3. En fait, il s'agit d'estimer les paramètres d'un modèle. Si l'on suppose que le modèle est linéaire du type $Y = a + bX$, ce sont les paramètres a et b que l'on doit estimer. S'il s'agit d'un modèle quadratique du type $Y = a + bX + cX^2$, ce sont les paramètres a, b et c que l'on doit estimer. Ce procédé est essentiellement le même dans le cas d'une CCI : un modèle est supposé (voir la section 4.2.4) et les paramètres de ce modèle doivent également être estimés.

régression polynomiale (figure 1.3), c'est la méthode d'estimation connue sous le nom du maximum de vraisemblance (et ses principales dérivées) qui est le plus souvent utilisée dans le cas d'une CCI, comme nous allons le voir au chapitre 6.

4.2.3. Exemples de CCI

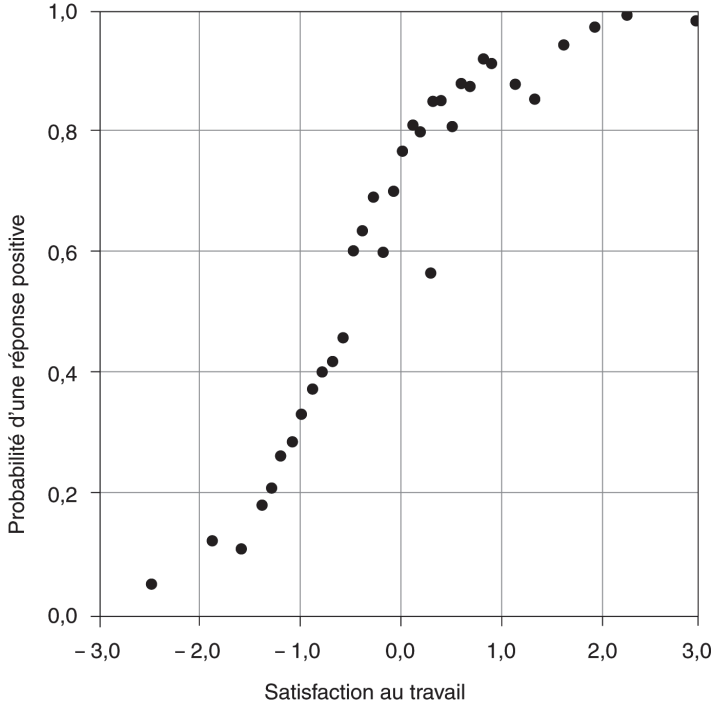
Voici quelques exemples qui montrent que le modèle en « S » d'une CCI n'émane pas d'une quelconque tour d'ivoire, mais est bel et bien observable depuis déjà fort longtemps. Hambleton *et al.* (1985, p. 6) de même qu'Embretson (1999) mentionnent que Binet et Simon sont parmi les premiers à avoir eu l'intuition d'une courbe caractéristique d'item. Regardons plutôt la figure 4.7. Elle représente la relation entre l'âge et la probabilité de réussir un item provenant d'un test d'intelligence. Manifestement, la probabilité de réussir cet item augmente avec l'âge, mais cette augmentation n'est pas linéaire : le modèle en « S » d'une CCI semble plus approprié pour expliquer la relation sous-jacente à ce nuage de points.

FIGURE 4.7
Relation entre l'âge et la probabilité de réussir un item du test d'intelligence de Binet et Simon (1916)



La figure 4.8 (Hulin *et al.*, 1983, p. 23) indique la relation entre la probabilité de répondre positivement à l'item (Trouvez-vous votre travail… « plaisant » ?) et la satisfaction générale au travail. Cet item provient de l'échelle *Job Descriptive Index* (JDI) de Smith *et al.* (1969). Le nuage de points a été obtenu à partir des réponses de 3812 personnes. Cette figure montre qu'environ 80 % des travailleurs qui ont une satisfaction moyenne à leur travail (score = 0) sont d'accord pour indiquer que leur travail leur semble plaisant. Encore ici on voit que plus les personnes interrogées trouvent leur travail satisfaisant en général et plus ils répondent positivement à l'item, c'est-à-dire qu'ils auraient recours au qualificatif « plaisant » pour décrire leur travail. Le modèle en « S » de la CCI semble bien approprié dans ce cas-ci également.

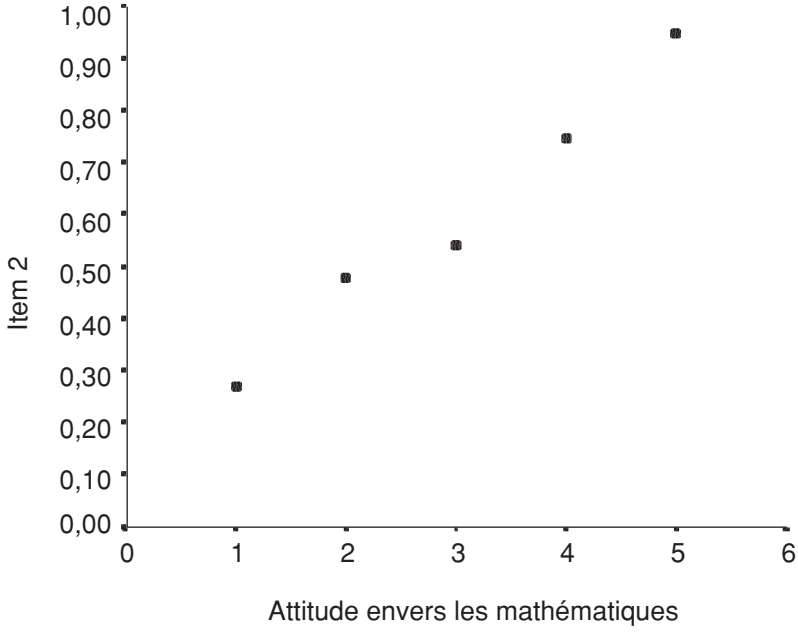
FIGURE 4.8
Relation entre la probabilité de répondre positivement à l'item (Trouvez-vous votre travail…) « plaisant » et la satisfaction générale au travail (Hulin *et al.*, 1983, p. 23)



La relation monotone croissante en forme de « S » est toujours perceptible dans le cas de l'item d'une échelle d'attitude envers les mathématiques observé à la figure 4.9. On y voit que plus l'attitude de l'élève envers les mathématiques est favorable, en considérant l'ensemble des items de l'échelle,

plus il a tendance à choisir la catégorie « tout à fait en accord » associée à un des items de l'échelle. L'échelle d'habileté dans ce cas-ci renvoie à l'attitude générale de l'élève envers les mathématiques, cette échelle étant divisée en cinq classes, la classe 1 regroupant les élèves les moins favorables et la classe 5 les élèves les plus favorables aux mathématiques.

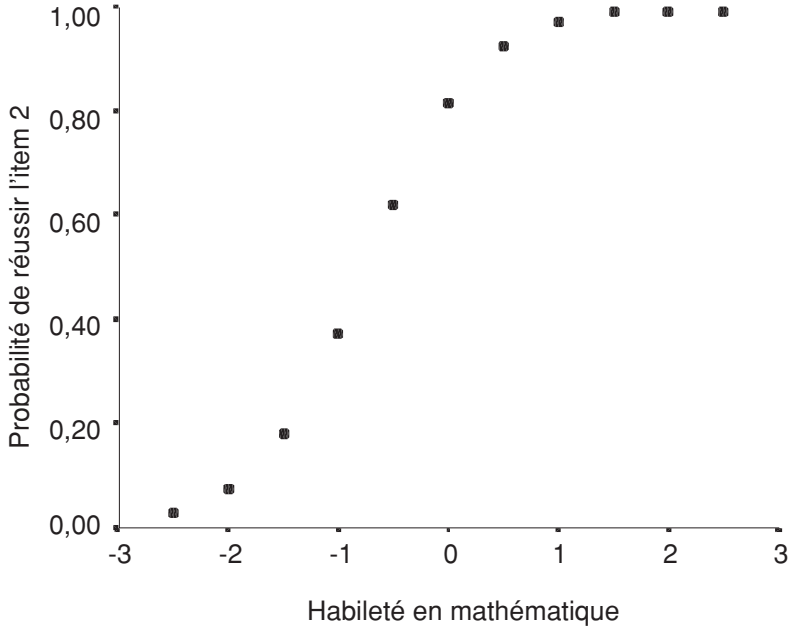
FIGURE 4.9
Relation entre l'attitude envers les mathématiques et la proportion d'élèves tout à fait en accord avec l'item 2 de l'échelle d'attitude



La figure 4.10 présente l'item 2 du test de mathématique administré à des étudiants québécois de 13 ans. La CCI a été tracée à partir des réponses données par 1000 étudiants regroupées en fonction de leur niveau estimé d'habileté en mathématique. On voit se profiler encore plus clairement que dans les autres cas le modèle en forme de « S ».

Tous ces exemples montrent que le modèle en forme de « S » est souvent approprié pour rendre compte d'une relation entre l'habileté et la probabilité de répondre correctement à un item visant cette habileté. Reste maintenant à caractériser formellement ce modèle. On ne peut tout de même pas toujours dire « un modèle en forme de S ». La prochaine sous-section présente des propositions de modèles qui ont vu le jour au fil des ans. Comme nous allons le voir, plusieurs de ces modèles dits historiques présentent des caractéristiques encore en vogue dans les modèles contemporains.

FIGURE 4.10
Proportions de réussite de 1000 étudiants québécois de 13 ans en fonction de leur niveau estimé d'habileté en mathématique pour l'item 2



4.2.4. CCI et modèles

Toute CCI est constituée à partir d'un modèle mathématique[4] défini a priori. Quoi de plus naturel alors que d'utiliser le modèle de l'ogive normale, dont on a si souvent fait l'éloge ? Nous avons vu à la figure 4.2 sa représentation géométrique, qui nous avait semblé plutôt sympathique ; surtout, elle est bien en forme de « S ». Regardons un peu à quoi ressemble la formulation algébrique de l'ogive normale.

$$P_i(\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_i(\theta - b_i)} e^{-\frac{1}{2}z^2} \, dz \qquad (4.1)$$

Bon, ce n'est pas trop rassurant à première vue : il s'agit d'une intégration (le grand S) de la fonction de densité de la loi normale. En fait, il s'agit de cumuler la proportion de surface sous la courbe normale qui se trouve

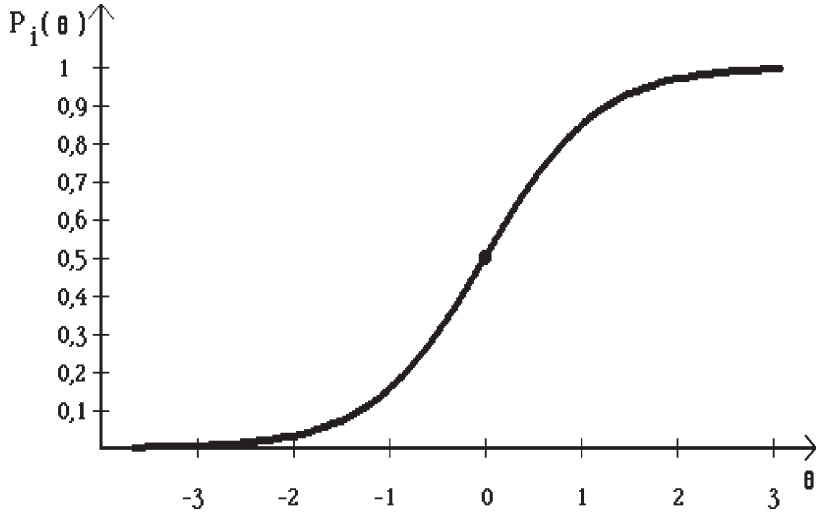
4. Excepté les modèles dits non paramétriques.

à gauche de $z = a_i (\theta - b_i)$. Cette proportion qui varie forcément entre 0 et 1 est de ce fait une bonne candidate pour représenter la probabilité de réussir un item i. Cependant, comme il s'avérait très laborieux d'effectuer plusieurs de ces intégrations, les pionniers de la théorie se sont tournés vers des modèles dont la formulation mathématique était moins complexe, mais dont la représentation graphique par ailleurs demeurait sensiblement la même. C'est à Birnbaum (1968) que revient l'honneur d'avoir montré que l'on pouvait approximer l'ogive normale (équation 4.1) par une fonction **logistique** du type

$$P_i(\theta) = \frac{1}{1 + e^{-D\, a_i (\theta - b_i)}} \tag{4.2}$$

Cette fonction ne requiert aucune intégration, en plus d'avoir une représentation graphique qui est sensiblement la même que l'ogive normale[5] pour peu qu'on fixe la constante D à 1,7. Comme on peut le voir à la figure 4.11, la courbe obtenue selon le modèle logistique est tout à fait semblable à l'ogive normale de la figure 4.2. C'est donc sans perte de généralité que cet ouvrage pourra s'en tenir à l'étude exclusive des modèles logistiques.

FIGURE 4.11
Courbe caractéristique d'item obtenue selon la fonction logistique (équation 4.2) où $a_i = 1$, $b_i = 0$ et D = 1,7



5. Birnbaum a même montré que si D = 1,7 alors l'écart entre l'ogive normale et l'ogive logistique ne dépasse jamais 1 %.

Notons par ailleurs qu'il est très facile de tracer une courbe caractéristique à partir d'une fonction logistique. Il n'est même pas nécessaire d'avoir recours à un logiciel dédié aux analyses statistiques comme SPSS ou à un tableur comme Excel : il suffit d'utiliser une calculatrice et du papier quadrillé. Le tableau 4.1 montre les valeurs de $P_i(\theta)$ correspondant à certaines valeurs de θ pour la fonction logistique de la figure 4.11. Lorsque $a_i = 1$, $b_i = 0$ et $D = 1{,}7$ la fonction logistique se réduit à $P_i(\theta) = \dfrac{1}{1 + e^{-1{,}7\theta}}$.

Tableau 4.1

Valeurs[6] de la fonction logistique $P_i(\theta) = \dfrac{1}{1 + e^{-1{,}7\theta}}$

θ	$P_i(\theta)$
–3	0,01
–2,5	0,01
–2	0,03
–1,5	0,07
–1	0,15
–0,5	0,30
0	0,50
0,5	0,70
1	0,85
1,5	0,93
2	0,97
2,5	0,99
3	0,99

Ce sont donc les modèles de la famille logistique que nous retiendrons ici ; ils sont d'ailleurs largement adoptés par les chercheurs qui s'appuient sur les modèles de réponses aux items. Il existe cependant des modèles plus simples qui peuvent, en quelque sorte, être considérés comme les précurseurs des modèles logistiques actuels. Nous décrivons trois de ces modèles, à savoir ceux présentant des caractéristiques qui ont pavé la voie aux modèles actuels. Nous renvoyons le lecteur intéressé à approfondir l'étude de ces modèles d'intérêt historique à Torgerson (1958) et à Hulin *et al.* (1983).

6. Il serait bien fastidieux de calculer les valeurs d'une fonction logistique à chaque fois que les valeurs de a_i et de b_i changent. C'est pourquoi, afin d'économiser le papier et d'épargner la patience des néophytes, nous avons produit un logiciel d'appoint nommé « LOG3 », qui permet de tracer les courbes à partir de la donnée des paramètres a_i et b_i. Ce logiciel peut être obtenu en s'adressant au premier auteur : Richard.Bertrand@fse.ulaval.ca.

Le modèle déterministe de Guttman (1950)

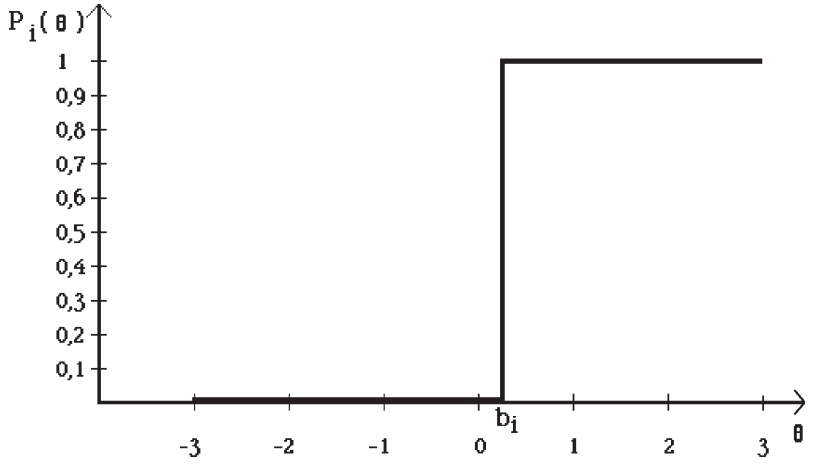
Début des années 1950, Guttman propose un modèle qui caractérise la relation entre l'habileté d'un individu, symbolisée par θ, et la probabilité de réussir un item noté i. Le modèle suppose un seul paramètre libre de varier d'un item à l'autre, b_i. La figure 4.12 présente un item reflétant le modèle de Guttman : on peut observer que, pour un item de paramètre b_i, si $P_i(\theta)$ est défini comme la probabilité d'un individu d'habileté θ de réussir l'item i,

$$P_i(\theta) = 0 \text{ , si } \theta < b_i$$

$$P_i(\theta) = 1 \text{ , si } \theta \geq b_i$$

Ce modèle renvoie à une règle de décision dichotomique : les individus dont l'habileté θ est inférieure à la valeur du paramètre b_i d'un item i n'ont, selon le modèle de Guttman, aucune chance de réussir l'item ($P_i(\theta) = 0$). Par contre, les individus dont l'habileté est supérieure ou égale à b_i sont certains de réussir l'item i ($P_i(\theta) = 1$).

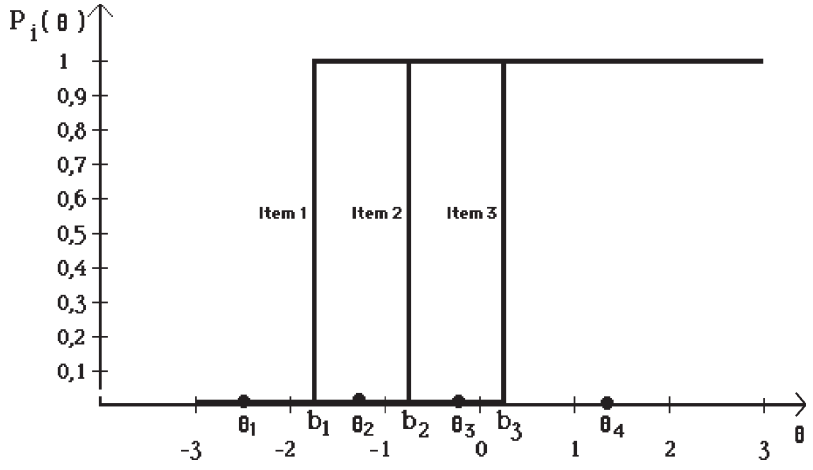
FIGURE 4.12
Courbe caractéristique d'item de paramètre b_i obtenue selon le modèle déterministe de Guttman



Bien qu'il y ait très peu de cas observables où l'on peut envisager la pertinence d'une situation si tranchée, ce modèle n'en demeure pas moins intéressant dans la mesure où plusieurs des caractéristiques des modèles couramment employés aujourd'hui en théorie des réponses aux items y sont présentes. Tout d'abord, le modèle donne lieu à une courbe monotone croissante : plus l'habileté d'un individu augmente, plus sa probabilité de réussir l'item

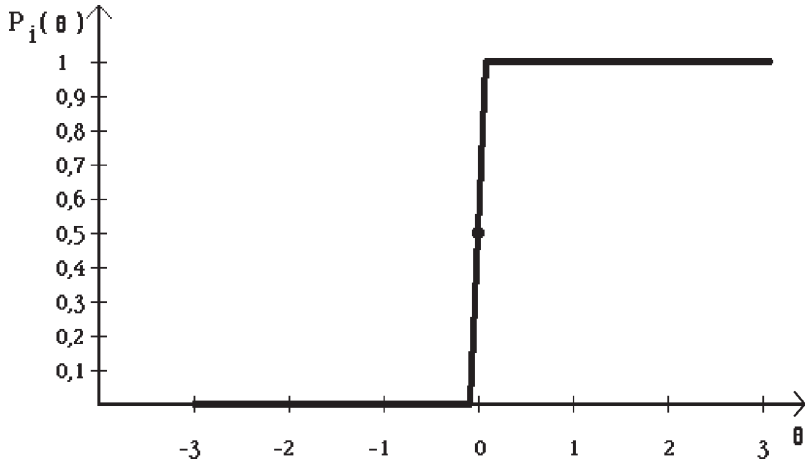
reste stable ou augmente. Ensuite, l'échelle d'habileté permet de comparer aisément l'habileté des individus à la valeur du paramètre d'item b_i. Mais que représente au juste ce paramètre b_i ? La figure 4.13 montre le tracé de trois CCI (avec leur paramètre respectif b_1, b_2 et b_3) obtenues à partir du modèle de Guttman. Sont aussi indiquées les valeurs de l'habileté pour quatre individus. On y voit que plus la valeur du paramètre b_i est élevée (c'est-à-dire plus l'item est situé vers la droite), plus l'item est difficile. En effet, un individu d'habileté θ_2 réussira l'item 1 mais pas les items 2 ou 3. De même, un individu d'habileté θ_3 réussira les items 1 et 2 mais pas l'item 3. Ainsi, plus la valeur du paramètre b_i est élevée, plus l'item i est difficile. De toute évidence, b_i joue le rôle d'un indice de difficulté de l'item i : l'item 3 est plus difficile que l'item 2, lui-même plus difficile que l'item 1.

FIGURE 4.13
Trois courbes caractéristiques d'item (de difficulté b_1, b_2 et b_3 respectivement) obtenues selon le modèle de Guttman : θ_1, θ_2, θ_3 et θ_4 représentent l'habileté de quatre individus



Il faut encore noter que le modèle de Guttman, bien qu'offrant très peu de possibilités sur le plan des applications, ne constitue en fait qu'un cas particulier d'un modèle logistique : celui où la pente au point d'inflexion (qui est proportionnelle au paramètre a_i) est infinie. Voici en effet, à la figure 4.14, une CCI provenant d'un modèle logistique dont la pente est extrêmement élevée. La ressemblance entre cette courbe, produite selon l'équation 4.2 d'un modèle logistique, et celle produite par le modèle déterministe de Guttman est patente.

FIGURE 4.14
Courbe caractéristique d'item produite à partir du modèle logistique, équation 4.2, où $b_i = 0$ et $a_i = 1000$



Le modèle de la distance latente de Lazarsfeld (1950)

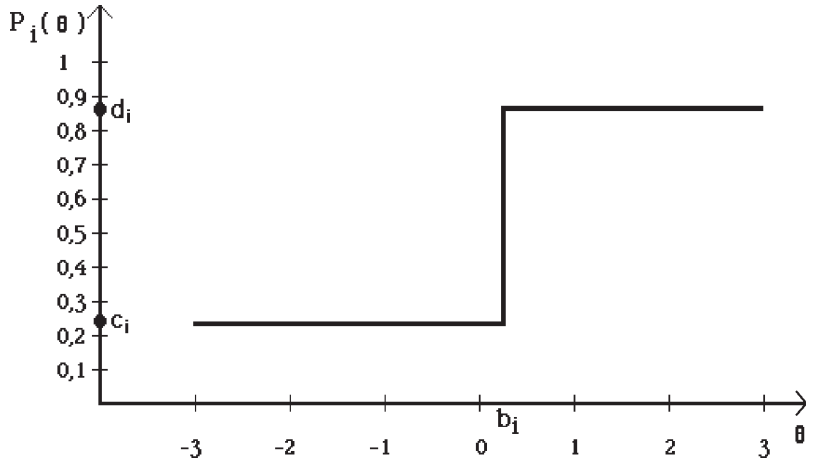
Le modèle dit de la distance latente de Lazarsfeld possède plusieurs des caractéristiques du modèle déterministe de Guttman, comme il est facile de le constater à la figure 4.15. Ce modèle respecte les conditions suivantes :

$$P_i(\theta) = c_i \text{ , si } \theta < b_i$$

$$P_i(\theta) = d_i \text{ , si } \theta \geq b_i$$

Le paramètre b_i joue le même rôle que le paramètre du même nom dans le modèle de Guttman. Cependant, le modèle de la distance latente utilise deux paramètres de plus que le modèle de Guttman : le paramètre c_i,, qui fixe une limite inférieure à la probabilité de réussir l'item i, et le paramètre d_i, qui indique la probabilité maximale de réussir l'item i. Bien sûr, ce modèle n'a pas eu beaucoup plus de succès que celui de Guttman compte tenu des restrictions qu'il impose aux observations. Nous élaborerons sur certaines caractéristiques de ce modèle, notamment la signification donnée au paramètre c_i, au moment de présenter le modèle logistique à trois paramètres.

FIGURE 4.15
Courbe caractéristique d'item produite à partir du modèle de la distance latente de Lazarsfeld



Le modèle linéaire de Lazarsfeld (1959)

Fin des années 1950, Lazarsfeld présente un modèle qui suppose une relation linéaire entre l'habileté d'un individu et la probabilité de réussir un item. Son modèle peut s'écrire comme l'équation d'une droite, soit

$$P_i(\theta) = k_i + a_i\theta$$

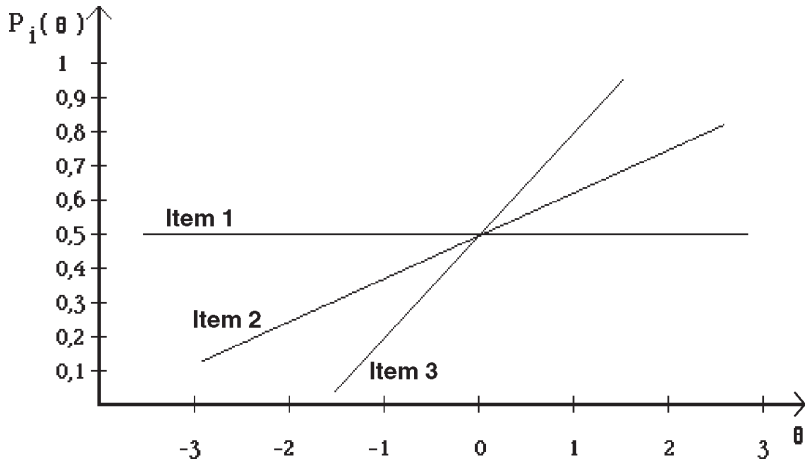
où k_i est l'ordonnée à l'origine (l'endroit où $\theta = 0$) et a_i est la pente de la droite.

Comme on peut facilement le constater en examinant la figure 4.16, le modèle linéaire n'est pas vraiment applicable : par exemple, la probabilité de réussir l'item 3 serait parfois inférieure à 0, pour une valeur suffisamment faible de θ, parfois supérieure à 1, pour une valeur suffisamment élevée de θ, une situation qui, comme on le sait, défie la règle limitant les valeurs d'une probabilité à l'intervalle [0 , 1].

Ce modèle n'est pas dénué d'intérêt pour autant. Attardons-nous d'abord au paramètre k_i, que nous avons décrit comme l'ordonnée à l'origine : il s'agit en réalité de la probabilité de réussir l'item i pour un individu d'habileté moyenne ($\theta = 0$). Ce concept a priori intéressant ne sera toutefois pas repris par les chercheurs qui ont élaboré les modèles contemporains de réponses aux items. Le rôle du paramètre a_i, la pente de la droite représentant l'item i, est cependant beaucoup plus important ; il aura des répercussions

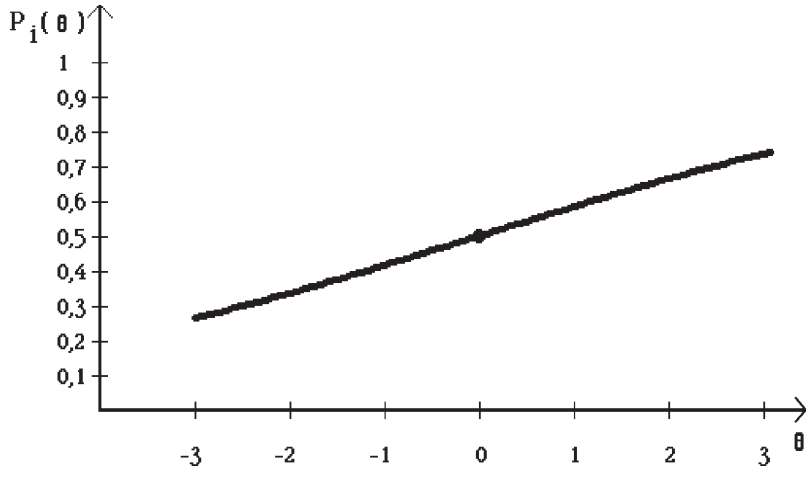
intéressantes au moment où nous présenterons les caractéristiques du modèle logistique à deux paramètres. Il n'est pas difficile de se rendre compte, en effet, que plus la pente a_i est élevée, plus l'item i discrimine. Considérons par exemple deux individus, le premier d'habileté $\theta_1 = 0$ et le second d'habileté $\theta_2 = 1$. L'item 1, de pente nulle, ne discrimine pas du tout ces deux individus, puisque $P_1(0) = P_1(1) = 0{,}50$. Par contre, l'item 2, de pente positive mais faible, discrimine légèrement ces deux individus puisque $P_2(0) = 0{,}50$ mais $P_2(1) = 0{,}62$. Enfin, l'item 3, de pente plus élevée que l'item 2, est celui qui discrimine le mieux ces deux individus puisque $P_3(0) = 0{,}50$ mais $P_3(1) = 0{,}80$: c'est-à-dire qu'un écart de 30 % distingue le premier du second individu quant à la réussite de l'item 3.

FIGURE 4.16
Courbes caractéristiques de trois items produite à partir du modèle linéaire de Lazarsfeld



Encore ici, le modèle linéaire de Lazarsfeld est un cas limite d'un modèle logistique défini par l'équation (4.2). Par exemple, la figure 4.17 présente une CCI obtenue à partir d'un modèle logistique où $b_i = 0$ et $a_i = 0{,}2$. On remarque que cette courbe est une assez bonne approximation du modèle linéaire. L'astuce ici consiste à s'en tenir à l'intervalle du tracé de la CCI entre $\theta = -3$ et $\theta = 3$. Ainsi, dans cet intervalle, la CCI ressemble à un modèle linéaire, alors que si nous pouvions voir la CCI sur toute l'étendue de l'échelle d'habileté θ, on verrait se profiler une courbe (quoique très couchée) en forme de « S », caractéristique maintenant familière des modèles logistiques. Notons que, contrairement aux courbes de la figure 4.16, jamais la CCI de la figure 4.17 ne prendra des valeurs $P_i(\theta)$ inférieures à 0 ou supérieures à 1.

FIGURE 4.17
Courbe caractéristique d'item produite à partir du modèle logistique, équation 4.2, où $b_i = 0$ et $a_i = 0{,}2$



4.3. LES TROIS MODÈLES LOGISTIQUES ET LES PARAMÈTRES D'ITEMS

La contrainte que nous nous sommes donnée de nous restreindre aux modèles logistiques ne nous autorise pas pour autant à faire l'économie du sens psychométrique à donner aux principales caractéristiques de la fonction logistique. C'est ici que sera explicitée l'interprétation des paramètres des modèles que nous avons qualifiés d'historiques et qui proviennent en bonne partie des travaux de Guttman et de Lazarsfeld.

Nous distinguerons les trois principaux modèles retenus suivant le nombre de paramètres considérés dans le modèle. Nous verrons que chacun des deux premiers modèles est en fait un cas particulier du modèle le plus complexe considéré ici, le modèle à trois paramètres. Mais l'objectif principal de cette section consiste beaucoup plus à donner une interprétation de chacun des paramètres des modèles unidimensionnels les plus couramment utilisés en théorie des réponses aux items.

4.3.1. Le modèle à un paramètre et le paramètre de difficulté

Le modèle logistique à un paramètre est obtenu en supposant que le seul paramètre d'item qui varie dans l'équation 4.2 est b_i. Ce modèle postule donc que, pour chaque item considéré, la valeur de a_i est constante. Un modèle

encore plus restrictif survient en posant $a_i = 1$ et $D = 1$. Ce dernier modèle, aussi appelé **modèle de Rasch** du nom de son concepteur, peut s'écrire

$$P_i(\theta) = \frac{1}{1 + e^{-(\theta - b_i)}} \tag{4.3}$$

Il s'agit d'un modèle très simple et aussi très populaire, du fait justement de sa simplicité, mais aussi des qualités particulières qui le caractérisent. Nous voulons ici mettre en lumière les principales caractéristiques de ce modèle, et notamment, l'interprétation du paramètre d'item b_i.

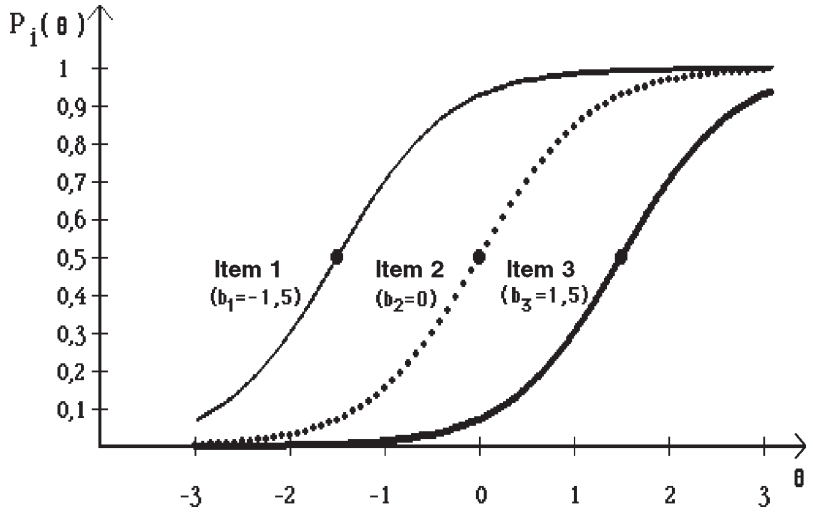
Souvenons-nous tout d'abord des courbes produites par le modèle de Guttman (figure 4.13). Un seul paramètre est libre de varier d'un item à l'autre, soit b_i. Nous avons vu que ce paramètre était un bon indicateur de la difficulté d'un item dans le cas du modèle de Guttman. Voyons ce qui en est de ce paramètre dans le cas d'un modèle logistique à un paramètre. La figure 4.18 présente les CCI de trois items tracés selon un modèle à un paramètre, a_i étant fixé à 1 et D à 1,7. Seul le paramètre b_i varie d'un item à l'autre : pour l'item 1, $b_1 = -1{,}5$; pour l'item 2, $b_2 = 0$ et pour l'item 3, $b_3 = 1{,}5$. En fait, tout comme pour les courbes tracées selon le modèle de Guttman, les CCI tracées selon le modèle de Rasch (ou tout modèle logistique à un paramètre) sont parallèles. Elles ne sont que des translations les unes des autres : il est possible d'obtenir, par exemple, l'item 2 (dans le même intervalle de l'échelle θ) en faisant glisser l'item 1 horizontalement vers la droite ou encore en faisant glisser l'item 3 vers la gauche[7].

Qu'en est-il maintenant du paramètre b_i ? Il est facile de se rendre compte qu'il joue le même rôle que le paramètre du même nom dans le modèle de Guttman, soit un indice de la difficulté de l'item i. En effet, un individu d'habileté moyenne, pour lequel $\theta = 0$, aura plus de difficulté à réussir l'item 3 où $b_3 = 1{,}5$ que l'item 2 où $b_2 = 0$; mais aussi plus de difficulté à réussir l'item 2 où $b_2 = 0$ que l'item 1 où $b_1 = -1{,}5$. On peut s'en rendre compte en calculant, pour cet individu, les probabilités $P_i(\theta)$ de réussir chacun des trois items à l'aide de l'équation 4.4 ou tout simplement en examinant la figure 4.19.

$$P_i(\theta) = \frac{1}{1 + e^{-1{,}7(\theta - b_i)}} \tag{4.4}$$

7. Incidemment, il est recommandé, à ce moment-ci, d'utiliser le logiciel LOG 3 pour illustrer cette translation : il s'agit de sélectionner une des courbes et de la faire glisser (*drag*) horizontalement (en gardant la touche majuscule enfoncée) jusqu'à ce qu'elle soit superposée à une des deux autres courbes.

FIGURE 4.18
Courbes caractéristiques de trois items selon le modèle à un paramètre : l'item 1 est plus facile que l'item 2, lui-même plus facile que l'item 3

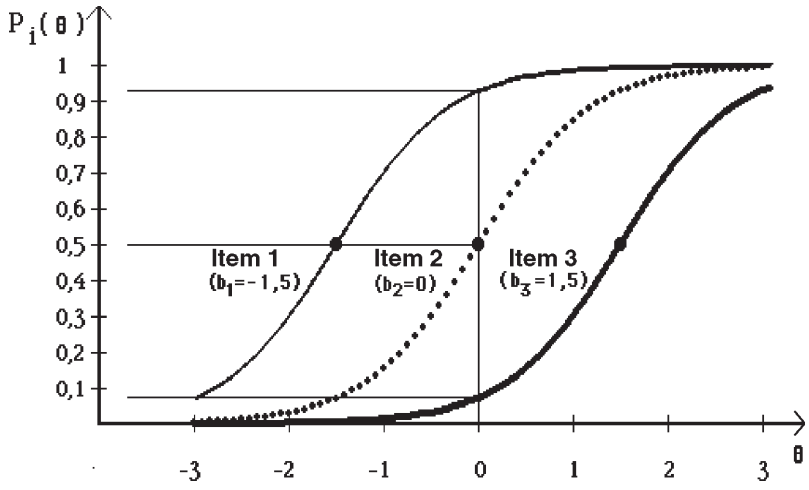


Cette figure nous montre que $P_1(0) = 0{,}93$, $P_2(0) = 0{,}5$ et $P_3(0) = 0{,}07$. Nous pourrions refaire ces mêmes calculs (ou tout simplement jeter un regard sur la figure 4.19) pour chacune des valeurs de l'échelle θ et nous obtiendrions le même verdict : plus la valeur du paramètre b_i d'un item i est élevée, plus il est difficile de le réussir. Dans une interprétation plus visuelle, on constate que plus la CCI se trouve à droite de l'échelle, plus l'item est difficile.

Quelques mises en garde sont cependant de rigueur. Premièrement, notons que la valeur du paramètre de difficulté peut être négative : ainsi, un item i où $b_i = -1$ sera plus difficile qu'un item j où $b_j = -2$. Deuxièmement, retenons que ce résultat général qui stipule que plus b_i est élevé, plus l'item i est difficile, n'est vrai de façon absolue que si l'on utilise un modèle à un paramètre : on montrera plus loin qu'il faut nuancer ce résultat lorsqu'il y a deux paramètres ou plus dans un modèle logistique. Troisièmement, il faut remarquer au passage que ce résultat fait de b_i un véritable indice de difficulté, ce qui contraste par exemple avec l'indice classique p_i qui est plutôt un indice de facilité tel qu'indiqué à la section 2.5. Finalement, il serait plus approprié de parler d'**un** modèle à un paramètre que **du** modèle à un paramètre. En effet, il faut se rappeler qu'il existe, en fait, plusieurs modèles à un paramètre, soit un pour chaque valeur fixée de a_i et chaque valeur de D, même si, la plupart du temps, on n'utilise que le modèle de Rasch ($a_i = 1$ et $D = 1$) ou le modèle normal à un paramètre ($a_i = 1$ et $D = 1{,}7$). Notons que le logiciel

BILOG 3 permet notamment de produire des modèles à un paramètre où a_i est fixé à une valeur différente de 1. Toutefois, lorsque nous parlerons **du** modèle à un paramètre, il faudra entendre le modèle décrit à l'équation 4.4 où $a_i = 1$ et $D = 1{,}7$. Dans le cas où $a_i = 1$ et $D = 1$, nous parlerons tout simplement du modèle de Rasch.

FIGURE 4.19
Représentation graphique de la probabilité ($P_i(\theta)$) de réussir les items 1, 2 et 3 pour un individu d'habileté moyenne ($\theta = 0$)



4.3.2. Le modèle à deux paramètres et le paramètre de discrimination

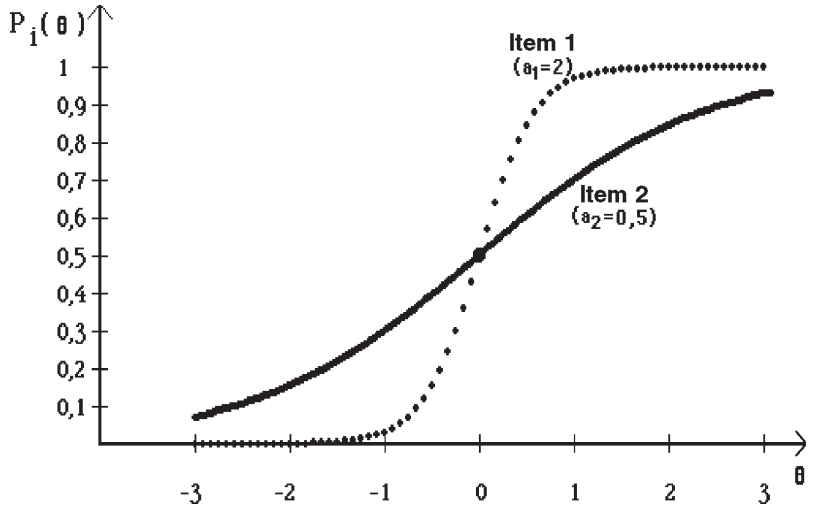
Revenons maintenant au modèle présenté à l'équation 4.2.

$$P_i(\theta) = \frac{1}{1 + e^{-D\, a_i (\theta - b_i)}} \qquad (4.2)$$

Il suppose deux paramètres libres de varier pour chacun des items, soit a_i et b_i. Le paramètre b_i joue, à quelques nuances près, le même rôle qu'au sein d'un modèle à un paramètre : c'est un indice de difficulté. Pour bien interpréter le paramètre a_i, il convient de se rappeler le double rôle que jouait le paramètre a_i dans le modèle linéaire de Lazarsfeld : une pente, mais aussi un indice de discrimination. Examinons plutôt la figure 4.20. Chacun de ces deux items possède le même indice de difficulté : $b_1 = b_2 = 0$. Cependant l'item 1

possède une CCI dont la pente au point d'inflexion[8], m_1, est beaucoup plus abrupte que la pente correspondante m_2 de l'item 2. La valeur du paramètre a_i est proprotionnelle à cette pente au point d'inflexion. Ainsi, plus la pente au point d'inflexion de la CCI est abrupte, et plus la valeur du paramètre a_i est élevée. Il peut être montré, en fait, que pour un modèle logistique à deux paramètres, $a_i = 4\ m_i/D$. En supposant, comme nous avons l'habitude de le faire, que D = 1,7 alors $a_i = 2{,}35\ m_i$.

FIGURE 4.20
Courbes caractéristiques de deux items dont les indices de discrimination diffèrent : l'item 1 discrimine plus que l'item 2 au point θ = 0



La définition même du paramètre a_i en fonction d'une pente nous amène, par analogie avec le modèle linéaire de Lazarsfeld, à définir a_i comme un paramètre de discrimination de l'item i. Pour soutenir plus formellement nos dires, supposons deux individus d'habileté distincte, l'individu 1 d'habileté $\theta_1 = -0{,}5$ et l'individu 2 d'habileté $\theta_2 = 0{,}5$. Examinons maintenant la figure 4.21. Il est clair que l'item 1 discrimine mieux entre ces deux individus que l'item 2. En effet, en s'appuyant sur l'équation 4.2, où D = 1,7 et $b_1 = b_2 = 0$, ou encore en se satisfaisant des approximations visuelles de la figure 4.21, on obtient

8. Plus précisément, nous parlons ici de la pente de la droite qui est tangente au point d'inflexion. Pour plus de détails techniques, voir l'annexe 4.1.

$$P_1(\theta_1) = P_1(-0,5) = 1/[1+e^{-1,7a_1(-0,5)}] = 0,154$$

$$P_1(\theta_2) = P_1(0,5) = 1/[1+e^{-1,7a_1(0,5)}] = 0,846$$

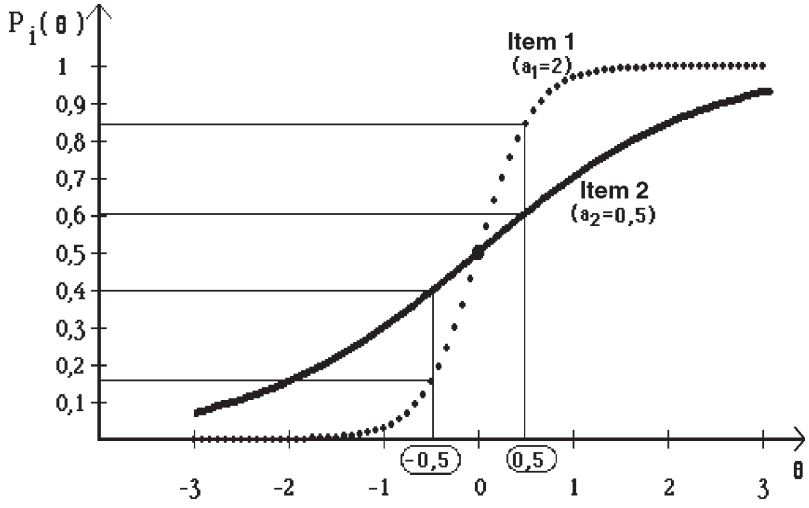
$$P_2(\theta_1) = P_2(-0,5) = 1/[1+e^{-1,7a_2(-0,5)}] = 0,395$$

$$P_2(\theta_2) = P_2(0,5) = 1/[1+e^{-1,7a_2(0,5)}] = 0,605$$

Ainsi, l'item 1 différencie beaucoup mieux les deux individus que l'item 2 puisque la différence de probabilité de réussite entre les deux individus est beaucoup plus accentuée dans le cas de l'item 1, comme en font foi les données suivantes :

- pour l'item 1, $P_1(\theta_2) - P_1(\theta_1) = 0,846 - 0,154 = 0,692$;
- pour l'item 2, $P_2(\theta_2) - P_2(\theta_1) = 0,605 - 0,395 = 0,210$.

FIGURE 4.21
Représentation graphique de la probabilité de réussir les items 1 et 2 pour des individus d'habileté $\theta_1 = -0,5$ et $\theta_2 = 0,5$



Mais attention, le paramètre a_i ne peut être considéré ici comme un indice global de discrimination de l'item i comme la corrélation bisériale l'était par exemple dans le cas du modèle classique. Au contraire, l'endroit où l'item i discrimine le plus, donc l'endroit où le paramètre a_i doit être interprété, dépend de la position du point d'inflexion de la CCI. Le paramètre a_i doit donc être interprété plus précisément comme un **indice de discrimination de l'item i dans le voisinage du point d'inflexion**. En outre, plus la valeur de a_i est élevée, plus ce voisinage est restreint. Il est clair, par exemple, à l'examen de

la figure 4.20, que même si a_2 est plus petit que a_1, l'item 2 discrimine mieux que l'item 1 les individus dont l'habileté se situe dans l'intervalle [–3 , –2] : tout simplement parce que, dans cet intervalle, la pente de l'item 2 est plus abrupte que la pente (pratiquement nulle) de l'item 1.

De plus, l'interprétation du paramètre b_i en tant qu'indice de difficulté de l'item i doit se faire avec beaucoup plus de prudence dans le cas d'un modèle à deux paramètres. En effet, contrairement au modèle à un paramètre, les CCI d'un modèle à deux paramètres ne sont pas nécessairement parallèles : en général, ces CCI se croisent puisque leurs pentes diffèrent. Ainsi, même si la valeur du paramètre b_i est la même ($b_1 = b_2 = 0$) pour les deux items de la figure 4.20, on peut se rendre compte sans peine que, pour des individus de faible habileté ($\theta < 0$), il est plus difficile de réussir l'item 1 que l'item 2 [$P_1(\theta) < P_2(\theta)$]. De façon symétrique, pour des individus d'habileté élevée ($\theta > 0$), il est plus difficile de réussir l'item 2 que l'item 1 [$P_2(\theta) < P_1(\theta)$]. Peut-on vraiment dire, en ce sens, que ce sont deux items de même difficulté ? Afin de déterminer la difficulté d'un item i à un endroit donné de l'échelle θ, il faut donc procéder à un examen visuel des CCI et interpréter la difficulté de l'item suivant la valeur de $P_i(\theta)$, la probabilité de réussite de l'item, à cet endroit de l'échelle : la valeur du paramètre b_i ne nous donne ici qu'une indication générale de la difficulté de l'item i. Strictement parlant, la valeur du paramètre b_i donne une indication de la position du point d'inflexion de la CCI le long de l'échelle θ.

Il est bon de souligner qu'il y a plusieurs modèles à deux paramètres : il suffit, notamment, de faire varier la valeur de D pour s'en rendre compte. Nous découvrirons plus loin, lors de l'étude du modèle à trois paramètres, que l'on peut fixer ce troisième paramètre de plusieurs façons, obtenant de la sorte autant de modèles à deux paramètres. Ainsi, lorsque nous mentionnerons **le** modèle à deux paramètres, il faudra tenir pour acquis qu'il s'agit du modèle logistique donné par l'équation 4.2 où D = 1,7.

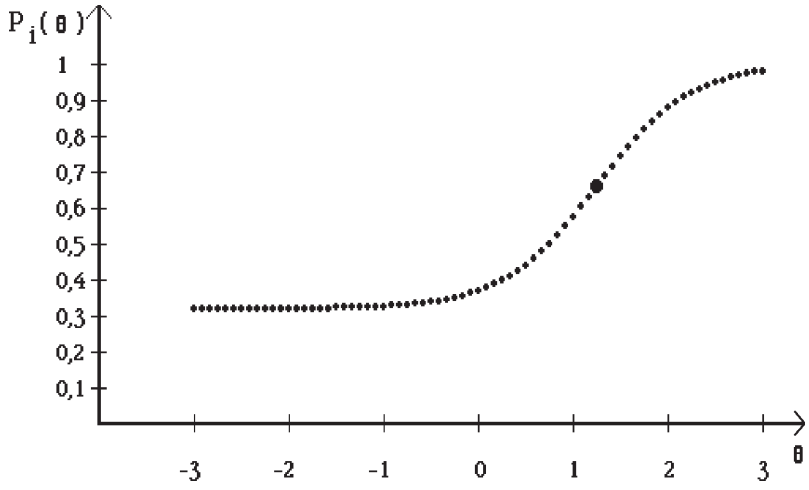
4.3.3. Le modèle à trois paramètres et le paramètre de pseudo-chance

Le modèle logistique à deux paramètres présenté à l'équation 4.2 ne suffit pas à rendre compte de toutes les formes de CCI couramment rencontrées. Considérons par exemple le cas de l'item 8 du test de mathématique obtenu à partir d'un échantillon de 1000 élèves québécois francophones de 13 ans (Lapointe *et al.*, 1992) : nous présentons la CCI de cet item à la figure 4.22.

Aucun des modèles logistiques présentés jusqu'ici ne peut produire une telle CCI. En effet, les modèles à un ou deux paramètres ont cette particularité de faire tendre l'asymptote inférieure de la CCI (la partie inférieure gauche de la CCI) vers 0 lorsque la valeur de θ est suffisamment petite. Pour s'en convaincre, il suffit de placer une valeur de θ, la plus petite possible, dans l'équation 4.2. Dans ce cas, il est toujours possible de choisir une valeur de θ

pour que la probabilité $P_i(\theta)$ de réussir un item i tende vers zéro. Considérons par exemple les deux CCI de la figure 4.20. Il n'est même pas nécessaire de calculer les valeurs de $P_i(\theta)$ pour se convaincre que l'asymptote inférieure de l'item 1 tend vers 0 : il n'y a qu'à regarder la CCI de cet item sur la figure. On voit bien qu'à partir de $\theta = -1{,}8$ environ, la CCI touche l'axe de l'échelle θ et donc que l'asymptote inférieure tend vers 0. C'est un peu moins clair dans le cas de l'item 2 puisque la figure ne montre pas toutes les valeurs de l'échelle θ. Cependant, sachant que $D = 1{,}7$, $b_2 = 0$ et $a_2 = 0{,}5$, l'équation 4.2 montre que pour les valeurs –3, –4 et –10 de θ, $P_2(-3) = 0{,}0724$, $P_2(-4) = 0{,}0323$ et $P_2(-10) = 0{,}0002$. Ainsi, à la limite, pour une valeur de θ aussi faible que possible, que l'on peut nommer $\theta = \theta_{inf}$, la valeur de $P_2(\theta_{inf})$ tendra bien vers zéro.

FIGURE 4.22
Item 8 du test de mathématique obtenu à partir d'un échantillon de 1000 élèves québécois francophones de 13 ans dans le cadre d'une enquête internationale



Or, il existe des items que même un individu d'habileté très faible a une chance non négligeable de réussir : c'est le cas, par exemple, des items à choix multiple. L'item 8 représenté par la CCI de la figure 4.22 est justement un item à choix multiple. On voit bien, dans ce cas, que la valeur de $P_i(\theta)$ ne tend pas et ne tendra pas vers 0 (il n'y a plus de pente !) même pour des individus d'habileté aussi faible que possible : la valeur de $P_i(\theta)$ se stabilise autour de 0,32 pour les individus qui ont une habileté $\theta < -1{,}5$. Ainsi, pour cet item, nous pouvons calculer $P_1(-1) = 0{,}3271$, $P_1(-2) = 0{,}3209$, $P_1(-3) = 0{,}3201$, $P_1(-4) = 0{,}320017$ et $P_1(-10) = 0{,}3200000001$.

À la limite, donc, pour une valeur de θ aussi faible que possible, $P_i(\theta_{inf})$ tendra vers 0,32 et non vers 0.

Comme nous l'avons déjà exprimé, le modèle logistique à deux paramètres représenté par l'équation 4.2 ne permet pas de rendre compte du comportement de ce type d'item. Une des caractéristiques de la CCI de la figure 4.22 n'est pas prise en compte par le modèle à deux paramètres : la hauteur minimale (non nulle) de la courbe. Or, cette caractéristique, qui peut aussi s'interpréter comme la probabilité minimale de réussir l'item, est présente dans le modèle de la distance latente de Lazarsfeld (figure 4.15) : le paramètre c_i, en effet, permet de hausser la courbe puisqu'il consiste en une probabilité minimale (non nulle) de réussite de l'item. Ce genre de paramètre de probabilité minimale de réussite peut aussi être incorporé à un modèle logistique de la façon suivante. Considérons l'équation 4.5, que nous appellerons dorénavant le modèle logistique à trois paramètres. Seul l'ajout du paramètre c_i distingue ce modèle du modèle logistique à deux paramètres déjà connu. Cette équation indique, en réalité, que la probabilité de réussir l'item i est composée de deux parties additives : c_i, la probabilité minimale de réussir l'item, et une valeur égale à $(1 - c_i)$ fois $P_i^*(\theta)$ où $P_i^*(\theta)$ n'est pas autre chose que la probabilité de réussir l'item i selon le modèle logistique à deux paramètres. C'est donc dire que, selon le modèle logistique à trois paramètres, la probabilité de réussir l'item i est égale à une valeur constante, notamment c_i, à laquelle nous devons ajouter la probabilité de réussir l'item selon le modèle à deux paramètres pour la partie de l'échelle des probabilités $P_i(\theta)$ supérieure à c_i, donc pour la partie restante, à savoir $1 - c_i$.

$$P_i(\theta) = c_i + (1 - c_i)\, P_i^*(\theta) = c_i + \frac{1 - c_i}{1 + e^{-D\, a_i (\theta - b_i)}} \qquad (4.5)$$

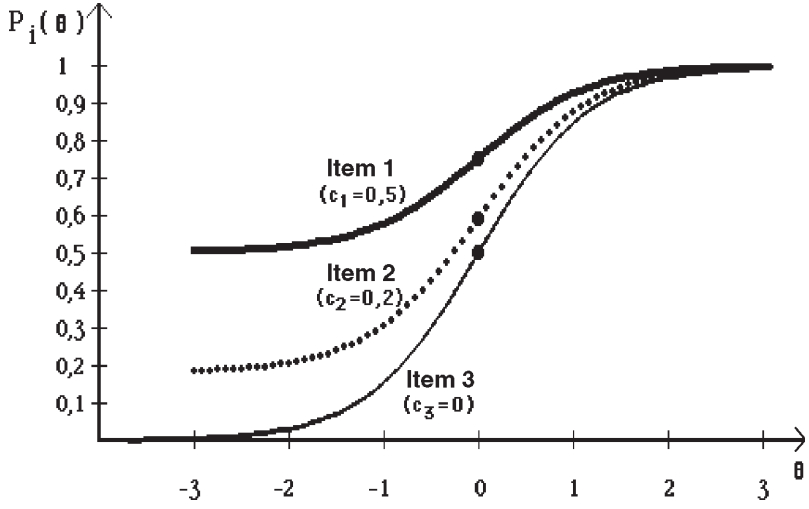
Le paramètre c_i du modèle logistique à trois paramètres doit s'interpréter comme la probabilité de réussir l'item i, pour un individu d'habileté θ aussi petite que l'on veut. Dans le cas de la CCI de la figure 4.22 par exemple, $c_i = 0{,}32$ signifie qu'un individu ayant peu ou pas d'habileté θ a tout de même une chance non négligeable de réussir cet item. Bien sûr, le modèle logistique à trois paramètres généralise le modèle logistique à deux paramètres puisqu'en posant $c_i = 0$ dans l'équation 4.5 on retrouve l'équation 4.2.

La figure 4.23 présente trois CCI obtenues selon le modèle logistique à trois paramètres. La valeur du paramètre de difficulté (b_i) de chacun de ces items est la même : $b_1 = b_2 = b_3 = 0$. En outre, la valeur du paramètre de discrimination de ces trois items est égale : $a_1 = a_2 = a_3 = 1$. Seule la valeur du paramètre c_i diffère d'un item à l'autre : $c_1 = 0{,}5$ alors que $c_2 = 0{,}2$ et $c_3 = 0$. L'impact du paramètre c_i est, toutes choses étant égales par ailleurs, de faire varier la probabilité de réussite de l'item i, surtout chez les individus de faible habileté : plus la valeur de c_i est élevée, plus les individus faibles voient

augmenter leurs chances de réussir l'item i. Ainsi, l'observation de cette figure montre que les individus d'habileté $\theta = -3$ ont une probabilité pratiquement nulle de réussir l'item 3, une probabilité d'environ 0,2 de réussir l'item 2 et une probabilité d'environ 0,5 de réussir l'item 1. Par contre, la probabilité de réussir l'un ou l'autre de ces items est essentiellement la même pour les individus dont l'habileté est $\theta = 3$.

Nous appellerons c_i le paramètre de pseudo-chance, signifiant par là que la probabilité minimale de réussir l'item i pour un individu d'habileté aussi faible que possible peut souvent être attribuée au hasard ou à la chance, mais qu'il pourrait aussi y avoir beaucoup d'autres facteurs explicatifs associés à cette probabilité (p. ex., la tricherie).

FIGURE 4.23
Courbes caractéristiques des items 1, 2 et 3 produites à partir du modèle à trois paramètres : la probabilité de réussir l'un ou l'autre item varie surtout pour les élèves très faibles



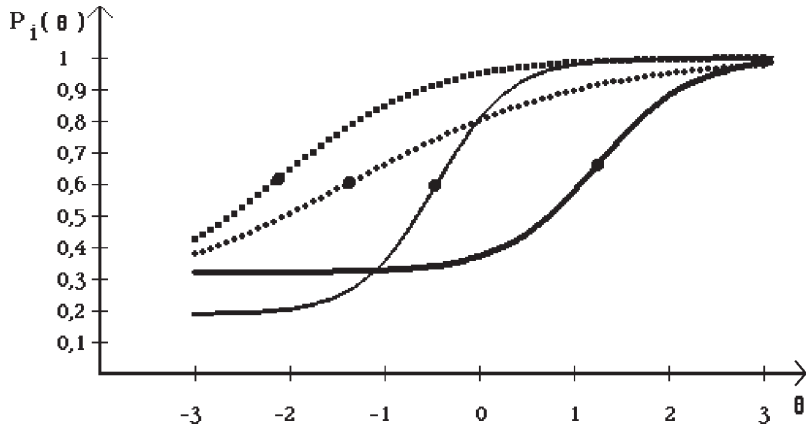
Il va sans dire que les mises en garde émises en rapport avec les nuances d'interprétation des paramètres de difficulté et de discrimination concernent aussi le modèle à trois paramètres. C'est pourquoi il est si important, voire crucial, de visualiser les CCI avant d'interpréter les qualités psychométriques d'un item modélisé selon la théorie des réponses aux items.

D'autres remarques spécifiques s'imposent dans le cas du modèle à trois paramètres. Soulignons d'abord que le paramètre c_i n'a que très peu d'influence sur la probabilité de réussite pour les individus d'habileté élevée. Un individu fort réussira l'item, que celui-ci ait un facteur de pseudo-chance

élevé ou non, comme en fait foi la figure 4.23. Il faut bien se rendre compte également que les items d'un modèle à trois paramètres ne se présentent pas vraiment tous comme à la figure 4.23 puisque, dans cette figure, seul le paramètre c_i est libre de varier. Une représentation plus réaliste d'items obtenus selon le modèle à trois paramètres, donc d'items dont les trois paramètres a_i, b_i, et c_i sont libres de varier, se trouve à la figure 4.24. Cette figure illustre en effet la représentation de la CCI de quatre items à choix multiples (4 choix chacun) d'un test de mathématique. Chacune de ces CCI a été obtenue à partir du modèle à trois paramètres. On voit bien que le fait de permettre la variation des trois paramètres à la fois a une influence sur la forme des CCI et qu'il est à toutes fins pratiques inutile d'essayer d'interpréter les CCI à partir de la seule donnée de la valeur des trois paramètres : **l'examen visuel** des courbes s'avère donc **nécessaire** afin d'obtenir une interprétation satisfaisante.

Il faut aussi noter que nous dirons toujours **le** modèle logistique à trois paramètres pour représenter n'importe quel modèle émanant d'une combinaison de paramètres de l'équation 4.5 où D est fixé à 1,7.

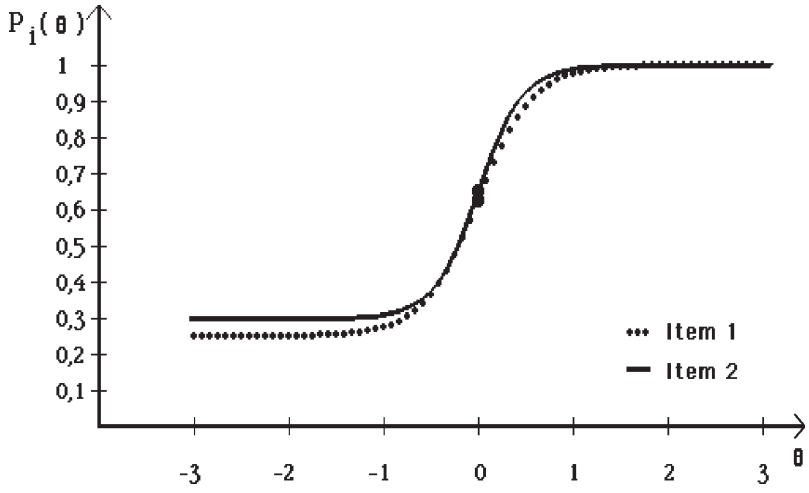
FIGURE 4.24
Courbes caractéristiques de quatre items à choix multiple produites à partir du modèle à trois paramètres (1000 individus, 76 items)



Autre remarque : les paramètres a_i et c_i sont sensibles à des changements subtils. En fait, il suffit de modifier un paramètre quelque peu pour que l'autre en soit aussi affecté. Par exemple, en changeant le paramètre a_i à la baisse (de $a_i = 2$ vers $a_i = 2{,}5$) et le paramètre c_i à la hausse (de $c_i = 0{,}25$ à $c_i = 0{,}3$), les deux CCI impliquées demeurent essentiellement superposées dans un intervalle donnée de l'échelle θ. La figure 4.25 montre que dans l'intervalle [–0,5 , 3], les deux CCI sont pratiquement superposées. Cela signifie

qu'au moment d'estimer les paramètres (voir le chapitre 6), l'algorithme pourrait avoir certaines difficultés à converger vers des estimés stables puisqu'à un nuage de points donné pourraient correspondre plusieurs possibilités de couplets de paramètres (a_i, c_i). Les paramètres a_i et c_i sont donc, en quelque sorte, dans une relation que l'on peut qualifier de compensatoire.

FIGURE 4.25
CCI de deux items montrant l'effet compensatoire entre les paramètres a_i et c_i : les paramètres sont, pour l'item 1, a_i = 2, b_i = 0, c_i = 0,25 et, pour l'item 2, a_i = 2,5, b_i = 0, c_i = 0,3.



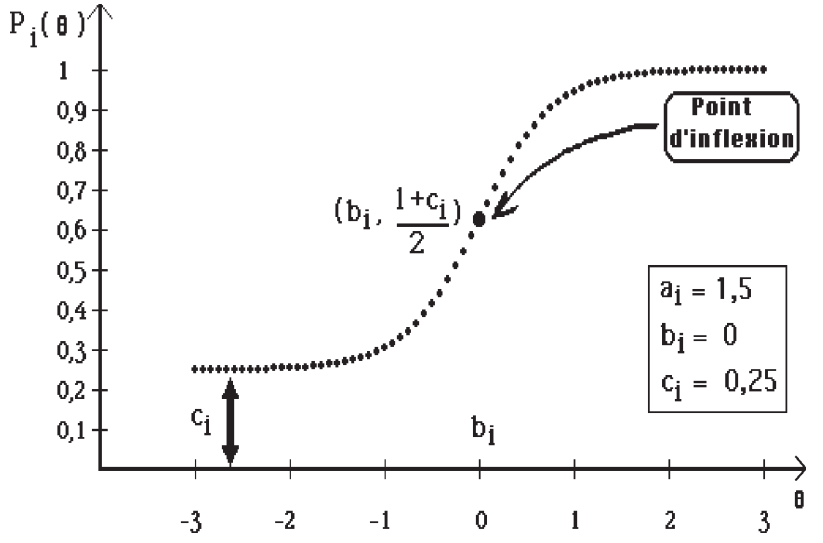
En guise de synthèse de la présentation des modèles logistiques de la théorie des réponses aux items, il nous a semblé opportun de revenir à un concept dont il a été question plutôt sommairement : le point d'inflexion d'une CCI. Examinons pour cela la figure 4.26. Le point d'inflexion représente, bien sûr, comme nous l'avons déjà souligné, le point où la CCI passe du concave au convexe, mais il y a plus. On pourrait dire que c'est à ce point de la courbe que tout se joue. En effet, les coordonnées du point d'inflexion sont égales à [b_i , (1+ c_i)/2]. Ainsi, l'abscisse du point d'inflexion donne la valeur du paramètre de difficulté b_i. L'ordonnée du point d'inflexion est située à mi-distance entre c_i et 1, où c_i peut être vue comme la probabilité minimale de réussir l'item et 1, la valeur maximale que peut prendre la probabilité de réussir l'item. Incidemment, lorsque c_i = 0, donc dans le cas d'un modèle à deux paramètres, l'ordonnée du point d'inflexion vaut tout simplement 1/2. Dans ce cas, on peut interpréter le point d'inflexion d'un item i comme l'endroit où l'on passe le cap psychologique du 50 % des chances de réussir l'item i. Le point d'inflexion est aussi l'endroit où la valeur maximale de la pente de la

courbe[9] est atteinte. Nous avons vu que le paramètre a_i est proportionnel à cette pente au point d'inflexion : en fait, pour le modèle à deux paramètres, $a_i = 2{,}35\ m_i$. Dans le cas du modèle à trois paramètres, on peut montrer[10] que $a_i = m_i(4/D)/(1 - c_i)$.

4.4. LA COURBE CARACTÉRISTIQUE DE TEST ET L'ÉCHELLE DES SCORES VRAIS

Le tracé de la CCI de chacun des items d'un test permet une étude locale du comportement de ces items à l'endroit voulu de l'échelle d'habileté θ. On pourra remarquer par exemple que tel ou tel item discrimine plutôt pour les individus d'habileté modeste ou encore que tel autre item est très difficile pour tous les individus. Ceci dit, il est souvent indispensable, dans certaines circonstances, d'étudier le comportement global du test le long de l'échelle d'habileté θ. Deux outils s'offrent à nous pour ce faire : la courbe caractéristique du test, à ne pas confondre avec celle de l'item, et la fonction d'information du test. La présente section s'intéresse au premier de ces outils. C'est au cours de la prochaine section que nous traiterons de la notion d'information.

FIGURE 4.26
Courbe caractéristique d'item produite à partir du modèle logistique à trois paramètres où D = 1,7



9. Encore ici, notons qu'il s'agit bien de la pente de la droite tangente à la courbe.
10. Voir la démonstration formelle à l'annexe 4.1.

La courbe caractéristique du test (CCT) s'obtient en additionnant chacune des CCI des items du test. Il s'agit donc de faire la somme des valeurs $P_i(\theta)$ à chaque niveau d'habileté θ.

La figure 4.27, à titre d'exemple, représente la CCT d'un test formé par les quatre items de la figure 4.24. Ce qui frappe d'abord, c'est que l'échelle des ordonnées n'est pas la même que celle des CCI. En effet, ce n'est plus la probabilité $P_i(\theta)$ qui se trouve en ordonnée, mais bien la somme de ces probabilités $\Sigma\ P_i(\theta)$ et ce, à chaque valeur de θ. Cette somme est prise sur les n items du test. La valeur maximale de cette échelle n'est donc plus 1, mais bien n. La valeur minimale[11] est 0. Puisqu'il y a dix petits traits horizontaux également espacés entre 0 et n sur cette échelle, le point milieu vaut donc 5n/10 ou encore n/2. Notons également que la forme typique d'une CCT ressemble grosso modo à celle d'une CCI, bien que certaines différences existent entre les deux types de courbes : par exemple, une CCT ne possède pas, à proprement parler, de point d'inflexion.

Il n'est pas difficile de se rendre compte que la CCT de la figure 4.27 est bien la somme des quatre CCI de la figure 4.24. Si l'on regarde d'un peu plus près la CCT, on voit qu'à $\theta = -3$ la valeur de la courbe est d'environ 3n/10, soit 12/10 ou 1,2 puisque n = 4. Or, cette valeur correspond approximativement à la somme des valeurs $P_i(-3)$ pour les quatre items de la figure 4.24 : en effet, à $\theta = -3$, il y a un item dont l'ordonnée est d'environ 0,2, un autre item dont l'ordonnée vaut environ 0,3 et deux autres items dont l'ordonnée vaut environ 0,4, pour un total de 1,3, soit une approximation assez bonne de la valeur observée pour la CCT à $\theta = -3$. Un examen plus minutieux, en effectuant par exemple les calculs à l'aide de l'équation 4.5, montrerait que chacun des points de la CCT est bel et bien la somme, à chaque valeur de θ, des points des quatre CCI.

La CCT du test de mathématique de 76 items administré à un échantillon de Québécois de 13 ans se trouve à la figure 4.28. La forme de cette CCT est très près de la forme classique d'une CCI obtenue à partir d'un modèle à trois paramètres. On y voit les principales caractéristiques du test administré à cet échantillon de 1000 Québécois. La forme de cette courbe montre qu'il s'agit d'un test qui discrimine bien des élèves d'habileté moyenne (dont l'habileté se situe dans le voisinage de $\theta = 0$) : c'est en effet dans le voisinage de $\theta = 0$ que la CCT possède une pente plus abrupte, c'est-à-dire qu'elle distingue le mieux les élèves se situant dans ce voisinage. Comme il s'agit d'items à choix multiple, la hauteur minimale de cette CCT n'atteindra pas 0 mais bien une valeur égale à Σc_i.

---

11. Strictement parlant, si l'on utilise un modèle à trois paramètres, la valeur minimale de chaque $P_i(\theta)$ étant de c_i , alors la valeur minimale de $\Sigma\ P_i(\theta)$ est donc de Σc_i.

FIGURE 4.27
Courbe caractéristique de test (CCT) pour les quatre items de la figure 4.24

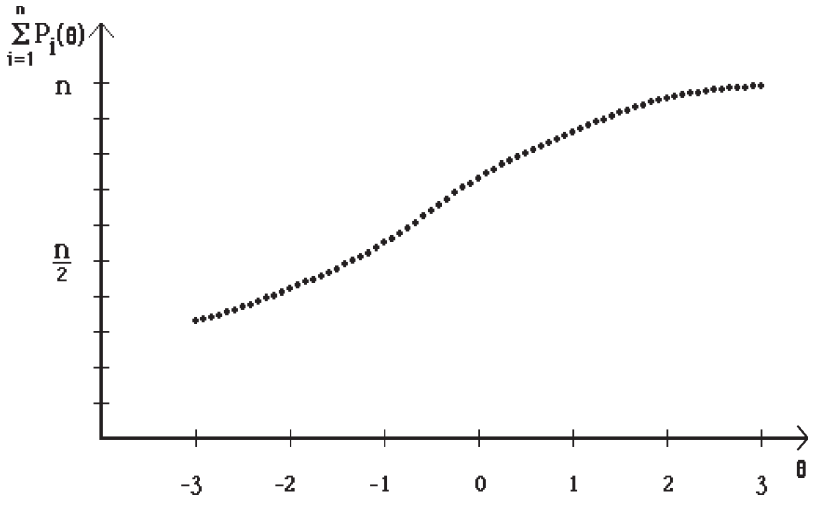
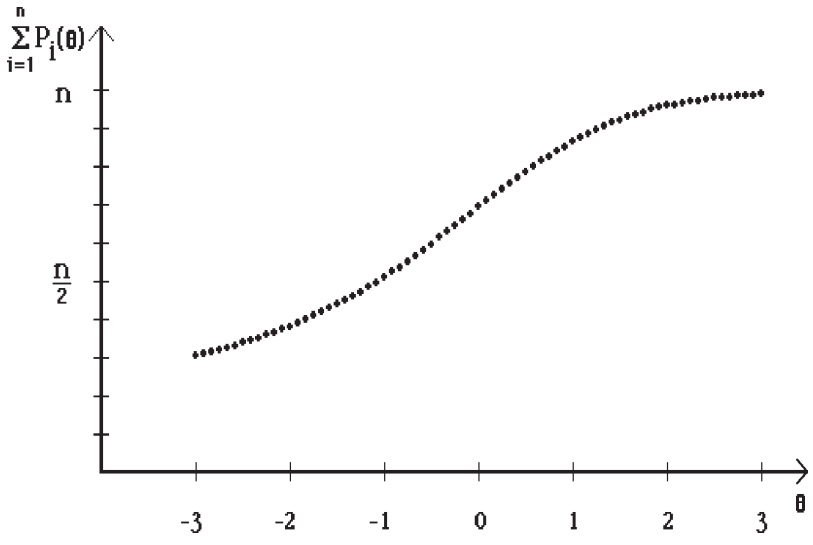


FIGURE 4.28
Courbe caractéristique de test (CCT) pour le test de 76 items de mathématique administré à un échantillon de 1000 Québécois de 13 ans



Mais l'avantage le plus important de la CCT est de permettre de changer d'échelle pour reporter les scores des individus de façon à les communiquer à des personnes du milieu, elles-mêmes non spécialistes des modèles de réponses aux items. Il n'est souvent pas très informatif d'apprendre à un enseignant, un employeur ou à un parent que Marie-Claude a obtenu un score de $\theta = -1{,}3$ à un test de mathématique ! La CCT, vue comme une transformation de l'échelle des thêtas à une échelle des $\Sigma P_i(\theta)$, permet justement de contourner ce problème de communication et de reporter les scores en utilisant l'échelle connue des pourcentages. Il s'agit de remarquer que l'échelle des $\Sigma P_i(\theta)$ est en fait l'échelle des scores vrais, c'est-à-dire que $V = \Sigma P_i(\theta)$. La preuve mathématique de cet énoncé se trouve à l'annexe 4.2. Notons, au passage, que le concept de score vrai dont il est question ici est le même que celui discuté lors de la présentation du modèle classique. Puisque les scores vrais sont situés sur la même échelle que les scores observés classiques, les $X = \sum_{i=1}^{n} U_i$, ce changement d'échelle permet, dès lors, de comparer la valeur du score classique X, soit la somme des items réussis, à la valeur du score vrai V, soit la somme des probabilités de réussir les items où les probabilités sont obtenues en tenant compte des paramètres d'items du modèle de réponses aux items.

Ce changement d'échelle que permet la CCT a des conséquences des plus intéressantes sur le plan pratique. Cela signifie en effet qu'à toute valeur de θ, variante pratique entre -3 et $+3$ (mais théoriquement entre $-\infty$ et $+\infty$) correspond une valeur de score vrai variant entre 0 et n, où n est le nombre d'items du test. Par la suite, on peut aisément transformer ce score vrai en pourcentage en divisant $\Sigma P_i(\theta)$ par n puis en multipliant par 100. Par exemple, à la figure 4.28, il est facile de se rendre compte qu'à $\theta = -1$ correspond la valeur n/2 sur l'échelle des scores vrais : c'est donc dire qu'un individu ayant une habileté de $\theta = -1$ en mathématique se verra attribuer un score vrai de 38/76, soit 50 %. Avouons qu'il est plus informatif d'apprendre que François a obtenu un score de 50 % qu'un score de -1.

Bien sûr, ce 50 % n'indique pas le pourcentage d'items réussis dans le test : cette interprétation revient au score classique. En réalité, avec 50 % de score vrai, le score classique peut être inférieur ou supérieur à 50 % suivant les caractéristiques des items réussis par François. Si celui-ci réussit des items très difficiles et discriminants, il obtiendra une valeur de θ plus élevée (donc un score vrai plus élevé) que s'il réussit le même nombre d'items faciles et tout aussi discriminants. Il est tout aussi possible que François obtienne le même score vrai de 50 % en réussissant soit 40 % d'items difficiles, soit 60 % d'items faciles. Vu de cette façon, l'écart entre le score vrai et le score observé permet de jauger l'apport de la prise en compte des caractéristiques des items (a_i, b_i, c_i) dans le calcul du score d'un individu. Mais nous n'avons pas encore une interprétation satisfaisante du score vrai, de ce 50 %. Il serait en effet intéressant de connaître 50 % de quoi François maîtrise vraiment !

Voici une proposition d'interprétation de ce score vrai, de ce 50 %. Pour un item i et un individu d'habileté θ, $P_i(\theta)$ peut être vu comme le pourcentage attendu d'items réussis parmi tous les items qui ont la même CCI que l'item i. Ainsi le score vrai V, la somme de ces $P_i(\theta)$, pourrait légitimement être interprété lui aussi comme un pourcentage attendu d'items réussis dans l'univers des items qui ont des CCI analogues aux items impliqués dans le calcul de la CCT. Pour compléter cette interprétation du score vrai, mentionnons que Hambleton *et al.* (1985, p. 67) parlent plutôt du score vrai comme du pourcentage de contenu maîtrisé. Ainsi, selon ces auteurs, il serait légitime de dire que François maîtrise 50 % du contenu du domaine visé par le test. Disons tout simplement qu'une telle interprétation nécessite un effort important d'analyse et de délimitation du contenu visé, du contexte et du format des items. La plupart du temps, une interprétation de ce type demeure une hypothèse.

4.5. LE CONCEPT D'INFORMATION

Si la courbe caractéristique du test est intéressante dans la mesure où elle permet de reporter les scores des individus sur l'échelle familière des pourcentages, ce sont véritablement les fonctions d'information d'item et de test qui permettront à la théorie des réponses aux items de s'ouvrir sur un potentiel d'applications somme toute inépuisable.

4.5.1. Information et erreur-type

Mais, avant tout, il faut spécifier ce que l'on entend par le concept d'**information**. Reportons-nous en l'an 2050. L'énoncé typique propre à un sondage d'opinion de cette époque pourrait s'exprimer de la façon suivante :

> « 45 % des électeurs ont l'intention de voter pour le parti des Verts, 19 fois sur 20, avec une marge d'erreur de 3 % ».

Cet énoncé signifie que l'intervalle de confiance à 95 % (19 fois sur 20 !) autour de la valeur observée de 45 % est [42 %, 48 %]. Autrement dit, à l'aide des résultats de ce sondage, nous sommes certains à 95 % que le pourcentage vrai d'électeurs qui voteraient pour le parti des Verts se situe entre 42 % et 48 %.

Supposons maintenant que nous ayions plutôt entendu un énoncé comme celui-ci :

> « 45 % des électeurs ont l'intention de voter pour le parti des Verts, 19 fois sur 20, avec une marge d'erreur de 5 % ».

Quelle différence fondamentale existe-t-il entre ces deux énoncés ? Le pourcentage observé est le même : 45 %. Dans les deux cas, nous aurons un intervalle de confiance à 95 %. Seule la marge d'erreur diffère : elle passe de 3 % à 5 %. Ainsi, l'intervalle de confiance à 95 % n'est plus [42 %, 48 %]

mais bien [40 %, 50 %]. L'impact net de cet accroissement de la marge d'erreur est donc d'augmenter l'empan de l'intervalle de confiance à 95 % et donc l'incertitude quant au pourcentage vrai d'électeurs qui voteraient pour le parti des Verts. Si la marge d'erreur augmentait encore à 10 %, l'incertitude s'accroîtrait aussi puisque l'intervalle de confiance serait alors [35 %, 55 %].

Ainsi, plus la marge d'erreur augmente, plus l'incertitude s'accroît. Mais il existe aussi une autre façon d'exprimer cette relation entre la marge d'erreur et l'incertitude face au pourcentage vrai que nous formulons comme ceci : plus la marge d'erreur diminue, plus l'information concernant le pourcentage vrai augmente. En effet, plus la marge d'erreur diminue, plus l'intervalle de confiance diminue et plus le pourcentage vrai est en quelque sorte cerné par les bornes de l'intervalle : c'est en ce sens que nous dirons avoir plus d'information sur le pourcentage vrai.

Considérons maintenant une situation propre à la mesure en éducation. Zoé a obtenu un score de 64 à l'examen de mathématique et l'erreur-type de mesure est de 5. L'intervalle de confiance à 95 % concernant le score vrai est donc $[64 - 1{,}96 \times 5,\ 64 + 1{,}96 \times 5] = [54{,}2,\ 73{,}8]$ que nous pouvons interpréter comme une certitude à 95 % que le score vrai de Zoé se trouve entre 54,2 et 73,8. Si l'erreur-type de mesure avait été de 2 alors l'intervalle de confiance à 95 % aurait été de [60,08, 67,92]. Pour une erreur-type plus petite que 2, l'intervalle de confiance aurait été encore plus petit. Ainsi, plus l'erreur-type de mesure diminue, plus l'intervalle de confiance diminue, plus le score vrai est cerné et plus nous avons donc de l'information concernant le score vrai (l'habileté) de Zoé en mathématique.

Le concept d'information défini plus haut n'est pas interprété de façon substantiellement différente dans le contexte de la théorie des réponses aux items. Avant de le définir de façon formelle, signalons que l'estimateur de l'habileté θ d'un individu sera noté $\hat{\theta}$.Cet estimateur s'obtient selon un procédé connu sous le nom de maximum de vraisemblance, procédé qui sera développé principalement dans le cadre du chapitre 6. La distribution de l'estimateur est supposée normale asymptotiquement[12] avec moyenne θ et écart-type $\sigma(\hat{\theta}|\theta)$. Cet écart-type $\sigma(\hat{\theta}|\theta)$ est appelé l'erreur-type de mesure associée à $\hat{\theta}$. Ceci dit, l'information sera définie comme l'inverse du carré de l'erreur-type de mesure. L'information concernant l'habileté θ sera donc définie comme

$$I(\theta) = \frac{1}{\sigma^2\left(\hat{\theta}|\theta\right)}$$

12. C'est-à-dire si le nombre d'items est suffisamment élevé.

Ainsi définie, l'information mais aussi l'erreur-type de mesure variera d'un niveau d'habileté θ à un autre, contrairement à ce qui se passe en théorie classique, où l'erreur-type de mesure est la même pour tous les individus d'un groupe, les forts comme les faibles[13]. La donnée d'une valeur d'information à chaque niveau d'habileté nous permet dès lors de construire un intervalle de confiance pour θ. Selon Hulin *et al.* (1983, p. 58), si le nombre d'items est suffisamment grand, le **plus petit intervalle de confiance** à 95 % pour θ est donné par $\left(\hat{\theta} - 1{,}96/\sqrt{I(\theta)}\,,\ \hat{\theta} + 1{,}96/\sqrt{I(\theta)}\right)$ où, ici, $\hat{\theta}$ indique une valeur particulière de l'estimateur de θ obtenue selon la méthode du maximum de vraisemblance ; c'est cet intervalle de confiance qui cerne le mieux l'habileté θ. Ainsi, plus l'information I(θ) sera élevée, plus on connaîtra avec précision l'habileté θ. Voilà un résultat d'une importance capitale.

Notons, au passage, que la notion d'erreur-type de mesure n'est pas définie de façon substantiellement différente en théorie classique et en TRI. À la section 2.3, nous avions constaté que l'erreur-type de mesure σ_{Ej} propre à un individu j consistait en l'écart-type de la distribution des scores observés X_{ij} autour de V_j, le score vrai de cet individu, où l'indice i indique les différentes répétitions de la mesure X à l'individu j. En théorie des réponses aux items, comme nous venons de le voir, l'erreur-type de mesure, $\sigma(\hat{\theta}|\theta)$, est l'écart-type de la distribution des valeurs de l'estimateur $\hat{\theta}$ obtenues par le principe du maximum de vraisemblance. Or, même si on peut constater une similarité entre ces deux définitions, donc une ressemblance théorique entre ces deux concepts, il n'en va pas de même quand vient le temps de trouver une statistique qui estime l'erreur-type de mesure. Dans le cas de la théorie des réponses aux items, en effet, nous verrons qu'il est possible d'estimer des valeurs d'information I(θ), donc aussi des valeurs d'erreur-type de mesure pour chaque θ. Il n'en est pas de même en théorie classique, puisque, comme on l'a déjà souligné, il est plutôt habituel d'estimer l'erreur-type de mesure propre à un individu, σ_{Ej}, en se rabattant sur l'erreur-type de mesure propre au groupe d'individus, σ_E, celle-ci étant interprétée, en quelque sorte, comme la moyenne des σ_{Ej}. Et comme si cela n'était pas assez, le concept d'erreur-type de mesure de groupe ne pourra être estimé qu'en passant par le concept polysémique de fidélité ! C'est donc le concept d'information qui, en TRI, prend plus ou moins la place du concept de fidélité en théorie classique et, par voie de conséquence, du concept de généralisabilité en théorie de la généralisabilité. Il serait donc légitime de dire qu'un test est informatif comme on dit qu'un test est fidèle. La différence capitale vient du fait que, contrairement à la théorie classique, en TRI on sait où, sur l'échelle d'habileté, un test est informatif (précis) : il

13. Sauf, bien sûr, si on décidait d'employer la méthode (non sans failles) de Woodruff présentée au chapitre 2.

n'y a qu'à examiner, comme nous allons le montrer à la section suivante, la courbe d'information ou, ce qui revient au même, la courbe d'erreur-type de mesure propre au test.

4.5.2. Fonctions d'information d'item et de test

Mais revenons plus spécialement au concept d'information lui-même. Lord (1980, chap. 5) a montré que la fonction d'information d'un item i à un niveau fixé θ de l'échelle d'habileté était fonction de la pente $P_i'(\theta)$ de la CCI à θ, soit

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)} \tag{4.6}$$

alors que la fonction d'information d'un test de n items est donnée par la somme des valeurs d'information des items, en se limitant toujours à un niveau fixé d'habileté θ.

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) = \sum_{i=1}^{n} \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)} \tag{4.7}$$

où $Q_i(\theta) = 1 - P_i(\theta)$ est la probabilité d'échouer l'item i pour un individu d'habileté θ et $P_i(\theta)Q_i(\theta)$ est la variance de l'item i à θ.

Puisque les valeurs de la partie de droite de l'équation 4.7 sont toutes positives, plus le nombre d'items sera élevé, plus l'information sera élevée et plus la précision relative à l'habileté θ sera donc élevée.

Si l'on combine maintenant l'équation 4.5 de la section 4.3, c'est-à-dire la formulation du modèle à trois paramètres, et l'équation 4.6 ci-dessus, nous pouvons enrichir et concrétiser un peu plus l'interprétation des fonctions d'information et obtenir (voir la démonstration à l'annexe 4.3) :

$$I_i(\theta) = D^2 a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{(1 - c_i)} \right]^2 \tag{4.8}$$

Cette dernière façon d'exprimer la fonction d'information d'un item permet d'identifier les principaux déterminants de la précision associée à un niveau d'habileté θ. Ainsi, la valeur d'information sera d'autant plus élevée que :

- ♦ les valeurs de l'indice de discrimination des items a_i seront élevées ;
- ♦ le nombre d'items sera élevé ;
- ♦ les valeurs de l'indice de pseudo-chance des items c_i seront faibles[14].

Comme la pente d'une CCI est maximale au point d'inflexion et que ce point a comme coordonnées $[b_i , (1+ c_i)/2]$, il ne sera pas étonnant d'apprendre que l'information maximale d'un item sera obtenue dans le voisinage de b_i. En réalité, pour les modèles à un et à deux paramètres, l'information maximale est obtenue exactement à b_i, alors que pour le modèle à trois paramètres, l'information maximale est obtenue à une valeur légèrement supérieure à b_i, soit au point $\theta_{max} = b_i + \frac{1}{D\,a_i} Ln\left(\frac{1+\sqrt{1+8\,c_i}}{2}\right)$ où Ln désigne le logarithme népérien. Ainsi, il faut tenir compte des trois paramètres a_i, b_i et c_i afin d'obtenir le point de l'échelle d'habileté θ où l'information est maximale.

Chaque fonction d'information (équation 4.6 ou équation 4.7) peut être représentée par une courbe, que nous nommerons respectivement la courbe d'information d'un item i et la courbe d'information du test. La forme de ces courbes ne suit pas du tout la forme des courbes caractéristiques d'item examinées précédemment. En effet, comme on peut s'en rendre compte à la figure 4.29, la courbe d'information d'un item i est non monotone, le sommet de la courbe indiquant l'information maximale obtenue à un endroit donné de l'échelle.

On peut montrer (Lord, 1980, p. 152) que, dans le cas du modèle à trois paramètres, l'information maximale relative à l'item i est donnée par[15]

$$I_i(\theta)_{max} = \frac{D^2 a_i^2}{8(1-c_i)^2}\left[1 - 20c_i - 8c_i^2 + (1+8c_i)^{\frac{3}{2}}\right] \qquad (4.9)$$

Cette formule peut paraître un peu rébarbative, mais dans le cas des modèles à un ou à deux paramètres, où $c_i = 0$, la formulation de l'information maximale se réduit à une expression beaucoup plus simple, soit $I_i(\theta)_{max} = \frac{D^2 a_i^2}{4}$ pour le modèle à deux paramètres et $I_i(\theta)_{max} = \frac{D^2}{4}$ pour le modèle à un paramètre.

---

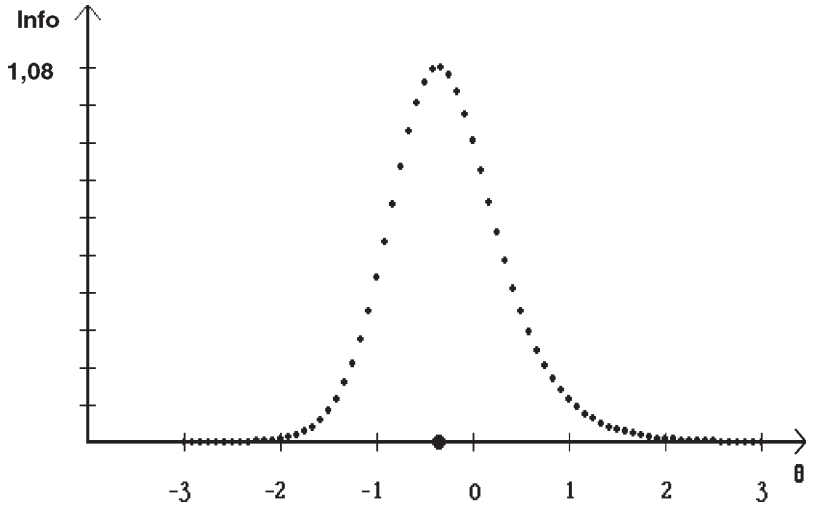
14. Il est facile de montrer que, pour une valeur fixe de $P_i(\theta)$, disons 0,5, plus la valeur de c_i est faible, plus la valeur du terme $(P_i(\theta) - c_i) / (1 - c_i)$ est élevée. Par exemple, si $c_i = 0{,}25$, ce terme égalera 1/3 ; si $c_i = 0{,}5$, ce terme vaudra 0.
15. Cette formule n'est pas sans rappeler l'équation 4.8. En effet, pour le modèle à deux paramètres par exemple, l'équation 4.8 devient D^2a^2PQ. Or cette valeur est maximale lorsque la variance PQ est maximale, soit lorsque P = 1/2. Dans ce cas, l'équation 4.8 devient $D^2a^2/4$, soit l'équation 4.9 pour le modèle à deux paramètres.

L'expression de l'information maximale d'un item montre que :

- dans le cas du modèle à trois paramètres, l'information maximale d'un item i dépend de la valeur de a_i et de la valeur de c_i ;
- dans le cas du modèle à deux paramètres, l'information maximale d'un item i ne dépend que de la valeur de a_i ;
- dans le cas du modèle à un paramètre, où $c_i = 0$ et a_i est constant, l'information maximale est constante ;
- dans le cas du modèle de Rasch, où $c_i = 0$, $D = 1$ et $a_i = 1$, l'information maximale égale ¼ pour chaque item.

Regardons encore la figure 4.29 : il s'agit de la courbe d'information de l'item 2 d'une enquête internationale dont les estimés des paramètres sont $a_i = 1{,}47$, $b_i = -0{,}46$ et $c_i = 0{,}19$. Comme il s'agit du modèle à trois paramètres, l'équation 4.9 nous indique que l'information maximale est égale à 1,08 au point $\theta_{max} = -0{,}36$.

FIGURE 4.29
Courbe d'information de l'item 2 d'une enquête internationale
($\theta_{MAX} = -0{,}36$; $I_i(\theta)_{MAX} = 1{,}08$)



Chacun des items d'un test génère une courbe d'information spécifique dont les particularités dépendent des valeurs des paramètres de l'item, un peu comme c'est le cas pour une courbe caractéristique d'item. Le tableau 4.2 présente les valeurs d'information maximale ($I_i(\theta)_{MAX}$) ainsi que le point (θ_{MAX}) de l'échelle θ où ce maximum d'information est atteint et ce, pour les 19 premiers items du test de 76 items de mathématique d'une enquête internationale (Lapointe *et al.*, 1992) déjà décrite plus haut. Notons

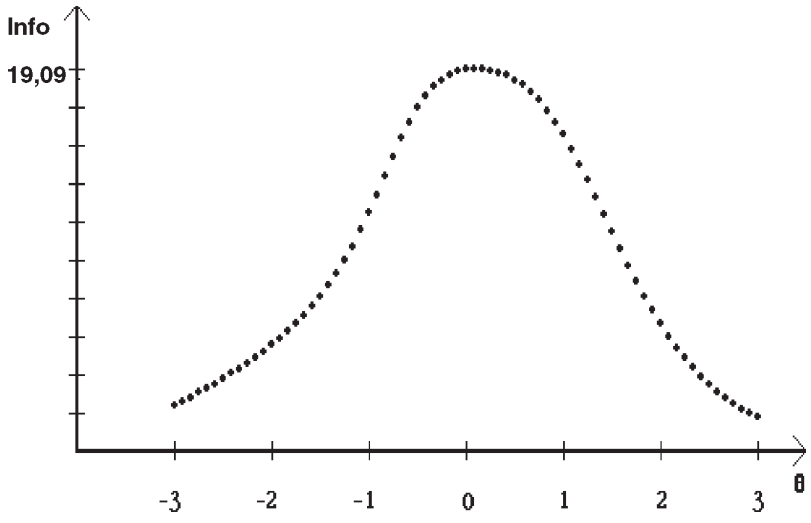
tout particulièrement les fluctuations importantes des valeurs d'information maximale des items en fonction des valeurs des paramètres a_i, b_i et c_i. Par exemple, on peut constater que l'information maximale attribuable à l'item 1, $I_1(\theta)_{MAX} = 0{,}11$, est près de 10 fois moins grande que l'information attribuable à l'item 2, $I_2(\theta)_{MAX} = 1{,}08$. Cet écart entre les valeurs de l'information maximale est principalement dû, dans ce cas, au fait que les valeurs de l'indice de discrimination sont fort différentes : en effet, $a_1 = 0{,}47$ alors que $a_2 = 1{,}47$. D'un autre côté, même si les valeurs des indices de discrimination des items 10 et 19 sont du même ordre ($a_{10} = 0{,}94$ alors que $a_{19} = 0{,}92$), les valeurs de l'indice de pseudo-chance sont suffisamment distinctes ($c_{10} = 0{,}35$ alors que $c_{19} = 0{,}13$), avec comme résultat que la valeur de l'information maximale de l'item 19 est de 50 % supérieure à la valeur de l'information maximale de l'item 10. Ainsi, tel que nous l'avons déjà indiqué avant, les valeurs de l'information maximale seront d'autant plus élevées que les valeurs de l'indice de discrimination seront élevées et les valeurs de l'indice de pseudo-chance seront faibles. Notons également que les précédentes remarques ne prévalent qu'en un seul point de l'échelle θ : le point que nous avons noté ici θ_{MAX}. Afin de connaître la valeur de la fonction d'information aux autres points de l'échelle, il faut soit les calculer grâce aux équations données plus haut, soit produire la courbe d'information et lire le résultat sur le graphique.

TABLEAU 4.2
Valeurs des maximums d'information ($I_i(\theta)_{MAX}$) et des points où ce maximum d'information (θ_{MAX}) est atteint pour chacun des 19 premiers items du test de 76 items de mathématique d'une enquête internationale (Lapointe *et al.*, 1992)

Nº de l'item	a_i	b_i	c_i	$I_i(\theta)_{MAX}$	θ_{MAX}
1	0,47	–1,36	0,21	0,11	–1,01
2	1,47	–0,46	0,19	1,08	–0,36
3	0,73	–2,12	0,23	0,25	–1,88
4	0,42	–1,38	0,28	0,07	–0,91
5	0,51	–1,15	0,26	0,11	–0,78
6	0,58	–1,24	0,28	0,14	–0,90
7	1,03	0,59	0,24	0,48	0,76
8	1,19	1,25	0,32	0,55	1,43
9	0,85	–0,6	0,22	0,34	–0,40
10	0,94	–0,29	0,35	0,32	–0,05
11	0,89	–0,63	0,15	0,43	–0,49
12	0,72	–1,13	0,15	0,28	–0,95
13	0,67	–0,95	0,18	0,23	–0,73
14	0,80	0,55	0,09	0,39	0,66
15	1,82	–0,55	0,16	1,76	–0,48
16	0,87	–0,60	0,22	0,36	–0,41
17	1,09	–0,04	0,28	0,50	0,14
18	1,16	0	0,15	0,73	0,11
19	0,92	2,05	0,13	0,48	2,17

La figure 4.30 présente la courbe d'information du test formé des 76 items. Suivant l'équation 4.7, cette courbe n'est pas autre chose que la somme des courbes d'information des 76 items, donc des courbes semblables à celles présentées à la figure 4.29. Ainsi, pour obtenir la courbe d'information du test, il faut additionner, à chaque point de l'échelle θ, les valeurs $I_i(\theta)$ où i varie de 1 à 76. La valeur maximale de l'information du test, soit ici 19,09, dépend de la combinaison des trois paramètres propres à chacun des items : comme nous l'avons déjà souligné, la valeur de l'information associée à un item i sera d'autant plus élevée que la valeur de a_i sera élevée et la valeur de c_i sera faible, l'information maximale étant obtenue, on s'en souvient, dans le voisinage de b_i. La forme de cette courbe est très typique des courbes d'information de plusieurs tests : plus ou moins symétrique, le mode se trouvant au centre de l'échelle, soit dans le voisinage de θ = 0. Ces caractéristiques montrent que le test de 76 items peut être considéré de difficulté moyenne par les 1000 élèves québécois qui l'ont passé. Comme on peut le constater en examinant le tableau 4.2, plusieurs des items de ce test sont de difficulté moyenne : autrement dit, plusieurs des valeurs de l'indice de difficulté b_i se trouvent entre –1 et +1, c'est-à-dire que, pour plusieurs items, le point d'inflexion, endroit où la pente est maximale, se trouvera entre –1 et +1. Puisque le maximum d'information d'un item est obtenu à un point de l'échelle θ proche de l'indice de difficulté b_i, il n'est pas étonnant d'observer que le maximum d'information de ce test se situe aussi entre –1 et +1.

FIGURE 4.30
Courbe d'information du test de 76 items d'une enquête internationale



Il faut cependant garder à l'esprit que les courbes d'information de test n'ont pas toutes la même forme « normale ». Examinons plutôt la figure 4.31 qui présente la courbe d'information d'un test de 4 items dont voici, au tableau 4.3, les estimés de paramètres et les valeurs d'information. Notons tout d'abord que la valeur de l'information maximale (0,99) est bien inférieure à celle que nous avions observée pour le test de 76 items (19,09). C'est naturel : la courbe de la figure 4.31 est la somme de 4 courbes d'information d'items alors que la courbe de la figure 4.30 était bâtie à partir d'un test de 76 items. Notons également et surtout que la forme de la courbe d'information de la figure 4.31 n'a plus rien à voir avec celle (d'allure gaussienne) de la courbe d'information de la figure 4.30. Nous avons maintenant deux modes, signifiant que le test comporte un maximum d'information autour de $\theta = -1$, puis un autre autour de $\theta = 1$. Ce n'est pas si étonnant puisque, selon le tableau 4.3, deux items, soit les items 1 et 4, donnent un maximum d'information autour de $\theta = -1$ et qu'un autre item, l'item 2, possède l'information maximale autour de $\theta = 1$. Observons aussi que l'item 2 informe beaucoup plus que chacun des items 1 et 4 : en effet, la valeur de son indice de discrimination est très élevée, soit $a_2 = 1{,}4$. Quant à l'item 3, son maximum d'information se situe dans le voisinage de $\theta = 0$: on peut donc croire qu'il contribue autant à l'un ou l'autre des deux modes.

FIGURE 4.31
Courbe d'information d'un test de 4 items

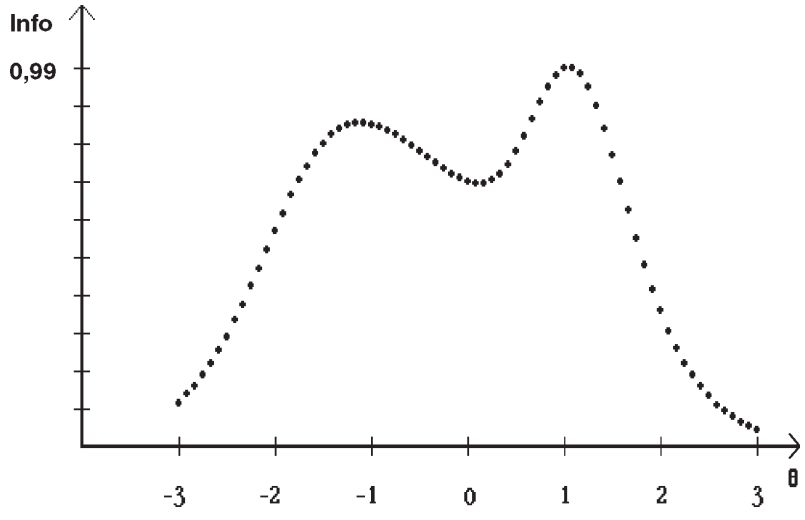


TABLEAU 4.3
Valeurs des maximum d'information ($I_i(\theta)_{MAX}$) et des points de maximum d'information (θ_{MAX}) pour chacun des items d'un test de 4 items

Nº de l'item	a_i	b_i	c_i	$I_i(\theta)_{MAX}$	θ_{MAX}
1	0,71	–1,7	0,29	0,21	–1,42
2	1,4	1	0,31	0,78	1,15
3	0,94	–0,28	0,29	0,37	–0,06
4	1	–1,5	0,16	0,53	–1,37

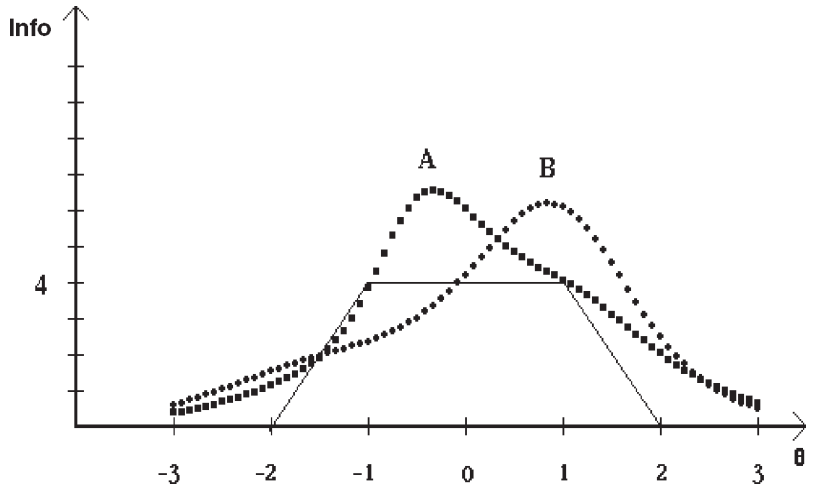
Comme nous le verrons au cours de la partie 2 de cet ouvrage, plusieurs des applications de la théorie des réponses aux items prennent appui, d'une façon ou d'une autre, sur le concept d'information. En plus d'être indispensable lors la mise sur pied d'un test adaptatif, application à laquelle nous consacrerons plus loin un chapitre complet, l'information, comme nous allons le présenter maintenant, est un concept clé pour des applications comme la construction de tests à référence normative, la construction de tests à référence critériée et le calcul de l'efficacité relative de deux tests.

Hulin *et al.* (1983, p. 90) discutent de l'utilisation de la fonction d'information de test dans la construction d'un test de compétences verbales. Voici la procédure qu'on pourrait suivre pour la construction d'un test de ce type, par analogie à la méthode présentée chez ces auteurs et en utilisant les données de notre banque de 76 items. Il faut d'abord tenir compte de l'objectif poursuivi par le test. Disons ici que nous voulons nous en tenir à la contrainte suivante : un test de 20 items qui mesure les élèves d'habileté moyenne avec un maximum de précision. Afin de tenir compte de cette contrainte, nous avons construit la courbe d'information-cible de la figure 4.32, formée de trois traits : un trait vertical à la hauteur de $I(\theta) = 4$ dans l'intervalle d'habileté moyenne [–1, 1] (ce qui équivaut à une erreur-type de mesure inférieure[16] ou égale à 0,5 entre $\theta = -1$ et $\theta = 1$), et deux traits diagonaux reliant respectivement les points –2 et +2 sur l'axe θ à partir de ce trait vertical. La forme finale de cette courbe d'information-cible rappelle celle d'un trapèze. Nous voulons choisir un test de 20 items dont la courbe d'information enveloppe, en quelque sorte, cette courbe-cible, c'est-à-dire une courbe dont l'information sera supérieure ou égale à la courbe-cible à chaque valeur de l'échelle θ. La figure 4.32 montre que le test A, formé des items 1 à 19 et de l'item 76, respecterait raisonnablement bien la contrainte que nous nous sommes donnée, alors que le test B, formé des items 57 à 76, ne respecterait pas cette contrainte.

16. Rappelons que, suivant la formulation de l'erreur-type en fonction de l'information, demander une erreur-type de mesure inférieure ou égale à 0,5 équivaut à demander une valeur d'information supérieure ou égale à 4, puisque $4 = 1 / (0,5)^2$. C'est ce que traduit la courbe-cible de la figure 4.32.

Par ailleurs, Hambleton *et al.* (1985, p. 257) présentent une façon de sélectionner des items autour d'un seuil de réussite θ_S lors de la construction d'un test à référence critériée. Plusieurs procédures peuvent servir à construire un tel test. Sans entrer dans les détails, indiquons simplement que le seuil de réussite, généralement déterminé sur l'échelle des pourcentages, peut facilement être transformé sur l'échelle θ en utilisant la courbe caractéristique de test. Comme nous voulons maximiser la précision autour du seuil θ_S, il s'agit alors de choisir des items qui informent le plus dans le voisinage de ce seuil en produisant un tableau contenant les valeurs d'information de chaque item au point θ_S ou encore en consultant les courbes d'information des items.

FIGURE 4.32
Courbe d'information-cible (trait plein) et deux courbes d'information empiriques pour les tests A et B : seule la courbe du test A enveloppe correctement la courbe-cible.



4.5.3. Efficacité relative

Hambleton *et al.* (1991, p. 95) traitent du concept d'efficacité relative entre deux tests mesurant le même concept. Il s'agit, en réalité, de comparer l'information des deux tests à chaque niveau d'habileté θ. Supposons, par exemple, deux tests de 19 items, nommés X et Y à la figure 4.33, où le test X informe surtout autour de θ = 0 alors que le test Y informe surtout autour de θ = 1. L'efficacité relative de X par rapport à Y est définie comme le rapport, à chaque valeur de θ, de l'information fournie par le test X et de l'information fournie par le test Y. C'est-à-dire

$$ER\ (X/Y) = I_X(\theta) / I_Y(\theta)$$

FIGURE 4.33
Courbes d'information de deux tests de 19 items chacun : le test X informe surtout autour de θ = 0 alors que le test Y informe dans le voisinage de θ = 1.

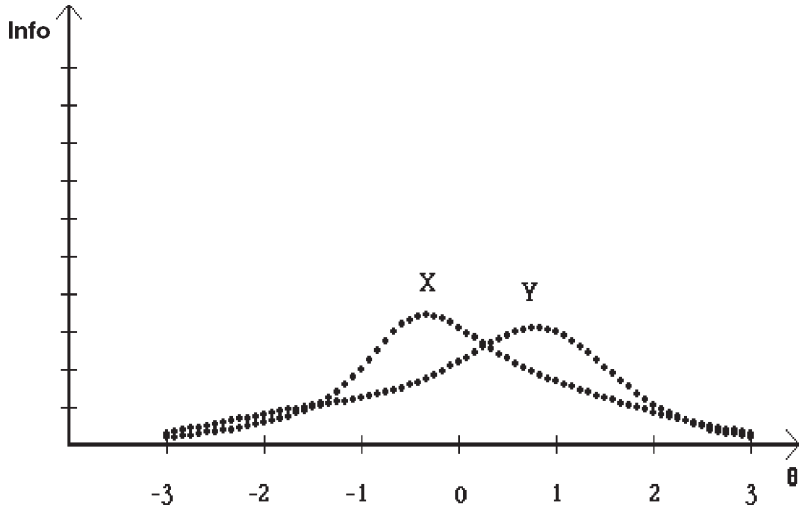
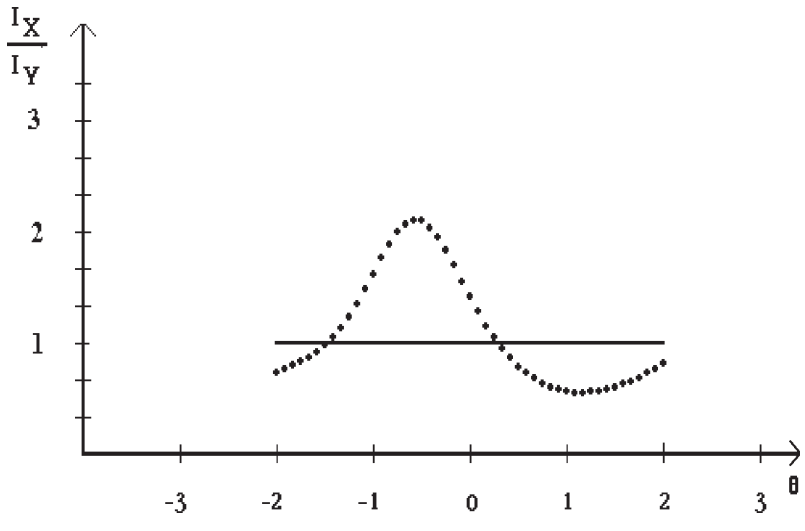


FIGURE 4.34
Efficacité relative de deux tests de 19 items chacun : le test X informe surtout autour de θ = 0 alors que le test Y informe dans le voisinage de θ = 1.



La courbe d'efficacité relative, donnée à la figure 4.34, montre que le test X donne plus d'information (est donc plus précis) dans l'intervalle d'habileté [–1,5, 0] parce que $I_X / I_Y > 1$ dans cet intervalle. Cependant, c'est le test Y qui donne plus d'information dans l'intervalle [0, 2]. Le test X serait plus approprié pour mesurer des individus d'habileté faible à moyenne, comme dans le cas d'un examen scolaire par exemple, alors que le test Y serait plus utile s'il devenait nécessaire de sélectionner les individus les plus habiles. Notons également que l'on devrait sélectionner le test Y à nouveau si l'on voulait mesurer les individus les plus faibles ($\theta < -1,5$) avec un maximum de précision.

4.6. AUTRES MODÈLES

Les modèles de réponses aux items parcourus jusqu'ici présentent tous le même profil : ils visent l'analyse de tests unidimensionnels et n'utilisent que l'information dichotomique obtenue pour chaque item : 1 pour une bonne réponse, 0 pour une mauvaise réponse. De plus, tous les modèles que nous avons présentés sont dits paramétriques dans le sens où il faut estimer les paramètres des items avant d'obtenir les courbes caractéristiques des items et les estimés d'habileté des sujets. Il ne fait pas de doute que certains résultats à des tests ne peuvent être réduits à une seule dimension et qu'il y aurait tout avantage à exploiter la nature multidimensionnelle des données dans ces situations. De même, il serait parfois approprié d'utiliser plus que l'information dichotomique des données, notamment lorsqu'il s'agit de tests à choix multiple ou d'échelles de type Likert. Enfin, les modèles paramétriques peuvent exiger des échantillons de sujets de taille très imposante, de l'ordre de 1000 sujets pour le modèle à trois paramètres par exemple, afin d'obtenir des estimés relativement stables des paramètres d'items : dans le cas où de tels échantillons ne sont pas disponibles, pourquoi alors ne pas avoir recours aux modèles non paramétriques qui permettent l'utilisation d'échantillons de taille beaucoup plus modeste ?

Les prochains paragraphes mettent en évidence des modèles qui n'ont pas encore retenu notre attention jusqu'ici. Plutôt que de s'en tenir à l'information dichotomique pour estimer les paramètres d'items et de sujets, les modèles polytomiques permettront en effet d'exploiter l'information nominale présente dans les différentes options de réponses des items à choix multiple ou encore l'information ordinale des échelles graduées associées aux items de type Likert. Plutôt que de supposer une seule dimension mesurant l'habileté des sujets, les modèles multidimensionnels tireront profit des multiples habiletés requises pour répondre à un item : il sera alors possible d'estimer non seulement les paramètres de l'item, mais aussi les vecteurs de paramètres d'habileté des sujets. Alors que les modèles paramétriques exigeaient des tailles d'échantillon très imposantes pour estimer les paramètres des items, les modèles

non paramétriques contourneront ce problème en estimant directement la courbe des items (d'ailleurs pas nécessairement monotone croissante) exprimant la relation entre l'habileté du sujet et la probabilité qu'il réussisse l'item.

4.6.1. Les modèles polytomiques

Modèle nominal de Bock (1972)

Les modèles logistiques à un, deux ou trois paramètres dont il a été question au cours de ce chapitre comportent une restriction majeure, à savoir l'utilisation exclusive d'informations du type bonne ou mauvaise réponse. Ainsi, l'estimation des paramètres d'item ou d'habileté n'est basée que sur des matrices de 1 (bonne réponse) et de 0 (mauvaise réponse). Or, il existe des formats d'items qui permettraient une exploitation différente des réponses exprimées par les répondants. C'est le cas notamment des items à choix multiple dont les différentes options constituent une échelle nominale. Prenons, par exemple, l'item 7 d'une enquête internationale (Lapointe, Mead et Askew, 1992) destinée aux élèves de 13 ans :

Un groupe d'élèves a 29 crayons en tout. Six élèves ont 1 crayon chacun, 5 élèves ont 3 crayons chacun, et les élèves qui restent ont 2 crayons chacun. Combien d'élèves ont seulement 2 crayons ?

A.	4
B.	6
C.	8
D.	9

Un modèle dichotomique comme le modèle logistique à trois paramètres n'exploite que l'information distinguant les élèves qui ont choisi la bonne réponse, l'option A, des élèves qui ont choisi l'un ou l'autre des trois leurres, à savoir les options B, C et D. Toutes choses étant égales par ailleurs, les élèves qui ont choisi la bonne réponse sont alors considérés plus habiles que les autres élèves qui, eux, sont considérés du même niveau d'habileté quel que soit le leurre choisi (si on ne se fie bien sûr qu'à cet item). Or, il est facile de se rendre compte que l'option C semble beaucoup plus près de la bonne réponse que l'option D par exemple. En conséquence, les élèves qui choisissent l'option C devraient, de façon générale, posséder des capacités mathématiques supérieures aux élèves qui choisissent l'option D.

Il existe des modèles, dits polytomiques, qui vont tenir compte du choix de réponse réellement fait par un élève pour attribuer un niveau d'habileté à cet élève. Ainsi, en utilisant ce genre de modèle pour analyser les scores et à partir du même item, on attribuerait un score d'habileté θ plus élevé à un élève qui aurait choisi l'option C qu'à un élève qui aurait choisi l'option D, l'idée étant d'attribuer un score d'habileté équivalent à tous les élèves qui ont choisi la même option de réponse.

Parmi les modèles polytomiques qui ont été développés par le passé pour analyser les items dont les choix de réponses sont placés sur une échelle nominale[17], comme les items à choix multiple par exemple, nous nous attarderons au modèle nominal de Bock (1972).

Le modèle nominal qui permet l'analyse des items à choix multiple est donné par :

$$P_{ix}(\theta) = \frac{e^{(a_{ik}\theta + c_{ik})}}{\sum_{k=1}^{m_i} e^{(a_{ik}\theta + c_{ik})}} \text{ pour } x = 1, 2, \ldots, m_i ,$$

où $P_{ix}(\theta)$ est la probabilité, pour un sujet d'habileté θ, d'endosser l'option x de l'item i et où les paramètres a_{ix} et c_{ix} permettent de caractériser l'allure de la courbe de l'option x de l'item i. Ainsi, chaque option de réponse d'un item à choix multiple pourra compter sur une courbe caractéristique d'option (CCO). En jetant un coup d'œil à la figure 4.35 qui correspond à l'item 7 de l'enquête internationale, il appert que

$$\sum_{x=1}^{m_i} P_{ix}(\theta) = 1$$

En effet, pour chaque valeur de thêta, la somme des courbes caractéristiques d'option est égale à 1 du fait que le sujet doit nécessairement endosser l'une ou l'autre des options[18]. De plus puisque, selon le tableau 4.4, la valeur de a_{ix} est relativement élevée et positive pour l'option A, la courbe caractéristique de cette option est monotone croissante. La valeur de a_{ix} pour l'option C est près de 0, reflétant le fait que la CCO sera monotone croissante pour un intervalle donné de l'échelle d'habileté et monotone décroissante pour un autre intervalle. Les valeurs négatives de ce paramètre pour les options B et D s'expliquent du fait que les courbes caractéristiques correspondantes sont monotones décroissantes. Alors que le paramètre a_{ix} peut être considéré analogue à la pente ou au pouvoir discriminant de la CCO, l'interprétation à donner au paramètre c_{ix} est beaucoup moins claire.

Si nous revenons au libellé de cet item, il est normal d'observer que la courbe de l'option A est monotone croissante puisque c'est la bonne réponse : en ce sens, elle se comporte comme la courbe caractéristique d'item d'un modèle

17. Thissen et Steinberg (1984) ont aussi proposé un modèle, dit à choix multiple, qui généralise le modèle nominal de Bock mais qui, par sa complexité, ne peut pas vraiment être utilisé à grande échelle.
18. Une valeur manquante peut aussi être considérée comme une option.

logistique connu à un, deux ou trois paramètres. Ainsi, plus un élève est habile en mathématique, plus il aura tendance à choisir l'option A. La courbe caractéristique de l'option C, par contre, ne se comporte pas du tout comme la CCI d'un modèle logistique connu : elle est non monotone. L'interprétation que l'on peut donner à ce genre de courbe est la suivante : les élèves dont le niveau d'habileté est inférieur à $\theta = 0$ ont plus de chances de choisir l'option C (une mauvaise réponse) que l'option A, la bonne réponse. Par contre, pour les élèves d'habileté supérieure à $\theta = 0$, le phénomène inverse prévaut : ils auront plus tendance à choisir l'option A que l'option C. On voit en outre que les options B et D sont beaucoup moins populaires : seuls des élèves très faibles ont une certaine attirance pour l'une ou l'autre de ces deux options. Cette observation est conforme à l'interprétation que nous avons proposée plus haut : à savoir que les élèves qui choisissent l'option D, par exemple, sont considérés, toutes choses étant égales par ailleurs, moins habiles que ceux qui choisissent l'option C. Comme on peut le voir à la figure 4.35, près de 40 % des élèves d'habileté moyenne ($\theta = 0$) choisissent l'option C alors qu'à peine 10 % des élèves de ce niveau d'habileté choisissent l'option D. Même à $\theta = 1$, plus de 25 % des élèves de ce niveau d'habileté choisissent encore l'option C alors qu'ils négligent à peu près tous l'option D.

FIGURE 4.35
Courbe caractéristique pour chacune des quatre options de l'item 7 de l'enquête internationale

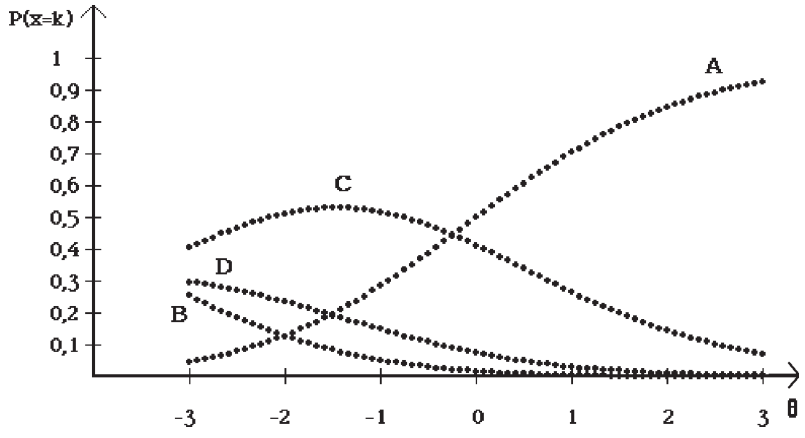
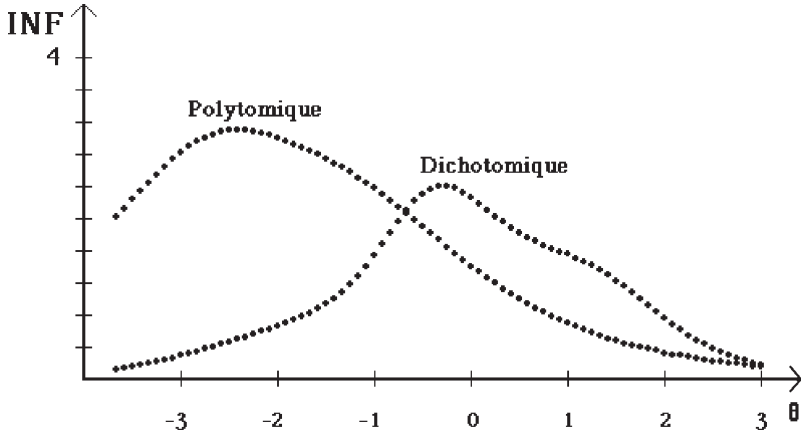


TABLEAU 4.4
Valeurs des estimés des paramètres a_{ix} et c_{ix} pour les quatre options de l'item 7 de l'enquête internationale : modèle nominal de Bock (1972)

Option	a_{ix}	c_{ix}
A	0,94	1,38
B	–0,77	–2,04
C	0,15	1,19
D	–0,32	–0,53

Qu'apporte donc ce modèle polytomique par rapport au modèle dichotomique ? Tout d'abord, le modèle polytomique permet de produire une courbe caractéristique pour chaque option de l'item et, partant, d'analyser le comportement de chaque choix de réponse, un peu comme nous l'avons fait plus haut. De plus, puisque la principale différence entre les deux modèles réside dans l'exploitation de l'information que donnent les leurres, il ne serait pas étonnant que cette différence se reflète auprès de ceux qui choisissent ces leurres. Or, les élèves plus habiles n'ont pas souvent recours aux leurres puisque, en général, ils choisissent la bonne réponse. C'est pourquoi, comme on le voit à la figure 4.36, même si on observe une légère perte d'information au niveau des élèves les plus habiles, c'est au niveau des élèves les moins habiles que l'exploitation de cette nouvelle information concernant l'option de réponse choisie est la plus visible.

FIGURE 4.36
Courbes d'information des 76 items de l'enquête de mathématique de l'IAEP 2 obtenues selon un modèle dichotomique et selon le modèle polytomique de Bock (1972)



Modèle gradué de Samejima (1969)

Plusieurs instruments de mesure de personnalité, d'attitude ou d'opinion renferment des items dont l'échelle de mesure est constituée de catégories graduées. Dans ces circonstances, il est approprié d'avoir recours à des modèles qui permettent une analyse plus subtile que les modèles dichotomiques.

Le modèle gradué (Samejima, 1969, 1997) constitue une généralisation du modèle dichotomique à 2 paramètres. Il s'agit d'un modèle qui semble tout particulièrement approprié pour analyser le comportement des items situés sur une échelle de Likert. Contrairement au modèle qui s'appuie sur une échelle d'évaluation (*rating scale model*), le modèle gradué peut très bien s'accommoder d'items comportant un nombre de catégories variable. Chaque item i est caractérisé par un seul paramètre de discrimination (*slope*) a_i et un certain nombre de paramètres de localisation (*thresholds*) entre les catégories, b_{ik} où $k = 0$ à $m_i - 1$.

Soit donc

$$P_i^*(k/\theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ik})}}$$

la probabilité d'adhérer à la catégorie k ou à une catégorie supérieure de l'item i, dans le cas où $k = 0, 1, 2, \ldots, m_i - 1$.

Étant donné qu'il est certain qu'une réponse à un item i qui contient les catégories 0, 1, 2… se retrouvera dans la catégorie 0 ou dans une catégorie supérieure, il vient que

$$P_i^*(0/\theta) = 1$$

Ainsi, nous définirons la probabilité d'adhérer à la catégorie k de l'item i par

$$P_i(k/\theta) = P_i^*(k/\theta) - P_i^*(k+1/\theta)$$

Pour chaque item i, il y a donc un seul paramètre a_i et autant de paramètres b_{ik} que le nombre de catégories de l'item moins une.

Chaque item représenté par $P_i^*(k/\theta)$ est traité comme une série de m_i contrastes dichotomiques (0 vs 1,2,3,4 ; 0,1 vs 2,3,4 ; 0,1,2 vs 3,4 ; 0,1,2,3 vs 4). Chaque contraste correspond à un modèle dichotomique à deux paramètres. De sorte que la différence entre deux $P_i^*(k/\theta)$ consécutifs, ce que nous avons noté $P_i(k|\theta)$, est la courbe caractéristique de la catégorie k de l'item i, celle qui représente la probabilité d'adhérer à la catégorie k.

La figure 4.37 présente les courbes caractéristiques de chacune des quatre catégories de l'item 13[19] de l'échelle de dépression de Beck[20]. Ces courbes ont été obtenues en utilisant la modélisation graduée proposée par Samejima. Le tableau 4.5 présente les valeurs des estimés des paramètres : comme il se doit, nous observons une seule valeur pour le paramètre a_i et une valeur pour chacun des 3 paramètres b_{ik}. Notons que le nombre total de catégories de cet item est bien 4.

Le paramètre a_i est une indication générale de la pente des courbes caractéristiques des catégories de l'item. La valeur du paramètre b_{i1} = 1,085, comme on peut le voir à la figure 4.37, représente sur l'axe d'habileté θ le point pour lequel la probabilité d'endosser la catégorie 1 ou une catégorie supérieure dépasse 50 % (Thissen et Wainer, 2001, p. 146). De même, la valeur du paramètre b_{i2} = 2,538 indique le point pour lequel la probabilité d'endosser la catégorie 2 ou une catégorie supérieure dépasse 50 %. Quant à la valeur du paramètre b_{i3} = 10,179, même si elle n'est pas visible sur la figure, elle correspond au point pour lequel la probabilité d'endosser la catégorie 3 dépasse 50 %. En réalité, si on pouvait observer tout l'axe d'habileté θ, on verrait bien que la courbe de la catégorie 3 continue de monter, tout comme c'est le cas pour la courbe de la catégorie 2, et qu'elle atteindra la marque de 50 % au point 10,179 de l'axe d'habileté θ.

Puisque, contrairement aux instruments analysés jusqu'ici, les items de l'échelle de Beck ne mesurent pas vraiment une habileté, une capacité ou un rendement quelconque, mais constituent plutôt un indice de dépression, il convient d'ajouter une interprétation aux courbes caractéristiques de la figure 4.37. Tout d'abord, l'axe θ ne constitue pas vraiment un axe d'habileté en tant que tel, mais peut s'interpréter plutôt comme un axe où plus une personne obtient une valeur de θ qui est faible, plus son indice de dépression est faible. De même, plus une personne obtient une valeur de θ qui est élevée, plus son indice de dépression est élevé. Il n'est donc pas étonnant de constater, à la figure 4.37, que toutes les personnes possédant une valeur de θ plus faible que 1 auront tendance à endosser la catégorie 0 (« Je prends des décisions avec autant de facilité qu'à l'habitude ») plutôt que n'importe quelle autre catégorie de cet item puisque ce sont les personnes considérées peu dépressives. De la même façon, les personnes dont la valeur de θ se trouve

---

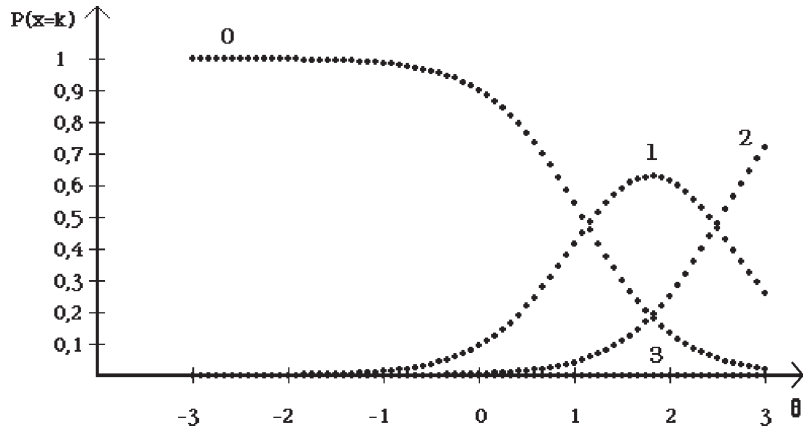
19. L'item 13 comporte 4 catégories : le sujet est requis de choisir l'une ou l'autre de ces catégories.
    0. *I make decisions about as well as I ever could.* (Je prends des décisions avec autant de facilité qu'à l'habitude.)
    1. *I put off making decisions more than I used to.* (Je remets mes décisions à une date ultérieure plus souvent qu'à l'habitude.)
    2. *I have greater difficulty in making decisions than I used to.* (J'éprouve plus de difficulté à prendre des décisions qu'à l'habitude.)
    3. *I can't make decisions at all anymore.* (Je ne peux plus prendre de décisions du tout.)
20. Beck, A.T., Rush, A., Shaw, B., et Emery, G. (1979). *Cognitive therapy of depression.* New York : Guilford Press.

entre 1 et 2,5 environ, donc les personnes révélant un certain degré de dépression, auront tendance à endosser la catégorie 1 (« Je remets mes décisions à une date ultérieure plus souvent qu'à l'habitude »). Enfin, les personnes possédant une valeur de θ plus élevée que 2,5, donc celles qui sont considérées les plus dépressives, auront tendance à endosser la catégorie 2 (« J'éprouve plus de difficulté à prendre des décisions qu'à l'habitude »). Il semble que la catégorie 3 (« Je ne peux plus prendre de décisions du tout ») soit si difficile à endosser qu'à peu près personne ne l'a fait. Soulignons que la population visée par cette analyse était constituée d'étudiants universitaires (Ramsay, 1993). Si une autre population était visée, par exemple des décrocheurs dont l'âge varie entre 15 et 19 ans ou des personnes psychiatrisées, sans doute que la catégorie 3 aurait tendance à être plus populaire.

TABLEAU 4.5
Valeurs des estimés des paramètres a_i et b_{ik} de l'item 13 de l'échelle de Beck ; comme cet item possède 4 catégories, il y a bien 3 valeurs pour b_{ik} selon le modèle gradué de Samejima

Option	a_i	b_{ik}
k = 1	2,032	1,085
k = 2		2,538
k = 3		10,179

FIGURE 4.37
Courbes caractéristiques de chacune des 4 catégories de l'item 13 de l'échelle de dépression de Beck



Autres modèles polytomiques

Les modèles nominal de Bock (1972) et gradué de Samejima (1969) ne constituent pas une liste exhaustive des modèles polytomiques, loin de là. Plusieurs autres modèles, paramétriques ou non, ont été développés au cours des 20 ou 25 dernières années. Pensons simplement au modèle d'échelle d'évaluation (*rating scale*) d'Andrich (1978), un modèle de la famille de Rasch, qui est considéré idéal pour traiter les échelles de type Likert (Embretson et Reise, 2000) ou encore le modèle à crédit partiel de Masters (1982), toujours dans la famille de Rasch, conçu pour analyser les items pour lesquels la valeur attribuée à une réponse est d'autant plus élevée que la réponse est exacte, ou même le modèle à crédit partiel généralisé de Muraki (1997), parfois utilisé pour analyser les items à réponse construite.

Dans le cas où l'échelle de mesure est nominale (p. ex., items à choix multiple), c'est le modèle nominal de Bock qui semble le bon choix (Embretson et Reise, 2000) même si le modèle à choix multiple de Thissen et Steinberg (1984) peut aussi être envisagé dans la situation où il y a peu d'items. Si l'échelle est graduée, mais que la discrimination semble différente d'un item à l'autre, il faut opter pour le modèle gradué de Samejima (1969, 1997). Il faut aussi savoir que le modèle d'échelle d'évaluation d'Andrich (1978) n'est pas applicable si l'échelle de mesure varie d'un item à l'autre. Par contre, Embretson et Reise (2000) suggèrent d'utiliser ce dernier modèle dans le cas de l'analyse d'une échelle d'attitude ou d'opinion de type Likert où chaque item comporte le même nombre de catégories.

4.6.2. Les modèles multidimensionnels

Bien que la présentation complète de modèles de réponses aux items multidimensionnels dépasse largement le cadre de cet ouvrage, nous désirons tout de même introduire, ne serait-ce que succinctement, les concepts inhérents à ces modèles. De façon générale, un modèle sera dit multidimensionnel s'il est nécessaire d'utiliser plusieurs paramètres d'habileté dans la modélisation. Ainsi, une façon bien naturelle de généraliser le modèle unidimensionnel logistique à trois paramètres à un modèle à m dimensions est de définir comme suit la probabilité de réussir un item i (Embretson et Reise, 2000) :

$$P(X_i = 1 | \theta_1, \theta_2, ..., \theta_m) = c_i + \frac{1 - c_i}{1 + e^{[-D(a_{i1}\theta_1 + a_{i2}\theta_2 + ... + a_{im}\theta_m + d_i)]}}$$

Dans ce cas, on dénombre m dimensions ou habiletés θ_1, θ_2, θ_3, …, θ_m ; le paramètre d_i est appelé l'intercept et est lié au paramètre de difficulté b_i qui sera défini plus bas ; les paramètres de discrimination sont donnés par a_{i1}, a_{i2}, a_{i3},…, a_{im} ; le paramètre de pseudo-chance est c_i.

Selon Reckase (1997, p. 276), le pouvoir discriminant de l'item i peut être défini comme suit :

$$\text{MDISC}_i = \sqrt{\sum_{k=1}^{m} a_{ik}^2}$$

De plus, le paramètre de difficulté multidimensionnelle b_i est donné par :

$$b_i = \frac{-d_i}{\text{MDISC}_i}$$

Il faut noter que, bien que ce modèle suppose m paramètres d'habileté et m paramètres de discrimination, on ne dénombre qu'un paramètre de difficulté et un paramètre de pseudo-chance. Notons aussi que rien dans ce modèle ne restreint les m dimensions à l'orthogonalité (à l'indépendance). Enfin, il est facile de se rendre compte que si m = 1, ce modèle revient exactement au modèle logistique unidimensionnel bien connu à trois paramètres.

Alors qu'on parlait de courbe caractéristique d'item dans le cas d'un modèle unidimensionnel, ici on traitera avec les surfaces caractéristiques des items modélisés avec un modèle multidimensionnel. La figure 1 de Reckase (1997, p. 274) illustre bien le concept de surface caractéristique d'item dans le cas où deux dimensions sont supposées. La projection de cette surface sur le plan (θ_1, Prob(θ_1,θ_2)) ou le plan (θ_2, Prob(θ_1,θ_2)), qui équivaut en réalité à éliminer une dimension, donc à revenir à un espace à une dimension, donnerait une courbe caractéristique en forme de « S » bien typique. Bien que ce modèle multidimensionnel semble attrayant, il n'est pas encore très utilisé compte tenu notamment de la difficulté d'estimer les paramètres des items et de la rareté des logiciels conçus pour estimer les paramètres.

Si le modèle précédent constituait une extension multidimensionnelle du modèle logistique à trois paramètres, le modèle de traits latents à plusieurs composantes (*multicomponent latent trait model*) MLTM, proposé par Embretson (1985, 1997), est une généralisation multidimensionnelle du modèle de Rasch. Il suppose que la réussite d'un sujet à une tâche (item total) donnée dépend de la difficulté des différentes composantes de cette tâche ainsi que de l'habileté du sujet à résoudre **chacune** de ces composantes de la tâche.

Le modèle MLTM est donné par

$$P(X_{ijT} = 1 \mid \theta_{jk}, b_{ik}) = (s - g) \prod_{k=1}^{m} \frac{e^{(\theta_{jk} - b_{ik})}}{1 + e^{(\theta_{jk} - b_{ik})}} + g$$

où X_{ijT} est la réponse du sujet j à la tâche T relative à l'item i,

θ_{jk} est l'habileté du sujet j à la composante k,

b_{ik} est la difficulté de la composante k associée à l'item i,

g est la probabilité de réussir l'item par pur hasard,

s est la probabilité de réussir la tâche étant donné la réussite à chacune des m composantes.

Notons que, contrairement au modèle multidimensionnel décrit par Reckase (1997), le modèle MLTM comprend autant de paramètres de difficulté que de dimensions.

Embretson (1983, p. 183) présente une application du modèle MLTM en prenant appui sur un item d'analogie verbale qui comprend m = 2 composantes. L'item total (la tâche) se lit comme suit :

CHAT : TIGRE : : CHIEN : _____
a) Lion b) Loup c) Aboiement d) Chiot e) Cheval

La première composante de cet item est la construction de la règle (*rule construction*). Il s'agit d'indiquer quelle est la règle sous-jacente à l'appariement des animaux cités.

CHAT : TIGRE : : CHIEN : _____
RÈGLE = _____

La deuxième composante de cet item est l'évaluation de la réponse (*response evaluation*). Il s'agit de compléter l'item d'analogie, une fois la règle connue.

Étant donné que la règle est « associer à de grands canidés sauvages », trouvez le mot manquant dans
CHAT : TIGRE : : CHIEN : _____
parmi les choix de réponses suivants :
a) Lion b) Loup c) Aboiement d) Chiot e) Cheval

Le modèle MLTM stipule que chacune de ces deux composantes devrait normalement être réussie pour que l'item total le soit. En modélisant les principales composantes qui influencent la réussite à une tâche, Embretson a défini en fait une procédure de validation de construit. Cette procédure a été employée par plusieurs dont Bertrand *et al.* (1993) et Janssen *et al.* (1991).

4.6.3. Les modèles non paramétriques

Bien que les modèles non paramétriques ne permettent pas d'applications aussi visibles et variées que les modèles paramétriques, comme le testing adaptatif par ordinateur (voir le chapitre 9) ou l'équilibrage des échelles (*equating*), ces

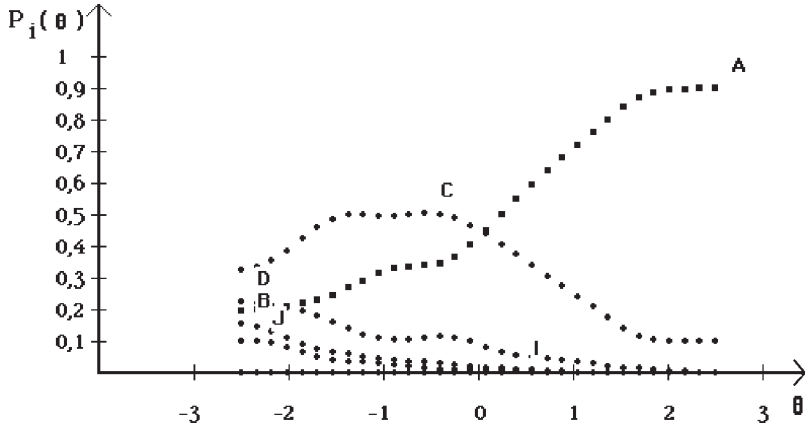
modèles sont toutefois appropriés pour l'analyse d'items ou encore pour vérifier les conditions d'application des modèles de la TRI comme l'unidimensionnalité (voir le chapitre 5).

Les modèles non paramétriques ne requièrent pas des échantillons de sujets aussi imposants que les modèles paramétriques pour obtenir un ajustement analytique du modèle aux données de même qualité.

Nous décrirons ici l'approche de modélisation non paramétrique proposée par Ramsay (1991), qui s'avère appropriée notamment pour l'analyse d'items[21]. Cette approche est tout particulièrement intéressante dans la mesure où elle allie le caractère non paramétrique au caractère polytomique (nominal ou ordinal). Il sera question des avantages mais aussi des inconvénients à utiliser ce modèle.

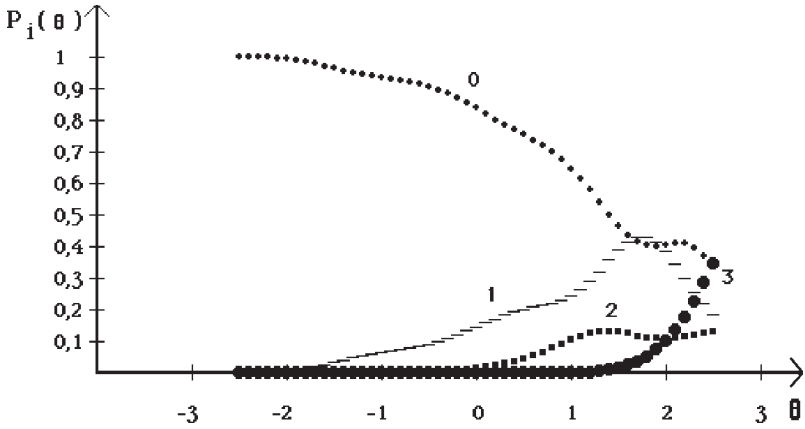
Une des caractéristiques des modèles non paramétriques proposés par Ramsay est de pouvoir tracer une courbe caractéristique d'option (CCO) sans devoir estimer les paramètres de l'item. La CCO est en effet lissée (*smoothed*) à partir des données : il n'y a aucune estimation de jeux de paramètres d'items. Une telle courbe épouse donc au mieux les données, mais elle a comme désavantage de ne pas être nécessairement monotone croissante, comme en font foi les figures 4.38 et 4.39.

FIGURE 4.38
Courbe caractéristique pour chacune des quatre options de l'item 7 d'une enquête internationale selon TESTGRAF (voir aussi la figure 4.35)



21. Cette approche non paramétrique est implantée dans le logiciel TESTGRAF.

FIGURE 4.39
Courbe caractéristique de chacune des quatre catégories de l'item 13 de l'échelle de dépression de Beck selon TESTGRAF (voir aussi la figure 4.37)



La similitude des courbes caractéristiques des figures 4.35 et 4.38 est particulièrement frappante. Ces deux figures représentent les courbes d'options du même item à choix multiple. Seule la modélisation change : paramétrique dans le cas de la figure 4.35, où le modèle nominal de Bock a été employé, non paramétrique dans le cas de la figure 4.38, où les courbes ont été lissées par l'approche non paramétrique de Ramsay.

Les figures 4.37 et 4.39 se rapportent au même item de l'échelle de Beck. On y voit la similitude des options 0 et 1 de l'item 13 pour les deux modélisations : le modèle gradué de Samejima dans le cas de la figure 4.37 et le modèle non paramétrique de Ramsay dans le cas de la figure 4.39. Notons toutefois un certain écart entre les deux modèles pour les courbes des options 2 et 3.

Exercices

1. Le paramètre b_i des modèles logistiques de la théorie des réponses aux items est communément appelé *threshold* ou encore indice de difficulté de l'item i. Pourquoi pensez-vous qu'il est plus approprié de considérer ce paramètre b_i comme un réel indice de difficulté dans le cas du modèle à un paramètre que dans le cas du modèle à deux paramètres ?

2. Donnez les paramètres en TRI de trois items qui sont tels que toute personne qui réussit l'item 1 et l'item 2 mais manque l'item 3 aura un estimé de type maximum de vraisemblance plus grand que b_1 mais plus petit que b_2 et plus petit que b_3.

3. L'examen d'histoire du ministère de l'Éducation contient sept problèmes à réponses longues. Les résultats de l'analyse d'items en TRI sont donnés plus bas : il s'agit d'une sortie informatique obtenue du logiciel BILOG- 3 (Mislevy et Bock, 1990).

   a) Quel modèle de la TRI a-t-on utilisé ? Pourquoi ?

   b) Selon les estimés de paramètres décrits plus bas, à quel genre de test a-t-on affaire ici ?

Item S.E.	Intercept S.E.	Slope S.E.	Threshold S.E.	Dispersn S.E.	Asymptote S.E.	Chisq (PROB)	Df
0001	1,961	0,490	–4,006	2,043	0,000	0,0	0,0
	0,274*	0,033*	0,560*	0,139*	0,000*	(10,0000)	
0002	1,658	0,490	–3,386	2,043	0,000	0,0	0,0
	0,223*	0,033*	0,455*	0,139*	0,000*	(10,0000)	
0003	0,894	0,490	–1,826	2,043	0,000	0,5	2,0
	0,145*	0,033*	0,296*	0,139*	0,000*	(0,7702)	
0004	1,176	0,490	–2,401	2,043	0,000	0,9	1,0
	0,174*	0,033*	0,354*	0,139*	0,000*	(0,3538)	
0005	1,962	0,490	–4,007	2,043	0,000	0,0	0,0
	0,279*	0,033*	0,569*	0,139*	0,000*	(10,0000)	
0006	0,551	0,490	–1,125	2,043	0,000	40,5	3,0
	0,127*	0,033*	0,260*	0,139*	0,000*	(0,2100)	
0007	0,608	0,490	–1,241	2,043	0,000	10,2	3,0
	0,128*	0,033*	0,261*	0,139*	0,000*	(0,7448)	

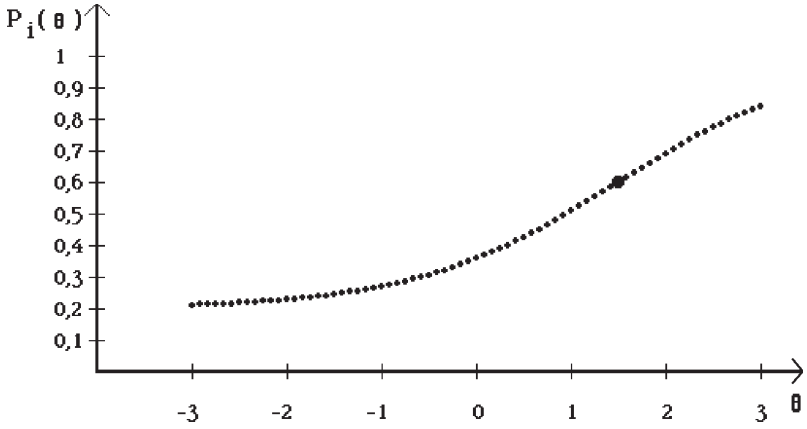
4. Quel est, parmi les 3 items suivants, celui qui discrimine le mieux les sujets d'habileté $\theta = -1$ par rapport aux sujets d'habileté $\theta = 0$? Expliquez votre réponse.

   item 1 : $a_1 = 0{,}2$, $b_1 = 0$, $c_1 = 0$

   item 2 : $a_2 = 0{,}5$, $b_2 = 0{,}5$, $c_2 = 0$

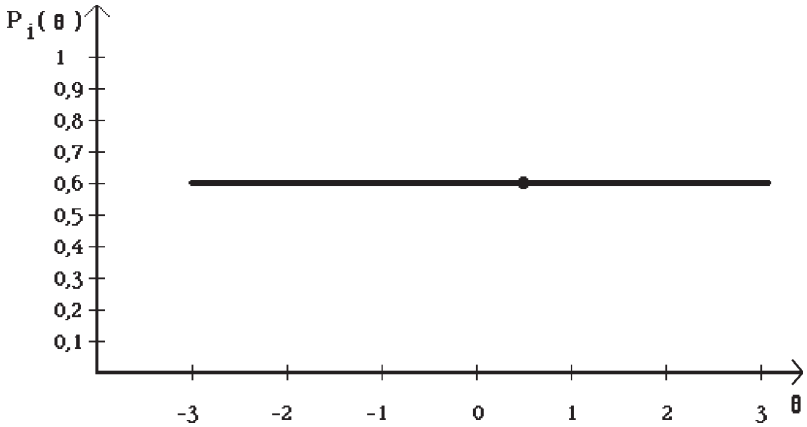
   item 3 : $a_3 = 2$, $b_3 = -2$, $c_3 = 0{,}2$

5. Trouvez, à une décimale près, les valeurs des paramètres (a_i, b_i, c_i) de l'item i représenté par la courbe caractéristique suivante.



6. Quelles sont les valeurs minimale et maximale de $P_i(b_i)$, si P_i est la fonction logistique à trois paramètres connue et b_i est l'indice de difficulté TRI ?

7. Donnez les valeurs des paramètres de difficulté, de discrimination et de pseudo-chance associés à la courbe caractéristique d'item suivante.



8. Lors de la calibration d'un test de sciences physiques, composé de 50 items à choix de réponse, on observe que l'item 18 a un comportement inusité. Voici les estimés TRI des paramètres de l'item 18 : $a_i = -1/2$, $b_i = 0$, $c_i = 0{,}35$. Qu'a donc l'item 18 de si inusité ?

Corrigé des exercices

1. Contrairement au paramètre classique p_i, qui est un indice de facilité, le paramètre b_i est bel et bien un indice de difficulté puisque plus l'item i est difficile, plus la valeur de b_i est élevée. Or, ceci est vrai seulement dans le cas du modèle à un paramètre, puisque dans le cas des modèles à deux ou trois paramètres les CCI peuvent se croiser comme à la figure 4.24 où, même si un indice b_i est supérieur à un indice b_j, l'item j peut être plus difficile à un endroit donné de l'axe θ.

3. a) Il s'agit d'un modèle à un paramètre puisque les valeurs du paramètre de discrimination sont toutes égales à 0,49 et que les valeurs du paramètre c_i sont égales à 0.

   b) Les items sont faciles puisque les estimés du paramètre b_i sont tous négatifs. Par ailleurs, il s'agit aussi d'un test peu discriminant puisque les estimés du paramètre a_i sont faibles, soit 0,49.

5. Les paramètres sont : $a_i = 1$, $b_i = 1{,}5$ et $c_i = 0{,}2$.

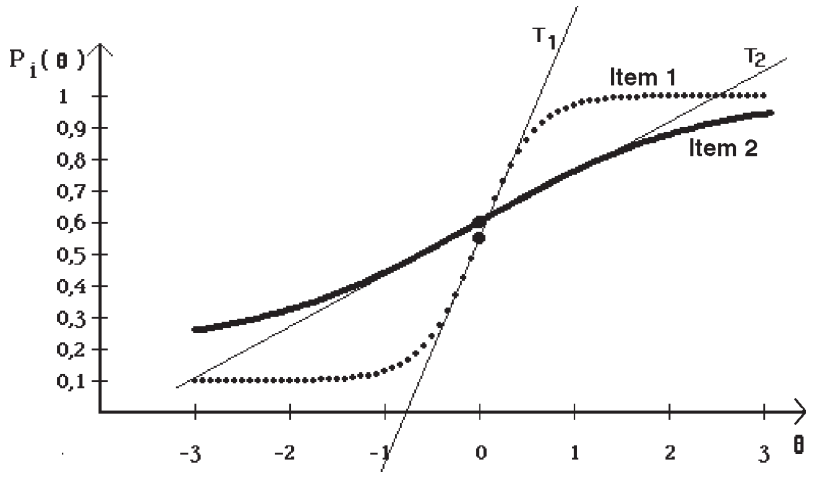
7. Les paramètres sont : $a_i = 0$, $b_i = 0{,}5$ et $c_i = 0{,}2$ [car $(1 + 0{,}2)/2 = 0{,}6$].

Annexe 4.1

Démonstration de la relation entre la pente et le paramètre de discrimination a_i

Examinons d'abord la figure A4.1, qui montre la CCI de deux items ainsi que la droite tangente (T_i) au point d'inflexion pour chaque CCI. On y voit que plus la courbe est abrupte au voisinage du point d'inflexion, plus la pente m_i de la droite qui est tangente (T_i) à ce point est élevée.

FIGURE A4.1
Courbes caractéristiques de deux items (modèle à trois paramètres) et les droites tangentes (T_1 et T_2) au point d'inflexion dans chaque cas



Nous allons montrer que

$$a_i = \frac{m_i\left(\frac{4}{D}\right)}{1-c_i} \text{ ou, ce qui revient au même, } m_i = \frac{(1-c_i)Da_i}{4}$$

où a_i est le paramètre de discrimination de l'item i,

D = 1,7 est la constante de normalisation,

m_i est la pente de la droite tangente (T_i) au point d'inflexion pour l'item i,

c_i est le paramètre de pseudo-chance de l'item i.

Rappelons d'abord les propriétés des dérivées que nous allons utiliser au cours de cette démonstration :

$$\frac{d}{d\theta}[\text{constante}] = 0 \tag{1}$$

$$\frac{d}{d\theta}[c\,f(\theta)] = c\frac{d}{d\theta}[f(\theta)] \tag{2}$$

$$\frac{d}{d\theta}\left[\frac{c}{f(\theta)}\right] = \frac{-c}{f(\theta)^2}\frac{d}{d\theta}[f(\theta)] \tag{3}$$

$$\frac{d}{d\theta}[e^{f(\theta)}] = e^{f(\theta)}\frac{d}{d\theta}[f(\theta)] \tag{4}$$

Rappelons aussi que le point d'inflexion a comme coordonnées [b_i , $(1 + c_i)/2$]. Ainsi, trouver la pente de la tangente au point d'inflexion revient à trouver la dérivée de la CCI, soit en fait $P_i(\theta)$, au point où $\theta = b_i$.

Ainsi, au point $\theta = b_i$, nous avons

$$m_i = \frac{d}{d\theta}[P_i(\theta)] = \frac{d}{d\theta}\left[c_i + \frac{1-c_i}{1+e^{-Da_i(\theta-b_i)}}\right]$$

et, en se servant des équations (1) et (2) il vient

$$m_i = (1-c_i)\frac{d}{d\theta}\left[\frac{1}{1+e^{-Da_i(\theta-b_i)}}\right]$$

et en utilisant l'équation (3),

$$m_i = \frac{-(1-c_i)}{\left[1+e^{-Da_i(\theta-b_i)}\right]^2}\frac{d}{d\theta}\left[1+e^{-Da_i(\theta-b_i)}\right]$$

puis à cause de l'équation (4),

$$, m_i = \frac{-(1-c_i)e^{-Da_i(\theta-b_i)}}{\left[1+e^{-Da_i(\theta-b_i)}\right]^2}\frac{d}{d\theta}[-Da_i(\theta-b_i)]$$

et, en utilisant l'équation (2),

$$m_i = \frac{-(1-c_i)e^{-Da_i(\theta-b_i)}}{\left[1+e^{-Da_i(\theta-b_i)}\right]^2}[-Da_i]$$

et enfin puisque $\theta = b_i$, il vient,

$$m_i = \frac{-(1-c_i)e^{0}}{\left[1+e^{0}\right]^2}[-Da_i] = \frac{(1-c_i)Da_i}{4}$$

Annexe 4.2

Démonstration de la relation entre le score vrai et Σ $P_i(\theta)$

Il faut d'abord se souvenir que le score observé X pour un test de n items peut se définir comme

$$X = \sum_{i=1}^{n} U_i$$

où U_i= 1 si l'item i est réussi et U_i = 0 si l'item i est échoué.

Le score vrai V est l'espérance mathématique du score observé X, soit V = E(X). Ainsi,

$$V = E(X) = E(\sum_{i=1}^{n} U_i) = \sum_{i=1}^{n} E(U_i)$$

Or, par la définition même d'une espérance mathématique,

$E(U_i) = [0 \times m_i(0)] + [1 \times m_i(1)] = m_i(1)$, où m_i est la fonction de masse associée à la variable de Bernouilli et donc définie par

$$m_i : D_i = \{0,1\} \mapsto \Re$$
$$0 \mapsto P(U_i = 0) = 1 - P_i(\theta)$$
$$1 \mapsto P(U_i = 1) = P_i(\theta)$$

où $P(U_i = 1)$ indique la probabilité que la variable U_i égale 1, c'est-à-dire la probabilité de réussir l'item i.

Ainsi $E(U_i) = m_i(1) = P_i(\theta)$

Enfin, $V = \sum_{i=1}^{n} E(U_i) = \sum_{i=1}^{n} P_i(\theta)$

Annexe 4.3

Démonstration de la formule de l'information (équation 4.8)

Il a été montré, à l'annexe 4.1, que la dérivée de $P_i(\theta)$ est égale à

$$P_i'(\theta) = \frac{d}{d\theta}[P_i(\theta)] = \frac{-(1-c_i)e^{-Da_i(\theta-b_i)}}{\left[1+e^{-Da_i(\theta-b_i)}\right]^2}[-Da_i]$$

dont nous pouvons simplifier temporairement l'écriture en posant

$$e = e^{-Da_i(\theta-b_i)}$$

On obtient alors

$$P_i'(\theta) = \frac{Da_i(1-c_i)e}{[1+e]^2}$$

Ainsi

$$P_i'(\theta)^2 = \frac{D^2a_i^2(1-c_i)^2e^2}{[1+e]^4} \qquad (1)$$

Souvenons-nous ensuite que

$$P_i(\theta) = c_i + \frac{(1-c_i)}{1+e} = \frac{c_i(1+e)+(1-c_i)}{1+e} = \frac{1+ec_i}{1+e} \qquad (2)$$

donc que

$$Q_i(\theta) = 1 - P_i(\theta) = 1 - \frac{1+ec_i}{1+e} = \frac{e(1-c_i)}{1+e} \qquad (3)$$

Ainsi,

$$P_i(\theta)Q_i(\theta) = \frac{1+e^{c_i}}{1+e^{\theta}} \frac{e^{\theta}(1-c_i)}{1+e^{\theta}} = \frac{e^{\theta}(1-c_i)(1+e^{c_i})}{(1+e^{\theta})^2} \tag{4}$$

À partir des équations (1) et (4) nous pouvons écrire, en simplifiant quelques termes,

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)} = \frac{\frac{D^2 a_i^2 (1-c_i)^2 e^{2\theta}}{[1+e^{\theta}]^4}}{\frac{e^{\theta}(1-c_i)(1+e^{c_i})}{(1+e^{\theta})^2}} = \frac{D^2 a_i^2 (1-c_i)\ e^{\theta}}{(1+e^{\theta})^2 (1+e^{c_i})} \tag{5}$$

Or d'après l'équation (2) nous pouvons déduire que

$$\frac{P_i(\theta)-c_i}{(1-c_i)} = \frac{1}{1+e^{\theta}} \tag{6}$$

De plus, en considérant simultanément les équations (2) et (3) nous avons

$$\frac{Q_i(\theta)}{P_i(\theta)} = \frac{\frac{e^{\theta}(1-c_i)}{1+e^{\theta}}}{\frac{1+e^{c_i}}{1+e^{\theta}}} = \frac{e^{\theta}(1-c_i)}{1+e^{c_i}} \tag{7}$$

En combinant les équations (5), (6) et (7), nous avons le résultat voulu, soit l'équation (4.8) $I_i(\theta) = D^2 a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{(1 - c_i)} \right]^2$

Pour un modèle à deux paramètres où $c_i = 0$, on obtient alors $I_i(\theta) = D^2 a_i^2 P_i(\theta)Q_i(\theta)$.

Enfin, pour un modèle à un paramètre où $a_i = 1$ et $c_i = 0$, on obtient $I_i(\theta) = D^2 P_i(\theta)Q_i(\theta)$.

CHAPITRE 5

Conditions d'application et critères d'adéquation des modèles

L'époque dans laquelle nous vivons est souvent décrite comme l'ère de l'information. À l'appui de cette épithète, nous constatons régulièrement que la quantité de données produites dans les différents domaines de la connaissance humaine ne cesse de croître et de se complexifier. L'entreprise de production, d'analyse et d'utilisation de ces données se révèle ainsi une entreprise fort délicate à réaliser avec doigté et discernement. Dans ce contexte, nous l'avons déjà mentionné au premier chapitre, le rôle d'un modèle consisterait principalement à représenter plus simplement une **réalité** complexe de façon à ce que les données recueillies illuminent la situation et permettent une intelligibilité du réel. La modélisation constitue donc une étape dans la description et la compréhension des données à notre disposition. Il s'agit d'un processus

permettant de rendre formel un cadre de référence en se soustrayant en partie à la complexité de la réalité. Mais comme cette modélisation est une simplification et une abstraction, il est possible qu'elle rende compte plus ou moins bien de ce qui est observé empiriquement. Ici, l'étiquette **plus ou moins bien** peut autant être associée à la qualité des données qu'à la qualité de l'adéquation du modèle, sans parler de la qualité du cadre conceptuel. L'étude de ce « plus ou moins bien » est ce qui permet simultanément un raffinement des instruments de recueil des données, de la stratégie de modélisation et du cadre conceptuel. Dans ce chapitre, c'est l'aspect de la qualité de l'ajustement du modèle qui retiendra notre attention.

La plupart des propositions de modélisation mathématique – c'est le cas des modèles de la TRI, mais aussi des modèles **classiques** pour l'analyse statistique des données – sont restrictives quant aux situations pour lesquelles elles sont considérées adéquates. Dans ces situations, nous disons que certaines conditions d'application du modèle doivent être satisfaites avant que celui-ci puisse démontrer sa pertinence.

À titre d'exemple, pensons à la situation relativement simple où nous désirons comparer deux groupes de sujets ayant fait l'objet d'une prise de mesure pour une variable donnée dans des contextes différents. Il est d'usage, dans plusieurs disciplines des sciences médicales et des sciences sociales, de faire appel à un modèle statistique-probabiliste pour comparer les moyennes des valeurs de la variable. Ces modèles, regroupés sous l'appellation de tests statistiques, peuvent se révéler utiles pour aider le chercheur à décider s'il existe une différence entre, par exemple, les moyennes des deux échantillons. Par la suite, le chercheur infère que les échantillons proviennent de la même population (égalité des moyennes et pas de différence entre les groupes) ou de deux populations différentes (inégalité des moyennes et différence significative entre les groupes) et relie cette conclusion à son hypothèse de recherche (par exemple, la méthode A d'enseignement des mathématiques est supérieure à la méthode B).

Le test statistique qui est sûrement le plus connu pour réaliser ce type d'analyse est le test **t** de Student pour échantillons indépendants. Outre le fait que les variables étudiées doivent être des variables aléatoires, donc issues d'une expérience où le hasard intervient, le test t exige que la distribution des variables dans les populations étudiées épouse la forme de la loi de probabilité normale (ou gaussienne). Lorsque cette condition est respectée, il est possible de démontrer que le test t est le test le plus puissant du point de vue de certaines propriétés statistiques. Lorsque la condition n'est pas respectée, d'autres approches sont plus optimales. Nous pensons ici à certaines procédures non paramétriques ou encore à l'utilisation d'approches robustes aux écarts à la normalité et à la symétrie de la distribution des données.

Ainsi, pour vérifier si la condition de normalité de la distribution est respectée, différentes propositions de tests statistiques ont vu le jour avec les années. Mentionnons entre autres le test du khi carré, celui de Kolmogorov-

Smirnov et celui de Shapiro-Wilks. Des procédures graphiques existent également (courbe Q-Q et courbe P-P, par exemple) pour élargir notre compréhension de l'adéquation du modèle au-delà du simple résultat du test statistique, c'est-à-dire rejeter ou ne pas rejeter l'hypothèse H_0 de la normalité de la distribution. Dans le même esprit, la version de base du test t demande que les variances des variables dans les populations soient égales. Nous disons donc que l'égalité des variances et la normalité sont deux conditions préalables à l'utilisation du test t de Student et qu'elles doivent faire l'objet de vérifications empiriques, elles doivent passer le test de l'analyse des données.

5.1. QUELLES CONDITIONS D'APPLICATION ?

Malgré leur apparente simplicité d'application et leur indéniable polyvalence, les modèles de la TRI sont eux aussi soumis à un certain nombre de conditions balisant les applications adéquates. Ces conditions sont de différents ordres et nécessitent la mise en place de démonstrations empiriques qui, idéalement, doivent être produites pour chacune des applications d'un modèle. L'écart entre les caractéristiques formelles du modèle et les données a fait l'objet de nombreuses publications, recherches et réflexions. La modélisation avec la TRI nous amène donc à tenir compte de quatre aspects qui forment la base de la vérification de la qualité de l'adéquation du modèle :

- le maintien de la propriété d'invariance des estimations des paramètres associés aux items et aux sujets ;
- les ajustements statistique et résiduel aux données d'un modèle ou de plusieurs modèle concurrents ;
- la dimensionalité de l'espace des variables latentes ;
- l'indépendance locale.

Ainsi, lorsque le modèle choisi pour une application donnée est un modèle **unidimensionnel**, il faut produire une preuve raisonnable de cette unidimensionalité. De plus, pour estimer les paramètres des modèles, nous posons comme condition qu'il y a indépendance, pour une valeur fixée sur le continuum de la variable habileté, entre les réponses à des items différents. Également, comme le modèle peut très bien ne pas être le bon modèle, il faut aussi montrer qu'il s'ajuste bien aux données, ou encore, si on est de l'école qui favorise la perspective sur la mesure fondamentale qu'offre le modèle de Rasch, que les données s'ajustent bien au modèle.

Enfin, les modèles de la TRI possèdent également une propriété théorique fondamentale, la propriété d'invariance, qui n'est ni plus ni moins qu'une propriété générique des modèles de régression qui sont exacts pour une population. Encore une fois, une opération de vérification du maintien de cette propriété dans les situations de modélisation est nécessaire. Cette propriété permet d'énoncer les affirmations suivantes :

- ♦ L'estimation de l'habileté d'un individu est indépendante des items auxquels celui-ci doit répondre.
- ♦ Les estimations des caractéristiques des items sont indépendantes des caractéristiques des individus qui répondent aux items.

5.2. DES CHOIX ÉCLAIRÉS

Évidemment, les choses seraient plutôt simples s'il existait une ou deux procédures clairement identifiées comme étant supérieures aux autres, avec des assises théoriques solides et une bonne performance dans la détection des écarts aux conditions d'utilisation. À l'heure actuelle, nous ne pouvons parler de procédures de vérification supérieures (même s'il existe des candidats à ce titre), et certaines définitions de concepts comme l'unidimensionalité, l'indépendance locale ou l'invariance peuvent varier à l'occasion d'un auteur ou d'une époque à l'autre.

Il faut bien saisir cependant qu'à l'instar de toutes les propositions de modélisation que l'on retrouve en science, la modélisation avec la TRI a passablement évolué depuis la publication des propositions de modélisation logistique par Birnbaum dans l'ouvrage de Lord et Novick (1968) et dans celui de G. Rasch paru au début de la même décennie. Bien sûr, il y a eu des précurseurs. Les premiers travaux sur le sujet sont ceux de Brogden (1946), Lawley (1943), Lazarsfeld (1950) et Lord (1952), qui jeta les bases de la TRI telle qu'on la connaît aujourd'hui et qui mit de l'avant le modèle basé sur la fonction de répartition de la loi normale (*ogive normal model*). Beaucoup d'eau a coulé sous les ponts depuis l'époque héroïque où les pionniers de la modélisation des scores aux tests devaient composer avec les limites de la technologie de calcul et d'estimation. Les choses ont en effet bien changé, surtout dans la dernière décennie.

Au premier plan en tant que responsable de cette évolution, nous retrouvons les développements fulgurants de la technologie qui a permis d'accroître grandement l'accès à la puissance et à la vitesse de calcul nécessaires aux modélisations. Les avancées technologiques ont stimulé le développement d'une foule de procédures comme la modélisation des patrons de réponses, l'utilisation de méthodes d'estimation complexes, le développement de tests d'ajustement, les modélisations multidimensionnelles, le testing adaptatif informatisé, etc. À titre d'exemple, il y a déjà presque vingt ans, Hattie (1984, 1985) recensait et étudiait plus de 80 approches et indices suggérés dans la littérature pour déterminer l'unidimensionalité de l'ensemble des scores à un test. Depuis la recension de Hattie, de nouvelles approches pour déterminer la dimensionalité ont été suggérées. Ces approches contrastent avec ce qui était sur la table au moment des travaux de Hattie en ce sens qu'elles sont plus adéquates théoriquement, qu'elles font appel à moins de raccourcis conceptuels par rapport au cadre de la théorie classique des tests, par exemple, et qu'elles s'alimentent à même la puissance de calcul des ordinateurs de notre époque.

Le travail du praticien ou du développeur qui désire appliquer les modèles et s'assurer qu'il le fait adéquatement se trouve compliqué pour la simple et bonne raison qu'il est difficile à l'heure actuelle de choisir entre plusieurs propositions qui ont les mêmes prétentions, mais pour lesquelles le consensus théorique et empirique ne se réalise point. Cette particularité de la prolifération des propositions de vérification de la qualité de l'adéquation des modèles, une caractéristique d'une science en train de se faire, comme dirait T. Kuhn dans son livre *La structure des révolutions scientifiques*, teinte l'ensemble du développement de la TRI, c'est-à-dire non seulement l'étude des conditions d'utilisation, mais également les propositions de **nouveaux** modèles de la TRI et les stratégies d'estimation des paramètres.

Nous devons également souligner l'état de dépendance à l'égard des logiciels de modélisation et d'analyse dans lequel se trouve le praticien. En effet, plus la modélisation et les stratégies d'estimations se complexifient, plus l'utilisateur doit faire confiance aux outils disponibles commercialement et espérer qu'il en existe pour résoudre le problème qui l'intéresse. Sinon, il faut programmer soi-même les procédures, ce qui, convenons-en, n'est pas à la portée de la grande majorité des étudiants et des chercheurs en sciences sociales. Actuellement, il n'existe pas de progiciel (*package*) statistique qui intègre une vaste gamme de procédures éprouvées et reconnues pour examiner simultanément, par exemple, la dimensionalité de l'ensemble des scores et la propriété d'invariance. Il existe cependant plusieurs petits logiciels qui accompagnent des procédures spécifiques et qui produisent des informations susceptibles de nous aider à tirer des conclusions. Étant donné la divergence existante quant aux performances des outils techniques disponibles sur le marché, nous suggérons fortement d'utiliser au moins deux procédures différentes et de comparer les résultats. Différents exemples qui utilisent cette perspective seront présentés à la section sur l'unidimensionalité et l'indépendance locale.

Les personnes qui désirent mener, dans un avenir plus ou moins rapproché, des analyses avec les modèles de la TRI seraient avisées de consulter le site Internet <www.assess.com> et celui de l'Institute for Objective Measurement (<www.rasch.org>) pour avoir un aperçu de ce qui est offert commercialement et à quel prix. À cet égard, à chaque fois que nous allons présenter une procédure pour vérifier les conditions d'utilisation, nous allons mentionner le ou les logiciels disponibles commercialement qui permettent de l'appliquer. Certaines des propositions de vérification des conditions demandent au préalable que les paramètres des modèles aient déjà été estimés, alors que les différentes stratégies d'estimation seront présentées au chapitre 6. Cela nous apparaît peu problématique dans le contexte actuel puisqu'il y a une certaine convergence des appréciations quant aux techniques à privilégier pour mener à bien l'opération d'estimation des paramètres.

Dans les deux sections qui suivent, nous allons donc présenter et définir certains concepts clés de la modélisation avec la TRI : l'invariance, l'ajustement, l'unidimensionalité et l'indépendance locale. Nous allons également y présenter certaines propositions visant à rendre opérationnelle la vérification des conditions d'application des modèles. Ces conditions apparaissent nettement distinctes lorsque nous les défilons ainsi les unes à la suite des autres ; toutefois, le lecteur observera un certain recouvrement entre les concepts et entre les façons de vérifier empiriquement que les conditions d'utilisation des modèles sont respectées.

À titre d'exemple, mentionnons que l'examen de la propriété d'invariance peut s'effectuer en étudiant l'ajustement du modèle avec différents sous-groupes de sujets. Il y a donc une relation étroite entre invariance et ajustement du modèle, mais ce n'est pas une relation d'équivalence parce que nous ne démontrons pas de cette manière que le modèle est exact dans la population.

Nous pourrions également croire que l'examen de l'ajustement d'un modèle unidimensionnel nous permet de conclure, par exemple, qu'une seule dimension est suffisante pour rendre compte des données. Mais, selon Van den Wollenberg (1988), il semble que certaines procédures d'examen de l'ajustement des données au modèle ne soient pas sensibles à la présence de plusieurs dimensions ou à des problèmes de dépendance entre les items. Il peut donc être nécessaire d'étendre la vérification en incluant d'autres procédures pour examiner spécifiquement l'unidimensionalité et l'indépendance locale.

De plus, comme la démonstration de l'unidimensionalité de l'ensemble des scores garantit théoriquement le respect de la condition d'indépendance locale, il semble superflu de vérifier la deuxième condition si la première est respectée. L'inverse ne serait cependant pas vrai : l'indépendance locale ne garantirait pas l'unidimensionalité. Toutefois, un auteur a proposé une approche où le respect de ce qu'il qualifie d'indépendance **essentielle** garantirait une unidimensionalité **essentielle** (Stout, 1987). D'autres auteurs ont proposé une approche pour vérifier si les données s'ajustent bien à un modèle monotone unidimensionnel avec indépendance locale (Holland et Rosenbaum, 1986). Il s'agirait donc d'une approche *omnibus* qui intègre à la fois l'ajustement, l'unidimensionalité et l'indépendance locale. Mais, cette approche n'est pas exempte de problèmes. Nous y reviendrons dans les exemples. Certains aspects sont aussi mieux documentés, car ils ont fait l'objet de nombreuses études et propositions étalées sur plusieurs décennies ; c'est le cas des études sur l'unidimensionalité et l'ajustement statistique, par exemple.

Rendu à cette partie de l'ouvrage, il apparaît important de préciser que le contenu des sections à venir de ce chapitre est celui qui, même s'il est difficile de prédire l'avenir, est le plus susceptible de devenir obsolète dans un avenir plus ou moins rapproché. Ce ne sont pas les modèles ou les concepts qui risquent de devenir obsolètes, mais bien les techniques développées pour vérifier le respect des conditions d'application et l'adéquation des modèles.

En effet, étant donné l'évolution rapide des technologies de traitement des données et l'augmentation de leur puissance, de même que l'accès à des banques de données de très grande taille, certains outils techniques qui apparaissent pertinents maintenant se retrouveront peut-être déclassés au profit d'outils plus conviviaux, mieux ancrés sur le plan théorique et plus performants sur le plan technique. Lorsque cela se produira, il s'agira d'une évolution normale des choses dans un secteur encore en plein développement.

Finalement, alors que le chapitre 4 présentait les modèles de base, les courbes caractéristiques et les principes généraux de la TRI, nous pouvons dire que cette section est celle de la confrontation de la modélisation à la réalité. Nous gardons toutefois en tête que la confrontation est réalisée avec certains des outils disponibles et, partant donc, avec les limites techniques et conceptuels qui s'y rattachent.

5.3. LA PROPRIÉTÉ D'INVARIANCE

La modélisation que l'on retrouve dans la TRI est comparable à une modélisation dans l'esprit de la régression en statistique. Dans ce cadre, elle procure une propriété théorique aux estimations des paramètres présents dans la représentation mathématique : la propriété d'invariance. Sous certaines conditions, les estimations du ou des paramètres associés aux items sont indépendantes du groupe de sujets qui est la cible de l'opération de mesure et les estimations du ou des paramètres associés aux sujets sont indépendantes du groupe d'items inclus dans l'opération de mesure.

Cette propriété existe pour tous les modèles de régression, mais les valeurs des coefficients de régression sont invariantes seulement si le modèle s'ajuste aux données pour l'ensemble de la population. Autrement dit, l'invariance est une propriété que l'on peut observer uniquement si nous avons accès à toute la population et si le modèle s'ajuste exactement pour la population.

Prenons le cas de deux variables aléatoires continues, X et Y, où X est la variable indépendante et Y la variable dépendante. La fonction de densité conjointe de ces deux variables est f(x, y) et la fonction de densité de Y conditionnelle à X est f(y | x). Pour une valeur de X donnée, x, la variable Y peut prendre plusieurs valeurs y. Un représentant possible de ces valeurs est l'espérance mathématique de Y étant donné la valeur x prise par X : $E(Y \mid X = x)$, c'est-à-dire la moyenne de Y étant donné une valeur x de X : $\mu_{Y \mid X}$.

L'ensemble des couples de points $[x_i, E(Y \mid X = x_i)]$, i = 1, …, n, décrit une courbe dans un espace à deux dimensions ; cette courbe est une courbe de régression. Elle représente la régression de Y sur X. Si le modèle est

exact dans la population, alors la régression de Y sur X est indépendante de la distribution de X et elle est invariante d'un groupe de valeurs x_{1i} à un autre ensemble de valeurs x_{2i} prises par la variable X.

Si nous supposons que la relation entre X et Y est linéaire, cette relation peut être représentée par le modèle $Y = \alpha X + \beta + \varepsilon$, où α et β sont les coefficients de régression et ε est une variable aléatoire représentant le résidu de l'ajustement du modèle linéaire avec les paramètres α et β. En supposant que $E(\varepsilon) = 0$, alors $E(Y \mid X = x) = \alpha X + \beta$. Dans ce cas, lorsque pour des valeurs estimées de α et β, donc pour les valeurs prises par $\hat{\alpha}$ et $\hat{\beta}$, l'hypothèse de la relation linéaire est confirmée pour la population, la relation sera la même peu importent les valeurs prises par la variable X. En d'autres mots, les valeurs pour $\hat{\alpha}$ et $\hat{\beta}$ posséderont la propriété d'invariance.

Pour la modélisation de la TRI, nous pouvons illustrer le tout en prenant le cas particulier de la situation où les scores sont dichotomiques (bonne réponse = 1, mauvaise réponse = 0, par exemple). Soit U_i la variable qui représente le score observé (1 ou 0) pour l'item i et $P_i(\theta)$, $Q_i(\theta)$, les probabilités respectives d'obtenir les résultats 1 ou 0 étant donné une position donnée sur le continuum d'habileté : θ $P_i(\theta) = P(U_i = 1 \mid \theta)$, $Q_i(\theta) = P(U_i = 1 \mid \theta)$.

Supposons que la fonction de probabilité de U_i est donnée par une loi de probabilité de Bernouilli : $f_i(U_i \mid \theta) = \begin{cases} P_i(\theta) \ldots si \ldots u_i = 1 \\ \\ Q_i(\theta) \ldots si \ldots u_i = 0 \end{cases}$.

Alors, la régression du résultat pour l'item i, U_i, sur l'habileté θ est donnée par : $E(U_i / \theta) = [P_i(\theta) \times 1] + [Q_i(\theta) \times 0] = P_i(\theta)$. La régression de U_i sur l'habileté θ est donc la fonction caractéristique (ou courbe caractéristique, voir le chapitre 4) de l'item i et, si le modèle est exact dans la population, les estimations des paramètres décrivant la fonction caractéristique sont invariantes.

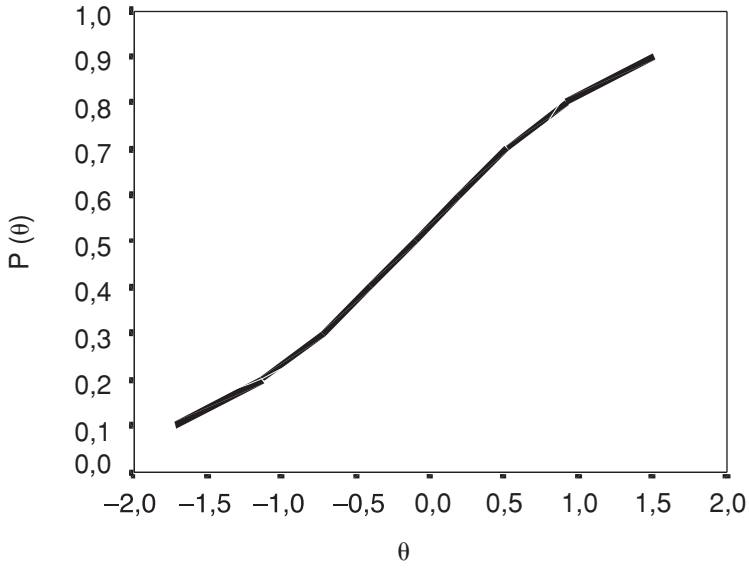
L'exemple suivant repris de Hambleton *et al.* (1991, p. 20-23) permettra au lecteur de mieux saisir la réalité de cette propriété. Supposons que nous connaissons exactement la probabilité de succès à un item pour différents niveaux d'habileté tel que présenté au tableau 5.2 et à la figure 5.1, et que nous désirons ajuster un modèle logistique avec deux paramètres à ces données[1].

Tableau 5.2
Niveaux d'habileté et probabilités correspondantes

θ	–1,716	–1,129	–0,723	–0,398	–0,100	0,198	0,523	0,919	1,516
$P(\theta)$	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9

1. Évidemment nous ne connaissons jamais θ avant d'estimer les paramètres du modèle. L'exemple n'est donc pas réaliste, mais nous croyons qu'il a une certaine utilité pédagogique.

FIGURE 5.1
Représentation graphique de la relation entre θ et P(θ).



Ainsi que nous l'avons vu au chapitre 4, le modèle logistique à deux paramètres nous est donné par l'équation suivante :

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

Par une simple manipulation, nous pouvons donc produire le rapport des chances (*odds ratio*) et son logarithme naturel :

$$\frac{P}{1-P} = e^{Da(\theta-b)}$$

$$ln\frac{P}{1-P} = Da(\theta-b) = \alpha\theta - \beta$$

où $\alpha = Da$ et $\beta = -Dab$.

Puisque nous connaissons P(θ) et θ, nous pouvons trouver facilement les valeurs de α et β en substituant, par exemple, les valeurs les plus extrêmes. Ainsi, en choisissant θ = –1,129 et θ = 0,919, avec les valeurs correspondantes pour P(θ) de 0,2 et 0,8 respectivement, nous obtenons :

$$ln\frac{0,2}{0,8} = \alpha(-1,129) + \beta \text{ et } ln\frac{0,8}{0,2} = \alpha(0,919) + \beta$$

En éliminant temporairement β nous obtenons :

$$ln\frac{0,8}{0,2} - ln\frac{0,2}{0,8} = \alpha(,919) - \alpha(-1,129)$$

En isolant α, nous obtenons α = 1,36 et en substituant cette valeur nous trouvons β = 0,136. Si nous procédons de la même manière, mais en utilisant des valeurs rapprochées pour P(θ) et θ, par exemple 0,1 et 0,2 pour P(θ), donc –1,716 et –1,129 pour θ, nous obtenons :

$$ln\frac{0,2}{0,8} = \alpha(-1,129) + \beta \text{ et } ln\frac{0,3}{0,7} = \alpha(-0,723) + \beta$$

De la même manière que précédemment, nous éliminons β pour obtenir :

$$ln\frac{0,3}{0,7} - ln\frac{0,2}{0,8} = \alpha(-,723) - \alpha(-1,129)$$

et nous trouvons à nouveau α = 1,36 et β = 0,136. À partir de ces valeurs de α et β, nous retrouvons facilement les valeurs des paramètres a et b du modèle logistique. Donc, peu importe l'endroit où nous nous situons sur le continuum de l'habileté, les estimations des paramètres seront toujours les mêmes.

En fait, nous avons simplement démontré que α et β sont respectivement la pente et l'ordonnée à l'origine de la droite qui décrit la relation entre $ln\frac{P}{1-P}$ et θ. Pour toutes les valeurs de θ la droite est la même ; donc, les valeurs de α et β, de même que celles de a et b, ne changent pas, peu importe l'endroit où le sujet se situe sur le continuum d'habileté θ.

Évidemment, la démonstration est quelque peu artificielle. Dans la réalité, si le modèle ne s'ajuste pas exactement dans la population (ce qui est le cas généralement), alors la relation entre $ln\frac{P}{1-P}$ et θ ne sera pas linéaire et nous obtiendrons différentes valeurs pour les estimations des paramètres.

La propriété d'invariance est principalement ce qui permet à la TRI d'étaler sa supériorité par rapport aux autres propositions de modélisation, comme la théorie classique des tests ou la théorie de la généralisabilité. En fait, la TRI et la propriété d'invariance qui l'accompagne constituent la modélisation la plus appropriée pour les développements contemporains du testing en éducation et en psychologie. Les deux dernières décennies ont en effet vu surgir ce que Van der Linden (1986) a qualifié de nouveau complexe dans le domaine du testing. En effet, ce que nous recherchons n'est plus lié à la mise au point d'un test constitué d'un groupe d'items fixes et invariables en contenu et en nombre. Nous cherchons plutôt à modéliser la rencontre entre un item et un sujet. Les préoccupations sont ainsi de l'ordre de la mise au point de banques d'items où chaque item d'une banque est en quelque sorte indépendant des autres items de la banque.

En principe, donc, n'importe quel item pourrait être extrait de la banque pour être inséré dans un test destiné à un sujet ou à un groupe de sujets. Il s'agit d'ailleurs du principe à la base du développement des tests adaptatifs informatisés. Lors d'un test adaptatif, chaque sujet est susceptible d'être exposé à des items différents, tant en nombre qu'en contenu ; il est donc nécessaire de recourir à une modélisation qui procure pour des sujets différents des estimations de l'habileté qui soient situées sur le même continuum. Il faut recourir à une modélisation où l'estimation de l'habileté est indépendante des items auxquels le sujet est exposé pendant la séance de testing (voir au chapitre 9 la description des principales caractéristiques du testing adaptatif informatisé et de l'application de la TRI à cette forme de testing). Il s'agit précisément d'un des avantages importants qu'offre la modélisation avec la TRI. Ainsi par exemple, après une opération de calibrage d'un groupe d'items avec un modèle logistique à deux paramètres, nous pouvons dire que l'item numéro 18 se situe au point 1,23 sur le continuum de la difficulté et, le cas échéant, au point 1,02 sur celui de la discrimination[2]. La propriété d'invariance permettrait d'affirmer que l'item 18 sera toujours situé à cette position sur les deux continuums peu importent les nouveaux sujets qui devront eux aussi fournir une réponse à l'item 18 lors d'une confrontation ultérieure et peu importe avec quels autres items l'item 18 sera utilisé.

Cependant, lorsque nous avons affaire à des échantillons, nous n'observons généralement pas une relation linéaire entre $\frac{P}{1-P}$ et θ. En effet, même si, selon la théorie, $P(\theta) = E(U \mid \theta)$, c'est-à-dire que $P(\theta)$ est la moyenne de

2. Avant d'intégrer un item à une banque d'items, il faut procéder à une opération de **calibrage** de l'item, c'est-à-dire qu'il faut étiqueter l'item avec des valeurs qui correspondent à des caractéristiques recherchées d'un point de vue métrique. Ainsi, il arrive souvent que les items soient étiquetés selon leur niveau de difficulté et/ou selon leur puissance de discrimination. Ces valeurs servent alors de référence pour toute nouvelle utilisation de l'item. Dans le cas de la TRI, les items sont **indexés** en fonction des estimations des paramètres du modèle privilégié et ce sont ces valeurs qui servent de référence lors de toute nouvelle utilisation de l'item.

toutes les réponses observées pour les sujets se situant au point θ, il serait exceptionnel dans des échantillons que la probabilité observée soit identique à $E(U \mid \theta)$ pour chacun des points sur le continuum de l'habileté.

La présence d'un modèle exact dans la population étant l'exception plutôt que la norme, il faut mettre en place différentes procédures de vérification de la propriété d'invariance avant de prendre des décisions suite à une application de la modélisation. La première condition pour que la propriété d'invariance puisse s'afficher est évidemment que le modèle s'ajuste aux données. Nous disons **évidemment**, parce qu'il est évident que si nous arrivons à la conclusion que le modèle n'est pas le bon pour un item donné, il ne sert à rien de poursuivre avec ce modèle pour cet item. Il faut soit changer de modèle, soit modifier l'item et reprendre l'opération de calibrage des paramètres du modèle pour cet item. Nous abordons d'ailleurs à la section suivante un certain nombre de suggestions pour la vérification de la qualité de l'ajustement du modèle aux données.

Si l'ajustement est acceptable pour l'ensemble des sujets et que nous réussissons à montrer que tout ajustement subséquent avec des sous-échantillons qui regroupent les sujets en fonction d'une variable ciblée (le sexe, par exemple) procure des valeurs des estimations des paramètres qui sont, à une transformation linéaire près, les mêmes que pour l'ensemble du groupe, alors nous confirmons, d'une certaine façon, la propriété d'invariance des paramètres du modèle. Il s'agit ici en fait de vérifier la correspondance entre des valeurs des paramètres lorsque ceux-ci sont estimés à partir de différents sous-échantillons d'individus. Si la relation linéaire entre deux ensembles de paramètres ne tient pas, cela sèmera peut-être le doute quant au respect de la propriété d'invariance, mais une condition plus faible, comme le respect de la relation d'ordre entre les deux ensembles de valeurs, pourrait tout de même être un argument en faveur d'une forme d'invariance.

Les divisions en sous-échantillons que nous pouvons étudier sont multiples et, de façon générale, il vaut mieux se concentrer sur des caractéristiques apparentes et pertinentes de nos sujets ou des scores. Ainsi, différentes estimations des paramètres du modèle employé pourraient être obtenues selon un regroupement des sujets en fonction des différents scores observés, du sexe des sujets, de l'origine ethnique, du milieu socioéconomique, etc. Cette stratégie est proposée par plusieurs auteurs, par exemple : Lord (1980), Wright et Masters (1982), Hambleton et Murray (1983), Van den Wollenberg (1988), Hambleton *et al.* (1991), Embretson et Reise (2000).

Une façon simple de vérifier si deux ensembles de paramètres sont en relation linéaire consiste à représenter la relation par un diagramme de dispersion et à résumer la relation par la valeur du coefficient de corrélation de Pearson. La figure 5.2 représente la relation (fictive) entre deux ensembles d'estimations du paramètre de difficulté du modèle de Rasch pour un ensemble de 50 items. Nous observons que la plupart des points se retrouvent sur la

droite et que très peu sont situés très à l'écart de la droite. La valeur du coefficient de corrélation entre les deux ensembles est de 0,99. Au contraire, si nous examinons la figure 5.3, nous observons un nuage de points plutôt distant de la droite et une corrélation de 0,52. Dans le premier cas nous pourrions conclure au respect de la propriété d'invariance, alors que dans le deuxième cas nous constaterions qu'il y a suffisamment de perturbations pour conclure que la propriété ne tient pas. À partir du moment où nous concluons que la propriété d'invariance ne tient pas, il faut emprunter la voie des études diagnostiques pour tenter de déterminer la cause du problème. Le problème peut être causé par le fait que l'ajustement des données n'est pas idéal et qu'il serait préférable de modéliser avec un modèle plus complet (avec deux ou trois paramètres par exemple). Des problèmes de multidimensionalité ou encore de dépendance entre les items, de biais de mesure, peuvent aussi compter parmi les raisons pouvant contribuer à l'explication de la situation.

FIGURE 5.2
Relation entre deux ensembles d'estimations d'un paramètre (r = 0,99).

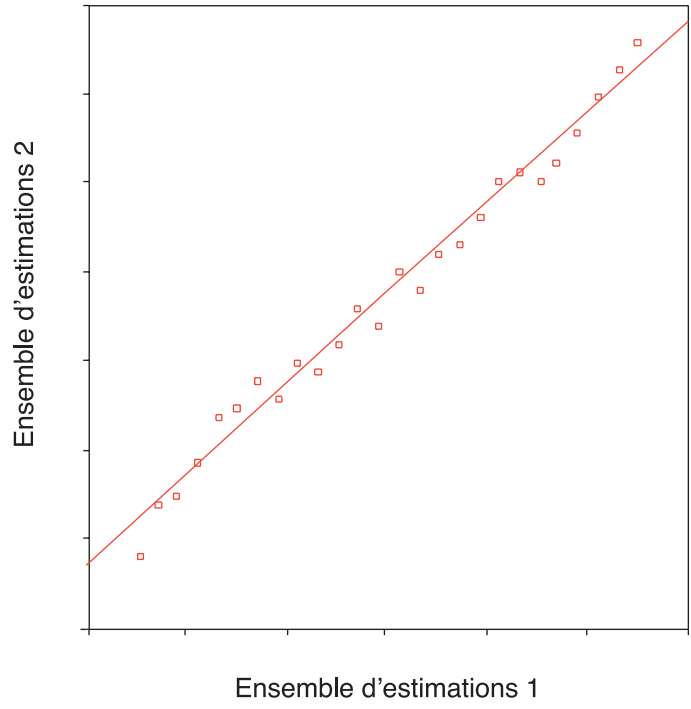
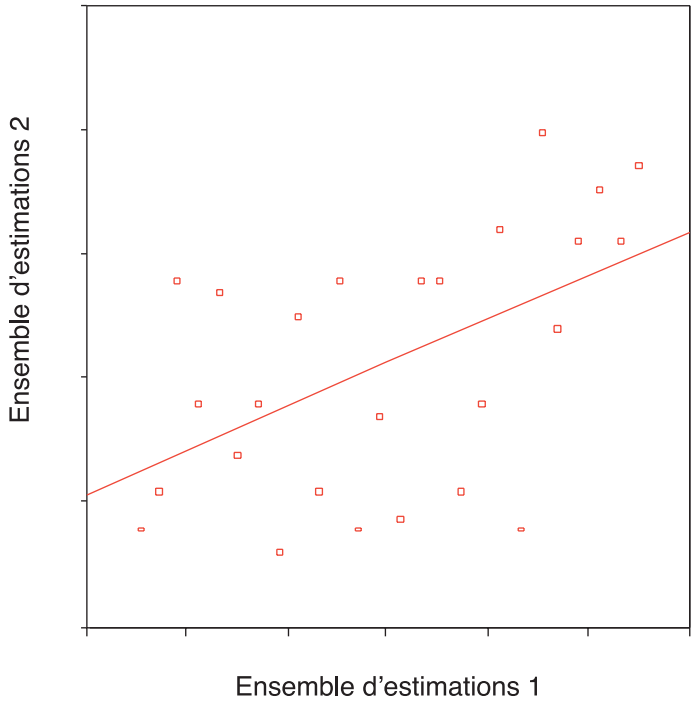


FIGURE 5.3
Relation entre deux ensembles d'estimations d'un paramètre (r = 0,52).



Parmi les approches qui s'éloignent de la description de la relation entre deux ensembles de paramètres, nous retrouvons la proposition de Van den Wollenberg (1988) pour tester statistiquement l'indépendance échantillonnale lorsque le modèle de Rasch est utilisé. Cet auteur a proposé une statistique Q_1 que l'on peut appliquer en séparant l'échantillon de départ en sous-échantillons en fonction de chacun des scores pour l'ensemble des items du test utilisé. La statistique Q_1 est construite à partir de la fonction de vraisemblance du test total et des fonctions de vraisemblance qui correspondent à chacun des r = 0, …, k – 1 scores observés. En effet, Andersen (1972) a montré que si le modèle de Rasch est adéquat pour un ensemble de k items, alors le rapport de vraisemblance λ, où L est la fonction de vraisemblance pour l'ensemble des sujets et L_r est la fonction de vraisemblance pour le sous-échantillon de sujets obtenant un score r pour l'ensemble des items, devrait être égal à 1 :

$$\lambda = \frac{L}{\prod L_r}$$

Nous pouvons montrer que la quantité empirique $Z = -2\ln\lambda$ suit une distribution du khi-carré avec $(k-1) \times (k-2)$ degrés de liberté. Comme le rapport de vraisemblance λ s'éloigne de 1 lorsque les estimations du paramètre de difficulté diffèrent d'un groupe à l'autre, la statistique Z permettrait de tester l'invariance des estimations du paramètre de difficulté en fonction des différents scores au test. En s'inspirant de ce résultat d'Andersen et des travaux de Martin-Lof (1974), Van den Wollenberg a proposé la statistique Q_1 qui suit également une distribution du khi-carré avec $(k-1) \times (k-2)$ degrés de liberté.

$$Q_1 = \frac{k-1}{k} \sum_i q_i$$

où

$$q_i = \sum_r^{k-1} \left\{ \frac{[n_{ri} - E(n_{ri})]^2}{E(n_{ri})} + \frac{[n_{ri} - E(n_{ri})]^2}{n_r - E(n_{ri})} \right\},$$

n_{ri} correspond au nombre de sujets qui ont obtenu un score r au test et qui ont répondu correctement à l'item i, $E(n_{ri})$ correspond à la valeur attendue pour n_{ri} et n_r correspond au nombre total de sujets ayant obtenu le score r.

La statistique du rapport de vraisemblance d'Andersen et la statistique Q_1 de Van den Wollenberg ont été intégrées au programme RSP (*Rasch Scaling Program*) mis en marché par la firme Assessment System Corporation, de sorte qu'elles pourraient être mises à contribution pour étudier la propriété d'invariance et nous permettre de compléter le diagnostic établi avec les représentations graphiques et la valeur du coefficient de corrélation.

5.4. L'AJUSTEMENT DU MODÈLE AUX DONNÉES

Nous avons mentionné à la section précédente que la vérification de la propriété d'invariance demande au préalable une étude de la qualité de l'ajustement du modèle aux données. En effet, le modèle choisi pour une application donnée ne constitue qu'une hypothèse parmi d'autres pour formaliser la relation entre la probabilité d'observer une réponse quelconque à un item et la position du candidat sur le continuum de l'habileté thêta. Évidemment, certains modèles possèdent des propriétés qui les rendent plus désirables que d'autres dans certaines situations, ce qui fait que nous ne pouvons pas a priori mettre tous les modèles sur le même pied.

À partir du moment où nous avons sélectionné le modèle qui, à la lumière de ses caractéristiques et de ses propriétés, nous apparaît le plus approprié pour une situation donnée, il faut faire la démonstration empirique

que celui-ci est effectivement approprié pour représenter les données. Cette démonstration s'inscrit dans le courant général des travaux de recherches sur l'ajustement statistique des modèles aux données (*goodness-of-fit*) et sur l'étude des résidus.

L'appréciation de l'ajustement d'un modèle de la TRI peut être dirigé vers un examen de l'ajustement global, de l'ajustement pour chaque item (*item fit*) ou de l'ajustement pour chaque sujet (*person fit*). Dans cette section, nous allons centrer notre propos sur l'ajustement du modèle aux items, non pas parce que l'étude de l'ajustement pour les sujets est dénuée d'intérêt, mais plutôt parce que, d'une part, les développements de l'ajustement en fonction des items sont plus complets et mieux documentés à l'heure actuelle (sans parler de la présence des procédures dans les logiciels d'analyse) et, d'autre part, parce que les travaux sur l'ajustement en fonction des personnes montrent beaucoup de parenté avec ce qui est fait pour vérifier l'ajustement au niveau des items. De plus, la question de l'ajustement global revient surtout lorsque nous désirons vérifier, par exemple, l'unidimensionalité de l'ensemble des scores ou l'indépendance locale qui sont l'objet de la section suivante.

L'étude de l'ajustement du modèle est un aspect important de la TRI et cet aspect est abondamment couvert dans les écrits. En fait, le lecteur averti aura observé que pratiquement toutes les fois qu'un auteur présente un modèle de la TRI, cette présentation s'accompagne automatiquement d'une description de la méthode d'estimation des paramètres du modèle (s'il s'agit d'un modèle paramétrique, évidemment) et d'une proposition d'estimation de la qualité de l'ajustement du modèle aux données. À titre de référence à ce sujet, nous pouvons mentionner le *Handbook* de Van der Linden et Hambleton (1997) où nous retrouvons 27 propositions de modèles, autant de propositions pour estimer les paramètres et presque autant de propositions pour apprécier l'ajustement.

Nous n'avons pas l'intention dans cette section de faire une description exhaustive des multiples propositions visant à documenter la qualité de l'ajustement. Nous allons plutôt présenter les directions générales qu'empruntent la plupart des auteurs pour développer les représentations graphiques et les indices statistiques de la qualité de l'ajustement.

Les approches pour vérifier l'ajustement du modèle au niveau des items sont des approches qui permettent de porter un jugement sur la qualité de la prédiction. Ce jugement se fonde généralement sur deux types d'analyse : d'abord, une première forme d'analyses qui sont graphiques et visuelles et qui reposent sur l'examen de la différence entre la courbe produite par le modèle et la courbe empirique observée à partir des données, de même que sur l'étude de la distribution des résidus (la différence entre le produit de la modélisation et ce qui est observé avec les données) ; ensuite, une deuxième forme d'analyses qui proposent des statistiques uniques, à la manière des tests statistiques, pour rejeter ou ne pas rejeter l'ajustement du modèle à chacun des items. Dans cette deuxième perspective nous retrouvons également des propositions dont

l'intérêt est orienté vers la comparaison de modèles et qui mènent à des tests statistiques globaux pour tous les items d'un test sur les gains dans l'ajustement d'un modèle par rapport à un autre. Nous aborderons quelques-uns de ces tests à la section suivante sur l'unidimensionalité, notamment dans le contexte de l'analyse factorielle complète du patron de réponse.

5.4.1. L'ajustement graphique

Les suggestions de représentations graphiques sont en droite ligne avec ce qui est prôné pour les modèles de régression. D'abord, il est suggéré d'examiner graphiquement la relation entre la courbe théorique produite par un modèle dont les paramètres ont été estimés et une courbe empirique construite à partir des résultats observés. Les valeurs qui correspondent à ce qui est observé sont en fait des proportions de sujets qui ont répondu correctement à l'item (ou qui se sont vu attribuer une cote donnée, si nous cherchons à mesurer les attitudes, par exemple) et qui ont été regroupés en fonction de leur position sur le continuum de l'habileté thêta. Nous devons dans un premier temps procéder à l'estimation des paramètres et ensuite estimer la position de chacun des sujets sur le continuum de l'habileté qui s'étend généralement de –3,00 à +3,00 (s'il n'y a qu'un seul paramètre associé aux sujets, évidemment[3]). Les sujets sont ensuite regroupés en catégories mutuellement exclusives et exhaustives dont les effectifs sont à peu près égaux. À cet effet, Hambleton (1989) a suggéré d'utiliser entre dix et quinze catégories, Hambleton *et al.* (1991) ont suggéré d'utiliser douze catégories et Embretson et Reise (2000) ont suggéré d'utiliser dix catégories. Kingston et Dorans (1985) ont exploré cette avenue avec des données provenant de la passation du test GRE (*Graduate Record Examination*) et ils ont décidé d'utiliser quinze catégories. Les suggestions quant au nombre de catégories ne constituent pas des règles fixes : elles dépendent en premier lieu du nombre total de sujets et du nombre de sujets qui se situent aux deux extrémités du continuum d'habileté thêta.

Les diagrammes de la figure 5.4 illustrent deux situations différentes où les données sont modélisées avec le modèle logistique à deux paramètres (diagramme a) et le modèle logistique à trois paramètres (diagramme b). Nous observons qu'en ajoutant un paramètre nous pouvons tenir compte du fait qu'avec les items à réponse choisie il est toujours possible de deviner la bonne réponse. Ainsi, la borne inférieure de $P(\theta)$ dans ces situations ne sera jamais égale à zéro, elle se situera plutôt près de 1/m, où m est le nombre de choix de réponses. Dans une situation de test avec des items à réponse choisie le modèle à trois paramètres produit donc une amélioration de l'ajustement.

3. Nous verrons au chapitre suivant que dans certaines situations les paramètres associés aux items et les paramètres associés aux sujets doivent être estimés simultanément, alors que dans d'autres situations les paramètres associés aux items sont déjà estimés et servent à estimer les paramètres pour les sujets.

Nous pouvons représenter graphiquement les valeurs de la proportion attendue et de la proportion observée de sujets, $E(p_{ij})$, et p_{ij} ayant répondu correctement à l'item i et se situant dans la catégorie j des résultats regroupés. Lorsque la distance entre $E(p_{ij})$ et p_{ij} est grande, nous pouvons être en présence d'un modèle qui s'ajuste mal ; notre regard devrait alors se tourner du côté des modèles plus complexes qui produiront un ajustement aux données supérieur. En effet, il peut s'agir d'un problème de multidimensionalité, d'un problème touchant la relation entre $P(\theta)$ et θ qui n'est pas monotone croissante, d'un problème causé par la présence d'un sous-échantillon de sujets trop différent de l'ensemble ou causé par un item mal conçu, etc. Cela ne veut pas dire qu'un modèle avec plus de paramètres doit nécessairement être envisagé si le modèle plus simple semble s'ajuster moins bien. D'autres considérations peuvent également entrer en ligne de compte. Ainsi, nous pouvons privilégier l'ajustement des données au modèle de Rasch, modèle à un paramètre, parce que nous voulons bénéficier de l'ouverture que permet ce modèle sur la mesure fondamentale et, ainsi, éliminer les items qui ne s'ajustent pas bien au modèle, plutôt que de changer de modèle.

Nous le rappelons, il y a plusieurs directions que peut prendre l'étude diagnostique de l'ajustement et il n'y a pas de recette permettant de jeter un seul regard sur le problème. La multiplicité des regards est la meilleure des garanties pour une utilisation adéquate des modèles de la TRI.

La différence entre $E(p_{ij})$ et p_{ij} est appelée le résidu de la modélisation ; c'est ce qui reste après l'ajustement du modèle : $r_{ij} = p_{ij} - E(p_{ij})$. Les résidus peuvent être étudiés graphiquement de la même manière qu'ils le sont dans le cadre des études de régression. Par exemple, ils peuvent être d'abord standardisés et ensuite confrontés aux valeurs prises sur le continuum d'habileté thêta. Le résidu standardisé RS_{ij} est donné par :

$$RS_{ij} = \frac{p_{ij} - E\left(p_{ij}\right)}{\sqrt{\frac{p_{ij}(1-p_{ij})}{N_j}}}$$

La figure 5.5 présente deux situations où, d'une part, la distribution des résidus standardisés indique qu'il y a un problème pour les valeurs élevées du continuum d'habileté thêta pour l'item 1 (diagramme a) et, d'autre part, où les résidus pour l'item 2 semblent bien répartis de part et d'autre de la moyenne des résidus (diagramme b). Le diagnostic serait donc qu'il y a un problème d'ajustement pour l'item 1 et que l'item 2 bénéficie d'un ajustement adéquat.

FIGURE 5.4
Représentations graphiques de situations où les modèles à un ou trois paramètres sont appropriés.

Diagramme A : deux paramètres

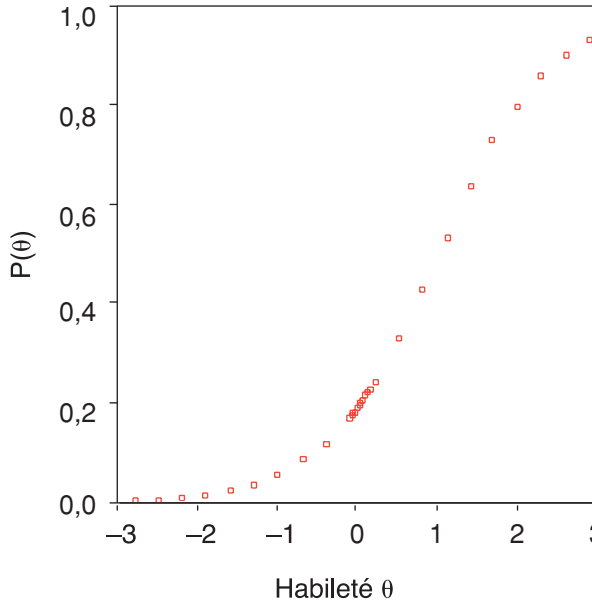


Diagramme B : trois paramètres

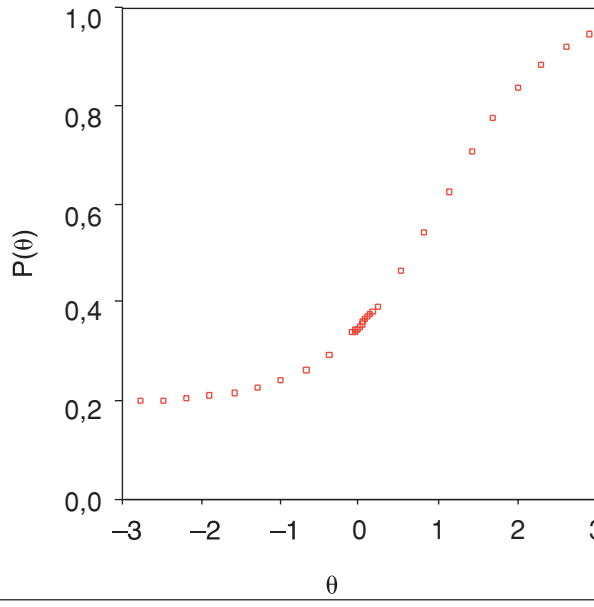
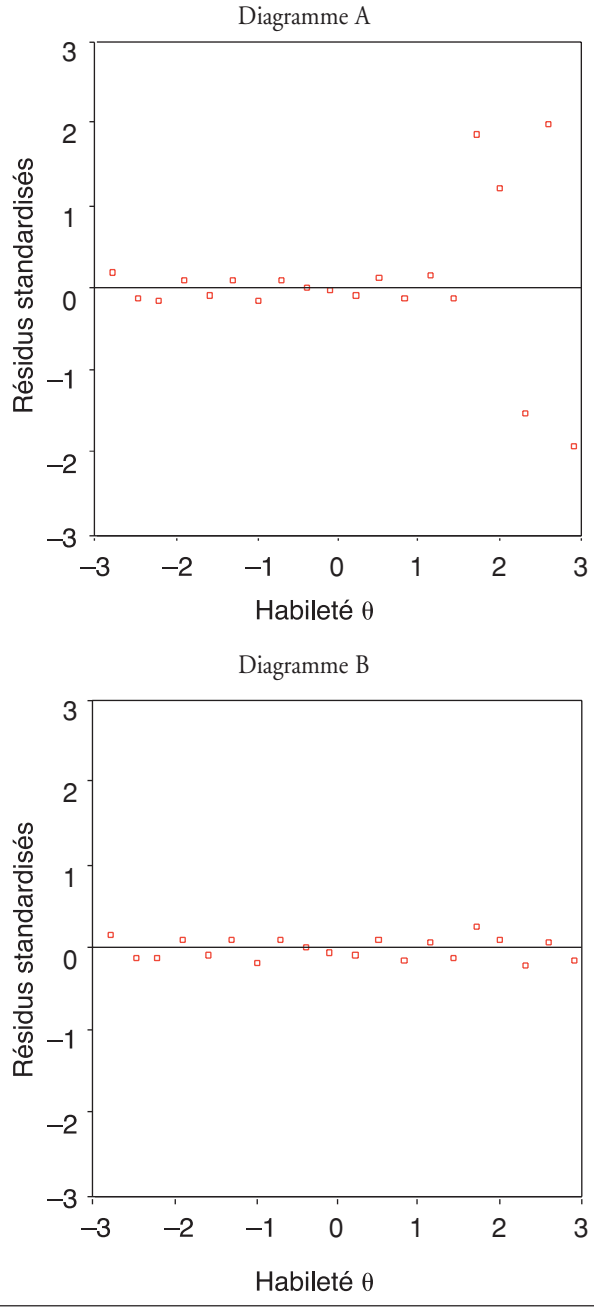


FIGURE 5.5
Diagrammes des résidus standardisés en fonction de l'habileté θ



5.4.2. L'ajustement statistique pour les items

Si nous désirons aller au-delà d'une exploration visuelle, nous pouvons exploiter certaines statistiques et les tests qui les accompagnent. Ces outils ont été mis au point par les chercheurs dans leur quête d'une approche plus formelle pour comparer les distributions théoriques et les distributions observées. Dans la plupart des approches, les tests statistiques font appel aux résidus, $r_{ij} = p_{ij} - E(p_{ij})$, et à un regroupement des sujets en fonction de leur position sur le continuum d'habileté thêta.

Ainsi, Bock (1972) accompagnait sa présentation du modèle nominal d'une statistique, BCHI, dont la distribution est celle du khi-carré pour tester l'ajustement du modèle :

$$\text{BCHI} = \sum_{j=1}^{J} \frac{N_j \left(O_{ij} - E_{ij}\right)^2}{E_{ij}\left(1 - E_{ij}\right)},$$

où O_{ij} et E_{ij} correspondent respectivement à la proportion observée de réponses endossées et à la proportion attendue selon le modèle pour l'item i et la catégorie j. N_j correspond au nombre de sujets dans la catégorie j. La statistique BCHI possède une distribution du khi-carré avec J – m degrés de liberté où J est le nombre de catégories et m le nombre de paramètres estimés. Yen (1981) a proposé la même statistique, appelée Q_1, mais avec une légère variante quant à la façon de regrouper les sujets dans les catégories.

Les logiciels BILOG-3 (Mislevy et Bock, 1990) et BILOG-MG (Zimowski *et al.*, 1996) proposent également des statistiques du même type. Nous y retrouvons trois statistiques dont la justesse dépend du nombre d'items. Les auteurs proposent ainsi une statistique lorsqu'il y a dix items ou moins, une statistique lorsqu'il y a entre dix et vingt items et une dernière statistique lorsqu'il y a plus de vingt items. Pour le cas où il y a plus de vingt items, la statistique suit une loi du khi-carré avec un nombre de degrés de liberté égal au nombre de catégories formées suite au regroupement des sujets ; elle est donnée par l'équation suivante :

$$G_i^2 = 2\sum_{j=1}^{J}\left[O_{ij}\, log_e \frac{O_{ij}}{N_j E_{ij}} + \left(N_j - O_{ij}\right) log_e \frac{\left(N_j - O_{ij}\right)}{N_j\left(1 - E_{ij}\right)}\right]$$

Les statistiques d'ajustement présentées ci-dessus sont valables pour les modèles à un, deux ou trois paramètres dans la situation où les réponses sont dichotomiques, mais certaines adaptations existent pour les situations où les réponses ne sont pas du type bonne ou mauvaise réponse (voir Van der Linden et Hambleton, 1997).

Pour le modèle de Rasch, Wright et Masters (1982), de même que Masters et Wright (1997), ont proposé une approche semblable à ce qui est présenté ci-dessus. La proposition ne concerne pas seulement le modèle le plus simple où les réponses sont dichotomiques, mais tous les types de modèles de la famille de Rasch, dont les modèles polytomiques. Nous retrouvons d'ailleurs une extension de cette approche intégrée aux logiciels QUEST (Adams et Khoo, 1992) et CONQUEST (Wu *et al.*, 1998) aussi mis en marché par la firme Assessment System Corporation. Seulement, au lieu de se servir du résidu formé par la différence entre la proportion observée et la proportion espérée de sujets ayant répondu dans une catégorie de réponse donnée, Wright et Masters se sont servis du résidu qui correspond à la différence entre x_{ni}, la réponse du sujet n à l'item i, et l'espérance de x_{ni}.

Le résidu est donc obtenu en calculant $r_{ni} = x_{ni} - E_{ni}$, avec $E_{ni} = \sum_{k=0}^{m} kP_{nik}$. La quantité P_{nik} correspond à la valeur attendue de x_{ni}, c'est-à-dire à la probabilité théorique obtenue à partir du modèle pour le sujet n de répondre dans la catégorie k de l'item i. La variance et le résidu standardisé pour x_{ni} sont données respectivement par :

$$\sum_{k=0}^{m} (k - E_{ni})^2 P_{nik}$$

et

$$z_{ni} = \frac{r_{ni}}{\sqrt{W_{ni}}}.$$

Une stratégie possible pour résumer l'ajustement consiste à produire la moyenne non pondérée des carrés des résidus (*mean square average*). Nous obtenons alors :

$$u_i = \sum_{n=1}^{N} \frac{z_{ni}^2}{N}$$

Lors d'applications cependant, cette somme non pondérée peut se révéler sensible aux valeurs extrêmes (*outliers*). Pour cette raison, Wright et Masters suggèrent qu'il peut être préférable d'utiliser une moyenne pondérée pour l'étude de l'ajustement. Cette moyenne pondérée et sa variance sont respectivement :

$$v_i = \frac{\sum_{n=1}^{N} W_{ni} z_{ni}^2}{\sum_{n=1}^{N} W_{ni}}$$

et

$$q_i^2 = \frac{\sum_{n=1}^{N} \left(C_{ni} - W_{ni}^2 \right)}{\left(\sum_{n=1}^{N} W_{ni} \right)^2}$$

avec[4]

$$\sum_{k=0}^{m} \left(k - E_{ni} \right)^4 P_{nik}$$

Les statistiques u_i et v_i sont distribuées approximativement comme des lois de probabilité du khi-carré et il est possible de les transformer pour obtenir des statistiques dont la distribution est approximativement normale. Ainsi, Wu (1997) a utilisé la procédure de Wilson-Hilferty pour produire les statistiques t_{out} et t_{in} qui sont approximativement distribuées selon une loi normale avec une moyenne près de 0 et une variance près de 1.

$$t_{out} = \frac{\left(u_i^{1/3} - 1 + \frac{2}{9rN} \right)}{\left(\frac{2}{9rn} \right)^{1/2}}$$

et

$$t_{in} = \left[\left(v_i^{1/3} - 1 \right) \left(\frac{3}{q_i} \right) \right] + \frac{q_i}{3},$$

où r est un nombre qui dépend de la procédure d'estimation de u_i.

Les valeurs des statistiques non standardisées u_i et v_i seraient susceptibles d'indiquer des problèmes d'ajustement lorsqu'elles sont plus grandes que 1,3 pour des échantillons de moins de 500 sujets, lorsqu'elles sont plus grandes que 1,2 pour des échantillons qui comprennent entre 500 et 1000 sujets, et lorsqu'elles sont plus grandes que 1,1 pour des échantillons de plus de 1000 sujets (selon Smith, Schumacher et Bush, 1998).

4. C_{ni} est la voussure (ou kurtose) de la distribution de x_{ni}.

Pour les statistiques standardisées t_{out} et t_{in}, il est recommandé d'étudier de plus près les items pour lesquels les valeurs des statistiques sont à l'extérieur de l'intervalle [–2, +2], qui correspond *grosso modo* à un intervalle avec un niveau de confiance de 95 %.

5.4.3. Des problèmes qui subsistent

Malgré ces développements intéressants du côté des statistiques et des tests d'ajustement, deux problèmes importants subsistent. D'une part, avec l'approche des tests statistiques, le modèle est toujours susceptible d'être rejeté si le nombre de sujets est assez élevé. Ce problème n'est pas unique aux tests d'ajustement pour les modèles de la TRI ; nous le retrouvons pour tous les tests statistiques pour lesquels nous ne pouvons formaliser la puissance du test et donc estimer si ce qu'on observe est trivial ou relié à un problème d'ajustement réel[5]. Il est donc important de ne pas se fier uniquement aux résultats des tests statistiques inclus dans les logiciels pour prendre une décision au sujet de la qualité de l'ajustement d'un modèle. D'autre part, si les catégories de sujets créées pour comparer les proportions observées et les proportions attendues le sont à partir de l'échelle thêta, cela signifie que la statistique doit être ajustée pour tenir compte du fait que l'habileté n'est pas connue, mais estimée (voir Orlando et Thissen, 2000). Il faut noter également que le nombre d'items peut aussi avoir un impact sur les valeurs prises par les statistiques d'ajustement.

De plus, lorsque nous utilisons x_{ni}, le score observé pour le sujet n et l'item i, pour produire les résidus, nous utilisons une variable discrète, avec 0 ou 1 comme événements possibles dans le cas dichotomique. La valeur attendue E_{ni} est toutefois une variable continue avec $0 < E_{ni} < 1$. Ainsi, les deux variables sont incompatibles théoriquement et la **vraie** distance entre les deux termes ne peut jamais, à strictement parler, être obtenue (voir à ce sujet Bond et Fox, 2001 ; Kabaratsos, 1999, 2000). Les tests statistiques ne seraient donc pas appropriés dans cette approche. Une solution pour ce problème, en lieu et place de l'étude des résidus, passe par une référence à la théorie axiomatique de la mesure et aux propriétés de la mesure conjointe additive. Ce sujet ne sera pas abordé dans le présent volume, car il en dépasse les objectifs. Le lecteur est plutôt invité à consulter à cet effet Kabaratsos (1999, 2000) ou Cliff (1992).

5. Les travaux de Klauer (1995) sur la puissance des tests sont susceptibles à notre avis de transformer les pratiques en la matière. Malgré tout cependant, à l'heure actuelle, aucun logiciel développé pour la TRI n'intègre des considérations de ce type.

5.5. LA DIMENSIONALITÉ D'UN ENSEMBLE DE SCORES ET L'INDÉPENDANCE LOCALE

Dans cette section, nous allons présenter différentes avenues qui s'offrent aux chercheurs pour vérifier techniquement la condition d'unidimensionalité de l'espace des traits latents et la condition d'indépendance, pour une valeur fixée de l'habileté, entre les réponses fournies aux items (c.-à-d. l'indépendance locale). Le lecteur qui s'interroge peut-être sur la pertinence de présenter ces deux concepts simultanément pourra observer que les définitions de ces deux concepts ont des liens étroits. De plus, comme nous l'avons déjà mentionné, étant donné la proximité des définitions, certaines suggestions de vérification de ces conditions proposent de faire d'une pierre deux coups en vérifiant les deux conditions simultanément. Ces raisons nous semblent justifier la présentation commune de ces deux notions dans une même section.

Le plus souvent, lors de la conception et de l'élaboration d'un test ou d'une épreuve standardisée, nous portons une attention particulière à ces deux conditions en ce sens que nous visons à bâtir un test qui mesure une seule habileté chez les sujets et que nous visons également à y intégrer des items qui ne donnent pas d'information permettant de répondre plus facilement aux autres items du test. Les hypothèses de l'unidimensionalité et de l'indépendance locale existent ainsi indépendamment du désir d'appliquer un modèle de la TRI. En fait, ce sont des hypothèses qui existent implicitement depuis que les tests existent. Les développements de la TRI leur ont donné une impulsion nouvelle et ont permis une réflexion actualisée à leur sujet.

En citant les travaux de Lazarsfeld et Henry (1968) sur les classes latentes et ceux de Lord et Novick (1968) sur la théorie des tests, des auteurs affirment que si le modèle est adéquat, donc avec ses paramètres déterminés, nous avons besoin uniquement de la valeur thêta de l'habileté d'un sujet pour déterminer sa probabilité de réussir un item donné. Si nous connaissons son habileté thêta, la réussite ou l'échec à d'autres items n'ajoute rien de plus à notre connaissance de son habileté. Si ce n'était pas le cas, alors nous pourrions dire que la performance à ces items est influencée par une autre habileté thêta (θ^*), ce qui serait en contradiction avec l'hypothèse que le modèle est adéquat (voir Lord, 1980, p. 19 ; McDonald, 1999, p. 255).

Selon la condition d'indépendance locale, pour un niveau d'habileté donné, la performance observée pour un item ne doit pas influencer la performance à un autre item. Le respect de cette condition équivaut à exiger que, pour une valeur de thêta fixée, toutes les corrélations soient nulles entre les réponses aux items pris deux à deux. Pour Lord et Novick (1968), « la performance d'un individu dépend d'un seul trait si, étant donné la valeur observée pour ce trait, rien d'autre ne peut contribuer à nous informer sur la performance

au test. Le trait latent est le seul facteur important et, lorsque la position de l'individu sur l'échelle de ce trait est connue, le comportement est aléatoire, au sens de l'indépendance statistique » (p. 538).

Ainsi, théoriquement, lorsqu'il n'y a qu'une seule dimension, l'indépendance locale devrait suivre parce qu'à part la dimension unique il n'y a rien d'autre qui puisse influencer directement la réponse à un item, notamment le fait de réussir ou d'échouer aux autres items. Par exemple, Ip (2001) affirme que c'est la multidimensionalité qui induit la dépendance entre les items pour une habileté thêta fixée.

Il nous semble raisonnable de penser que ces raisons expliquent en partie pourquoi les efforts des chercheurs se sont surtout concentrés sur la démonstration de la présence d'une seule dimension et moins sur la démonstration de l'absence de dépendance entre les items pour une valeur thêta fixée. Nonobstant ce qui vient d'être dit et comme nous l'avons mentionné, nous constatons également que certaines approches *omnibus* ont été proposées et qu'elles permettraient de vérifier les deux conditions simultanément. Le lecteur pourra consulter les travaux de Yen (1984, 1993), Leary et Dorans (1985), Thompson et Pommerich (1996), Chen et Thissen (1997) et Ip (2001) pour une meilleure compréhension des enjeux et des difficultés rencontrées lorsque les items d'un test ne sont pas indépendants l'un de l'autre et que nous voulons le détecter.

5.5.1. L'unidimensionalité : une préoccupation qui n'est pas nouvelle

En éducation et en psychologie, les travaux de modélisation ont adopté les expressions habileté et compétence (*ability* et *proficiency*) pour désigner les traits génériques à l'œuvre dans les tests d'habileté intellectuelle comportant de bonnes et de mauvaises réponses. Les scores au test représenteraient ainsi une manifestation de la mise en action de un ou plusieurs traits latents, c'est-à-dire non observables, des candidats dans des conditions données et en fonction de certains items.

Les préoccupations concernant la dimensionalité de l'ensemble des scores à un test ne sont pas nouvelles dans le domaine du développement des tests en éducation et en psychologie. Aux États-Unis, Thurstone proposait déjà en 1925 une méthode visant la construction d'échelles de mesure unidimensionnelles. Thurstone s'appuyait alors sur des travaux de l'époque en psychophysique, principalement la loi du jugement comparatif qui, en quelque sorte, régirait le processus de réponse à un certain type de tâche (Martin, 1999). C'est d'ailleurs Thurstone qui, dans son entreprise de théorisation de la mesure, donna naissance au concept de continuum ou d'échelle. À la suite des travaux de Thurstone, d'autres chercheurs (Likert, Coombs, Guttman) proposèrent différentes méthodes de construction d'échelles unidimensionnelles. En fait, ces méthodes ont été élaborées pour vérifier l'existence d'un continuum, pour légitimer la mesure à partir de ce continuum ou,

encore, pour vérifier l'unidimensionalité de certaines variables psychologiques (Martin, 1999, p. 22). Thurstone participa également aux développements de la méthode de l'analyse factorielle de la matrice des coefficients de corrélation et ses travaux accompagnent ceux de Spearman comme initiateur des développements dans ce champ de recherches.

5.5.2. Pourquoi étudier le nombre de dimensions ?

L'hypothèse d'une seule **dimension**, ou d'un seul trait latent, n'a pas toujours été présentée d'une façon aussi formelle que ce que nous retrouvons dans la TRI, c'est-à-dire à l'intérieur d'un modèle, mais elle a toujours été une hypothèse fondamentale dans la théorie des tests[6]. En effet, on affirme souvent que les scores possèdent une certaine pertinence dans la mesure où les items du test ne mesurent qu'une seule habileté (Hattie *et al.*, 1996 ; McNemar, 1946, p. 298), tout à fait comme pour un instrument de mesure dans les sciences de la nature où nous privilégions les instruments qui donnent de l'information sur une seule quantité à la fois (la longueur, le poids, la vitesse, la densité, la luminosité, etc.).

Si le test est composé d'items qui mesurent différents traits, différentes habiletés, nous pensons qu'il est alors difficile d'interpréter les scores, de les mettre en relation avec d'autres scores ou de mettre en relief les différences individuelles. De plus, la conclusion au sujet de la dimensionalité d'un test pourrait avoir un impact non négligeable sur l'utilisation de ce test parce que nous pourrions être contraints de travailler avec deux ou trois scores différents qui représenteraient des dimensions différentes.

Dans une situation réelle, plusieurs éléments entrent en jeu qui influencent toujours, jusqu'à un certain point, la performance au test, donc les scores observés. Ces éléments sont associés autant aux items (individuellement et collectivement) qu'aux candidats et au contexte, et aux interactions entre chacun. Il semble alors que, face à une situation aussi complexe, il soit beaucoup plus réaliste et pratique de considérer que l'hypothèse d'unidimensionalité est vérifiée lorsqu'on peut montrer qu'une dimension dominante explique ou est responsable de la performance et des réponses des candidats (Humphreys, 1984). D'autre part, même si l'analyse des scores peut permettre certaines vérifications techniques de l'hypothèse d'unidimensionalité, ils peuvent rester difficiles à interpréter conceptuellement si le trait visé par le test ne peut pas être défini clairement et sans ambiguïté, d'autant plus s'il n'est pas observable directement.

---

6. Le premier modèle d'analyse factorielle, celui de Spearman au début du XX^{e} siècle, postulait aussi l'existence d'un seul facteur pouvant expliquer les résultats.

Même si l'expression **dimension dominante** peut rendre la vie plus facile du point de vue de la prise de décision, elle ne dispense pas la définition du concept d'unidimensionalité, en terme de variable latente unique, d'être concrète et relativement opérationnelle.

Il faut rappeler, à la suite de Reckase (1990), qu'il existe une distinction entre le construit psychologique non observable faisant l'objet de la prise de mesure et les outils statistiques employés pour confirmer l'existence du construit. Dans le cadre particulier de l'étude de la dimensionalité, il faut faire une distinction entre dimensionalité conceptuelle (ou psychologique) et dimensionalité statistique. La première étiquette nous renvoie à la définition et aux assises théoriques du trait mesuré par les items – le fameux **construit** –, tandis que la deuxième se veut en quelque sorte une proposition de définition opérationnelle de la première.

Il faut souligner que la dimensionalité conceptuelle est l'objet d'attentions avant et après la cueillette de données, alors que la seconde n'est possible que s'il existe des données. En effet, même si on attribue souvent le qualificatif undimensionnel au test ou à l'ensemble d'items après une analyse de données avec une technique particulière, c'est l'analyse de l'ensemble des scores qui est décisive à cet égard. Les scores constituent la trace de la rencontre entre les candidats et les items ; ce sont eux qui sont l'objet direct de la modélisation, pas le contenu des items. La démonstration de l'unidimensionalité d'un test repose donc en bonne partie, comme nous le verrons plus loin, sur l'étude de l'ensemble de scores qui est généré par la rencontre entre des items et des candidats, dans des conditions données et selon une appréciation précise. En contrepartie, la dimensionalité conceptuelle est intimement liée à la validité globale du test. Elle englobe donc l'étude de la dimensionalité statistique, mais également la solidité du cadre conceptuel et les qualités métriques de l'instrument utilisé pour recueillir les données. Les stratégies dont nous allons maintenant faire part au lecteur sont donc toutes des stratégies qui s'intéressent à la dimensionalité statistique.

5.5.3. Différentes avenues pour étudier l'unidimensionalité

L'importance dans le cadre de la théorie des réponses aux items de pouvoir démontrer qu'une dimension dominante est responsable de la performance des candidats est accentuée par le fait qu'un certain nombre de simulations et d'études avec des données réelles ont démontré que les paramètres de différents modèles unidimensionnels sont mieux estimés lorsqu'il n'y a qu'une seule dimension présente dans les données (pour les simulations) ou dans la structure conceptuelle du test (pour les données réelles). Nous retrouvons des problèmes d'estimation des paramètres lorsque d'autres dimensions que celle visée prennent de l'importance (pour ces résultats, voir Reckase, 1979 ; Drasgow et Parsons, 1983 ; Doody-Bogan et Yen, 1983 ; Harrison, 1986 ; Wang, 1988 ; Blais, 1987 ; Greaud, 1988 ; Kim et Stout, 1993).

En outre plusieurs propositions de modélisation multidimensionnelle pourraient être plus rentables pour représenter les données (la rentabilité d'une procédure étant évidemment liée à l'objectif poursuivi). Toutefois, l'utilité des modèles multidimensionnels reste encore à illustrer à l'extérieur d'études où les données sont simulées, c'est-à-dire dans des situations faisant une place importante à l'interprétation. Ainsi, les applications multidimensionnelles se sont faites plutôt rares jusqu'à maintenant, mais le lecteur peut tout de même consulter les travaux de Luecht (1996), Van der Linden (1996) et McDonald (1997).

Plusieurs suggestions ont été mises de l'avant pour élaborer une méthode statistique qui fournisse une définition opérationnelle efficiente de l'unidimensionalité statistique. L'approche que l'on pourrait qualifier de classique fait appel aux procédures associées à l'analyse factorielle dans la lignée des travaux de Spearman. Même si le chapitre 7 sur la validité présente les bases du modèle de l'analyse factorielle multiple, nous allons nous attarder brièvement à ce modèle qui, lorsqu'on pose l'hypothèse de normalité de la distribution du trait latent et la présence d'un seul facteur, est équivalent au modèle normal de la TRI. Nous y reviendrons plus loin dans cette section.

Hattie (1984, 1985) a produit une étude détaillée de certaines procédures statistiques ayant été suggérées pour déterminer si l'ensemble des scores à un test est unidimensionnel. Les différentes procédures recensées par Hattie peuvent être classées selon qu'elles étudient les patrons de réponse, qu'elles sont issues de la théorie classique des tests ou de la théorie des réponses aux items, ou qu'elles ont des liens avec des techniques de réduction des données comme l'analyse en composante principale et l'analyse factorielle.

Un premier groupe de propositions recensées par Hattie reposent sur l'idée que la dimensionalité d'un test dépend de la distance entre l'ensemble des réponses observées et un schéma de réponse idéal qui produirait une échelle parfaite des sujets et des items. Ainsi, lorsque les items sont ordonnés selon leur degré de difficulté, les scores provenant d'un test unidimensionnel devraient permettre d'obtenir une échelle de Guttman, c'est-à-dire une hiérarchie chez les items selon la difficulté et une hiérarchie chez les répondants selon leur nombre de bonnes réponses. Le degré d'adéquation entre les items et une échelle de Guttman peut être apprécié à l'aide de différents indices, dont ceux de **reproductibilité** (Cliff, 1983). D'autres auteurs tels Loevinger (1947), Green (1956) ou Cliff (1977), ont également proposé des indices d'unidimensionalité reposant sur les schémas de réponse. Il semble que les indices de cette famille soient plus utiles pour détecter les patrons de réponses anormaux (comme nous allons le voir au chapitre 7) que la présence de plusieurs dimensions (Hattie, 1985 ; Wise, 1983). Pour Hattie, ces approches confondent les méthodes pour vérifier la dimensionalité avec l'identification de la dimension mesurée.

Un deuxième groupe de propositions associées à l'estimation de la fidélité en théorie classique des tests, comme l'indice alpha de Cronbach et tous ses dérivés, constituent une des approches qui a été abondamment utilisée comme indicateur de l'unidimensionalité. L'utilisation d'alpha correspond à la perspective que la dimensionalité d'un test est reliée au rang de la matrice des corrélations entre les items (p. ex., Lumsden, 1957). Si cette matrice est de rang un alors elle s'ajuste au modèle à un facteur de Spearman et le test serait alors unidimensionnel. Par ailleurs, Green *et al.* (1977) ont montré avec une simulation de type Monte-Carlo que le principal problème du coefficient alpha comme indice d'unidimensionalité est que sa valeur augmente en même temps que le nombre d'items augmente. Ce qui signifie que tous les tests très longs, sans distinction, seraient « plus unidimensionnels » que les tests plus courts. Plusieurs exemples pourraient illustrer que la réalité s'accommode difficilement de ce genre de simplification.

Un troisième groupe de propositions s'inspirent des résultats d'analyses en composante principale ou d'analyses factorielles. Par exemple, il a été recommandé d'examiner la proportion de la variance expliquée par la première composante d'une analyse en composante principale comme indice de dimensionalité. Plus cette proportion est grande, plus le test serait unidimensionnel. Malheureusement, il n'est pas évident d'établir le seuil que cette proportion doit franchir avant que le test puisse être considéré unidimensionnel. Certains auteurs ont suggéré 40 % ou 20 %, mais il n'existe pas d'arguments solidement étayés dans cette direction. On a aussi proposé de considérer le nombre de composantes/facteurs dont les valeurs propres associées sont plus grandes que 1 (Kaiser, 1970). D'autres encore ont proposé d'examiner le rapport entre les deux valeurs propres les plus élevées (Lumsden, 1957 ; Hutten, 1980) ou encore le rapport de la différence entre les deux plus grandes valeurs propres et de la différence entre les deuxième et troisième valeurs propres (Lord, 1980 ; Divgi, 1980). Hattie (1984) donne un exemple simple où un indice de ce type pourrait faillir à la tâche s'il est utilisé pour déterminer la dimensionalité d'un test. Ainsi, s'il existe quatre composantes et que les deuxième et troisième valeurs propres sont presque égales, alors la valeur de l'indice pourrait être élevée. Au contraire, s'il existe trois composantes et que la différence entre les deuxième et troisième valeurs propres est élevée, alors la valeur de l'indice pourrait être faible. L'indice identifierait donc la situation avec quatre composantes comme étant unidimensionnelle et celle avec trois composantes comme étant multidimensionnelle.

Finalement, un quatrième groupe de propositions ont été développées en parallèle avec la TRI. Ainsi, pour le modèle de Rasch, Wright et Panchapakesan (1969) ont affirmé que si les scores s'ajustent bien au modèle unidimensionnel de Rasch, alors tout indique qu'il n'y a qu'une seule habileté en action, une seule dimension. Cette affirmation ramènerait l'examen de la dimensionalité de l'ensemble de scores à la comparaison de l'ajustement des données pour un modèle unidimensionnel versus un modèle multidimen-

sionnel. Comme nous l'avons déjà mentionné cependant, la statistique du khi-carré est utilisée la plupart du temps comme indicateur de la qualité de l'ajustement et celle-ci est grandement influencée par le nombre de sujets dans l'échantillon. Bejar (1980) a proposé de comparer les estimations des paramètres obtenues d'abord avec le test complet, puis avec des sous-ensembles d'items regroupés selon la pertinence du contenu, exactement dans le sens de ce qui est suggéré pour examiner le maintien de la propriété d'invariance. La procédure a été utilisée pour soutenir l'hypothèse d'unidimensionalité dans des tests d'habileté langagière (Henning *et al.*, 1985), mais elle a donné de moins bons résultats lors de certaines simulations (Hambleton et Rovinelli, 1986).

En résumé, nous pouvons dire d'une part que plusieurs techniques, indices et approches qui ont fait partie de l'étude de Hattie sont en quelque sorte des indices d'une autre époque, celle d'avant l'accès à la puissance de calcul des ordinateurs actuels. À la limite, comme le mentionne McDonald (1999), elles sont d'un intérêt historique ou didactique. Ces indices ont pu être appropriés à une époque où la modélisation devait faire plusieurs concessions et se restreindre à des situations hypothétiquement idéales (par exemple, la présence de tests parallèles ou équivalents) pour faciliter les calculs. Ce n'est plus le cas et de nouvelles approches attirent l'attention, notamment une procédure d'analyse factorielle de l'information complète telle que développée par Bock, Gibbons et Muraki (1988), les travaux de Rozenbaum (1984) et Holland et Rozenbaum (1986), et la procédure non paramétrique développée par Stout (1987, 1990) qui a été améliorée par Nandakumar et Stout (1993). Il faut également mentionner la résurgence de la procédure d'analyse factorielle non linéaire polynomiale de McDonald (1967, 1982, 1999) telle qu'implantée dans le logiciel NOHARM.

Nous allons donc présenter brièvement ces quatre approches. Elles sont d'une part plus contemporaines, pourrions-nous dire, et d'autre part, il existe des logiciels commerciaux disponibles pour ces approches (sauf à notre connaissance pour le test de Mantel-Haenszel). Ainsi, pour l'analyse factorielle de l'information complète le logiciel TESTFACT est disponible, pour l'analyse factorielle non linéaire polynomiale le logiciel NOHARM peut être utilisé et le logiciel DIMTEST est disponible pour l'analyse non paramétrique. Cependant, avant de nous lancer dans la description de ces approches, nous présenterons une proposition de définition formelle de la dimensionalité qui nous apparaît bien articulée et qui s'inscrit dans la lignée des approches que nous décrirons à la section 5.5.5.

5.5.4. Définir la dimensionalité

Évidemment, lorsque nous disons qu'un test est unidimensionnel si les items qui le composent ne mesurent qu'un seul et même trait (habileté ou performance), nous restons dans une définition de principe plutôt vague qui, comme

nous l'avons vu, a été apprêtée de plusieurs façons pour rendre le concept opérationnel, c'est-à-dire pour trouver une façon de démontrer statistiquement la dimensionalité. Peu de ces stratégies ont des bases analytiques solides, mais des efforts spécifiques ont été réalisés dans cette direction depuis quelques années.

Le principe sous-jacent à plusieurs des approches utilisées pour déterminer l'unidimensionalité statistique est celui de l'appréciation de la covariation des scores aux items du test et, lorsque le concept de dimensionalité est replacé dans le contexte des théories du trait latent, il est possible d'élaborer une définition formelle de la dimensionalité. Cette voie a été empruntée par Stout (1987, 1990), par Holland et Rozenbaum (1986) et par Chen et Thissen (1997). Elle demande de considérer la dimensionalité en parallèle avec le concept d'indépendance locale tel que l'a suggéré McDonald (1981).

Ainsi, pour Lord et Novick (1968, p. 531-541), McDonald (1981) et Stout (1990), la notion de dimensionalité est régie par le principe d'indépendance locale. Le nombre de dimensions k d'un ensemble de n mesures est le nombre minimal de traits latents produisant des réponses indépendantes pour ces n mesures. Il y a indépendance locale si, étant donné un ensemble de traits latents, n mesures sont indépendantes en probabilité dans une sous-population de candidats se situant au même endroit sur le continuum des valeurs prises par chaque trait latent.

Pour le cas particulier où les réponses observées sont notées de façon dichotomique, soit $\mathbf{U_n} = (U_1, U_2, \ldots, U_n)$, le vecteur des variables identifiant les scores aux items (par exemple, $U_i = 0$ ou 1), et $\boldsymbol{\Theta}$, le vecteur des traits latents $(\theta_1, \theta_2, \ldots, \theta_k)$. Soit $P_i(\theta) = P_i[U_i = u_i \mid \boldsymbol{\Theta} = \theta]$ la probabilité qu'un candidat choisi aléatoirement dans un groupe de candidats d'habileté $\boldsymbol{\Theta} = \theta$ réussisse l'item i et se voie attribuer un score $U_i = u_i$, la condition d'indépendance locale exige que pour chaque schéma de réponses $(u_1, u_2, \ldots, u_n)$ et pour chacune des valeurs de $\theta = \boldsymbol{\Theta}$:

$$P\left[U_1 = u_1, U_2 = u_2, \ldots, U_n = u_n \,/\, \Theta = \theta\right] = \prod_{i=1}^{n} P_i\left[U_i = u_i \,/\, \Theta = \theta\right]$$

Le nombre d de dimensions de l'ensemble des scores sera la dimensionalité minimale requise du vecteur $\boldsymbol{\Theta}$ pour produire l'indépendance des fonctions $P_i(\theta)$. Ainsi, selon cette définition, l'indépendance locale peut tenir même s'il y a plusieurs dimensions. Dans cette perspective, il n'est donc pas exact de dire que l'unidimensionalité et l'indépendance locale sont équivalentes. Cependant, il est vrai que ces deux notions s'équivalent dans le cas où l'unidimensionalité est avérée. En effet, lorsqu'il y a une seule dimension, il y a automatiquement indépendance locale pour chaque thêta du continuum d'habileté.

Cette condition à l'indépendance locale est très stricte parce qu'elle exige non seulement que les covariances entre les scores soient nulles, mais également que tous les moments supérieurs soient des produits des moments univariés (Hattie *et al.*, 1996 ; McDonald, 1981). Ainsi, par exemple, il faudrait également que :

$$P[U_1 = u_1, U_2 = u_2, U_3 = u_3 / \Theta = \theta]$$
$$= P(U_1 = u_1 / \Theta = \theta)P(U_2 = u2 / \Theta = \theta)P(U_3 = u_3 / \Theta = \theta)$$

Une définition moins stricte demanderait de vérifier uniquement si les covariances entre les scores sont nulles (McDonald, 1981), c'est-à-dire que pour toutes les paires d'items i et j :

$$\mathrm{Cov}(U_i, U_j / \Theta = \theta) = 0$$

Il serait donc possible de faire la distinction entre une condition stricte d'indépendance locale, la condition forte, et une condition moins stricte, la condition faible. Les procédures de McDonald, de Stout et de Holland et Rozenbaum vérifient en quelque sorte que le modèle est adéquat étant donné que la condition faible d'indépendance s'avère raisonnable. La procédure de l'analyse factorielle de l'information complète, quant à elle, vérifie l'adéquation du modèle étant donné que la condition forte d'indépendance locale est rencontrée. D'après McDonald (1999), il y a peu de différences observées dans les résultats lorsqu'on utilise les deux approches pour modéliser un ensemble de scores à un test.

5.5.5. L'analyse factorielle et la modélisation de la dimensionalité

Même si le modèle de l'analyse factorielle sera présenté en détail au chapitre 7, il nous apparaît judicieux d'y jeter brièvement un coup d'œil maintenant étant donné son utilisation dans plusieurs études sur la dimensionalité d'ensembles de scores.

L'analyse factorielle a d'abord été développée en fonction de l'étude des scores à différents test ; elle a ensuite été appliquée à des situations où ce sont les scores aux items qui sont étudiés. Au fur et à mesure que la perspective de la modélisation s'est déplacée des scores aux tests vers les scores aux items, les méthodes de l'analyse factorielle des scores à des tests ont été importées pour réaliser des analyses factorielles sur les scores aux items. Rapidement, les chercheurs se sont rendu compte que ces emprunts ne constituaient pas la voie la plus adéquate. Ainsi, dans les situations où les scores aux items sont dichotomiques, les chercheurs ont longtemps pensé qu'il existait une relation entre la solution factorielle et la distribution de la difficulté des items,

laquelle a un impact sur les coefficients de corrélation entre les items. Lors de certaines analyses de données, il était possible en effet d'observer des regroupements d'items sur les facteurs en fonction de la difficulté des items.

Pour contourner ce problème, on a examiné l'effet d'un changement de la mesure d'association entre les items. Ainsi, des chercheurs proposèrent des analyses factorielles des scores dichotomiques à partir de la matrice des coefficients de corrélations tétrachoriques. Mais, lorsque les réponses sont issues d'items à réponse choisie, le phénomène du hasard peut jouer un rôle déterminant dans le comportement de ce coefficient. Lord (1980) considère que la distorsion introduite est trop grande et il suggère de ne pas utiliser ce type de coefficient de corrélation dans une situation où le hasard peut intervenir dans le choix de la réponse. Cependant, McDonald et Ahlawat (1974) ont montré que ce facteur n'était pas dû à la distribution de la difficulté, mais plutôt au fait qu'on tentait d'appliquer un modèle linéaire plutôt qu'un modèle non linéaire. Suite à ces observations, McDonald (1981, p. 14-15) concluait qu'il est relativement raisonnable d'affirmer qu'un ensemble de n tests ou un ensemble de n items dichotomiques est unidimensionnel si et seulement si on peut lui ajuster un modèle factoriel non linéaire avec un facteur commun. Bock *et al.* (1985) ont aussi proposé d'utiliser une procédure d'analyse factorielle complète de l'information (*full-information factor analysis*) qui ne fait pas intervenir la matrice des corrélations et évite ainsi les problèmes associés à son utilisation avec des scores dichotomiques. Nous explorerons ces deux dernières pistes plus à fond et présenterons des exemples d'utilisation de ces analyses dans une section ultérieure.

Le modèle d'analyse factorielle le plus simple est le modèle linéaire à un facteur de Spearman, qui peut prendre la forme suivante :

$$x_i = \lambda\ \theta_i + \varepsilon_i$$

Le modèle linéaire multiple de Thurstone peut aussi être présenté de façon semblable :

$$x_i = \lambda_{i1}\theta_1 + \lambda_{i2}\theta_2 + \lambda_{i3}\theta_3 + ... + \lambda_{ik}\theta_k + \varepsilon_i$$

Pour ces deux modèles, x représente ce qui est observé comme résultat pour l'item i, λ est la saturation associée au facteur, θ représente le ou les facteurs et ε représente la partie unique ou le résidu.

Selon la tradition de l'analyse factorielle, si le modèle à un facteur est celui qui s'ajuste le mieux dans une situation donnée, cela nous amène à conclure qu'il n'y a qu'une seule dimension (un seul trait latent) et si c'est le modèle multiple qui s'ajuste le mieux, nous concluons à la présence de plusieurs dimensions (plusieurs traits latents).

Bock et Aitkin (1981) ont adapté de la façon suivante le modèle factoriel multiple pour le cas où les réponses sont dichotomiques. Soit Y_i le processus non observable qui détermine quelle réponse sera donnée à l'item i. Ce processus est relié linéairement à un ensemble de M traits latents :

$$Y_i = \lambda_{i1}\theta_1 + \lambda_{i2}\theta_2 + ... + \lambda_{im}\theta_m + \varepsilon_i$$

où θ_m représente l'habileté sur la dimension m et λ_m symbolise le poids de la dimension m pour l'item i. Nous supposons que chaque item est caractérisé par une constante γ_i de telle sorte que :

- Si $Y_i \geq \gamma_i$, alors la réponse est bonne ($u_i = 1$).
- Si $Y_i < \gamma_i$, alors la réponse est mauvaise ($u_i = 0$).

Si les résidus, ε_i, possèdent une distribution normale, alors la probabilité de répondre correctement à l'item i étant donné thêta peut être représentée par la fonction de répartition de la distribution normale :

$$P(u_i = 1 | \Theta) = \Phi\left[\frac{\lambda_{i1}\theta + \lambda_{i2}\theta + ... + \lambda_{im}\theta - \gamma_i}{\sigma_i}\right]$$

où $\sigma_i^2 = 1 - \sum\lambda_i^2$.

Après une transformation, l'expression $P(u_i = 1 | \Theta)$ peut aussi s'écrire sous la forme d'une modélisation multidimensionnelle :

$$P(u_i = 1 | \Theta) = \Phi[a_{i1}\theta_1 + a_{i2}\theta_2 + ... + a_{im}\theta_m + d_i]$$

Il y a autant de paramètres a associés à la discrimination qu'il y a de dimensions et un seul paramètre d associé à la difficulté. S'il n'y avait qu'une seule dimension nous pourrions facilement retrouver le modèle normal[7] :

$$P(u_i = 1/\theta) = \Phi[a_i(\theta - b_i)]$$

Si nous supposons que la relation est logistique plutôt que normale, alors :

$$P(u_i = 1/\theta) = \Psi[Da_i(\theta - b_i)]$$

7. Si les paramètres a_i sont égaux, alors nous retrouvons le modèle de Rasch.

La fonction Ψ représente la fonction de répartition de la loi de probabilité logistique :

$$\Psi(x) = \frac{1}{1+e^{-x}}$$

Dans de nombreux écrits, McDonald a qualifié ces deux adaptations de l'analyse factorielle pour les variables dichotomiques d'analyses factorielles non linéaires[8]. De plus, il a également développé une autre perspective par rapport à l'analyse factorielle non linéaire en substituant une fonction polynomiale aux fonctions de répartition ϕ et Ψ (voir entre autres McDonald, 1967, 1982). Par exemple, il a suggéré d'ajuster le modèle cubique à un facteur suivant :

$$P(U_i = u_i / \theta) = \lambda_{i^1}\theta + \lambda_{i^2}\theta^2 + \lambda_{i^3}\theta^3 + \varepsilon_i$$

Dans un des exemples d'études de l'ajustement des modèles de la TRI que nous aborderons à la section 5.6, Nandakumar a ajusté un modèle quadratique et un modèle cubique à un facteur pour des données undimensionnelles et un modèle quadratique à deux facteurs pour des données bidimensionnelles. Le modèle quadratique ajusté à deux facteurs était :

$$P(U_i = u_i / \theta_1, \theta_2) = \lambda_{i1^1}\theta_1 + \lambda_{i1^2}\theta_1^2 + \lambda_{i2^1}\theta_2 + \lambda_{i2^2}\theta_2^2 + \varepsilon_i$$

La démarche d'analyse factorielle de l'information complète selon la proposition de Bock et Aitkin (1981) a l'avantage de ne pas reposer sur l'analyse de la matrice des corrélations entre les items. Nous l'avons mentionné, des difficultés surgissent lorsque les coefficients de corrélation tétrachorique sont utilisés dans les situations où les scores sont dichotomiques[9]. Pour régler ce problème, l'analyse de Bock et Aitkin utilise plutôt les fréquences pour chaque patron de réponse observé et modélise le tout avec une distribution multinomiale[10]. L'adéquation du modèle peut être vérifiée en utilisant la statistique G^2, qui est une approximation du test du rapport de vraisemblance. Si la taille de l'échantillon est suffisamment élevée et que les 2^n patrons de réponses ont une espérance minimale d'apparaître, alors : $G^2 = 2\sum r_l \, ln\left(r_l / N\tilde{P}_l\right)$ est

8. Pour un traitement élaboré des liens entre la théorie classique des tests, l'analyse factorielle et la théorie des réponses aux items, le lecteur est invité à consulter le plus récent ouvrage de McDonald (1999) sur la théorie des tests.
9. Les propositions de Christoffersson (1975) et Muthen (1978) visaient à contourner ce problème, mais elles étaient difficilement applicables à des tests de plus de vingt items (Muthen, 1984).
10. À noter que Muraki et Carlson (1995) ont généralisé le modèle pour les scores dichotomiques aux situations où les scores sont polytomiques.

une statistique qui possède une distribution du khi-carré avec $(2^n - (k + 1) + (k\,(k-1)/2))$ degrés de liberté (Bock, Gibbons et Muraki, 1988). La modélisation débute avec l'ajustement du modèle à un facteur et est répétée en ajoutant un facteur à la fois. Pour vérifier la dimensionalité d'un ensemble de scores, nous devons examiner si la contribution du dernier facteur ajouté est significative, c'est-à-dire si le gain de la statistique G^2 est statistiquement significatif lorsque nous passons d'un modèle à un facteur à un modèle à deux facteurs par exemple. Il est également suggéré, comme l'a fait Zwick (1987), d'explorer la dimensionalité de façon plus traditionnelle techniquement en étudiant les contributions respectives des facteurs en terme de pourcentage de la variance expliquée.

Pour l'analyse factorielle non linéaire polynomiale, McDonald (1985) avait suggéré d'utiliser les moyennes des valeurs absolues des covariances résiduelles après l'ajustement d'un modèle. Hattie (1984, 1985) avait d'ailleurs montré dans une simulation que cet indice était un de ceux qui discriminait le mieux certains ensembles de scores undidimensionnels d'autres ensembles multidimensionnels. McDonald et Mok (1995) et McDonald (1999) ont proposé plusieurs autres statistiques d'ajustement dont la statistique GFI qui est intégrée au logiciel NOHARM. Pour calculer la statistique GFI (*general fit index*), il faut d'abord calculer q_u, qui sert à estimer la distance entre les covariances échantillonnales et les covariances obtenues avec le modèle ajusté :

$$q_u = \left(1/m^2\right)\sum_i \sum_k \left(s_{ik} - \sigma_{ik}\right)^2$$

où m est le nombre d'items s_{ik}, la covariance échantillonnale entre les items i et k et σ_{ik} la covariance avec le modèle ajusté. Il faut également calculer $c = \left(1/m^2\right)\sum_i \sum_k s_{ik}^2$; la valeur de l'indice est alors donnée par :

$$GFI = 1 - q_u / c$$

Lorsque l'ajustement est bon, la valeur de GFI devrait être près de 1, sa valeur maximale. McDonald (1999, p. 84) considère que l'ajustement est bon si la valeur de la statistique GFI est supérieure à 0,95 et acceptable lorsque la valeur est supérieure à 0,90.

L'indice GFI peut être ajusté en fonction du nombre m d'items et du nombre d de degrés de liberté dans le modèle et son interprétation est semblable à la version non ajustée (Swygert, McLeod et Thissen, 2001) :

$$AGFI = 1 - \frac{m(m+1)}{2d}(1 - GFI)$$

5.5.6. La statistique T de Stout

Stout (1987, 1990) a proposé de définir opérationnellement le concept de dimension dominante. Stout a ainsi suggéré d'examiner l'unidimensionalité ***essentielle*** d'un ensemble de scores en faisant intervenir le test statistique :

$$H_0 : d_E = 1$$

$$H_1 : d_E > 1,$$

où d_E est l'unidimensionalité essentielle de l'ensemble des scores au test. La proposition a par la suite été améliorée par Nandakumar et Stout (1993). Cette proposition vise exclusivement les ensembles de scores dichotomiques constitués de 0 et de 1, mais d'autres propositions pour analyser des ensembles de scores polytomiques existent.

La condition faible d'indépendance locale pourrait s'exprimer, comme le propose Stout (1990), en définissant une condition d'indépendance essentielle qui tiendrait si :

$$\lim_{n\to\infty} \frac{\sum_{1\le i<j\le n} \left|\mathrm{Cov}(U_i, U_j) \mid \Theta = \theta\right|}{\binom{n}{2}} \to 0$$

La condition tiendrait donc si la moyenne des covariances entre toutes les paires d'items tend vers zéro lorsque le nombre d'items tend vers l'infini. Le nombre de dimensions essentielles d_E, que Stout appelle l'unidimensionalité essentielle, de l'ensemble des scores au test serait donc le nombre minimal de traits nécessaire à la réalisation de cette expression du principe faible d'indépendance locale. La définition de l'indépendance essentielle est élaborée en fonction d'un vecteur Θ de traits latents. Elle rejoint l'idée de la présence d'une dimension dominante puisque même si plusieurs dimensions ou traits contribuent à la production des réponses observées, cela n'empêche nullement la réalisation de la condition d'indépendance essentielle. Cette approche ne permet pas la démonstration directe de la présence d'une seule dimension, mais plutôt la démonstration de l'existence d'une représentation adéquate des scores par un modèle unidimensionnel monotone pour lequel la condition faible d'indépendance locale tient (Stout, 1987, 1990). L'approche est élaborée dans le contexte d'un nombre d'items infini où les propriétés des estimateurs statistiques sont asymptotiques. Cependant, dans des situations où le nombre d'items est réduit, il semble que ces propriétés asymptotiques tiennent plus ou moins (voir de Champlain et Gessaroli, 1991).

La procédure menant au test de l'hypothèse ci-dessus a été décrite par Stout (1987), Nandakumar et Stout (1993) et Blais et Laurier (1997) et intégrée au logiciel DIMTEST. Elle se déroule selon les étapes suivantes :

1. M items sont sélectionnés pour faire partie du premier sous-test de vérification, STV1. Pour des considérations de robustesse de l'estimation, le nombre d'items composant STV1 ne devrait pas dépasser le quart du nombre total d'items. Deux stratégies sont suggérées pour constituer STV1 : a) une analyse conceptuelle de l'ensemble des items par un ou des experts pour produire une sélection d'un sous-ensemble d'items le plus unidimensionnel possible ; b) une analyse en composantes principales de la matrice des corrélations tétrachoriques où ce sont les M items ayant les saturations les plus élevées sur le deuxième facteur (avant rotation) qui sont sélectionnés pour faire partie de STV1.
2. Un second ensemble de M items est sélectionné à partir des items restants de façon à ce que la difficulté et la dimensionalité de l'ensemble d'items ressemblent à ce qu'on retrouve pour STV1. Cet ensemble constitue le deuxième sous-test de vérification, STV2. Ce sous-test sera utilisé pour apporter une correction à la statistique issue de STV1.
3. Les items non utilisés pour STV1 et STV2, les N – 2M items qui restent, forment le sous-test de répartition, STR. Les scores au sous-test de répartition servent à regrouper les candidats selon le résultat obtenu. Ainsi, en excluant les sujets qui n'ont que de bonnes ou de mauvaises réponses, le sous-test de répartition permet de former au plus N – 2M – 2 = R regroupements. Pour conserver les propriétés asymptotiques de la statistique, il est suggéré de former des regroupements d'au moins vingt sujets.
4. On estime la variance des scores pour chacun des sous-tests de vérification, $\hat{\sigma}_r^2$, et la variance unidimensionnelle, $\hat{\sigma}_{U,r}^2$, pour chacun des r regroupements de candidats produits par le sous-test de répartition. On calcule une statistique T_i pour les deux sous-tests de vérification (voir l'annexe 5.1) :

$$T_i = \frac{1}{\sqrt{r}} \sum_{r=1}^{R} \left[\frac{\hat{\sigma}_r^2 - \hat{\sigma}_{U,r}^2}{S_r} \right]$$

Finalement, on calcule la statistique $T = \frac{T_1 - T_2}{\sqrt{2}}$ et on vérifie l'hypothèse $H_0 : d_E = 1$, en profitant du fait que la distribution de T est asymptotiquement normale avec une moyenne 0 et une variance 1 (Stout, 1987). C'est-à-dire que H_0 est rejetée si $T \geq Z_\alpha$, où Z_α est le 100 $(1-\alpha)$ centile supérieur de la distribution normale standard et α le niveau de signification désiré.

Essentiellement, la procédure de Stout vérifie le degré de proximité entre un modèle unidimensionnel et le modèle qui a généré les scores observés. La statistique T_1 est une information sur le degré de multidimensionalité que l'on retrouve localement pour le regroupement r. Elle est sensible à la multidimensionalité et au biais de l'estimation. La statistique T_2 est calculée à partir d'un ensemble d'items, STV2, que l'on considère équivalent à l'ensemble STV1 et elle est utilisée pour corriger le biais d'estimation de la statistique T_1.

5.5.7. Le test de Mantel-Haenszel

Rosenbaum (1984, 1985) et Holland et Rosenbaum (1986) ont démontré que les scores entre les items sont en relation positive s'ils sont localement indépendants et unidimensionnels, et que les courbes caractéristiques sont monotones croissantes. Ils ont proposé de tester l'hypothèse statistique suivante :

$$H_0 : \mathrm{Cov}\left(X_i , X_j \mid \sum_{i,j \neq k} X_k \right) \geq 0$$

$$H_1 : \mathrm{Cov}\left(X_i , X_j \mid \sum_{i,j \neq k} X_k \right) < 0$$

Cette hypothèse vérifie l'association pour chaque paire d'items étant donné le score obtenu pour les items restants. Pour mettre cette hypothèse à l'épreuve, c'est le test de Mantel-Haenszel (1959) qui est suggéré et il faut examiner toutes les tables (2×2) de contingence pour deux items donnés, pour chacun des scores aux items restants (voir le tableau 5.2).

TABLEAU 5.2
Scores aux items i et j étant donné un score total k pour les items restants

		Score à l'item j		
		1	0	Total
Score à l'item i	1	n_{11k}	n_{10k}	n_{1+k}
	0	n_{10k}	n_{00k}	n_{+0k}
	Total	n_{1+k}	n_{+0k}	n_{++k}

Étant donné deux items i et j, n_{11k} représente le nombre de personnes ayant choisi la bonne réponse aux items i et j et ayant un score de k = 1, 2, …, K pour les items restants. Définissons de la même façon n_{00k} comme le nombre de personnes ayant fourni des réponses erronées aux items i et j, n_{10k} le nombre

de personnes ayant fourni une bonne réponse à l'item i et une réponse erronée à l'item j, n_{01k} le nombre de personnes ayant fourni une réponse erronée à l'item i et une bonne réponse à l'item j.

La statistique de Z du test de Mantel-Haenszel est donnée par :

$$Z = \frac{n_{11+} - E(n_{11+}) + 1/2}{\sqrt{V(n_{11+})}}$$

L'espérance mathématique $E(n_{11+})$ et la variance $V(n_{11+})$ de n_{11+} sont respectivement :

$$E(n_{11+}) = \sum_{k=1}^{K} \frac{n_{1+k} n_{+1k}}{n_{++k}}$$

et

$$V(n_{11+}) = \frac{\sum_{k=1}^{K} n_{1+k} n_{0+k} n_{+1k} n_{+0k}}{n_{++k}^2 (n_{++k} - 1))}$$

De plus, $n_{11+} = \sum_{k=1}^{K} n_{11k}$, c'est-à-dire que n_{11+} représente le nombre total de personnes ayant fourni une bonne réponse au items i et j (la somme sur k = 1, 2, …, K).

Comme la statistique Z de Mantel-Haenszel possède une distribution normale, un test de signification statistique peut être réalisé pour chacune des N(N – 1)/2 paires d'items. Un résultat statistiquement significatif implique que les items de la paire étudiée sont conditionnellement associés, étant donné le score total pour les items restants, et ne sont donc pas cohérents avec une modélisation unidimensionnelle et monotone croissante. Si on observe un grand nombre de paires d'items qui sont conditionnellement associées, alors la modélisation unidimensionnelle n'est pas appropriée. Chacun des tests de Mantel-Haenszel est effectué avec un niveau de signification alpha donné et l'inférence à partir de l'ensemble des paires d'items peut être réalisée en faisant appel à une procédure du type Bonferonni (Holland et Rosenbaum, 1986 ; Zwick, 1987). Ainsi, selon cette procédure, l'hypothèse H_0 est rejetée si au moins un des tests produit un résultat statistiquement significatif avec un niveau de signification α/t, où t représente le nombre de tests effectués (i.e. N(N – 1)/2 tests), et l'hypothèse n'est pas rejetée si le nombre de tests

significatifs au niveau alpha est près de $t\alpha$. Notons que l'approche de Mantel-Haenszel est aussi employée dans l'identification d'items présentant un biais d'ordre linguistique, culturel ou sexiste (voir le chapitre 8).

5.6. EXEMPLES D'ÉTUDES DE L'UNIDIMENSIONALITÉ ET DE L'INDÉPENDANCE LOCALE

5.6.1. Premier exemple

Pour étudier la dimensionalité des scores aux épreuves 1983-1984 de lecture du National Assessment of Educational Progress (NAEP), Zwick (1987) a comparé les résultats d'analyses selon trois méthodes : l'analyse en composantes principales, l'analyse factorielle complète du patron de réponses de Bock et ses collaborateurs (1985), et le test de Mantel-Haenszel proposé par Holland et Rosenbaum (1986). Les analyses ont été réalisées pour trois niveaux/âges différents : 4e/9 ans ; 8e/13 ans ; 11e/17 ans. Pour chacun de ces niveaux, il y avait respectivement 108, 100 et 95 items, de même que 26 087, 28 405 et 28 861 sujets. Environ 25 items parmi l'ensemble se retrouvaient dans tous les cahiers et ont donc été passés par l'ensemble des sujets, indépendamment du niveau.

Les cahiers contenant les items des épreuves 1983-1984 de lecture du NAEP ont été assemblés selon un design particulier, soit un design en spirale avec des blocs équilibrés et incomplets (voir Beaton, 1987). Les candidats ne passaient pas tous une épreuve identique. Les items ont d'abord été regroupés en blocs de six à douze items, puis répartis dans les cahiers selon le design. Chaque item et chaque paire d'items étaient ainsi administrés un nombre précis de fois. Le design a permis de créer 60 cahiers différents par niveau/âge. La plus grande partie (95 %) des items des épreuves du NAEP étaient des items à réponse choisie. Le reste des items étaient constitués d'items à réponse construite notés sur une échelle de 1 à 5. Tous les items ont été classés par des experts en lecture sur la base des objectifs, du type de tâche et du contenu.

En guise d'exploration de la dimensionalité, des analyses en composante principale ont été réalisées pour chaque niveau avec les matrices des coefficients de corrélation phi et les matrices des coefficients de corrélation tétrachorique. Deux analyses incluant les répondants des trois niveaux et les 25 items communs ont aussi été réalisées. Pour chacune des huit matrices ainsi analysées, le poids d'une première valeur propre constituait entre 17 % et 25 % de la trace pour les matrices des corrélations phi et entre 30 % et 40 % pour les matrices des corrélations tétrachoriques. La deuxième valeur propre représentait toujours moins du quart de la première valeur propre. À la lumière de ces analyses, il ne semblait pas déraisonnable de penser à l'existence d'une dimension dominante.

L'analyse du patron de réponses avec TESTFACT a été réalisée avec un nombre réduit d'items et un seul groupe de sujets : 42 items et 2 020 sujets du niveau 8e/13 ans. En effet, au moment de la réalisation de l'étude de Zwick, des coûts élevés étaient associés à l'application de l'analyse factorielle complète des patrons de réponses (n'oublions pas que les analyses ont été réalisées en 1986-1987). À l'heure actuelle, ces problèmes n'existent plus et l'utilisateur de TESTFACT (ou de tout autre logiciel) a beaucoup de latitude quant au nombre d'items et de sujets qu'il peut inclure à une analyse. Les 42 items furent choisis avec l'objectif de maximiser la possibilité de détecter une situation multidimensionnelle. La modélisation normale à trois paramètres a été privilégiée pour réaliser les analyses. Une solution à un seul facteur produisait l'émergence d'un facteur qui comptait pour 39 % de la variance totale. Une solution à deux facteurs produisait un premier facteur comptant pour 36 % et un deuxième facteur comptant pour 4 %. Ces résultats semblaient confirmer la présence d'une seule dimension importante dans les données.

Zwick a également étudié le gain que procure l'ajustement d'un modèle à deux facteurs par rapport à un modèle à un facteur. La différence entre les valeurs de la statistique G^2 de TESTFACT permettrait en effet de conclure à la prépondérance de la solution à un facteur par rapport à une solution à plus d'un facteur. Cependant, il semble qu'une certaine prudence soit de mise envers la statistique G^2. En effet, selon des études avec des données réelles (Dorans et Lawrence, 1987) et des données simulées (Zwick, 1987), la différence entre deux valeurs de cette statistique pour deux modèles différents pourrait mener à une surestimation du nombre de facteurs.

La méthode suggérée par Holland et Rosenbaum (1986) a également été appliquée à un sous-ensemble des données. La raison était la même que précédemment : le coût élevé du temps-machine. Zwick a retenu 56, 53 et 56 items respectivement pour les trois niveaux. Ainsi le nombre de tests du khi-carré pour chacun de ces niveaux s'élève à 1 540, 1 378 et 1 540 respectivement (le nombre de paires d'items). En choisissant un niveau de signification α de 0,01, le nombre de tests statistiquement significatifs s'élevait à 4, 4 et 6 pour les niveaux 4e, 8e et 11e. En établissant α à 0,05 plutôt qu'à 0,01, le nombre de tests significatifs s'élevait respectivement à 31, 29 et 26.

Une dernière analyse a été réalisée avec les 25 items communs aux épreuves passées par les sujets des trois niveaux. Avec un α de 0,05, aucun des tests réalisés n'était statistiquement significatif. À partir de ces résultats, Zwick a conclu qu'il était raisonnable de retenir l'hypothèse que les réponses aux items de l'épreuve de lecture peuvent être modélisées avec un modèle monotone unidimensionnel conditionnellement indépendant étant donné une valeur de thêta fixée.

En bout de ligne, les différentes analyses indiquaient qu'il était raisonnable de considérer les ensembles de données étudiés comme des ensembles unidimensionnels. L'analyse préliminaire avec la méthode de l'analyse en

composante principale permettait de constater que le poids respectif des premières valeurs était toujours plus important que celui des valeurs propres suivantes. L'analyse factorielle des patrons de réponses produisait l'émergence d'un premier facteur comptant pour 39 % de la variance totale. Finalement, l'approche des tables de contingence avec la procédure Mantel-Haenszel menait à la rétention de l'hypothèse que les données sont adéquatement représentées par un modèle monotone unidimensionnel conditionnellement indépendant. D'après Zwick, les trois méthodes fournissaient des résultats s'accordant assez bien.

5.6.2. Deuxième exemple

En utilisant les programmes TESTFACT et DIMTEST, Blais et Laurier (1995, 1997) ont mis en parallèle la procédure de l'analyse factorielle complète du patron de réponses et la procédure non paramétrique de Stout pour confirmer ou infirmer l'unidimensionalité d'ensembles de scores provenant d'une version expérimentale d'un test de placement en français langue seconde. Le test a été administré à des étudiants canadiens-anglais de différents collèges et universités inscrits à des cours d'été de français dans le cadre d'un programme de bourse pour l'apprentissage d'une des langues officielles du Canada. Le test était divisé en trois sous-tests de 50 items chacun. Les ensembles de scores étaient constitués d'un noyau de réponses de 348 étudiants qui ont répondu à l'ensemble des questions du test. Pour les sous-tests, ce nombre initial a été augmenté à 694 étudiants pour le premier sous-test, à 681 étudiants pour le deuxième sous-test et à 661 étudiants pour le troisième sous-test. Les deux procédures ont été appliquées à l'ensemble des scores du test complet et aux ensembles de scores provenant des sous-tests.

Pour la procédure d'analyse factorielle complète du patron de réponse pour l'ensemble du test, une solution de TESTFACT à trois facteurs indiquait un premier facteur comptant pour 25 % de la variance observée et des deuxième et troisième facteurs comptant respectivement pour 2,4 % et 1,4 % de la variance observée. Un examen des saturations a permis de constater que les 50 premiers items étaient surtout associés au premier facteur. Ce premier facteur retenait aussi des items du troisième sous-test qui étaient clairement associés au troisième facteur. Pour les items du deuxième sous-test, les saturations étaient généralement peu élevées, ce qui pouvait indiquer que ce sous-test était moins cohérent et/ou mesurait plusieurs habiletés.

Pour l'analyse par sous-test, une solution à trois facteurs de TESTFACT pour le premier sous-test indiquait la présence d'un premier facteur dominant comptant pour 33 % de la variance observée. Avec une solution à deux facteurs, ce pourcentage a augmenté très légèrement pour atteindre 34 % avec un deuxième facteur à 2,3 %. En examinant les saturations pour ces deux facteurs, les auteurs ont conclu que ces facteurs font ressortir deux habiletés cognitives différentes mais fortement corrélées, soit une habileté

(dominante) à reformuler une information dans une deuxième langue et une habileté (secondaire) à faire des inférences à partir d'une information donnée. Pour le deuxième sous-test, le pourcentage de variance expliqué par le premier facteur d'une solution à trois facteurs était de 24 % avec seulement six items dont les saturations sur le premier facteur étaient élevées. De plus, onze items présentaient des saturations faibles pour les trois facteurs simultanément. Pour le troisième sous-test, une solution à deux facteurs ne confirmait pas la présence d'une distinction entre les compétences en grammaire et les compétences en vocabulaire. De surcroît, une solution à trois facteurs augmentait de 23 % à 28 % le pourcentage de la variance expliquée que l'on peut associer au premier facteur.

Après ces analyses, les auteurs ont conclu : 1) que le premier sous-test était unidimensionnel avec un premier facteur nettement dominant et deux premiers facteurs fortement corrélés ; 2) que le deuxième sous-test était **multidimensionnel**, mais sans vraiment pouvoir préciser tout à fait pourquoi ; 3) que le troisième sous-test était multidimensionnel.

L'approche non paramétrique de Stout a ensuite été appliquée aux mêmes ensembles de scores. Cette approche consiste en la production d'une statistique T dont la distribution asymptotique est celle d'une loi de probabilité normale de moyenne zéro et de variance un. Pour produire cette statistique, l'utilisateur du programme DIMTEST doit d'abord sélectionner un premier sous-test de vérification, STV1, contenant environ le quart des items. Comme nous l'avons déjà mentionné, cette sélection peut se faire de deux manières : suite aux résultats d'une analyse en composantes principales ou suite aux résultats d'une analyse conceptuelle de la part du chercheur. Blais et Laurier (1995, 1997) ont mis à contribution ces deux façons de faire pour analyser la dimensionalité de chacun des trois sous-tests. Suite à une analyse d'un expert, les items de chaque sous-test ont été regroupés en deux domaines, A et B.

Pour chaque domaine d'un sous-test, une sélection de 12 items, soit environ le quart des 50 items, a été effectuée pour satisfaire les recommandations de Nandakumar et Stout (1993). Nous devons souligner que le processus de sélection des items pour la constitution de STV1 peut poser des difficultés lorsqu'il faut choisir des items pour former cet ensemble dans un bassin qui dépasse largement le nombre suggéré. Certaines stratégies de sélection peuvent être plus intéressantes que d'autres ; il faut donc explorer cet aspect de la modélisation avant de tirer des conclusions.

Les auteurs ont constaté que la procédure d'analyse en composantes principales a proposé des sous-tests de vérification pour lesquels la valeur de la statistique T n'était jamais statistiquement significative. Cette première procédure pouvait donc conduire à conclure que les trois sous-tests étaient unidimensionnels. Le regroupement des items en domaines A et B a produit une statistique T qui n'était pas statistiquement significative pour les deux regroupements du premier sous-test, mais qui était statistiquement

significative pour les deux regroupements des deuxième et troisième sous-tests. Les deux procédures de constitution des sous-tests de vérification ne menaient donc pas à des constatations tout à fait convergentes quant à l'unidimensionalité des ensembles des scores issus des sous-tests.

Les auteurs ont donc observé certaines convergences entre les résultats, mais également que des approches utilisées isolément pouvaient mener à différentes décisions concernant l'unidimensionalité d'un ensemble de scores (voir le tableau 5.3). Ainsi, l'analyse avec le logiciel TESTFACT pouvait mener à conclure à l'existence de deux dimensions plus importantes, le premier sous-test constituant la dimension dominante et le troisième sous-test, une dimension secondaire où on retrouverait deux composantes. Cependant, l'existence d'une dimension associée directement aux scores du deuxième sous-test ne pouvait pas être confirmée. Pour ce qui est de l'analyse avec les scores de chacun des sous-tests, Blais et Laurier ont conclu à partir des résultats de TESTFACT que le premier ensemble de scores était unidimensionnel et que les deux autres étaient multidimensionnels. Cependant, du point de vue conceptuel et étant donné les items qui possèdent les saturations les plus élevées, les auteurs ajoutaient qu' « il serait difficile de déterminer la nature de la multidimensionalité que l'on retrouve dans ces deux sous-tests ». À partir des résultats d'une analyse avec DIMTEST qui intégrait le jugement d'un expert, le premier ensemble de scores était considéré unidimensionnel et les deuxième et troisième étaient considérés multidimensionnels. Les deux façons de constituer les sous-tests de vérification avec DIMTEST ne produisaient pas les mêmes résultats.

TABLEAU 5.3
Comparaison des décisions quant à la dimensionalité des ensembles de scores

	Sous-test 1	Sous-test 2	Sous-test 3	Test complet
Analyse factorielle avec TESTFACT	d = 1	d > 1	d > 1	d = 2, 3
DIMTEST Analyse en composantes	d = 1	d = 1	d = 1	?
DIMTEST Expert	d = 1	d > 1	d > 1	?

5.6.3. Troisième exemple

En utilisant des données simulées et des données réelles, Nandakumar (1994) a comparé la performance de différentes procédures pour déterminer l'unidimensionalité d'un ensemble de scores. Elle a comparé la procédure non paramétrique de Stout qui est intégrée au logiciel DIMTEST, le test de Mantel-Haenszel proposé par Holland et Rosenbaum (1986) et l'analyse factorielle non linéaire de McDonald avec le logiciel NOHARM.

Pour la simulation, trois ensembles de scores unidimensionnels ont été générés à partir du modèle logistique à trois paramètres et quatre ensembles bidimensionnels ont été générés à partir du modèle multidimensionnel compensatoire de Reckase et McKinley (1983). Ce modèle bidimensionnel intègre un paramètre de pseudo-chance, deux paramètres de discrimination, deux paramètres de difficulté pour les items et deux paramètres pour rendre compte des deux dimensions associées à l'habileté chez les sujets :

$$P_i(\theta_1,\theta_2) = c_i + \frac{1-c_i}{1+\exp\left\{-1,7\left[a_{1i}(\theta_1 - b_{1i}) + a_{2i}(\theta_2 - b_{2i})\right]\right\}}$$

Les trois ensembles unidimensionnels représentaient les scores de 2 000 sujets à des tests de 25, 40 et 50 items respectivement. La valeur du paramètre de pseudo-chance a été fixée à 0,20 pour tous les items. Les valeurs des paramètres de discrimination et de difficulté du modèle ont été simulées pour différents intervalles se rapprochant de situations réelles. Par exemple, pour l'ensemble 1, le paramètre de difficulté prenait des valeurs dans l'intervalle [–1,39 , 1,27] avec une moyenne de 0,09 et un écart-type de 0,72, et pour l'ensemble 2 le paramètre de difficulté prenait des valeurs dans l'intervalle [–3,11 , 2,07] avec une moyenne de 0,03 et un écart-type de 0,96. Les 2 000 valeurs du paramètre associé aux sujets (l'habileté des sujets) ont été générées selon une distribution normale standard (moyenne de 0 et écart-type de 1).

Les quatre ensembles bidimensionnels représentent les réponses de 2 000 sujets à des tests de 25 et 50 items. Les 2 000 valeurs des deux paramètres de l'habileté des sujets ont été générées selon une distribution binormale avec des moyennes de 0 et des variances de 1 et avec des corrélations entre les deux dimensions de 0,3 et 0,7 respectivement. Les valeurs des deux paramètres de discrimination et des deux paramètres de difficulté ont été générées selon des distributions normales indépendantes (corrélation de 0). La valeur du paramètre de pseudo-chance a également été fixée à 0,20.

Les ensembles de données réelles provenaient de deux sources. D'abord, des données tirées des épreuves d'histoire et de littérature (niveau 11/âge 17) du NAEP 1986 et, ensuite, des données provenant de la passation de la batterie de tests d'orientation professionnelle des forces armées américaines (l'ASVAB) pour le raisonnement arithmétique et les sciences (niveau 10^{e} année). De plus, pour produire une situation où il y avait deux dimensions, deux ensembles de réponses ont été construits en combinant histoire et littérature à partir des résultats du NAEP et deux ensembles ont été construits en combinant raisonnement arithmétique et sciences pour l'ASVAB.

Nandakumar (1994) a trouvé que pour les données simulées la procédure DIMTEST se révélait efficace pour détecter les ensembles de données unidimensionnels et bidimensionnels. La procédure Mantel-Haenszel de

Holland et Rosenbaum et la procédure d'analyse factorielle non linéaire se sont révélées efficaces pour détecter les deux ensembles de données unidimensionnels et les deux ensembles de données bidimensionnels où la corrélation entre les habiletés est faible (r = 0,3), mais moins efficace pour détecter les ensembles de données bidimensionnels où la corrélation simulée entre les habiletés est plus élevée (r = 0,7).

Pour les ensembles de données réelles, l'analyse avec DIMTEST a suggéré la présence d'une seule dimension pour les épreuves d'histoire du NAEP et de raisonnement arithmétique et de sciences de l'ASVAB. L'analyse a aussi suggéré la présence de plus d'une dimension pour l'épreuve de littérature du NAEP et pour les deux ensembles issus de la combinaison des données. L'analyse avec la procédure Mantel-Haenszel suggérait de son côté que les huit ensembles de données réelles étaient unidimensionnels. Finalement, la procédure d'analyse factorielle non linéaire a produit des résultats semblables à la procédure de Mantel-Haenszel, sauf en ce qui a trait à une des combinaisons visant à produire une épreuve avec deux dimensions où un modèle quadratique avec deux facteurs s'ajustait mieux aux données qu'un modèle avec un facteur.

En conclusion, Nandakumar suggérait d'être prudent avec DIMTEST lorsque le nombre d'items et de sujets sont petits (25 items et 500 sujets). De plus, il semble que la procédure dérivée de Mantel-Haenszel soit fortement influencée par le nombre d'items et le nombre de sujets. En effet, selon Ben-Simon et Cohen (1990), des tests plus longs et des échantillons de grande taille facilitent la détection de la multidimensionalité avec la procédure suggérée par Holland et Rosenbaum (1986).

5.7. QUELLE PROCÉDURE CHOISIR POUR DÉMONTRER L'ADÉQUATION D'UN MODÈLE

Les stratégies à suivre dans une étude de l'adéquation d'un modèle de la TRI sont toujours particulières au contexte et les outils techniques mis à contribution ne sont pas exempts de problèmes spécifiques. Dans n'importe quelle étude, la stratégie suivie est tributaire des ressources disponibles et de la profondeur de l'élaboration conceptuelle qui a présidé à la mise au point des tests et des items. Les outils privilégiés sont d'abord ceux qui sont disponibles à une époque donnée et pour lesquels il existe très souvent une procédure d'utilisation suffisamment conviviale.

De plus, les techniques mises en œuvre pour déterminer si les conditions d'utilisation des modèles sont respectées ne sont pas non plus à l'abri de problèmes de conception ou d'interprétation. Par exemple, une règle non écrite recommande qu'il est plus prudent, pour des raisons de robustesse de l'estimation des coefficients de corrélation, d'utiliser une procédure d'analyse factorielle uniquement si le nombre de sujets est dix fois plus élevé que le nombre

d'items. Même si la procédure d'analyse factorielle complète du patron de réponse n'utilise pas de matrice de corrélations, nous pouvons nous demander si elle est aussi efficace peu importe le ratio nombre de sujets/nombre d'items.

Voici un autre exemple de problème technique. Lors d'une étude de la dimensionalité avec l'approche non paramétrique de Stout, la constitution des sous-tests de vérification à partir du jugement d'un expert ne se fait pas aussi automatiquement que la procédure semble le suggérer. Il y a des décisions à prendre qui peuvent influencer la conclusion quant au statut de la dimensionalité de l'ensemble des scores. Par exemple, comme les sous-tests contiennent chacun 50 items, les sous-tests de vérification devraient intégrer environ 12 items. Cette condition d'utilisation rend le test statistique plus robuste, mais la plupart du temps elle impose au chercheur de faire le choix de ces items parmi un ensemble plus grand que douze. En effet, dans une situation réelle, l'expert réussira probablement à diviser l'ensemble des items du test en deux ou trois ensembles relativement homogènes. Il peut ainsi diviser un test de 50 items en trois sous-tests de 25, 15 et 10 items respectivement. Lequel de ces sous-tests devrait être mis à contribution et lesquels douze items devraient être choisis parmi les ensembles où on en retrouve plus de douze ? Est-ce qu'un choix aléatoire convient et est-ce que tous les choix aléatoires donnent des réponses indiquant des tendances similaires ?

Annexe 5.1

Le calcul de la statistique T de Stout

Notons d'abord U_{ijr}, le score du sujet j à l'item i pour le regroupement r et J_r le nombre de sujets dans le regroupement r.

L'estimation de la variance des scores pour le regroupement r est :

$$\hat{\sigma}_r^2 = \sum_{j=1}^{J_r} \frac{\left(Y_{jr} - \overline{Y}_{jr}\right)^2}{J_r}, \text{ où } Y_{jr} = \sum_{i=1}^{M} \frac{U_{ijr}}{M} \text{ et } \overline{Y}_{jr} = \sum_{j=1}^{J_k} \frac{Y_{jr}}{J_r}.$$

L'estimation de la variance unidimensionnelle pour le regroupement r est :

$$\hat{\sigma}_{U,r} = \sum_{i=1}^{M} \frac{\hat{p}_{ir}\left(1-\hat{p}_{ir}\right)}{M^2}, \text{ où } \hat{p}_{ir} = \sum_{j=1}^{J_r} \frac{U_{ijr}}{J_r}$$

$$\text{Soit : } \hat{\mu}_{4,r} = \sum_{j=1}^{J_r} \frac{\left(Y_{jr} - \overline{Y}_{jr}\right)^4}{J_r}$$

$$\hat{\delta}_{4,r} = \sum_{i=1}^{M} \hat{p}_{ir}\left(1-\hat{p}_{ir}\right)\left(1-2\hat{p}_{ir}\right)^2$$

$$S_r^2 = \frac{\left(\hat{\mu}_{4,r} - \hat{\sigma}_r^4\right) + \frac{\hat{\delta}_{4,r}}{M^4} + 2\left(\frac{\left(\hat{\mu}_{4,r} - \hat{\sigma}_r^4\right)\hat{\delta}_{4,r}}{M^4}\right)^{1/2}}{J_r}$$

$$\text{Alors : } T_i = \frac{1}{\sqrt{r}} \sum_{r=1}^{R} \left[\frac{\hat{\sigma}_r^2 - \hat{\sigma}_{U,r}^2}{S_r}\right]$$

$$\text{et } T = \frac{T_1 - T_2}{\sqrt{2}}$$

CHAPITRE 6

L'estimation des paramètres associés aux items et aux sujets

Dans les modèles de la TRI, la probabilité d'observer une réponse pour un item dépend de la position du sujet sur le continuum d'habileté thêta et de la position de l'item sur le ou les continuums des paramètres associés aux items. Ces positions sont en fait les valeurs qui sont attribuées aux paramètres qui caractérisent l'item, ou le sujet, et l'attribution de ces valeurs constitue l'étape de l'estimation des paramètres du modèle. Sans l'obtention d'estimations qui rencontrent des critères statistiques de qualité bien précis, les applications trouvent difficilement une légitimité dans la modélisation.

Dans la théorie statistique classique, il existe des solutions relativement simples pour une grande partie des problèmes d'estimation. Les estimateurs possèdent plusieurs qualités statistiques intéressantes : exhaustivité, absence de biais, convergence, efficacité, etc. Les méthodes du maximum de vraisemblance (*maximum likelihood*, *ML*) et des moindres carrés réussissent le

plus souvent dans les situations classiques à fournir des estimateurs rencontrant ces exigences. Pour les modèles de la TRI, le contexte d'estimation se révèle plutôt complexe : chaque sujet se voit attribuer un paramètre et chaque item est modélisé avec un, deux, trois paramètres, sans parler des modélisations polytomiques ou multidimensionnelles qui nécessitent encore plus de paramètres. Il n'existe pas de solution analytique et les estimations sont issues de procédures d'analyse numérique itérative (comme la procédure de Newton-Raphson ou encore l'algorithme EM) et posent des problèmes inhérents à ce type de solution, comme la non-convergence des estimations qui se manifeste principalement avec le modèle logistique à trois paramètres.

Un des bénéfices recherchés dans l'application des modèles de la TRI est de pouvoir travailler avec une banque d'items dont les paramètres ont déjà été estimés. En effet, selon la théorie, ces estimations possèdent la propriété d'invariance et les items peuvent donc être réutilisés dans de nouvelles situations de testing en conservant leur position respective sur le continuum des paramètres. Chaque item de la banque est donc indexé en fonction des valeurs des paramètres du modèle privilégié et la sélection ultérieure d'un item dépend des exigences spécifiques de chaque situation. Mais les estimations présentes dans la banque d'items ne tombent pas du ciel ; elles doivent être obtenues à partir des réponses observées dans une situation réelle de testing.

Ainsi, lorsqu'il s'agit d'estimer des paramètres dans les modèles de la TRI, deux situations se présentent généralement :

- les estimations pour les items sont connues et il faut obtenir les estimations pour les sujets à partir des valeurs des paramètres et des données ;
- les estimations pour les sujets et les estimations pour les items sont toutes les deux inconnues et il faut les obtenir simultanément à partir des données.

La première situation correspond à l'opération d'estimation de la valeur du paramètre thêta pour un sujet étant donné qu'il a répondu à un ensemble d'items dont la valeur des paramètres est connue. La deuxième situation correspond à l'opération de calibrage des items, c'est-à-dire à l'opération qui permettra de donner des valeurs aux paramètres de façon à ce que les items soient identifiés dans la banque en fonction de ces paramètres. Du point de vue des procédures statistiques d'estimation des paramètres, la deuxième situation est plus complexe parce que le nombre de paramètres croît rapidement en fonction du nombre d'items et de sujets. La première situation est plus simple, justement parce que le nombre de paramètres à estimer est réduit (seulement un pour chaque sujet si le modèle appliqué est unidimensionnel).

Plusieurs procédures différentes ont été proposées pour produire des estimations adéquates dans l'une ou l'autre situation. Ainsi, nous retrouvons des méthodes heuristiques (Urry, 1974 ; Jensema, 1976 ; Cohen, 1979), des méthodes robustes (Wainer et Wright, 1980 ; Mislevy et Bock, 1982a), des méthodes bayesiennes (Birnbaum, 1969 ; Meredith et Kearns, 1973 ; Owen,

1975 ; Swaminathan et Gifford, 1982, 1985, 1986). La méthode du maximum de vraisemblance a aussi été proposée sous différentes formes dont celle du maximum de vraisemblance conditionnelle (Andersen, 1972, 1973 ; Wright et Masters, 1982) et celle du maximum de vraisemblance marginale (Bock et Lieberman, 1970 ; Bock et Aitkin, 1981 ; Thissen, 1982).

À l'heure actuelle, une personne qui examine les différents programmes informatiques disponibles sur le marché constate que certaines procédures semblent recueillir la faveur de ceux qui développent les outils informatiques. Nous allons donc nous limiter dans ce chapitre à présenter les procédures les plus souvent mentionnées dans les écrits et les plus souvent présentes dans les programmes informatiques. De plus, nous ne désirons pas traiter en profondeur les différentes méthodes d'estimation, car cela dépasse les objectifs du présent ouvrage. Pour un traitement assez exhaustif du sujet, nous renvoyons le lecteur à Baker (1992).

6.1. L'ESTIMATION DE L'HABILETÉ LORSQUE LES ESTIMATIONS DES PARAMÈTRES DES ITEMS SONT CONNUES

Selon ce que nous avons vu au chapitre 5, le principe de l'indépendance locale nous permet de présenter la probabilité conjointe d'observer le patron de réponses (U_1, U_2, …, U_i, …, U_n) comme le produit des probabilités d'observer chaque réponse :

$$P(U_1, U_2, ..., U_n \mid \theta) = \prod_{i=1}^{n} P(U_i \mid \theta)$$

Dans le cas particulier où les variables U_i sont dichotomiques, l'équation précédente devient :

$$P(U_1, U_2, ..., U_n \mid \theta) = \prod_{i=1}^{n} P(U_i \mid \theta)^{U_i} \left[1 - P(U_i \mid \theta)\right]^{1-U_i}$$

ou encore :

$$P(U_1, U_2, ..., U_n \mid \theta) = \prod_{i=1}^{n} P_i^{U_i} Q_i^{1-U_i}$$

Lorsque les variables U_i se réalisent, c'est-à-dire lorsque $U_1 = u_1$, $U_2 = u_2$, …, $U_n = u_n$, la fonction obtenue est appelée la fonction de vraisemblance :

$$L(u_1, u_2, ..., u_n \mid \theta) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i}$$

Il est cependant plus pratique de considérer le logarithme naturel de la fonction de vraisemblance :

$$ln\,L(u|\theta) = \sum_{i=1}^{n} \left[u_i ln P_i + (1-u_i) ln(1-P_i) \right]$$

La valeur de thêta pour laquelle la fonction de vraisemblance (ou son logarithme naturel) est maximale s'appelle tout simplement l'estimation du maximum de vraisemblance. Généralement, le maximum de cette fonction peut être trouvé à l'aide de la procédure itérative de Newton-Raphson. Étant donné que les valeurs des paramètres des items sont connues, le maximum de la fonction de vraisemblance sera obtenu en utilisant la dérivée première et la dérivée seconde de la fonction *ln* L pour le sujet j. L'estimation $\hat{\theta}$ de θ est obtenue par itération successive en corrigeant la valeur de l'estimateur $\hat{\theta}_t$ à chaque itération t par la soustraction du rapport de la dérivée première sur la dérivée seconde. Les valeurs des dérivées dépendent du modèle privilégié (normal, logistique ou autre) et du nombre de paramètres que celui-ci intègre (un, deux ou trois). Ainsi, l'algorithme de Newton-Raphson utilise les fonctions :

$$\frac{\partial \ln L}{\partial \theta_j} = \sum_{i=1}^{n} u_{ij} \frac{1}{P_{ij}} \frac{\partial P_{ij}}{\partial \theta_j} + \sum_{i=1}^{n} (1-u_{ij}) \frac{1}{Q_{ij}} \frac{\partial Q_{ij}}{\partial \theta_j}$$

et

$$\left(\hat{\theta}_j\right)_{t+1} = \left(\hat{\theta}_j\right)_t - \left(\frac{\partial^2 \ln L}{\partial \theta_j^2} \right)_t^{-1} \left(\frac{\partial \ln L}{\partial \theta_j} \right)_t .$$

Les estimations du maximum de vraisemblance possèdent des propriétés intéressantes lorsque le nombre d'items augmente. Dans ces situations, l'estimateur $\hat{\theta}$ du maximum de vraisemblance possède une distribution normale avec une moyenne θ et une variance $1/I(\theta)$[1]. L'estimateur du maximum de vraisemblance est donc un estimateur non biaisé lorsque le nombre d'items est élevé. Étant donné que l'estimateur $\hat{\theta}$ possède une distribution normale, il est possible de construire un intervalle de confiance au niveau α pour θ de la façon classique suivante :

$$\left[\hat{\theta} - z_{\alpha/2} \frac{1}{\sqrt{I(\hat{\theta})}}, \hat{\theta} + z_{\alpha/2} \frac{1}{\sqrt{I(\hat{\theta})}} \right]$$

1. I(θ) est la fonction d'information présentée au chapitre 4.

La procédure du maximum de vraisemblance donne d'assez bons résultats lorsque les modèles normal et logistique à un paramètre sont appliqués et que le nombre d'items est supérieur à vingt. Cependant, il existe des problèmes avec tous les modèles lorsque des sujets ont répondu correctement ou incorrectement à tous les items. Dans ces situations $\hat{\theta} = +\infty$ ou $\hat{\theta} = -\infty$. Le problème est le même lorsque, face à des énoncés jumelés à des échelles du type de celles étudiées par Likert, un sujet choisit la même option pour tous les items. De plus, il y a également des problèmes avec les modélisations à trois paramètres lorsqu'il y a moins de vingt items. Il est possible en effet que dans ces situations la fonction de vraisemblance possède plusieurs maximums (Samejima, 1973). Cette dernière remarque est importante pour les situations qui prévalent en testing adaptatif (voir le chapitre 9), puisqu'un des objectifs d'une approche adaptative est de mieux cibler les items administrés à un sujet, ce qui entraîne la plupart du temps une diminution du nombre d'items administrés.

Pour pallier à ces difficultés, certains auteurs ont suggéré d'utiliser des méthodes d'estimation bayesiennes, c'est-à-dire des méthodes qui incorporent des informations préalables modifiant la fonction de vraisemblance. Par exemple, certaines méthodes posent comme hypothèse a priori que la distribution de thêta est une distribution normale (voir Swaminathan et Gifford, 1982).

En utilisant les propriétés du théorème de Bayes, nous pouvons écrire que la distribution de probabilité $f(\theta \mid u_1, u_2, \ldots, u_n)$ est :

$$f(\theta / u_1, u_2, ..., u_n) = L(u_1, u_2, ..., u_n / \theta) f(\theta)$$

ou encore :

$$f(\theta \mid u_1, u_2, ..., u_n) = \prod_{i=1}^{n} (P_i^{u_i} Q_i^{1-u_i}) f(\theta)$$

La distribution de probabilité $f(\theta \mid u_1, u_2, \ldots, u_n)$ est appelée la distribution a posteriori de θ et son mode est l'estimateur le plus probable pour θ. Cet estimateur est appelé l'estimateur modal a posteriori ou estimateur MAP.

Cependant, le mode n'est pas la seule statistique qui permet de décrire la distribution a posteriori de θ. La moyenne de la distribution a posteriori peut aussi être utilisée à cet effet. Bock et Mislevy (1982) ont ainsi proposé une façon de calculer l'espérance a posteriori de thêta en s'appuyant sur une distribution a priori obtenue à partir des données. Ils ont appelé l'estimateur découlant de cette procédure l'estimateur de l'espérance a posteriori ou estimateur EAP (*expected a posteriori*). La moyenne peut être obtenue en utilisant l'opérateur mathématique de l'espérance de la façon habituelle, soit :

$$E(\theta \mid u_1, u_2, ..., u_n) = \frac{\sum_{j=1}^{k} \theta_j f(\theta_j \mid u_1, ..., u_n)}{\sum_{j=1}^{k} f(\theta_j \mid u_1, ..., u_n)}$$

où k représente le nombre de points choisis sur l'échelle de thêta pour constituer et informer la distribution a priori de thêta.

Un avantage important des méthodes bayesiennes est qu'elles permettent l'estimation du paramètre thêta même lorsque les réponses des sujets aux items sont identiques pour tous les items. Même si certains chercheurs n'apprécient pas cette idée typiquement bayesienne de préciser des informations a priori, on voit mal comment la modélisation peut s'en tirer sans utiliser de l'information préalable, puisqu'il y a indétermination des scores parfaits ou nuls. Les auteurs du logiciel LOGIST (Wingerski, Barton et Lord, 1982), qui fait appel à la procédure du maximum de vraisemblance, ont ainsi intégré des limites inférieures et supérieures de –7 et +3 respectivement. Les auteurs du logiciel BIGSTEPS (Linacre et Wright, 1995) pour le modèle de Rasch proposent un système d'extrapolation pour produire des estimations des scores parfaits ou nuls. D'un autre côté, des logiciels comme BILOG (Mislevy et Bock, 1990) et MULTILOG (Thissen, 1991) offrent des options bayesiennes permettant de sélectionner les procédures d'estimation MAP et EAP.

6.2. L'ESTIMATION SIMULTANÉE DE L'HABILETÉ DES SUJETS ET DES PARAMÈTRES DES ITEMS

Nous l'avons mentionné, un des problèmes rencontrés lorsqu'il faut estimer simultanément le paramètre associé aux sujets et les paramètres associés aux items est que le nombre de paramètres à estimer augmente très rapidement. Neyman et Scott (1948) ont utilisé les termes « paramètre accidentel » (*incidental parameter*) et « paramètre structurel » (*structural parameter*) pour identifier les paramètres actifs, dont le nombre augmente avec le nombre d'observations, et les paramètres passifs, dont le nombre reste stable.

Lorsque le nombre de sujets et d'items augmente simultanément, il semble que les estimateurs de l'approche du maximum de vraisemblance convergent vers les valeurs réelles des paramètres (Hambleton et Swaminathan, 1985, p. 129). Cependant, le nombre d'items et le nombre de sujets doivent être relativement élevés lorsque nous désirons utiliser cette stratégie d'estimation dans le cas des modèles à deux ou trois paramètres. Ainsi, pour que les estimateurs du maximum de vraisemblance soient adéquats, il est suggéré de les utiliser uniquement si le nombre d'items est supérieur à 30 et le nombre de sujets supérieur à 500 avec le modèle à deux paramètres, et uniquement si le nombre d'items est supérieur à 50 et le nombre de sujets supérieur à 1000

avec le modèle à trois paramètres (Hulin *et al.*, 1982). De plus, alors que le nombre d'items ne peut augmenter indéfiniment, le nombre de sujets peut atteindre plusieurs dizaines de milliers dans certaines applications.

Un autre problème surgit lorsque nous désirons estimer simultanément le ou les paramètres associés aux sujets et le ou les paramètres associés aux items : il s'agit du fait que l'unité de mesure du continuum d'habileté thêta n'est pas déterminée d'une seule et unique manière. En effet, si nous ajoutons par exemple une constante à chaque estimation de thêta et la même constante à chaque estimation de la difficulté, la quantité $(\theta - b_i)$ demeure inchangée et la fonction caractéristique $P_i(\theta)$ du modèle à un paramètre également. Cela signifie, comme le précise Lord (1980, p. 36), que l'origine pour l'échelle d'habileté thêta peut être fixée arbitrairement. La procédure la plus commune consiste à fixer la moyenne de thêta à 0 et son écart-type à 1. Cependant, comme cette unité est fixée ainsi à chaque occasion où il y a estimation simultanée des paramètres des sujets et des items, l'utilisateur doit en tenir compte lorsqu'il veut comparer les estimations des paramètres des items pour des groupes différents. Un regard trop strict sur la propriété d'invariance pourrait en effet nous amener à penser que celle-ci ne tient pas, alors qu'il s'agit plutôt du problème de localiser l'origine de l'échelle d'habileté thêta, qui est indéterminée dans cette situation.

Lorsqu'un groupe de N sujets se voit administrer un ensemble de n items, la fonction de vraisemblance qui sert à estimer les paramètres θ et, disons, b devient (l'indépendance locale étant postulée) :

$$L(u_1, u_2, ..., u_n \mid \theta, b) = \prod_{i=1}^{n} \prod_{j=1}^{N} P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}}$$

Une façon simple d'estimer les paramètres consiste à procéder en deux étapes. D'abord, il faut choisir des valeurs initiales de thêta. Par exemple, nous pouvons calculer et standardiser les valeurs du logarithme du rapport entre le nombre de réponses correctes et le nombre de réponses incorrectes. Ces valeurs peuvent ensuite être utilisées comme si elles étaient connues pour estimer les valeurs des paramètres des items. La procédure en deux étapes est répétée jusqu'à ce que nous observions une certaine stabilité et que les valeurs des estimations ne changent pas trop entre deux cycles d'estimation. Cette procédure en deux étapes est celle que nous retrouvons dans le logiciel LOGIST pour les modèles à un, deux ou trois paramètres, et dans certains logiciels dédiés aux modèles de la famille de Rasch, comme BICAL, BIGSCALE, FACETS, etc. La procédure est généralement appelée procédure d'estimation conjointe du maximum de vraisemblance (*joint maximum likelihood*) et quelquefois, par certains auteurs préférant le modèle de Rasch, procédure d'estimation non conditionnelle.

La procédure d'estimation conjointe du maximum de vraisemblance ne permet pas non plus d'estimer la valeur de thêta pour un sujet qui a répondu correctement à tous les items ou incorrectement à tous les items, pas plus qu'elle ne permet d'estimer les paramètres d'un item pour lequel les réponses de tous les sujets sont identiques. Encore une fois, il faut regarder du côté des procédures bayesiennes pour obtenir une solution qui soit théoriquement plus satisfaisante à cet égard.

La procédure bayesienne la plus populaire à l'heure actuelle est la procédure dite du maximum marginal de vraisemblance (MML)[2]. Cette procédure d'estimation a été développée à l'origine par Bock et Lieberman (1970) et raffinée par Bock et Aitkin (1981). Nous la retrouvons intégrée notamment dans les logiciels TESTFACT, BILOG, BILOG-MG, MULTILOG, PARSCALE, QUEST et CONQUEST. Pour le modèle de Rasch, une autre procédure attire la faveur d'un certain nombre de chercheurs. Il s'agit de celle du maximum de vraisemblance conditionnelle (CML) qui a été suggérée à l'origine par Rasch (1960) et raffinée par Andersen (1972, 1977). Nous la retrouvons notamment intégrée aux logiciels BIGSCALE, BIGSTEPS et FACETS.

La principale différence entre ces deux approches d'inspiration bayesienne est que l'approche CML utilise le fait que pour le modèle à un paramètre de Rasch il existe une statistique exhaustive pour estimer la valeur de thêta pour un sujet et pour estimer la valeur du paramètre de la difficulté pour un item. Ainsi, la procédure MML est conditionnelle à la distribution $f(\theta)$ spécifiée par l'utilisateur, alors que la procédure CML est conditionnelle aux différents scores r observés[3]. Pour Masters et Wright (1997), le fait de ne pas avoir besoin de spécifier une distribution pour thêta rend la procédure d'estimation CML plus robuste (p. 111). Il faut évidemment vérifier dans ce dernier cas que c'est bien le modèle à un paramètre qui convient.

6.3. LA MODÉLISATION NON PARAMÉTRIQUE DE LA COURBE CARACTÉRISTIQUE D'UN ITEM

Nous avons déjà mentionné qu'il existait présentement une certaine unanimité au sujet des procédures d'estimation à privilégier dans les applications de la TRI. Alors qu'au début des années 1980, plusieurs études comparant différentes méthodes d'estimation étaient régulièrement publiées, il faut constater qu'à l'heure actuelle les nouvelles candidates ne semblent pas ajouter d'une

2. Une consultation rapide du *Handbook* de Van der Linden et Hambleton (1997) permettra au lecteur de constater que la méthode MML est effectivement celle privilégiée par un grand nombre de chercheurs.
3. S'il y a n items, il y a n+1 scores possibles.

façon significative à la qualité des méthodes d'estimation les plus populaires, c'est-à-dire les méthodes MML et CML. Cependant, nous devons rappeler, ême brièvement, l'existence de modèles non paramétriques qui, bien que ne comportant aucun paramètre comme leur appellation l'indique, n'en demandent pas moins une forme d'estimation de la courbe caractéristique de l'item.

La modélisation non paramétrique des réponses aux items d'un test remonte bien avant la popularité actuelle de la TRI. Par leurs travaux, Guttman (1950) et Lazarsfeld (1950) font figure de pionniers en la matière (Mokken, 1997). La modélisation non paramétrique s'appuie sur le fait que, pour plusieurs des variables rencontrées en sciences sociales, il est difficile de vérifier l'hypothèse que l'échelle de mesure possède les propriétés d'une échelle d'intervalle[4]. Il est donc préférable dans ces circonstances de se rabattre sur l'hypothèse de la présence d'une échelle ordinale. Ainsi, il est possible de vérifier la pertinence d'ordonner les sujets selon leurs réponses aux items et la pertinence d'ordonner les items selon les réponses données par les sujets. Pour Ramsay (1997), la modélisation non paramétrique a l'avantage de permettre un examen de l'ensemble de la fonction caractéristique et une prise de distance par rapport à un examen de la modélisation uniquement centrée sur les paramètres de difficulté et de discrimination.

Différentes propositions de modélisation non paramétrique ont été passées en revue au chapitre 4 ; on trouve des présentations intéressantes dans Sijtsma (1998) et dans certains chapitres de Van der Linden et Hambleton (1997). Des propositions de modélisation non paramétrique peuvent être appliquées en utilisant les logiciels TESGRAF (Ramsay, 1993) et MSP (Molenaar *et al.*, 1994). Les travaux inspirés des propositions de Stout (1987, 1990) sont également à ranger dans la catégorie des approches non paramétriques ; ils sont notamment intégrés aux logiciels DIMTEST, SIBTEST, POLY-SIBTEST et DETECT. Finalement, les travaux de Rosenbaum (1984) et de Holland et Rosenbaum (1986), dont nous avons parlé au chapitre précédent, doivent être aussi placés dans la catégorie des modélisations non paramétriques.

4. Certaines personnes soutiennent que les variables « humaines » des sciences sociales ne peuvent être considérées comme possédant les caractéristiques d'une échelle d'intervalle. Nous ne désirons pas aborder cette polémique dans le cadre de cet ouvrage le lecteur qui désire en avoir un aperçu pourra consulter Michell (1999).

CHAPITRE 7

Du concept de validité

Ce chapitre constitue une coupure par rapport aux précédents, du moins quant au niveau de technicité qu'il véhicule. Après avoir insisté sur les tenants et les aboutissants des modèles de mesure, principalement des modèles de la théorie des réponses aux items, présentation qui nous a tout de même permis de quantifier l'erreur de mesure aléatoire en nous appuyant sur des concepts comme la fidélité, la généralisabilité ou l'information, nous porterons notre intérêt sur un concept de la plus haute importance mais parfois galvaudé, souvent mal défini ou défini sommairement, sans véritable effort d'opérationnalisation : il s'agit, on l'aura compris, du concept de validité.

Nous donnerons une définition de la validité qui s'appuie sur les propositions tout aussi élégantes qu'opérationnelles de Samuel Messick. Cette approche s'éloigne de la conception traditionnelle de la validité (un test est valide s'il mesure bien ce qu'il prétend mesurer), que nous trouvons un tantinet simpliste et pas suffisamment opérationnelle. De notre point de vue, ce sont les interprétations des scores au test qui doivent être considérées valides

ou non, pas le test en lui-même. Sans proposer une procédure fixe et structurée pour valider les interprétations des scores à un test, procédure qui risquerait d'être prise pour une recette, nous suggérons d'avoir recours à des méthodes éprouvées comme l'analyse factorielle et de procéder à une étude approfondie des biais qui limitent la validité.

Il ne sera donc pas trop étonnant de constater que l'analyse factorielle fait partie de ce chapitre. La présentation de quelques concepts de base en analyse factorielle permettra d'expliquer en quoi cette procédure tout usage, d'ailleurs souvent utilisée à mauvais escient, sera employée pour autopsier des construits complexes et isoler leurs principales composantes.

Plusieurs types de biais peuvent affecter l'interprétation des scores : nous avons cru bon de distinguer les biais liés à l'instrument lui-même et à son administration (ou à la façon de l'utiliser) des biais liés à la façon de répondre des sujets testés. Parmi les méthodes qui ont été développées pour identifier les biais qui touchent la façon d'utiliser un instrument, l'analyse factorielle a encore ici une place prépondérante. Nous pourrons apprécier enfin jusqu'à quel point les concepts de la théorie des réponses aux items sont utilisés pour développer des méthodes visant à détecter des biais produits par la façon de répondre des sujets.

7.1. RÉFLEXIONS CONCEPTUELLES

Plutôt que, d'entrée de jeu, plaquer une définition de la validité, il nous paraît plus à propos de discuter, à l'instar de Suen (1990, p. 134) de ce que la validité n'est pas, du moins à nos yeux, c'est-à-dire des fausses conceptions généralement entretenues en rapport avec ce concept. Ainsi, il est plus ou moins vrai de dire que nous allons valider un test ou un instrument. En ce sens, il n'est pas réellement approprié de brandir un test en alléguant qu'il est valide. Il est beaucoup mieux de préciser qu'il s'agit d'accumuler des évidences de validité à propos des interprétations faites à partir des scores à un test. Ainsi, un test n'est pas valide en soi, de façon absolue : il faut apporter et analyser des preuves empiriques avant de parler de validité. Il ne s'agit donc pas de valider un test, mais bien les interprétations ou inférences faites à partir des scores à ce test et ce, dans un contexte donné. D'ailleurs, plusieurs évidences empiriques doivent être collectées à cette fin. La validité ne se résume pas non plus à un coefficient ou à un indice. S'il est envisageable, comme on l'a vu, de résumer la fidélité, la généralisabilité ou même l'information à un indice, à une valeur, et donc de parler de coefficient de fidélité ou de courbe d'information, il n'en est pas du tout de même pour la validité. Parler d'un coefficient de validité peut même entretenir l'idée que la validité peut se résumer à une seule valeur. Avant de parler de validité, il faudra faire l'effort d'accumuler plusieurs indicateurs, plusieurs évidences qui permettront d'étayer les interprétations ou inférences alléguées à partir des scores au test.

Explicitement ou non, plusieurs laissent entendre aux personnes en charge de développer le test qu'il revient exclusivement aux personnes en charge de développer le test, de chercher et de publier des évidences de validité. Bien sûr, les personnes en charge du développement doivent fournir ces évidences probants, mais ce ne sont pas les seules à devoir en fournir. Les utilisateurs du test ont aussi cette responsabilité. Ils doivent aussi fournir des évidences sur la base de données collectées selon leur protocole de recherche. Il n'est pas garanti qu'une échelle d'anxiété envers les ordinateurs développée aux États-Unis donnera, une fois traduite et adaptée pour le Québec, les mêmes évidences de validité que le suggèrent les concepteurs américains. C'est donc (aussi) à l'utilisateur de test que revient la responsabilité de chercher ces évidences, de fournir des preuves de validité, des preuves qui justifient l'interprétation que l'on fait à partir des scores provenant du test.

Soutenir qu'il existe plusieurs sortes de validité, à savoir la validité de contenu, la validité liée à un critère et la validité conceptuelle est en quelque sorte hérétique, du moins selon notre conception de la validité. Émanant en bonne partie des *Standards for educational and psychological testing* (American Psychological Association, 1985, 1992, 1999), notre conception de la validité soutient plutôt qu'il existe plusieurs façons ou stratégies visant à collecter des évidences de validité. Nous donnerons un aperçu sommaire de ces stratégies de validation un peu plus loin dans le texte.

Autre fausse conception : la validité relative à ce test, à cet instrument a été établie une fois pour toutes. Au contraire, la validité tient d'un processus qui n'est jamais fini. Les résultats d'un processus de validation établis à un moment donné peuvent varier avec le temps, se raffiner à mesure que les évidences collectées s'additionnent. Par exemple, il est bien possible, compte tenu de la démocratisation croissante des outils informatisés, que les évidences de validité associées à une échelle d'anxiété face aux ordinateurs développée dans les années quatre-vingt évoluent avec le temps. Il faut tenir compte du contexte, de la population visée par un test (un test de résolution de problèmes peut être valide pour une sous-population de bons lecteurs, mais non valide pour une sous-population de mauvais lecteurs ; un test d'items à choix multiple peut être valide pour des Canadiens, mais non pour des Africains ; une procédure de sélection comme l'appréciation par simulation (APS) peut être considérée très utile pour choisir des cadres intermédiaires dans une PME, mais beaucoup moins pertinente pour sélectionner les meilleurs candidats aux études en médecine).

Tous ces propos nous amènent à tenter une définition de la validité, moins classique mais aussi plus opérationnelle et moins facile que « un test est dit valide s'il mesure bien ce qu'il prétend mesurer ». La définition de la validité que nous proposons prend appui pour une bonne part sur les définitions proposées par Messick (1988, 1995) :

> *La validité consiste en un jugement basé sur des preuves empiriques et sur une argumentation de nature théorique qui vise à justifier l'interprétation des scores obtenus à la suite de l'administration d'un test dans un contexte donné.*

Nous n'avons pas voulu présenter, et encore moins définir, trois ou quatre types de validité comme on le fait souvent parce que, comme nous l'avons affirmé avant, il n'existe tout simplement pas trois ou quatre types de validité, du moins pas selon notre conception. Nous n'avons pas voulu non plus trop insister sur la validité conceptuelle en tant que concept unitaire, bien que cette idée soit beaucoup plus près de ce que nous pensons vraiment lorsqu'on fait référence à la validité. Nous pensons que les propos de Cronbach et Meehl[1] (1955, p. 300), de Loevinger[2](1957, p. 636), de Messick[3] (1980, p. 1015), des *Standards*[4] (American Psychological Association, 1985, p. 9) puis de Messick[5] (1988, p. 33) font partie de l'évolution du concept de validité : c'est cependant grâce à ce genre de propos, émanant d'éminents chercheurs et penseurs, que nous avons pu arrêter notre propre définition de la validité.

Si la validité est une évaluation, un jugement, la validation est un processus, à savoir le processus qui mène à valider les interprétations faites en prenant en compte les scores émanant du test. Suivant la conception de validité que nous avons exposée, s'il n'existe qu'un seul type de validité, il y a cependant plusieurs façons ou stratégies utilisées pour valider les interprétations ou inférences faites à partir des scores à un test. Pour autant, ce ne serait pas exact de soutenir qu'il faille utiliser toutes ces stratégies de validation simultanément : le contexte de la recherche déterminera le type de stratégie le plus approprié. Une première stratégie consiste à déterminer la pertinence du contenu à l'aide d'un panel d'experts : ainsi, s'il s'agit d'un examen de rendement scolaire, il serait approprié de collecter des évidences de validité fondées sur l'analyse du contenu de l'examen par des experts de ce contenu. Une seconde stratégie consiste à vérifier si le test prédit bien un ou des critères que ce test prétend prédire : ce serait le cas, par exemple, pour un instrument utilisé à des fins de sélection des étudiants en médecine. Une troisième stratégie, enfin,

---

1. « *Construct validity cannot generally be expressed in the form of a single simple coefficient.* »
2. « *Since predictive, concurrent and content validies are all essentially ad hoc, construct validity is the whole of validity.* »
3. « *Construct validity is indeed the unifying concept of validity that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships.* »
4. « *Validity always refers to the degree to which evidence supports the inferences that are made from the scores.* ».
5. « *Validity is an overall evaluative judgement, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of inferences and actions based on test scores.* »

vise à faire l'autopsie du ou des concepts mesurés par le test et de le relier à d'autres concepts bien connus. Ainsi, dans le cas d'une échelle d'anxiété face aux ordinateurs, il pourrait être légitime de chercher les principales composantes de cette forme d'anxiété (p. ex., composante 1 : peur de se blesser avec des appareils électroniques ; composante 2 : peur de se perdre dans la structure des systèmes d'exploitation ; etc.), ou encore des corrélations avec d'autres échelles mesurant parfois cette forme d'anxiété ou une forme similaire, parfois d'autres formes d'anxiété (envers les chats, les ascenseurs, etc.).

Les deux premières stratégies visant à valider les interprétations faites à partir des scores à un test ne feront pas l'objet de plus amples développements. Nous suggérons au lecteur intéressé par ce sujet de consulter Laveault et Grégoire (2002). C'est la troisième stratégie qui sera l'objet de nos propos au cours des prochains paragraphes. Du moins, les méthodes dont nous allons discuter renvoient principalement à cette troisième stratégie, que nous appellerons la **validation conceptuelle**. Celle-ci implique, d'une part, d'identifier les principales composantes du concept véhiculé par le test et, d'autre part, de situer le concept dans un réseau nomologique. C'est notamment par le biais de l'analyse factorielle que nous pourons y parvenir : c'est pourquoi nous trouvons raisonnable de consacrer une section de ce chapitre à cette technique si importante.

En concordance avec ce que nous avons déjà dit de la validité, le processus de validation d'un instrument est une aventure complexe qui, comme on l'a déjà indiqué, ne s'arrête jamais. Il sera rarement suffisant d'appliquer, même très bien, l'une ou l'autre des trois stratégies décrites plus haut. Il faut également, comme le précise Messick (1988, p. 39), « écarter les hypothèses rivales[6] » relatives à l'interprétation des scores au test. En d'autres mots, il est aussi nécessaire d'étudier puis, éventuellement, de mettre de côté les interprétations qui ne sont apparemment pas pertinentes. Disons, par exemple, qu'un test de résolution de problèmes mathématiques est administré à des élèves de 6e année en adaptation scolaire. Ce test a été choisi parce qu'il comportait des indications claires sur sa fidélité et sa validité, du moins tel qu'exposé par les concepteurs. Or, ce test n'avait jamais été utilisé avec des élèves en adaptation scolaire, donc des élèves qui risquent d'avoir de la difficulté à lire le texte inhérent aux problèmes et à comprendre la tâche, ce qu'on attend d'eux. Bref, les scores au test seraient tout autant dus à l'habileté à lire qu'à l'habileté mathématique en elle-même. Dit autrement, l'interprétation devrait tenir compte, au minimum, de deux aspects non négligeables : l'habileté à lire (pourtant négligeable dans le cas des bons lecteurs) et l'habileté mathématique. Par exemple, affirmer que Zoé est plutôt faible en mathématique parce

6. « *discounting plausible rival hypotheses* ».

qu'elle n'a obtenu que 64 % à l'examen de résolution de problèmes mathématiques du Ministère (qui ne comportait que des problèmes longs) pourrait être un énoncé valide si l'hypothèse rivale « Zoé est faible en lecture » avait été écartée. Autre exemple : nous voyons de plus en plus de tests papier-crayon adaptés pour qu'ils soient administrés à l'aide d'un logiciel quelconque, de façon adaptative ou non (voir le chapitre 9). Or, il n'est pas du tout certain que le test informatisé mesurera exactement le même concept que la version originale papier-crayon ; du moins, cela peut dépendre des circonstances, des populations de personnes visées par le test. Si ces personnes (comme des chômeurs d'âge mûr ou des ressortissants de pays en voie de développement) sont intimidées par un ordinateur, il y a menace que le test ne mesure pas seulement « ce qu'il prétend mesurer » ou que les interprétations faites à partir des scores à ce test soient, au moins en partie, erronées. Ce genre de biais est une menace à la validité. Vu comme cela, cette facette du processus de validation des inférences faites à l'égard d'un test n'est pas très différente de ce que Campbell et Stanley (1963) ou Cook et Campbell (1979) ont proposé pour assurer la validité interne d'une recherche : étudier puis, éventuellement, contrôler les biais qui menacent l'interprétation des résultats de la recherche, donc la validité interne. Nous allons étudier, au cours d'une prochaine section de ce chapitre, la nature de certains biais associés au processus de mesure.

Enfin, il ne faut pas se limiter aux biais liés à l'administration ou à l'utilisation de l'instrument lui-même, mais discuter également des biais liés à la façon de répondre de la personne à qui est destiné le test : ce qui pourrait être appelé une façon stéréotypée de répondre ou un *response set*. Pensons une minute à un élève de 6e année qui réussit l'essentiel des items difficiles d'un test, mais qui échoue la plupart des items faciles. Bien que nous puissions trouver, à l'aide d'une modélisation classique ou TRI, un score pour cet élève, il n'en demeure pas moins que son patron de réponses doit être considéré atypique, voire bizarre. Toute interprétation faite à partir d'un tel score risquerait d'être peu valide. Après tout, l'élève qui a produit ce patron aberrant est peut-être un tricheur ou un chanceux, à moins qu'il ne maîtrise pas bien la langue employée dans les questions du test.

7.2. L'ANALYSE FACTORIELLE

Ce n'est aucunement notre intention ici de prétendre à une présentation exhaustive de l'analyse factorielle. Il s'agit d'une méthode beaucoup trop élaborée pour qu'on puisse la cerner dans une section, voire un chapitre complet. Des dizaines et des dizaines de volumes ont déjà été consacrés exclusivement à la présentation des tenants et aboutissants de cette méthode à plusieurs facettes. Nous voulons tout au plus en faire une présentation qui, quoique sommaire, demeure selon nous incontournable lorsqu'il est question de valider un instrument de mesure, d'en trouver le nombre de dimensions (comme

en TRI), de les interpréter ou encore d'étudier les biais de concept (voir la section 7.3). Plutôt que d'en faire une présentation technique (incluant les très nombreuses procédures d'extraction et de rotation), ce qui est archi-facile en plus d'avoir déjà été fait ailleurs, nous adoptons une approche conceptuelle visant, en quelques paragraphes, à donner à l'utilisateur quelques notions de base de cette méthode. Le lecteur intéressé à poursuivre son étude de l'analyse factorielle pourra toujours consulter l'un ou l'autre des ouvrages suivants : Harman (1976), Nunnaly (1978), Kline (1994).

7.2.1. Un premier exemple : le *Thurstone box problem*

L'analyse factorielle regroupe un ensemble impressionnant de procédures visant à réduire les contours d'un problème ou d'une situation de façon à mieux l'étudier. Partant d'une situation définie par plusieurs dimensions (p. ex., l'étude des sous-concepts d'un test défini par plusieurs items), il s'agit de ramener ce nombre de dimensions à un nombre plus petit de dimensions significatives. L'analyse factorielle du fameux *Thurstone box problem* permettra de bien saisir cette procédure de réduction du nombre de dimensions (Harman, 1976, p. 156).

Imaginons que nous devions classer 20 boîtes de carton, lesquelles ont été mesurées sur neuf variables[7] : x^2, y^2, z^2, exp(x), exp(y), exp(z), Log(x), Log(y) et Log(z), où x est la longueur, y la largeur et z la hauteur de la boîte. Quel est le plus petit nombre de variables nécessaires pour classer ces boîtes ? En d'autres termes, comment réduire l'ampleur de ce problème initialement représenté dans un espace à neuf dimensions ? Quel est le nombre de dimensions, manifestement entre 1 et 9, qui permettraient de classer ces boîtes en perdant le moins d'information possible, sachant qu'en conservant les neuf dimensions pour classer les boîtes signifie ne pas perdre d'information du tout ? Il s'agit, en d'autres termes, d'un problème de parcimonie : faire aussi bien avec un dispositif beaucoup moins lourd.

Le tableau 7.1 indique les corrélations entre les mesures des boîtes prises deux à deux. On y voit par exemple que les corrélations entre les trois mesures associées à la longueur d'une boîte, x^2, exp(x) et Log(x), sont très élevées. Sont aussi très élevées les corrélations entre les trois mesures de largeur, y^2, exp(y) et Log(y), et les corrélations entre les trois mesures de hauteur, z^2, exp(z) et Log(z). On peut observer ce phénomène en examinant les trois parties ombrées de cette matrice de corrélations : toutes ces valeurs de corrélations sont particulièrement élevées. Ces parties ombrées formées de corrélations particulièrement élevées forment ce que nous appellerons des regroupements de variables initiales.

7. Dans la version originale, 20 variables ont été utilisées.

L'analyse factorielle, qui est en fait basée sur l'analyse de la matrice de corrélations, vise à extraire autant de facteurs (nouvelle variable définie à l'aide de corrélations entre les variables initiales) qu'il y a de tels regroupements de variables initiales. Puisque l'analyse de la matrice de corrélations présentée au tableau 7.1 montre trois regroupements de variables, on devrait logiquement extraire trois facteurs de cette matrice.

TABLEAU 7.1
Matrice de corrélations entre neuf variables extraites du *Thurstone box problem* (Harman, 1976, p. 156)

	x^2	Log(x)	exp(x)	y^2	Log(y)	exp(y)	z^2	Log(z)	exp(z)
x^2	1,000	0,987	0,980	0,262	0,213	0,295	0,098	0,104	0,093
Log(x)	0,987	1,000	0,937	0,288	0,237	0,322	0,097	0,101	0,092
exp(x)	0,980	0,937	1,000	0,220	0,175	0,250	0,097	0,105	0,090
y^2	0,262	0,288	0,220	1,000	0,978	0,984	0,247	0,198	0,260
Log(y)	0,213	0,237	0,175	0,978	1,000	0,924	0,299	0,246	0,312
exp(y)	0,295	0,322	0,250	0,984	0,924	1,000	0,194	0,151	0,206
z^2	0,098	0,097	0,097	0,247	0,299	0,194	1,000	0,949	0,991
Log(z)	0,104	0,101	0,105	0,198	0,246	0,151	0,949	1,000	0,898
exp(z)	0,093	0,092	0,090	0,260	0,312	0,206	0,991	0,898	1,000

L'examen au tableau 7.2 de la matrice des saturations (corrélations entre une variable initiale et un facteur), produit de l'analyse factorielle, révèle effectivement trois facteurs, chacun renvoyant à l'un ou l'autre des trois regroupements déjà observés. Chaque facteur est défini en tenant compte des variables initiales dont les saturations sont les plus élevées. Dans le cas du facteur 1, par exemple, trois variables sont impliquées : x^2, Log(x) et exp(x). Les saturations faibles (p. ex., < 0,3) indiquent que la variable n'est pas vraiment reliée au facteur : par exemple, puisque la variable Log(z) n'a qu'une corrélation très faible (0,054) avec le facteur 1, elle ne sera pas considérée comme faisant partie de ce facteur, c'est-à-dire qu'elle ne servira pas à définir ce facteur.

Le résultat de cette analyse factorielle montre que seulement trois dimensions, plutôt que neuf, pourraient suffire pour traiter le problème du classement des boîtes. L'examen de la matrice des saturations révèle que ces trois dimensions ou facteurs peuvent être définis comme suit : le facteur 1 correspond à la longueur x, le facteur 2 à la largeur y et le facteur 3 à la hauteur z. Ce résultat n'a bien sûr rien d'étonnant. Cependant, l'exercice en soi est d'un intérêt certain puisqu'il a pour effet de montrer que les résultats d'une analyse factorielle peuvent aussi être conformes à une certaine réalité. Même s'il s'agit d'un exercice un peu trivial, il n'en demeure pas moins qu'il exprime bien l'objectif propre à l'analyse factorielle : réduire l'ampleur d'un problème ou d'une situation de façon à mieux en étudier les contours. La plupart des problèmes rencontrés en éducation et en psychologie sont cependant beaucoup plus complexes et comportent des solutions beaucoup moins triviales, comme nous allons le voir prochainement.

Tableau 7.2
Matrice des saturations après rotation des trois facteurs pour les neuf variables extraites du *Thurstone box problem* (Harman, 1976)

	Facteur 1	Facteur 2	Facteur 3
x^2	0,991	0,129	0,043
exp(x)	0,979	0,086	0,048
Log(x)	0,972	0,159	0,038
y^2	0,131	0,984	0,118
exp(y)	0,170	0,968	0,065
Log(y)	0,081	0,963	0,177
z^2	0,040	0,127	0,989
exp(z)	0,032	0,144	0,970
Log(z)	0,054	0,077	0,962

7.2.2. Quelques concepts nécessaires à la compréhension du déroulement d'une analyse

Notre objectif ici est de présenter les concepts nécessaires à la compréhension des bases de l'analyse factorielle de façon à pouvoir juger de l'à-propos du nombre de facteurs extraits (dans une situation non triviale) et de l'interprétation donnée à ces facteurs. Il est clair que cette présentation des concepts de base ne suffira pas à former des spécialistes de l'analyse factorielle. Nous avons cependant l'ambition de permettre au lecteur de développer un œil plus critique par rapport aux résultats d'une analyse factorielle.

Le point de départ d'une analyse factorielle est, la plupart du temps, une matrice de corrélations entre des variables que nous qualifierons de **variables initiales**. Disons, pour fixer les idées, que nous allons désormais nous intéresser à l'analyse de la structure interne d'un instrument de mesure. Ainsi, les variables initiales seront les items de l'instrument. Le point de départ de l'analyse sera donc la matrice des corrélations interitems.

Gardons à l'esprit que l'analyse factorielle vise à extraire autant de facteurs qu'il y a de regroupements entre les variables initiales, c'est-à-dire de groupes de variables initiales bien corrélées entre elles. Ces facteurs, appelés aussi variables latentes, sont en général inconnus, sauf dans des cas limites comme celui du *Thurstone box problem,* où nous savions à l'origine qu'il y avait trois regroupements, donc trois facteurs, puisque les boîtes de carton sont des entités repérables dans un espace à trois dimensions. En théorie, donc, une analyse factorielle pourrait se faire uniquement en examinant les regroupements de variables initiales dans la matrice de corrélations car c'est exactement de cela qu'il s'agit. Or, sauf dans des cas très simples comme celui du *Thurstone box problem*, l'examen d'une matrice de corrélations dans l'espoir d'y extraire les regroupements voulus s'avère extrêmement onéreux sinon impossible, en plus de comporter une composante subjective qui risque fort de biaiser l'opération. C'est pourquoi plusieurs techniques ont été mises de

l'avant pour identifier le nombre et la nature des facteurs. Nous n'entrerons pas dans le détail de ces techniques, ce qui constitue souvent l'essentiel des volumes sur l'analyse factorielle ; plutôt, nous les tiendrons pour acquises et tenterons de concentrer nos efforts sur l'interprétation des résultats. La trousse statistique SPSS servira à produire les statistiques et graphiques nécessaires aux interprétations.

Green, Salkind et Akey (2000, p. 302) traitent d'une échelle d'humour en cinq points (1= complètement en désaccord ; 5 = complètement en accord) administrée à 100 universitaires et comprenant les dix items que nous avons traduits de la façon suivante :

Item01 – J'aime rire des autres
Item02 – Je fais rire les gens en riant de moi-même
Item03 – Les gens me trouvent drôle quand je fais des blagues sur les autres
Item04 – Je parle de mes problèmes pour faire rire les gens
Item05 – Les autres font souvent l'objet de mes blagues
Item06 – Les gens me trouvent drôles quand je leur parle de mes défauts
Item07 – J'aime bien faire rire les gens en étant sarcastique
Item08 – Je suis plus comique quand je parle de mes propres faiblesses
Item09 – Je fais rire les gens en montrant les bêtises des autres
Item10 – Je suis comique quand je dis aux gens quelles sottises j'ai pu faire

La matrice des corrélations interitems est présentée au tableau 7.3. Une étude sommaire mais avertie de ce tableau montre qu'un premier regroupement pourrait être formé par les items impairs puisque les items 1, 3, 5, 7 et 9 montrent des corrélations plus importantes entre eux qu'avec les autres items : de la même façon, un second regroupement pourrait être caractérisé par les items pairs. Notons aussi, au passage, que les valeurs des corrélations sont beaucoup moins élevées que dans le cas du *Thurstone box problem*, rendant l'identification visuelle des regroupements d'autant plus périlleuse et le recours à des techniques sophistiquées d'autant plus nécessaire.

TABLEAU 7.3
Matrice de corrélations entre les dix items de l'échelle d'humour présentée par Green, Salkind et Akey (2000, p. 302)

	Item01	Item03	Item05	Item07	Item09	Item02	Item04	Item06	Item08	Item10
Item01	1,000	0,268	0,221	0,233	0,071	–0,218	–0,116	–0,089	–0,024	–0,136
Item03	0,268	1,000	0,591	0,420	0,338	–0,100	0,116	0,174	0,041	–0,056
Item05	0,221	0,591	1,000	0,297	0,345	–0,024	0,163	0,071	0,038	–0,042
Item07	0,233	0,420	0,297	1,000	0,274	–0,165	–0,058	0,056	–0,146	–0,010
Item09	0,071	0,338	0,345	0,274	1,000	–0,130	–0,010	0,057	–0,070	0,098
Item02	–0,218	–0,100	–0,024	–0,165	–0,130	1,000	0,575	0,294	0,245	0,284
Item04	–0,116	0,116	0,163	–0,058	–0,010	0,575	1,000	0,261	0,434	0,182
Item06	–0,089	0,174	0,071	0,056	0,057	0,294	0,261	1,000	0,381	0,554
Item08	–0,024	0,041	0,038	–0,146	–0,070	0,245	0,434	0,381	1,000	0,369
Item10	–0,136	–0,056	–0,042	–0,010	0,098	0,284	0,182	0,554	0,369	1,000

L'analyse en composantes principales est une des multiples techniques employées en analyse factorielle. Elle vise à expliquer les corrélations entre les variables initiales, telles qu'elles sont présentées dans une matrice de corrélations, en définissant un certain nombre de nouvelles variables appelées dimensions ou facteurs qui expliqueront tour à tour le maximum des covariations entre les variables initiales. Le premier facteur sera celui qui expliquera le plus ces corrélations ; le second facteur expliquera le plus les corrélations partielles, c'est-à-dire une fois seulement que nous aurons tenu compte du premier facteur (il expliquera donc au maximum les résidus), etc. Si le premier facteur explique presque toutes les covariations entre les variables initiales, il n'y aura qu'un seul facteur significatif, donc un seul regroupement. Ce sera le cas si toutes les variables initiales sont suffisamment reliées entre elles deux à deux, comme dans l'exemple du *Saxon Carreer Scale* traité par Green, Salkind et Akey (2000, p. 311), dont la matrice de corrélations se trouve au tableau 7.4.

TABLEAU 7.4
Matrice de corrélations entre les dix items de l'échelle de choix de carrière présentée par Green, Salkind et Akey (2000, p. 311)

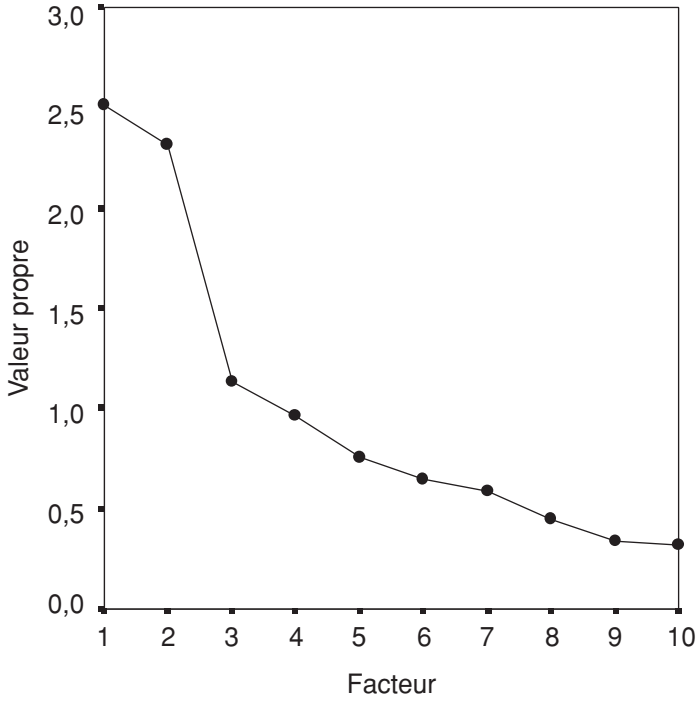
	Item a	Item b	Item c	Item d	Item e	Item f	Item g	Item h	Item i	Item j	Item k	Item l
Item a	1,000	0,512	0,448	0,452	0,468	0,470	0,392	0,498	0,365	0,440	0,534	0,496
Item b	0,512	1,000	0,450	0,485	0,524	0,545	0,427	0,555	0,506	0,532	0,437	0,427
Item c	0,448	0,450	1,000	0,343	0,460	0,399	0,362	0,391	0,549	0,355	0,259	0,510
Item d	0,452	0,485	0,343	1,000	0,437	0,398	0,416	0,525	0,509	0,448	0,376	0,526
tem e	0,468	0,524	0,460	0,437	1,000	0,390	0,440	0,412	0,507	0,622	0,415	0,560
Item f	0,470	0,545	0,399	0,398	0,390	1,000	0,369	0,517	0,433	0,500	0,476	0,393
Item g	0,392	0,427	0,362	0,416	0,440	0,369	1,000	0,512	0,417	0,473	0,435	0,488
Item h	0,498	0,555	0,391	0,525	0,412	0,517	0,512	1,000	0,431	0,578	0,477	0,461
Item i	0,365	0,506	0,549	0,509	0,507	0,433	0,417	0,431	1,000	0,467	0,303	0,413
Item j	0,440	0,532	0,355	0,448	0,622	0,500	0,473	0,578	0,467	1,000	0,488	0,503
Item k	0,534	0,437	0,259	0,376	0,415	0,476	0,435	0,477	0,303	0,488	1,000	0,392
Item l	0,496	0,427	0,510	0,526	0,560	0,393	0,488	0,461	0,413	0,503	0,392	1,000

Par contre, si les corrélations entre les variables initiales sont très faibles, il pourrait être nécessaire d'extraire autant de facteurs significatifs que de variables initiales : on conclurait dans ce cas au manque d'utilité de l'analyse factorielle puisque le principe de parcimonie n'aurait pas été respecté.

Un des bons outils pour décider du nombre de facteurs significatifs est le graphique des éboulis (figure 7.1) qui présente, en ordre d'importance, les facteurs successivement extraits de la matrice de corrélations. La valeur propre (*eigenvalue*) précise l'importance du facteur : c'est en quelque sorte la part de variance expliquée par le facteur. Ce graphique montre que la solution à deux facteurs est très plausible puisque les deux premiers facteurs semblent posséder plus ou moins la même importance (valeur propre du facteur 1 = 2,5 ; valeur propre du facteur 2 = 2,3) alors que les autres facteurs (3 à 10) ont des valeurs sensiblement semblables et nettement plus faibles que les deux premiers facteurs. L'éboulis (une chute soudaine de pente dans la ligne brisée suivie d'une absence de pente) semble s'être produit après le deuxième facteur. Il existe bien sûr beaucoup d'autres critères que le graphique des éboulis pour choisir le nombre de facteurs à extraire d'une matrice de corrélations. Un des critères les plus populaires est celui de la valeur propre (part de variance) supérieure à 1. En effet, puisque la somme des valeurs propres est égale à n, le nombre de variables[8] initiales, on pourrait dire que chaque variable initiale possède la même valeur (la même importance), soit 1. Par conséquent, si un facteur obtient une valeur propre supérieure à 1, c'est que sa part de variance est supérieure à celle de chacune des variables initiales, donc que le facteur fait mieux que l'une ou l'autre des variables initiales. C'est en ce sens que, selon ce critère, nous dirons qu'un facteur est significatif. Or, il a été montré (Kline, 1994, p. 75) que ce critère est parfois trop libéral : l'utiliser résulte souvent en l'extraction d'un trop grand nombre de facteurs. Nous proposons, en conséquence, d'utiliser conjointement les deux critères : la valeur propre supérieure à 1 et les valeurs propres les plus élevées situées en haut de l'éboulis lors de l'examen du graphique des éboulis[9].

8. En analyse factorielle, il est de coutume de supposer que les variables initiales sont standardisées et donc que la variance de chacune est de 1.
9. Avouons qu'il entre une bonne part de subjectivité dans l'examen du graphique des éboulis, mais l'expérience montre que, combiné au critère de valeur propre supérieure à 1, il s'agit tout de même d'un outil très utile lorsque vient le temps de choisir le nombre de facteurs. Soulignons qu'il est possible de produire sans peine les valeurs propres et le graphique des éboulis en employant l'une ou l'autre des trousses statistiques connues comme SPSS ou SAS.

FIGURE 7.1
Graphique des éboulis des facteurs extraits de la matrice de corrélations entre les dix items de l'échelle d'humour présentée par Green, Salkind et Akey (2000, p. 302)



Une fois fixé le nombre de facteurs, il faut définir, c'est-à-dire interpréter, chacun de ces facteurs. Afin d'y parvenir, il est généralement admis que nous devons en arriver à une structure simple, à savoir que la matrice des saturations, c'est-à-dire la matrice des corrélations entre chaque variable initiale (V_i) et chaque facteur (F_j), ait plus ou moins l'allure que l'on peut observer au tableau 7.5 : chaque variable est saturée sur au plus un facteur et chaque facteur comporte un nombre restreint de saturations élevées. Les 1 dans le tableau signifient des corrélations qui approchent la valeur 1 et les 0 des corrélations qui approchent la valeur 0. Bien sûr, nous n'aurons jamais affaire à un tel cas idéal, mais il s'agit de la matrice-cible, celle que nous visons. Soulignons qu'il est souvent d'usage en analyse factorielle d'employer un seuil de 0,3 pour distinguer les saturations faibles, tendant vers 0, des saturations élevées, tendant vers 1.

TABLEAU 7.5
Matrice des saturations pour dix variables initiales et trois facteurs présentant une structure simple

	F_1	F_2	F_3
V_1	1	0	0
V_2	1	0	0
V_3	1	0	0
V_4	1	0	0
V_5	0	1	0
V_6	0	1	0
V_7	0	1	0
V_8	0	0	1
V_9	0	0	1
V_{10}	0	0	1

Le tableau 7.6 présente la matrice des saturations obtenues après une analyse en composantes principales de la matrice de corrélations des dix items de l'échelle d'humour. La structure factorielle qui se dégage de cette matrice des saturations ne permet pas une interprétation aussi claire que la matrice-cible.

TABLEAU 7.6
Matrice des saturations avant rotation pour les dix items de l'échelle d'humour

	F1	F2
Item02	0,718	0,006
Item04	0,663	0,266
Item08	0,659	0,190
Item10	0,642	0,184
Item06	0,634	0,361
Item01	–0,366	0,358
Item03	–0,171	0,817
Item05	–0,132	0,760
Item07	–0,309	0,593
Item09	–0,173	0,568

C'est souvent le cas lorsqu'on emploie des techniques comme l'analyse en composantes principales : le premier facteur semble prendre toute la place. C'est pourquoi l'emploi de cette technique ne constitue vraiment qu'un premier pas, une première étape de l'analyse factorielle, à savoir celle qui permet d'identifier le nombre de facteurs considérés significatifs en ayant recours à des critères comme la valeur propre supérieure à 1 et le graphique des éboulis. Il faut par la suite interpréter ces facteurs dits significatifs. Une deuxième étape consiste alors à procéder à une rotation[10] de façon à générer de nouvelles saturations. Le tableau 7.7 montre le résultat que nous avons obtenu après rotation des facteurs 1 et 2 présentés au tableau 7.6.

10. Voir la sous-section suivante.

TABLEAU 7.7
Matrice des saturations après rotation pour les dix items de l'échelle d'humour

	Facteur	
	1	2
Item06	0,716	0,139
Item04	0,713	0,039
Item08	0,685	–0,032
Item02	0,682	–0,224
Item10	0,667	–0,031
Item03	0,100	0,828
Item05	0,118	0,763
Item07	–0,103	0,661
Item09	0,018	0,593
Item01	–0,232	0,456

Cette matrice des saturations après rotation n'est pas tout à fait identique à la matrice-cible qui est le reflet de la structure simple idéale, mais nous nous en sommes tout de même approchés suffisamment. En effet, les cinq premiers items (les items pairs) ont des saturations élevées (> 0,3) sur le facteur 1 mais des saturations faibles (< 0,3) sur le facteur 2. Par contre, les cinq derniers items (items impairs) possèdent des saturations élevées sur le facteur 2 mais faibles sur le facteur 1. C'est cette structure, dite simple, qui clarifie l'interprétation de ces deux facteurs : en effet, ici, puisque les items pairs réflètent un humour axé sur l'autodérision et que les items impairs traitent d'un humour fondé sur la dérision des autres, nous pouvons interpréter le facteur 1 comme de l'humour en riant de soi et le facteur 2 comme de l'humour en riant des autres.

Dans la prochaine sous-section nous présenterons quelques-unes des considérations techniques essentielles à la compréhension des bases de l'analyse factorielle. Nous sommes bien conscients, soulignons-le encore, que ces considérations ne constituent en quelque sorte que la partie émergée de l'iceberg, mais il nous semble tout de même utile de maîtriser ces concepts pour permettre une compréhension minimale qui mette davantage l'accent sur l'interprétation des résultats que sur des considérations procédurales.

7.2.3. Aspects techniques

Chacun des k facteurs retenus[11] F_j, où j varie entre 1 et k, peut être vu comme une variable indépendante dans un contexte de régression multiple (Nunnaly, 1978, p. 334) où les n variables initiales V_i, i variant de 1 à n, sont considérées dépendantes :

11. Cette façon de présenter les équations de régression ne permet pas de distinguer la technique d'extraction des facteurs qu'est l'analyse en composantes principales (où k = n) de la technique connue notamment sous le nom de l'analyse en facteurs communs et spécifiques (où k < n). Nous en sommes bien conscients. Nous avons voulu nous concentrer sur les facteurs reconnus comme significatifs au terme d'une analyse factorielle, indépendamment de la technique d'extraction.

$$V_1 = a_{11}F_1 + a_{12}F_2 + a_{13}F_3 + \ldots + a_{1k}F_k$$

$$V_2 = a_{21}F_1 + a_{22}F_2 + a_{23}F_3 + \ldots + a_{2k}F_k$$

$$V_3 = a_{31}F_1 + a_{32}F_2 + a_{33}F_3 + \ldots + a_{3k}F_k$$

$$\ldots$$

$$V_n = a_{n1}F_1 + a_{n2}F_2 + a_{n3}F_3 + \ldots + a_{nk}F_k$$

Bien sûr, le principe de parcimonie suppose que, au terme de l'analyse, k soit plus petit que n. Dans le cas du *Thurstone box problem*, les V_i seraient les n = 9 variables notées x^2, y^2, z^2, exp(x), exp(y), exp(z), Log(x), Log(y) et Log(z). Tandis que les facteurs au nombre de k = 3 seraient les trois dimensions : longueur, largeur et hauteur. Les valeurs a_{ij} sont appelées les saturations et constituent les corrélations entre la variable V_i et le facteur F_j.

Dans le cas de l'échelle d'humour, où nous avons compté n = 10 items V_i et k = 2 facteurs F_j, les équations de régression (après rotation), d'après le tableau 7.7, se résument à

$$V_1 \text{ (Item 6)} = 0{,}716F_1 + 0{,}139F_2$$

$$V_2 \text{ (Item 4)} = 0{,}713F_1 + 0{,}039F_2$$

$$V_3 \text{ (Item 8)} = 0{,}685F_1 + (-0{,}032)F_2$$

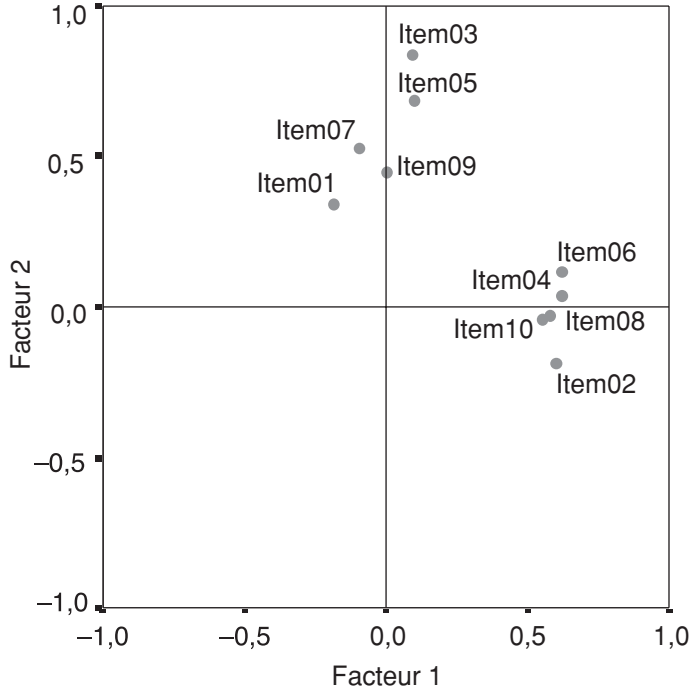
.
.
.

$$V_{10} \text{ (Item 1)} = -0{,}232F_1 + 0{,}456F_2$$

Ce qui peut se représenter dans un plan cartésien comme à la figure 7.2, où les axes indiquent les facteurs et les points dans le plan constituent les variables initiales, c'est-à-dire ici les items. On y voit que les cinq items pairs se regroupent autour d'un axe, le facteur 1, et les cinq items impairs se regroupent autour d'un autre axe, le facteur 2. Les saturations sont donc les coordonnées des variables, chacun des facteurs représentant un axe du système.

Strictement parlant, comme nous l'avons déjà souligné au passage, deux étapes sont nécessaires avant d'obtenir les saturations après rotation. La première étape, l'extraction (à l'aide d'une technique comme l'analyse en composantes principales), consiste à obtenir un premier jeu de saturations, les a_{ij}, et à déterminer ainsi le nombre de facteurs significatifs ; les valeurs propres servant à déterminer ce nombre sont définies comme la somme des carrés des saturations associées à ce facteur. Par exemple, dans le cas de l'échelle d'humour, le premier jeu de saturations était donné par les valeurs obtenues au tableau 7.6.

FIGURE 7.2
Représentation dans un plan cartésien des saturations (coordonnées) après rotation des dix items en fonction des deux facteurs extraits de l'échelle d'humour présentée par Green, Salkind et Akey (2000, p. 302)



En se rappelant qu'une saturation est une corrélation, on peut considérer le carré d'une saturation a_{ij} comme un pourcentage de variance commun entre la variable V_i et le facteur F_j. La somme des carrés des saturations indique donc jusqu'à quel point le facteur explique la variance des variables initiales, reflet de son importance globale. Dans le cas de l'échelle d'humour,

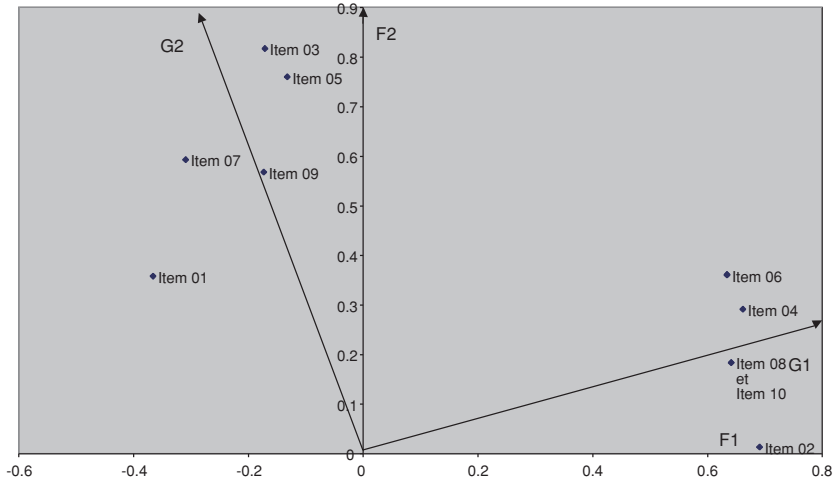
$$(0{,}718)^2 + (0{,}663)^2 + \ldots + (-0{,}173)^2 = 2{,}51 = \text{première valeur propre.}$$

Une fois complétée la première étape, le nombre de facteurs est déterminé[12] : par exemple deux facteurs, dans le cas de l'échelle d'humour. Or, comme on l'a vu, ces saturations ne satisfont que très rarement à une structure simple et il faut procéder à une rotation des axes (facteurs), le résultat étant un autre système d'axes, donc un autre jeu de facteurs G_j et un autre jeu

12. Gardant bien à l'esprit le côté subjectif de l'affaire.

de saturations b_{ij} (coordonnées). Une pléthore de procédures[13] de rotation ont été développées au cours des décennies, la plupart visant l'obtention d'une sorte de structure simple, définie un peu comme nous l'avons fait plus haut, c'est-à-dire une structure qui permettrait de mieux interpréter les facteurs. Dans le cas qui nous occupe, présenté à la figure 7.3, nous voyons bien que les axes (facteurs) F_1 et F_2 avant rotation ne représentent que bien chichement les deux regroupements d'items. Par contre, chacun des axes (facteurs) après rotation, G_1 et G_2, transperce littéralement l'un ou l'autre des deux regroupements des variables initiales. En général, plus les variables seront situées près de l'axe (facteur) et à l'extrémité de celui-ci (donc, près de la valeur maximale 1), plus elles serviront à définir ce facteur. C'est pourquoi nous dirons que le facteur 1 est défini par les items pairs de l'échelle d'humour et que le facteur 2 est défini par les items impairs.

FIGURE 7.3
Plan factoriel représentant la matrice des saturations avant rotation (facteurs F_1 et F_2) et après rotation (G_1, G_2) des dix items en fonction des deux facteurs extraits de l'échelle d'humour présentée par Green, Salkind et Akey (2000, p. 302)



7.2.4. Validation conceptuelle

Tel que souligné précédemment, l'analyse factorielle peut servir à accumuler des évidences de validité en exposant, par l'analyse de la matrice des corrélations interitems, la structure interne d'un test et en le situant dans un réseau

13. La procédure *varimax* (Kaiser, 1958) est de loin la plus populaire.

nomologique, par l'analyse de la matrice des corrélations entre ce test et d'autres tests mesurant un concept similaire. L'analyse de la matrice des corrélations interitems permettra d'identifier, comme nous l'avons fait plus haut avec l'échelle d'humour, les sous-concepts impliqués dans la définition du concept sous-jacent à un test. L'analyse de la matrice des corrélations entre le test étudié et les autres tests permettra de comparer le comportement du test étudié en fonction des autres tests.

Prenons le cas fictif d'un test de résolution de problèmes mathématiques destiné à des élèves du premier cycle de l'enseignement secondaire. Imaginons que plusieurs chercheurs soutiennent que les examens de mathématique de fin d'année du ministère de l'Éducation font appel autant à une habileté de compréhension en lecture qu'à une habileté mathématique. Ils choisissent de mettre leur hypothèse à l'épreuve en administrant à un échantillon d'élèves les trois derniers examens de mathématique en plus d'un test de substitution algébrique, d'un test de reconnaissance des formes en géométrie, donc deux tests faisant appel à des habiletés typiquement mathématiques, et de deux tests bien connus de compréhension en lecture. À chaque élève sont donc associés dix scores. Chacun des trois examens de mathématique du Ministère génère deux scores : un score relatif aux items à choix multiple (formés d'énoncés courts et de 4 choix de réponses) et un score relatif aux items à réponse élaborée (formées d'énoncés longs). Ajoutons à cela un score associé au test de substitution algébrique, un score associé au test de géométrie et un score pour chacun des deux tests de compréhension en lecture. L'analyse factorielle de la matrice de corrélations de ces dix variables révèle deux facteurs. Comme en fait foi le tableau 7.8, le premier facteur est défini par les trois sous-tests formés des items à choix multiple, le test d'algèbre et le test de géométrie. Le second facteur, d'une importance pratiquement aussi grande que le premier, est composé des trois sous-tests formés des items à réponse élaborée et des deux tests de compréhension en lecture. Les conclusions de cette analyse constituent une preuve à l'appui de l'hypothèse des chercheurs voulant que les examens de mathématique du Ministère exigent principalement deux habiletés de la part des élèves : une habileté proprement mathématique (F_1) et une habileté de compréhension en lecture (F_2).

Tableau 7.8
Résultat d'une analyse factorielle montrant deux facteurs : F_1, un facteur mathématique, et F_2, un facteur de compréhension en lecture

F_1	F_2
Examen1-choix multiple	Examen1- réponse élaborée
Examen2-choix multiple	Examen2- réponse élaborée
Examen3-choix multiple	Examen3- réponse élaborée
Algèbre	Compréhension en lecture 1
Géométrie	Compréhension en lecture 2

7.3. BIAIS LIÉS À L'ADMINISTRATION DE L'INSTRUMENT

7.3.1. Types de biais

Au chapitre 2, nous avons défini l'erreur de mesure aléatoire comme la différence entre un score observé et un score vrai. Plus cette différence est élevée, plus la fidélité est faible. C'est en ce sens que l'on peut dire que l'erreur de mesure aléatoire constitue un écart à la fidélité, à la généralisabilité ou à l'information. D'un autre côté, le biais, qui est une erreur systématique, peut être perçu comme un écart à la validité. Zoé, on s'en souvient a obtenu un score de 64 % à un examen de mathématique. Supposant, pour la forme, que l'examen ait eut lieu un mardi, rien n'indique qu'elle aurait pu obtenir le même score si le test avait été administré le lundi ou le vendredi. Zoé aurait pu être malade le lundi mais en pleine forme le vendredi. La variation (potentielle) aléatoire de la note de Zoé à cet examen de mathématique est une manifestation de l'erreur de mesure **aléatoire**. Par contre, si on avait administré à Zoé une version informatisée de l'examen, sachant que Zoé a toujours été anxieuse devant un ordinateur, il faudrait alors parler d'erreur de mesure systématique, de biais. En effet, le score de Zoé ne reflèterait pas seulement son habileté en mathématique mais aussi son anxiété face aux ordinateurs. En ce sens, on dirait que le test est biaisé envers un sous-groupe de la population visé à savoir, ici le sous-groupe de personnes qui se sentent moins à l'aise devant un ordinateur. Remarquons que l'erreur de mesure aléatoire s'ajoute au biais, car Zoé n'aurait probablement pas obtenu le même score biaisé si le test informatisé avait été administré le lundi ou le vendredi.

Identifier puis contrôler les biais constitue une partie importante du processus de validation des interprétations liées à un instrument de mesure. Un biais peut être inhérent à l'administration de l'instrument en tant que tel, comme l'emploi d'un mot mal placé ou ambigu dans un item d'un test, ou être généré par la façon de répondre du sujet comme un *response set* par exemple. La présente section traitera de biais liés particulièrement à l'instrument ou encore aux caractéristiques liées à l'administration de l'instrument. Nous discuterons à la section suivante des biais liés à la façon de répondre du sujet. À l'instar de Van de Vijver et Leung (1997), nous distinguerons trois types de biais liés à l'instrument : le biais de concept, le biais de méthode et le biais d'item. Le dernier type de biais est si important que nous lui réservons le prochain chapitre. Nous aborderons donc plus particulièrement ici les biais de concept et les biais de méthode. Notons que peu importe le type de biais auquel nous faisons référence, ce sera toujours à partir des scores au test que nous les étudierons. Selon notre conception, donc, un biais devra toujours être, d'une façon ou d'une autre, quantifiable. C'est cette quantité qui servira à prendre une décision à savoir si l'on peut décréter qu'il y a ou non un biais.

Deux éléments de contexte justifient plus particulièrement l'étude des biais liés à un instrument de mesure : la mondialisation, qui a permis la traduction/adaptation de tests dans différentes cultures, et l'informatisation des tests, conséquente d'une démocratisation des micro-ordinateurs et de l'Internet. En conformité avec la mondialisation des échanges et des marchés, les enquêtes menées à l'échelle internationale et qui visent la comparaison des programmes et des capacités moyennes des jeunes dans plusieurs pays, comme la *Third International Mathematics and Science Study* (Martin *et al.*, 1997) par exemple, ont recours à des instruments de mesure standardisés traduits dans plusieurs langues ou adaptés à plusieurs cultures. Même si des précautions sont prises pour s'assurer que l'instrument traduit ou adapté est en tous points conforme à l'original, il n'est pas certain que ce soit toujours le cas. La traduction ou l'adaptation d'un test peut faire en sorte que le ou les concepts initialement visés par l'instrument original n'aient pas le même sens dans la version traduite ou adaptée. Cela risquerait d'être le cas, par exemple, d'un test américain mesurant l'importance de la filiation et traduit en mandarin (Ho, 1996) ou encore d'un test canadien de personnalité comportant une échelle de féminité et adapté pour les pays de l'Afrique francophone. Ce glissement du concept initialement visé par le test et possédant un autre sens dans la version traduite ou adaptée est nommé **biais de concept**. Un tel glissement du concept d'une version à l'autre d'un test peut aussi survenir dans le cas suivant. Plusieurs tests autrefois offerts dans une version papier-crayon sont maintenant disponibles également sous un format informatisé (Gauthier, 2003). Il n'est cependant en aucun cas garanti que le test informatisé conserve exactement les mêmes caractéristiques que le test papier-crayon. Par exemple, les graphiques peuvent être plus difficiles à examiner sur un écran ou, encore, il n'est peut-être pas possible de réviser les réponses à un test informatisé à la fin de la session de testing (voir le chapitre 9). Nous parlerons donc encore de biais de concept si la version informatisée d'un test ne mesure pas le même concept que la version originale administrée de façon traditionnelle à l'aide d'un papier et d'un crayon.

Nous réservons l'appellation **biais de méthode** aux biais qui peuvent survenir au moment de l'administration de l'instrument. De tels biais peuvent être générés par le format d'item, la procédure d'échantillonnage, les consignes écrites pour les administrateurs, les conditions physiques ou matérielles de l'administration, etc. Pensons, à titre d'exemple, à un test constitué d'items à choix multiple administré à un échantillon d'élèves québécois rompus à ce format d'items et à un échantillon d'élèves africains qui n'ont jamais eu à se frotter à ce format d'items : ces derniers risquent d'être défavorisés non pas par le contenu du test, mais par le format des items. Un biais de méthode peut aussi être observé dans le cas d'un manque de conformité à certaines conditions liées à l'administration des tests : par exemple, si certaines consignes standardisées ne sont pas lues ou sont mal comprises par certains administrateurs, les scores des élèves visés risquent d'être empreints de biais de

méthode. Il faut remarquer que les élèves peuvent être aussi bien favorisés (p. ex. l'administrateur s'arrange pour, en quelque sorte, suggérer quelques-unes des réponses) que défavorisés (p. ex., l'administrateur ne se fait pas bien comprendre par les élèves). Notons que, contrairement à Van de Vijver et Leung (1997), nous ne considérons pas les particularités inhérentes à la façon de répondre des sujets (p. ex., *response set*, tricherie, etc.) dans la catégorie des biais de méthode. Compte tenu de la grande spécificité de ce genre de biais, nous avons préféré en faire une catégorie à part et réservé une section complète de ce chapitre à décrire les contours de ce problème et à présenter les façons de le résoudre.

7.3.2. Comment les identifier

Ce n'est pas tout de connaître et de donner un nom aux divers types de biais liés à un instrument de mesure. Il faut aussi pouvoir les identifier et les contrôler. Nous consacrons une bonne partie du prochain chapitre aux méthodes, dont celles qui s'appuient sur la TRI, permettant d'identifier puis de contrôler les biais d'item, de loin le type de biais sur lequel les chercheurs se sont le plus penchés. Pour l'heure, discutons un peu de quelques procédures mises de l'avant pour contrer les biais de concept et les biais de méthode.

Si l'analyse factorielle peut servir à identifier les composantes internes propres à un test, pourquoi ne serait-elle pas aussi utile pour comparer la structure interne de ce test obtenue à partir de deux ou plusieurs groupes, donc pour détecter un biais de concept ? La procédure consiste tout d'abord à identifier le nombre de facteurs significatifs pour chaque groupe (p. ex., anglophones vs francophones, test papier-crayon vs test informatisé, etc.). Si le nombre de facteurs est différent d'un groupe à l'autre, il faut déjà suspecter la présence d'un biais. Par contre, même si le nombre de facteurs est identique pour les deux groupes, il faut s'interroger sur la nature de chacun des facteurs en comparant les valeurs des saturations des facteurs correspondant du premier et du second groupes. C'est seulement lorsque les différences de saturations entre les deux groupes seront minimes que nous pourrons éviter de parler de biais de concept. Nous retiendrons deux coefficients pour comparer les saturations entre les groupes : le coefficient de linéarité (corrélation de Pearson) et le coefficient d'identité (Van de Vijver et Leung, 1997, p. 92).

Les façons d'examiner les biais de méthode ne sont ni aussi naturelles ni aussi directes que celles utilisées dans le cas des biais de concept. En effet, un biais de méthode peut se présenter sous plusieurs formes, chacune étant si particulière qu'elle requerra une procédure d'identification différente. Dans certains cas, il sera nécessaire d'obtenir des mesures répétées du test avant de pouvoir se prononcer sur un éventuel biais : si les scores de sujets de groupes distincts sont équivalents au départ, mais qu'ils n'ont pas du tout la même progression d'une répétition à l'autre, il faut envisager la possibilité de biais de méthode (van de Vijver et Leung, 1997, p. 17). Dans d'autres cas, il s'agira de

mesurer une variable potentiellement nuisible comme la motivation, puis de la contrôler statistiquement à l'aide de l'analyse de la covariance. Il n'est cependant pas nécessaire d'utiliser des méthodes statistiques sophistiquées pour tenir compte d'un biais de méthode. Par exemple, dans le cadre du contrôle de la qualité d'un projet visant à collecter des données dans plusieurs écoles, il peut être nécessaire d'observer les administrateurs de tests sur le terrain, surtout si nous soupçonnons que les consignes pourtant mises à l'essai ne sont pas réellement suivies : il faudra alors exclure certains sujets ou écoles des analyses pour limiter le biais de méthode. De la même façon, dans le cadre d'une évaluation internationale des acquis des élèves (ex. TIMSS), le plan d'échantillonnage proposé peut ne pas être suivi par l'un ou l'autre des pays participants, pavant la voie à un éventuel biais de méthode. Certains (p. ex., IEA) ont proposé de former deux groupes de pays pour reporter les résultats d'une telle évaluation, de telle sorte que seuls les pays qui ont suivi rigoureusement le plan initial soient comparés entre eux.

7.3.3. Une application

Dans le cadre d'une étude visant à identifier les biais de concept, les biais de méthode et les biais d'item associés à une enquête à grande échelle impliquant plusieurs juridictions canadiennes, Bertrand *et al.* (2001) ont proposé d'utiliser l'analyse factorielle complète du patron de réponses (Bock, Gibbons et Muraki, 1988) pour quantifier la part de biais de concept et l'analyse de la covariance pour quantifier la part de biais de méthode.

Cette étude prévoyait des comparaisons deux à deux de la plupart des juridictions canadiennes pour chacun des deux groupes linguistiques, les anglophones et les francophones. Après avoir effectué une analyse factorielle complète du patron de réponses à l'aide de TESTFACT (Zimowski *et al.*, 1996) pour chacune des juridictions, les auteurs ont convenu, pour les comparaisons affichant une structure factorielle composée d'un seul facteur statistiquement significatif, d'analyser les saturations sur ce facteur et, pour les comparaisons affichant une structure factorielle composée de deux facteurs statistiquement significatifs, d'analyser les saturations sur chacun des deux facteurs.

Afin d'apprécier l'ampleur du biais de concept associé ici au manque d'équivalence factorielle, les auteurs ont eu recours à la différence entre les saturations des facteurs correspondants à une comparaison donnée. C'est la racine carrée moyenne (*root mean square*) qui a servi à quantifier la différence entre les saturations : elle est donnée par

$$RMS_{G_1-G_2} = \sqrt{\frac{\sum (sat_{G_1} - sat_{G_2})^2}{m*n}}$$

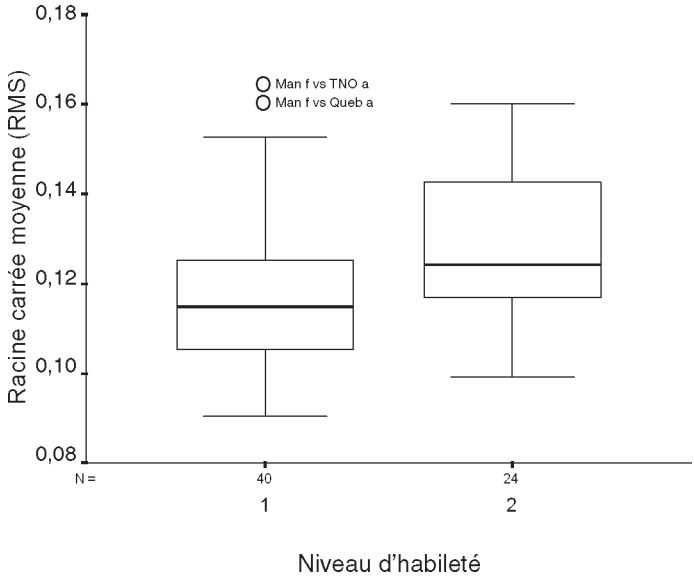
Dans le contexte où les deux groupes comparés sont notés G_1 et G_2, sat_{G1} représente une saturation pour un item donné du groupe G_1 tandis que sat_{G2} réfère à la saturation de l'item correspondant pour le groupe G_2. La somme est prise sur m, le nombre de facteurs et n, le nombre d'items. Le nombre de facteurs retenus est égal au nombre de facteurs considérés statistiquement significatifs. Plus la valeur de cet indice était faible, plus nous avons considéré qu'il y avait équivalence factorielle. Afin de comparer les valeurs de la racine carrée moyenne associées aux diverses comparaisons, Bertrand *et al.* (2001) ont proposé d'utiliser le diagramme en boîte et moustaches (Bertrand et Valiquette, 1986).

Par exemple, en examinant le diagramme en boîte et moustaches de la racine carrée moyenne pour l'ensemble des comparaisons impliquant les différentes provinces canadiennes, deux comparaisons se distinguent des autres : la comparaison impliquant les Territoires du Nord-Ouest anglophones (TNO a) et le Manitoba francophone (Man f) ainsi que la comparaison impliquant le Québec anglophone (Queb a) et le Manitoba francophone. Les valeurs de la racine carrée moyenne associées à ces deux comparaisons sont anormalement élevées par rapport aux autres valeurs. Les comparaisons impliquant ces provinces génèrent donc moins d'équivalence factorielle, donc plus de biais de concept que les comparaisons impliquant les autres provinces canadiennes.

C'est à l'analyse de la covariance que nous avons eu recours pour quantifier les éventuels biais de méthode impliqués dans les comparaisons des juridictions canadiennes. C'est la variable touchant la confiance que les étudiants ont de leur habileté en sciences qui a servi de covariable. Un biais de méthode était d'autant plus important que cette covariable affectait les scores en sciences.

Plus précisément, voici comment cette méthode d'analyse a été utilisée pour quantifier un biais de méthode. Nous voulions étudier l'ampleur du biais entre deux groupes dont les moyennes non ajustées en sciences étaient de 521,52 pour le premier groupe formé des étudiants francophones du Québec (QuFr) et de 505,95 pour le second groupe formé des étudiants anglophones du Québec (QuAn). La différence (en valeur absolue) entre ces deux moyennes non ajustées, D_{NAJ}, était donc de 15,57. Les moyennes en sciences ajustées par la covariable étaient respectivement de 516,39 pour le groupe QuFr et de 509,60 pour le groupe QuAn. La différence (en valeur absolue) entre ces deux moyennes ajustées, D_{AJ}, était donc de 6,79. On voit bien que le fait de tenir compte de la covariable a réduit la différence de scores en sciences entre les deux groupes. Cette réduction de la différence, nous l'avons nommée l'écart E, $D_{AJ} - D_{NAJ} = -8{,}78$. Selon les groupes comparés, certains de ces écarts étaient plus faibles et d'autres, plus élevés. Plus la valeur de cet écart était élevée (et négative), plus l'ajustement des scores moyens en sciences par la covariable était grand donc plus le biais de méthode était considéré important.

FIGURE 7.4
Diagramme en boîte mettant en évidence les valeurs extrêmes de la racine carrée moyenne (RMS) de la structure factorielle du test de mathématique pour l'ensemble des comparaisons interlinguistiques-interculturelles impliquant les différentes provinces



Notre graduation de l'importance des biais de méthode est fondée sur le rapport de la valeur de l'écart E à la valeur de l'écart-type de la distribution générale des scores en sciences, soit 100. En ce sens, le biais de méthode est considéré **inexistant** si la valeur absolue de E est plus petite que 5 (soit moins d'un vingtième d'écart-type). Le biais de méthode est dit **faible** si la valeur absolue de E est plus petite que 10 mais supérieure ou égale à 5. Le biais est **modéré** si la valeur absolue de E est plus petite que 25 mais supérieure ou égale à 10. Enfin, le biais est **élevé** si la valeur absolue de E est supérieure ou égale à 25. Ainsi, un écart supérieur à un quart d'écart-type génère ce que nous avons convenu d'appeler un biais de méthode élevé.

Dans le cas qui nous intéresse, nous avons observé que toutes les valeurs de l'écart E se situaient entre 5 et 10, ce qui signifie que tous les biais de méthode observés sont considérés faibles.

7.4. BIAIS LIÉS À LA FAÇON DE RÉPONDRE DES SUJETS

Que ce soit dans le cadre de la mesure d'une habileté (p. ex., examen de rendement scolaire, test d'aptitude) ou d'un comportement typique (p. ex., test de personnalité, échelle d'attitude), il est toujours possible de rencontrer des patrons de réponses aberrants, provenant de sujets qui peuvent avoir triché, répondu au hasard ou tout simplement répondu suivant une façon stéréotypée (*response set*). Dans le cas où les scores à un test sont utilisés pour attribuer éventuellement un diplôme ou un emploi, il est impératif de détecter les sujets qui ont répondu de façon aberrante puisque, le cas échéant, leur score risque de ne pas refléter correctement leur niveau d'habileté. Un étudiant faible qui triche sur son voisin plus habile obtiendra un score qui surestimera son habileté. Le score d'un sujet qui, candidat à un poste, répond de façon malhonnête à un test de personnalité en donnant les réponses qui le feront bien paraître ne convient pas plus. De même, on risque d'attribuer un score qui sous-estimera l'habileté d'un étudiant très habile, mais de nature plutôt paresseuse, qui décide de choisir les réponses au hasard.

Plusieurs stratégies peuvent être employées pour identifier ces personnes qui adoptent un comportement aberrant au cours d'une session de testing. Un local bien éclairé, le recours à plusieurs surveillants dans le local de testing et l'emploi d'échelles de désirabilité sociale comptent parmi ces stratégies. Bien sûr, le recours à plusieurs surveillants dans un local très éclairé peut minimiser les risques de tricherie. Mais ce n'est pas toujours possible, faute d'un local adéquat, d'un budget suffisant ou encore de surveillants disponibles. L'emploi de mesures externes (p. ex., désirabilité sociale, *faking bad*, *faking good*) peut sembler fort approprié mais, tel que le stipulent Zickar et Drasgow (1996), comportent certaines limites. En effet, selon ces auteurs, ces tests supplémentaires sont dispendieux et consomment beaucoup de temps autant pour ceux qui doivent les administrer, les corriger et les interpréter que pour ceux à qui ils sont destinés. À l'instar de plusieurs autres (Levine et Rubin, 1979 ; Levine et Drasgow, 1982 ; Drasgow, Levine et Williams, 1982 ; Hulin, Drasgow et Parsons, 1983 ; Levine et Drasgow, 1988), ils proposent d'utiliser une méthode interne de détection des réponses atypiques, fondée sur l'observation des patrons de réponses. Par exemple, il serait possible, semble-t-il, de détecter des sujets du genre petit malin voulant délibérément choisir la mauvaise réponse puisque ceux-ci opteraient pour des leurres pratiquement jamais choisis par les sujets choisissant des réponses typiques (Levine et Drasgow, 1988).

Cependant, une des stratégies les plus répandues, tout en étant moins dispendieuse et tout aussi efficace qu'une stratégie externe, consiste à étudier les patrons de réponses des sujets en faisant une analyse détaillée des séquences des valeurs 0 (pour mauvaise réponse) et 1 (pour bonne réponse). Un sujet qui a reluqué les réponses à des items difficiles sur un voisin plus habile présentera un patron atypique où il aura échoué un certain nombre d'items faciles

(son patron présentera une séquence de 0) mais réussi plusieurs items difficiles (son patron présentera une séquence de 1). Ceci dit, toutes les formes de patrons atypiques ne sont pas aussi facilement détectables. Par exemple, Meijer, Molenaar et Sijtsma (1994) ont montré que les sujets qui copiaient les réponses aux items difficiles sur leurs voisins plus habiles étaient beaucoup plus faciles à repérer que les sujets répondant au hasard.

Plusieurs formes de patrons peuvent être identifiées selon Wright et Stone (1979). Au tableau 7.9, nous avons tant bien que mal tenté de qualifier sept patrons de réponses provenant de sujets à qui on avait administré un test de huit items. Nous supposons que ces items sont classés, de gauche à droite, du plus facile au plus difficile. Cette façon de nommer les sujets à partir de leurs patrons de réponses est à la limite du caricatural, mais elle exprime tout de même une différence notable entre les façons de répondre des sujets. Par ailleurs, signalons que cette terminologie ne signifie pas que tous ces patrons soient atypiques. Par exemple, le sujet qualifié de consciencieux-lent est aussi tout à fait conforme au modèle de Guttman (voir plus loin). Par contre, les patrons des sujets parfaitement forts et parfaitement faibles pourraient aussi être considérés comme trop beaux pour être vrais (Wright et Stone, 1979). L'assignation du qualificatif chanceux à un des sujets pourrait être une façon polie de le qualifier de tricheur. Toujours est-il qu'il est bien difficile, et c'est ce que nous voulons faire ressortir de cette brève présentation, d'accorder la même crédibilité à tous ces patrons de réponses et aux scores qui en découlent.

TABLEAU 7.9
Exemples de qualificatifs de patrons de réponses : les items sont classés, de gauche à droite, du plus facile au plus difficile

Patron de réponses	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8
Parfait-fort	1	1	1	1	1	1	1	1
Normal	1	1	1	1	1	0	1	0
Endormi	1	1	1	0	0	1	0	1
Aléatoire	1	0	1	0	1	0	1	0
Consciencieux-lent	1	1	0	0	0	0	0	0
Chanceux	1	0	0	0	0	1	1	1
Parfait-faible	0	0	0	0	0	0	0	0

Une foule d'indices ont déjà été mis de l'avant pour quantifier le degré d'aberrance des patrons de réponses. Suivant Hulin, Drasgow et Parsons (1983), nous avons choisi de classifier ces indices en deux catégories : les indices heuristiques forment la première catégorie alors que ceux basés sur un modèle de la TRI constituent la deuxième catégorie. Les indices heuristiques ne font pas explicitement appel à un modèle et sont par le fait même beaucoup plus faciles à employer. D'ailleurs, ils n'exigent pas non plus un nombre important de sujets comme le font les indices basés sur un modèle de la TRI. Il semble

cependant qu'en général, les indices de la première catégorie, de nature empirique, soient moins puissants (Levine et Drasgow, 1988) que les indices de la deuxième catégorie quand vient le temps de détecter les patrons aberrants.

Donlon et Fischer (1968) furent parmi les premiers à proposer un indice heuristique : la corrélation bisériale de personne. Il s'agit de la corrélation entre, d'une part, le patron de réponses d'un sujet formé de 1 (pour réussite de l'item) et de 0 (pour échec de l'item) et, d'autre part, les valeurs de l'indice de difficulté associées à chaque item. C'est le pendant de la corrélation bisériale bien connue entre l'item et le total sauf qu'ici, c'est au sujet que sera attribuée une valeur associée à cette corrélation bisériale de personne. Si cette valeur est élevée, le patron de réponses sera jugé typique puisque les items réussis par le sujet seront associés aux items faciles et les items échoués aux items difficiles. Par contre, si un sujet se voit attribuer une valeur de corrélation de personne faible, voire négative, c'est que son patron comporte des séquences de réponses aberrantes comme la réussite à des items difficiles et l'échec à des items faciles : ce patron sera dès lors considéré comme atypique.

L'indice de Sato (1975), revu et corrigé par Harnisch et Linn (1981), doit aussi être considéré comme heuristique, bien qu'il fasse explicitement référence au modèle de Guttman. Cet indice, en effet, est une mesure de l'écart qui existe entre le patron de réponses observé et un patron dit de Guttman, soit un patron qui exige qu'à partir du moment où un item est réussi, tous les items plus faciles doivent l'être aussi. Pour bien identifier un tel patron, il suffit de classer les items, de gauche à droite, du plus facile au plus difficile. Trois types de patrons de Guttman peuvent alors être envisagés : le patron de réussite parfaite (1111111…), le patron d'échec parfait (00000…) et le patron parfaitement cohérent (111100000). L'obtention de l'un ou l'autre de ces trois types de patrons de Guttman générera une valeur (parfaite) de 0 pour l'indice de Sato. Tout écart à l'un ou l'autre de ces trois types de patrons parfaits résultera en une valeur supérieure à 0. Plus le patron de réponses comportera d'items difficiles réussis combiné à des items faciles échoués, plus la valeur de cet indice tendra vers 1. Cette valeur maximale de 1, reflet d'un maximum d'aberrance dans le patron, sera observée si un sujet réussit tous les items les plus difficiles tout en échouant tous les items les plus faciles. Thibault (1992) a montré que l'indice de Sato était très puissant pour détecter des patrons atypiques et ce, même si on le comparait aux indices basés explicitement sur un modèle de la TRI. Nous présenterons un exemple d'application de cet indice un peu plus loin.

Hulin, Drasgow et Parsons (1983, p. 110) proposent toute une série d'indices basés sur les modèles de la TRI, notamment ceux émanant de l'idée émise par Levine (p.122) : l'estimé d'habileté thêta, obtenu par la méthode du maximum de vraisemblance, est la valeur sur l'échelle thêta qui maximise les chances d'observer un patron de réponses donné. En d'autres termes, c'est la valeur pour laquelle la probabilité d'observer un tel patron (la fonction de vraisemblance) est maximale. Or, ce maximum peut être très faible, c'est-à-

dire qu'il n'y a peut-être pas de valeur de thêta qui rende la fonction de vraisemblance réellement élevée. L'indice proposé par Levine est directement relié à la valeur de cette fonction puisqu'il est basé sur le logarithme du maximum de la fonction de vraisemblance. La valeur de cet indice L_0 sera d'autant plus faible que les patrons seront peu vraisemblables : par exemple, elle sera plus particulièrement faible pour un sujet qui a réussi des items difficiles et échoué des items faciles. C'est sur la base des faibles valeurs de cet indice que les patrons seront classés atypiques. En réalité, c'est souvent à partir de la transformation de cet indice en une version standardisée, notée L_Z, que la décision se prendra, celle-ci étant moins sensible que L_0 aux items sans réponse et aux différents niveaux d'habileté des sujets.

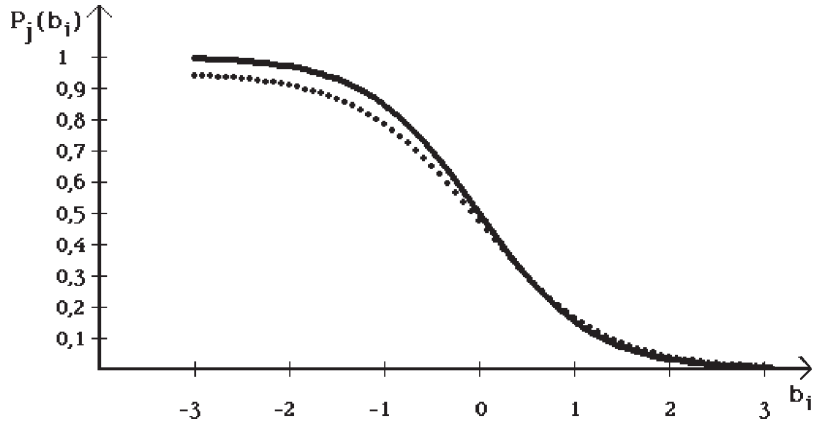
Plusieurs autres indices ont été développés en exploitant l'idée originale de Levine. C'est le cas des indices polytomiques, notés P_0 et P_Z, basés sur les patrons des choix de réponses (Drasgow, Levine et Williams, 1982), c'est-à-dire sur les séquences des choix de réponses comme AAAMMDCDCAB ou 334234413 plutôt que sur les items corrigés (les 0 et les 1) en tant que tels. Hulin *et al.* (1983, p. 142), en effet, discutent d'indices permettant d'identifier les patrons de réponses atypiques provenant d'une échelle d'attitude. Ces indices permettraient de détecter les patrons décrits par van de Vijver et Leung (1997, p. 15), où les sujets d'une culture particulière choisissent plus souvent les extrémités que le centre de l'échelle (p. ex., 11112455555). Il serait du même coup possible de détecter des sujets qui décident de ne pas trop se fatiguer et de choisir plus souvent qu'autrement le centre de l'échelle (p. ex., 3333313333433333). Zickar et Drasgow (1996) nous mettent en garde contre un emploi abusif de ces indices polytomiques. Il peut arriver, en effet, que ces indices, beaucoup plus coûteux, ne soient pas plus efficaces que les indices dichotomiques L_0 et L_z : ce serait le cas, notamment, si le concept mesuré par le test amenait les sujets à ne choisir que les catégories extrêmes (p. ex., tout à fait en accord, tout à fait en désaccord) au détriment des autres catégories de l'échelle de mesure.

Selon Schmitt, Cortina et Whitney (1993), comme les indices basés sur un modèle de la TRI sont d'autant plus efficaces que le test est long, il serait avisé d'employer un indice multitest (Drasgow, Levine et McLaughlin, 1991) noté L_{zm} et permettant de tenir compte d'un patron de réponses provenant de plusieurs tests (en fait T tests) relativement courts. Ceci dit, bien sûr, après avoir administré tous ces tests !

$$L_{zm} = \frac{\sum_{t=1}^{T} [L_{0t} - E(L_{0t})]}{\sqrt{\sum_{t=1}^{T} Var(L_{0t})}}$$

Un dernier indice fondé sur un modèle de la TRI mérite d'être présenté. Trabin et Weiss (1983) utilisent le concept de courbe caractéristique de sujet (CCS) pour identifier des patrons atypiques. Cette courbe est le pendant de la CCI. Sauf que $P_i(\theta_j)$, la probabilité de réussir l'item i étant donné l'habileté θ_j , est remplacé par $P_j(b_i)$, la probabilité pour un individu j (d'habileté θ_j) de réussir un item de difficulté b_i. Dans ce cas, la courbe caractéristique de sujet est tracée en considérant $P_j(b_i)$ en fonction de b_i. Contrairement à une CCI, une CCS devrait être monotone décroissante puisque la probabilité de réussir un item facile ($b_i < 0$) devrait être plus grande que la probabilité de réussir un item difficile ($b_i > 0$), comme le révèle la figure 7.5. À l'instar des indices d'ajustement des items, une valeur de khi-carré est calculée entre la CCS, basée sur un modèle de la TRI, et la courbe empirique correspondante qui représente, elle, la proportion d'items réussis en fonction de la difficulté de l'item. Plus la courbe empirique (en pointillés) s'éloignera de la courbe caractéristique (en trait plein), plus la valeur du khi-carré sera élevée et plus le sujet sera considéré atypique.

FIGURE 7.5
Courbe caractéristique de sujet (CCS) modélisée à l'aide la TRI (en trait foncé) et courbe empirique de sujet (en pointillé)



Pour avoir une meilleure idée du comportement des indices L_0 et Lz, nous discutons maintenant d'un exemple qui émane de la présentation de Hulin *et al.* (1983). Nous avons enrichi cet exemple en calculant aussi les valeurs de l'indice de Sato sur les mêmes patrons de réponses. Il s'agit de quatre sujets de même habileté, notée $\hat{\theta}$, qui ont répondu à un test de cinq items. Chacun de ces sujets a réussi trois items ($X_j = 3$) mais le patron de réponses diffère d'un sujet à l'autre, comme on peut le voir au tableau 7.10.

TABLEAU 7.10
Patrons de réponses de 4 sujets à un test de 5 items auxquels nous avons associé les valeurs de l'indice de Sato et des indices L_0 et Lz

	Item 1	Item 2	Item 3	Item 4	Item 5	X_j	Sato	L_0	L_z
Sujet 1	1	1	1	0	0	3	0	–1,62	0,88
Sujet 2	1	1	0	1	0	3	0,33	–2,46	0,10
Sujet 3	0	1	1	0	1	3	0,67	–6,01	–3,18
Sujet 4	1	0	1	1	0	3	0,33	–3,31	–0,68
n_i	3	3	3	2	1			$E(L_0) = -2,57$	
$P_i(\hat{\theta})$	0,9	0,7	0,5	0,3	0,1			$Var(L_0) = 1,17$	

Notons u_{ij} la variable qui est égale à 1 si le sujet j réussit l'item i et à 0 si le sujet j échoue l'item i. Le score d'un sujet X_j est nul autre que son score classique, soit la somme des items réussis. La valeur n_i, directement proportionnelle à l'indice de difficulté classique p_i, est définie comme le nombre de sujets ayant réussi l'item i (il faut noter que les items sont toujours indicés de 1 à I selon leur niveau de difficulté, l'item 1 étant le plus facile et l'item I le plus difficile.). Dans ce cas, si I indique le nombre d'items du test, alors, pour un sujet j, la valeur de l'indice de Sato est donnée (MacArthur, 1987, p. 83) par :

$$S_j = \frac{\sum_{i=1}^{X_j}(1-u_{ij})n_i - \sum_{i=X_j+1}^{I} u_{ij}n_i}{\sum_{i=1}^{X_j} n_i - \sum_{i=I+1-X_j}^{I} n_i}$$

Cette formule peut sembler un tantinet rébarbative, mais, dans les faits, le calcul des valeurs de cet indice s'obtient plutôt facilement. Considérons par exemple le patron du sujet 1. Il s'agit d'un patron de Guttman du type parfaitement cohérent. En principe, donc, la valeur de l'indice de Sato devrait être nulle. En effet, cette hypothèse est confirmée puisque :

$$S_j = \frac{\sum_{i=1}^{X_j}(1-u_{ij})n_i - \sum_{i=X_j+1}^{I} u_{ij}n_i}{\sum_{i=1}^{X_j} n_i - \sum_{i=I+1-X_j}^{I} n_i} = \frac{\sum_{i=1}^{3}(1-u_{ij})n_i - \sum_{i=3+1}^{5} u_{ij}n_i}{\sum_{i=1}^{3} n_i - \sum_{i=5+1-3}^{5} n_i}$$

$$= \frac{(1-1)3+(1-1)3+(1-1)3-(0*2)-(0*1)}{(3+3+3)-(3+2+1)} = 0$$

Le patron du sujet 3 est moins typique, dans le sens qu'il correspond moins à un patron de Guttman. Ce sujet, en effet, a réussi l'item le plus difficile tout en manquant un des trois items les plus faciles. La valeur de l'indice de Sato dans son cas est donnée par :

$$S_j = \frac{\sum_{i=1}^{X_j}(1-u_{ij})n_i - \sum_{i=X_j+1}^{I} u_{ij}n_i}{\sum_{i=1}^{X_j} n_i - \sum_{i=I+1-X_j}^{I} n_i} = \frac{\sum_{i=1}^{3}(1-u_{ij})n_i - \sum_{i=3+1}^{5} u_{ij}n_i}{\sum_{i=1}^{3} n_i - \sum_{i=5+1-3}^{5} n_i}$$

$$= \frac{(1-0)3+(1-1)3+(1-1)3-(0*2)-(1*1)}{(3+3+3)-(3+2+1)} = \frac{2}{3}$$

Les valeurs relatives à l'indice de Sato présentées au tableau 7.10 pour les autres sujets peuvent être calculées de la même façon.

Le calcul des indices L_0 et L_z demande une plus grande élaboration. La valeur de L_0 pour un sujet j est donnée par le logarithme de la fonction de vraisemblance associée à son patron de réponses.

Rappelons que les valeurs $P_i(\hat{\theta})$ correspondent à la probabilité de réussir l'item i pour un sujet d'habileté $\hat{\theta}$. Dans le cas qui nous intéresse, les quatre sujets possèdent le même niveau d'habileté. Ainsi, pour le sujet 1, compte tenu de son patron de réponses (1 1 1 0 0) :

$$L_{0(\text{sujet 1})} = \sum_{i=1}^{n} \left\{ u_i \ln[P_i(\hat{\theta})] + (1-u_i)\ln[1-P_i(\hat{\theta})] \right\}$$

$$= \ln[0,9]+\ln[0,7]+\ln[0,5]+\ln[1-0,3]+\ln[1-0,1]$$

$$= -1,62$$

Dans le cas du sujet 3 dont le patron est (0 1 1 0 1), nous avons :

$$L_{0(\text{sujet 3})} = \sum_{i=1}^{n} \left\{ u_i \ln[P_i(\hat{\theta})] + (1-u_i)\ln[1-P_i(\hat{\theta})] \right\}$$

$$= \ln[1-0,9]+\ln[0,7]+\ln[0,5]+\ln[1-0,3]+\ln[0,1]$$

$$= -6,01$$

Il est clair, suivant les valeurs de L_0 calculées plus haut, que le sujet 3 possède un patron beaucoup moins vraisemblable que le sujet 1. Ce résultat est conforme à ce qui avait été observé en utilisant l'indice de Sato.

Drasgow, Levine et Williams (1982) ont introduit l'indice L_z pour tenir compte, notamment, de sujets d'habileté distincte. Ils ont montré, en effet, que l'indice L_0 pouvait mener à des interprétations erronées quant à l'aberrance de patrons de réponses si ceux-ci provenaient de sujets qui n'avaient pas la même habileté. De plus, l'indice L_z suit approximativement une loi normale[14] centrée et réduite et permet donc une interprétation bien connue qui n'est pas accessible en employant l'indice L_0. On dira donc qu'un patron est atypique si $|L_z| > 2$. La valeur de l'indice L_z est obtenue à l'aide de la formule suivante :

$$L_z = \frac{L_0 - E(L_0)}{\sqrt{Var(L_0)}}$$

où

$$E(L_0) = \sum_{i=1}^{n} \left\{ P_i(\hat{\theta}) \ln[P_i(\hat{\theta})] + (1 - P_i(\hat{\theta})) \ln[1 - P_i(\hat{\theta})] \right\}$$

$$Var(L_0) = \sum_{i=1}^{n} \left\{ P_i(\hat{\theta}) \, [1 - P_i(\hat{\theta})] \, \ln \left[\frac{P_i(\hat{\theta})}{1 - P_i(\hat{\theta})} \right]^2 \right\}$$

On pourra vérifier que, dans le cas du tableau précédent, $E(L_0) = -2{,}57$ et $Var(L_0) = 1{,}17$. Les valeurs de l'indice L_z dans le cas des sujets 1 et 3 sont donc de :

$$L_{z(\text{sujet}1)} = \frac{L_0 - E(L_0)}{\sqrt{Var(L_0)}} = \frac{-1{,}62 - (-2{,}57)}{\sqrt{1{,}17}} = 0{,}88$$

$$L_{z(\text{sujet}3)} = \frac{L_0 - E(L_0)}{\sqrt{Var(L_0)}} = \frac{-6{,}01 - (-2{,}57)}{\sqrt{1{,}17}} = -3{,}18$$

Les valeurs calculées plus haut constituent une autre justification pour qualifier d'atypique le patron du sujet 3 et de typique le patron du sujet 1 : c'est aussi à cette conclusion que nous en étions arrivés en interprétant l'indice de Sato. Notons également que la valeur $E(L_0) = -2{,}57$ signifie la valeur auquel on est en droit de s'attendre de la fonction de vraisemblance, donc de L_0,

14. Certains chercheurs dont Nering (1995) ont critiqué la proposition que L_z était distribué selon la loi normale.

pour un sujet d'habileté donnée $\hat{\theta}$. La valeur de L_z est donc une mesure de l'écart entre la valeur de L_0 et la valeur attendue (cet écart étant pondéré par la variance). Puisque la valeur attendue est –2,57, le patron du sujet 2 (L_0 = –2,46) est donc plus proche de ce à quoi on devrait s'attendre que le patron du sujet 1 (L_0 = –1,62), qui serait vu ici comme trop beau pour être vrai, compte tenu du niveau d'habileté considéré ici, soit $\hat{\theta}$. On se souviendra que le verdict serait différent s'il se fondait uniquement sur l'indice L_0 qui ne tient pas compte du niveau d'habileté : dans ce cas, suivant le tableau 7.10, le patron du sujet 2 paraît plus atypique que celui du sujet 1.

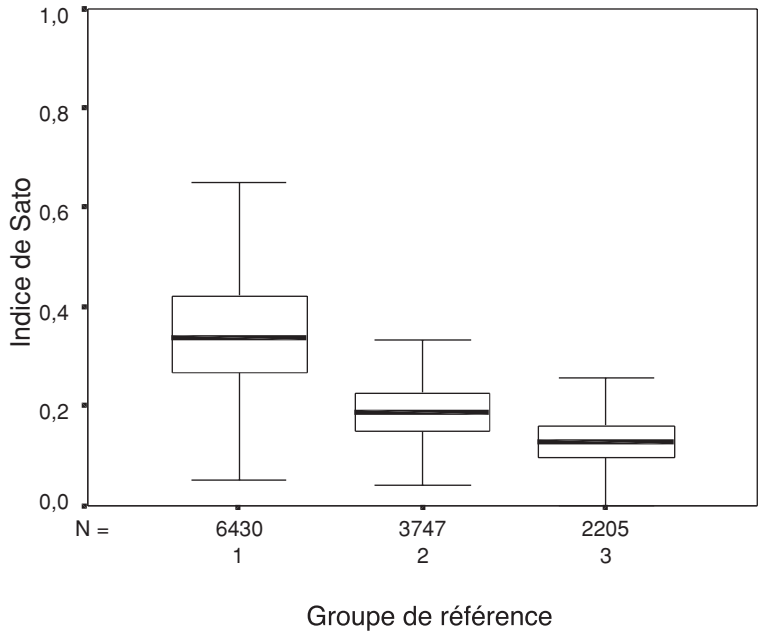
Nous avons décidé[15] de présenter un exemple détaillé de détection de patrons aberrants en utilisant l'indice de Sato plutôt qu'un indice basé sur la TRI comme L_0 ou L_z. La décision de présenter une application de l'indice de Sato va de soi puisque cet indice est simple et peu dispendieux en plus de présenter un grand pouvoir de détection des patrons atypiques (Thibault, 1992) et d'être, selon notre expérience, souvent concordant avec l'indice L_z basé sur un modèle à un paramètre de la TRI.

En 1997, le Conseil des ministres de l'Éducation du Canada, dans le cadre de son Programme d'indicateurs du rendement scolaire (PIRS), a collecté des données auprès de plus de 25 000 élèves de 13 et de 16 ans pour connaître leur niveau d'habileté en mathématique. Cette enquête s'est déroulée dans toutes les juridictions canadiennes et dans les deux langues officielles, le français et l'anglais. Un test de 125 items a été préparé pour l'occasion, ces items étant répartis également dans l'un ou l'autre des cinq niveaux de difficulté, les items du niveau de difficulté 1 étant les plus faciles tandis que les items du niveau 5 étaient les plus difficiles. Ainsi, chaque niveau de difficulté comptait 15 items à choix multiple et 10 items à réponse brève ou élaborée. Pour des questions d'efficacité, les élèves ont été répartis en trois groupes suivant leur performance à un test de classement de 15 items : les élèves les plus faibles, ceux du groupe 1, devaient débuter le test par l'item le plus facile, soit le premier item du niveau de difficulté 1 ; les élèves d'habileté moyenne, appartenant au groupe 2, débutaient pour leur part au premier item du niveau de difficulté 2 tandis que les élèves considérés les plus habiles commençaient le test par le premier item du niveau de difficulté 3. La consigne donnée aux administrateurs du test voulait que les élèves exécutent le plus d'items possibles, les items étant alors placés en ordre de difficulté. Ainsi, tous les élèves pouvaient théoriquement exécuter les 125 items. Avant d'estimer l'habileté de chacun des élèves en utilisant le modèle à trois paramètres de la TRI, les auteurs

15. Nous ferons de même au chapitre suivant pour identifier les biais d'item en présentant un exemple détaillé par la méthode de Mantel-Haenszel, beaucoup plus simple à utiliser, mais basée sur des idées similaires à celles qui ont présidé aux méthodes de détection des biais par la TRI (p. ex., aire entre les CCI). La méthode de Mantel-Haenszel prend en quelque sorte la méthode TRI de l'aire entre les CCI comme modèle. Nous en dirons plus au prochain chapitre.

(Bertrand et Laroche, 1999) ont voulu vérifier jusqu'à quel point les conditions d'application de cette théorie étaient satisfaites. Toutefois, comme condition préalable à la vérification de ces conditions, ceux-ci voulaient savoir jusqu'à quel point les élèves des groupes 1 et 2 avaient véritablement essayé les items à choix multiple difficiles, soit ceux des niveaux de difficulté 3, 4 et 5. En effet, ils avaient un doute sur le sérieux qu'avaient mis les élèves à répondre à ces items à choix multiple puisque les items à réponse brève ou élaborée associés à ces trois niveaux de difficulté comportaient un taux impressionnant de valeurs manquantes. Il a donc été décidé de calculer les valeurs de l'indice de Sato pour les élèves des trois groupes. Le résultat, qui se trouve à la figure 7.6, étaye en partie les doutes des auteurs puisque plus les élèves appartiennent à un groupe d'élèves faibles (particulièrement le groupe 1), plus la valeur moyenne de leur indice de Sato est élevée, une indication d'un plus grand taux de patrons atypiques chez ces élèves. Un regard sur la forme que prennent ces patrons de

FIGURE 7.6
Diagramme en boîte des valeurs de l'indice de Sato pour les élèves des trois groupes qui ont répondu au test de mathématique proposé par le PIRS en 1997 (seuls les sujets qui ont répondu aux 45 items à choix multiple des niveaux 3, 4 et 5 ont été retenus)



réponses est encore plus révélateur. Comme on peut le voir en examinant les patrons de réponses du tableau 7.11, les élèves des groupes 1 et 2 semblent avoir répondu aux 45 derniers items à choix multiple (les plus difficiles) sinon de façon complètement aléatoire[16] du moins avec bien peu de constance : les items étant placés, de gauche à droite, en ordre croissant de difficulté, il appert que ces sujets ont réussi des items plus ou moins indépendamment de leur niveau de difficulté : on retrouve, en effet, dans ces patrons, des 1 et des 0 aussi bien à gauche du patron qu'au centre ou à droite de celui-ci. Soulignons que ces élèves ont tous obtenu une valeur de plus de 0,5 à l'indice de Sato, le calcul de cet indice ayant été fait en considérant tous les sujets et non pas seulement les quelques-uns présentés ici.

Tableau 7.11
Patrons de réponses de quelques élèves des groupes 1 et 2 (les plus faibles) qui ont répondu aux 45 derniers items à choix multiple (les plus difficiles) : toutes les valeurs de l'indice de Sato sont supérieures à 0,5.

111000100011100001101000000100010101010111011
010101001000001010110101100000010001011110100
011001100010000001110101001000000011110100011
111111110010110000101000000000111001010101000
100010101010000001100100000100011100001000111
100001001011000110110100000011010000000100011
010111011100000000000100110101100000000000010
011000010001000011101000000100111100001000010
001101000000000101010001000011011110100000010
110000110000000000010001000111010101000100110
111111111101111010100110000010101010110000111
101000000000110001101111100100001010000010000
001010000001111100010001001000000110000011000
100100011000000010001100110000010001011000100
011000001000001010001100001010011000010010100
001100110000000000101010000101110100001000010
100110101000000010000000101100001001000000111

16. Puisque ce test ne compte pas vraiment pour les élèves, nous considérons comme peu vraisemblable l'hypothèse de la tricherie massive.

prise sur la foi du score de chaque sujet. Dans le cas qui nous intéresse, l'exemple du test de mathématique du PIRS, le résultat de l'analyse des patrons de réponses[17], a amené les auteurs à ne considérer, pour chacun des élèves, que les 75 items les plus pertinents afin de calculer leur score d'habileté TRI. Ainsi, pour les élèves du groupe 1, les plus faibles, seuls les 75 items des niveaux de difficulté 1, 2 et 3 ont été retenus ; pour les élèves du groupe 2, seuls les 75 items des niveaux de difficulté 2, 3 et 4 ont été retenus alors que pour les élèves du groupe 3, seuls les 75 items des niveaux de difficulté 3, 4 et 5 ont été considérés valides. Puisque la détection de patrons atypiques ne peut pas toujours aboutir à une solution aussi élégante, nous n'insisterons jamais assez sur la préparation de consignes claires, mises à l'essai chez un échantillon de sujets de la population visée, et sur le contrôle de la qualité lors de l'administration d'un instrument de mesure. Le vieux dicton « mieux vaut prévenir que guérir » prend ici tout son sens.

17. D'autres considérations ont aussi été prises en compte dont, bien sûr, la proportion d'items non atteints.

Exercices

1. Dites pourquoi il n'est pas approprié d'affirmer qu'un test de chimie organique est valide du seul fait qu'il est formé d'items de chimie organique.

2. Expliquez, en vos propres mots, pourquoi il est plus approprié de parler de méthodes de validation que de types de validité.

3. Dans le cadre d'une réplication du *Thurstone box problem*, seulement huit variables sont retenues : x, y, x^2, y^2, log(x), log(y), exp(x), exp(y). Combien de facteurs devrions-nous logiquement obtenir à la suite d'une analyse en composantes principales de ces huit variables ?

4. Un test de mathématique de 20 items a été administré à 600 individus de 18 ans. Tous les items dont les numéros sont pairs (I2, I4, etc.) portent sur la résolution de problèmes en algèbre alors que les autres items traitent des figures géométriques. Par ailleurs, les dix premiers items (I1 à I10) sont des problèmes de la vie courante alors que les autres sont des problèmes de portée purement mathématique. Lors d'une étude de validité, on veut vérifier la structure interne de ce test. Comment interpréter la structure interne de ce test si, au terme d'une analyse factorielle, trois facteurs ont été retenus et les saturations observées après rotation sont les suivantes ?

	Facteur 1	Facteur 2	Facteur 3
I19	0,83861	0,11021	0,10405
I7	0,82752	0,01035	0,00284
I11	0,73207	0,00238	0,10003
I5	0,50563	0,09260	0,12031
I3	0,50498	0,11063	0,00274
I17	0,41746	0,08345	0,10006
I4	0,00129	0,87544	0,01004
I10	0,06004	0,78072	0,10082
I8	0,00004	0,85074	0,00054
I2	0,10368	0,76016	0,12043
I20	0,11008	0,01923	0,77526
I1	0,00236	0,00003	0,65727
I6	0,10934	0,12001	0,84019

5. Donnez cinq procédures qui permettraient de déterminer le nombre de facteurs émanant d'une analyse factorielle.

6. En relation avec l'exemple proposé au tableau 7.10 (p. 268), supposons qu'un cinquième sujet, de patron 1 0 1 0 1 mais de même habileté que les quatre sujets déjà présents, ait passé ce test de cinq items. Calculez l'indice de Sato et l'indice L_z de ce cinquième sujet.

Corrigé des exercices

1. La validité d'un instrument est liée à l'interprétation qui est faite des scores et non à l'allure que semble avoir l'instrument. Après tout, même si un test comporte des problèmes mathématiques écrits, il peut être tout à fait inapproprié d'interpréter les scores de ce test comme indiquant une habileté en calcul s'il est administré à de mauvais lecteurs.

3. Deux facteurs, car seulement deux dimensions, x et y, sont prises en compte.

5. Plusieurs procédures peuvent être considérées pour déterminer le nombre de facteurs : l'inspection des regroupements de variables dans une matrice de corrélations ; l'inspection visuelle du graphique des éboulis ; le nombre de valeurs propres supérieures à 1 ; le test du khi-carré dans le cas de la procédure d'extraction par le maximum de vraisemblance ; l'utilisation du pourcentage de corrélations résiduelles.

PARTIE 2

APPLICATIONS

CHAPITRE 8

Détection des biais d'item

Il y a quelques années, un organisme américain très connu, l'Educational Testing Service (ETS), lançait une série d'enquêtes appelée International Assessment for Educational Progress (IAEP), le pendant international des enquêtes nationales américaines connues sous le nom de National Assessment for Educational Progress (NAEP). Les résultats de ces enquêtes (Lapointe, Mead et Askew, 1992), par ailleurs très défavorables aux élèves américains de 13 ans, ont suscité des commentaires d'un grand nombre de chercheurs. Parmi ceux-ci, Howard Wainer, employé par ETS, tentant d'expliquer la contre-performance des jeunes Américains, avançait l'idée que ces enquêtes pouvaient difficilement être considérées justes et équitables, car les jeunes Coréens, grands vainqueurs de ce concours, étaient honorés d'avoir été choisis pour défendre la gloire de leur pays et, de ce fait, beaucoup plus motivés que les jeunes Américains qui, pour leur part, étaient plus ou moins tirés à contrecœur de leur cours d'éducation physique pour répondre aux questions de mathématique et de sciences du test de l'IAEP (Bertrand et Jeanrie, 1995). Cet argument,

par contre, n'explique pas tout : il n'il y pas que la Corée qui ait déclassé les États-Unis, mais aussi la plupart des autres pays d'Europe qui ont participé aux enquêtes ! Pour ajouter à la controverse, quelques années plus tard, était publié un tableau d'honneur des prix Nobel décernés par pays au cours du XX^{e} siècle. Si, au début du siècle, les Européens obtenaient la plupart des prix, cette tendance s'est inversée depuis la fin de la Deuxième Guerre mondiale, si bien qu'au cours des vingt dernières années les Américains ont obtenu ce prestigieux prix deux fois plus souvent que les Européens. Plusieurs facteurs peuvent expliquer cette inversion de tendance, dont les vagues successives d'immigration d'Européens aux États-Unis, notamment depuis le milieu du XX^{e} siècle ; n'empêche que la situation est tout de même un peu curieuse. Lors de la conférence de presse, tenue en 1990, au cours de laquelle étaient annoncés les résultats désastreux qu'avaient obtenus les jeunes Américains aux enquêtes de l'IAEP, le président George Bush (père) avait promis à ses compatriotes que les jeunes Américains feraient beaucoup mieux au tournant du siècle. Malheureusement pour nos voisins du sud, la situation n'a guère changé depuis. Faudrait-il accuser le système scolaire américain de favoriser une élite au détriment de la masse ? Nous n'avons ni le temps ni la compétence pour répondre à cette question. Demandons-nous plutôt dans quelle mesure ces comparaisons peuvent être considérées valides, car au-delà de ce genre de réflexions, somme toute anecdotiques, se trouve un besoin de justice et d'équité lorsqu'il s'agit de comparer des groupes et d'établir un palmarès des écoles, des universités, des juridictions d'un pays ou, plus globalement, des pays. En ce sens, il semble tout à fait légitime de s'interroger sur l'équivalence des critères de comparaison, des plans d'échantillonnage ou encore des différentes versions des instruments de mesure utilisés pour faire ces comparaisons. Plus spécifiquement, il semble bien légitime de se poser la question de la présence de biais dans les instruments de mesure employés lors de ces comparaisons.

C'est au chapitre précédent que nous avons distingué les trois types de biais associés à la construction, à l'administration ou à la traduction d'un instrument de mesure : nous avons alors présenté les notions de biais de concept, de biais de méthode et de biais d'item. Le biais de concept, nous l'avons vu, est engendré par un glissement du concept lorsque l'instrument est traduit ou adapté dans une autre langue ou une autre culture. Nous avons élargi le sens initialement donné au biais de concept par Van de Vijver et Leung (1997) en considérant aussi le changement de modalité d'administration du test, par exemple de la modalité papier-crayon à la modalité informatisée. Le biais de méthode, pour sa part, concerne tout ce qui a trait aux caractéristiques touchant l'administration du test : format d'item, respect des consignes, etc. Nous avons volontairement exclu de ce type de biais celui qui provient de la façon de répondre du sujet : nous en avons fait une catégorie de biais à part, compte tenu de l'importance que nous lui accordons. Enfin, le biais d'item, celui sur lequel se portera notre attention au cours de ce chapitre, origine du préjudice que la formulation de l'item peut porter à certains sujets, particulièrement à

la suite de la traduction ou de l'adaptation du test. De nature différente, ces biais ont cependant la même conséquence, à savoir défavoriser un sous-groupe de sujets à qui le test est destiné, que ce sous-groupe soit généré par des différences culturelles, socioéconomiques ou linguistiques ou encore par une différence de sexe. Les biais doivent être détectés de façon à ce que les valeurs que sont la justice et l'équité, auxquelles renvoient souvent les documents officiels du ministère de l'Éducation portant sur les politiques d'évaluation des apprentissages, ne restent pas lettre morte.

Après avoir distingué les concepts de biais d'item et de biais de test, nous présenterons un florilège des méthodes récentes de détection de biais d'item et de test en distinguant les méthodes fondées sur les modèles de réponses aux items de celles qui ne sont pas fondées sur ces modèles. Cette présentation sera suivie par la description détaillée d'un exemple mettant en scène tantôt une méthode non fondée sur un modèle de la TRI, tantôt une méthode fondée sur un tel modèle. Une dernière section portant sur les limites associées aux méthodes de détection de biais d'item conclura ce chapitre.

8.1. VERS UNE DÉFINITION DU CONCEPT DE BIAIS D'ITEM

L'objectif de ce chapitre n'est pas de présenter de façon exhaustive toutes les méthodes de détection de biais d'item apparues depuis cinquante ans. Ce n'est pas non plus, pour autant, de donner une recette gagnante pour détecter des biais d'item, car il n'y en a tout simplement pas. Nous pensons cependant que les très importants travaux effectués dans ce domaine de la mesure appliquée en éducation et en psychologie, plus particulièrement au cours des quinze dernières années, méritent d'être connus. Il ne sera donc pas question des méthodes qualitatives basées sur le jugement d'experts (Berk, 1982), non seulement parce que ces méthodes commencent à prendre de l'âge, mais aussi et surtout parce que ces méthodes n'ont tout simplement pas fait leurs preuves (Camilli et Shepard, 1994, p. 136). Nous nous concentrerons donc sur des méthodes dites empiriques, à savoir celles qui se basent sur les scores au test pour déterminer a posteriori les items jugés biaisés envers un sous-groupe de la population visée par un test. Il est certain, par contre, que nous ne dénigrons pas la méthode qui consiste à effectuer un examen attentif du contenu des items avant que le test ne soit administré. Il serait bien légitime, en effet, de rejeter des items a priori sur la foi d'un contenu trop chargé culturellement ou encore parce que ces items font appel à des propos sexistes. Bien que cette méthode soit bien légitime, nous considérons qu'il ne s'agit pas là d'une méthode de détection de biais d'item en bonne et due forme, en tout cas pas au sens où nous l'entendons ici, mais bien d'une étape dans la construction d'un instrument de mesure. Notons que le rejet d'un item a priori par un groupe d'experts ne se fonde pas du tout sur les scores au test. Nous verrons par contre que, selon la nature libérale ou conservatrice de l'approche adoptée

par le chercheur, une méthode de détection de biais peut être ou non suivie d'un examen, par un comité d'experts, des items jugés potentiellement biaisés suite aux résultats obtenus à l'aide d'une méthode empirique. Il faut préciser que, contrairement à l'examen attentif de tous les items du test fait a priori par un groupe de spécialistes, le comité d'experts dont il est question ici n'examinera, le cas échéant et a posteriori, que les items jugés potentiellement biaisés par une méthode empirique.

La notion de biais d'item a beaucoup évolué depuis 25 ans, en tout cas bien autant que les méthodes permettant de le détecter qui ont évolué en parallèle. C'est pourquoi il apparaît important de suivre l'évolution de cette notion de façon à la distinguer de notions non équivalentes, mais tellement parentes qu'elles pourraient être prises à tort pour un biais d'item. Nous pensons que ce n'est qu'en définissant cette notion de façon suffisamment rigoureuse que nous pourrons suggérer de meilleures méthodes pour détecter ce genre de biais.

Plaçons-nous dans la situation où on a administré, à un groupe d'étudiants anglophones et à un groupe d'étudiants francophones, un test de mathématique comprenant notamment des items à choix multiple d'algèbre et de géométrie et des items à réponse construite de résolution de problèmes. Appelons groupe de référence le groupe d'anglophones et groupe focal le groupe de francophones. Nous voulons savoir si le fait de traduire ce test dans une autre langue (imaginons que le test a d'abord été conçu en anglais puis traduit en français) aurait pu engendrer des biais d'item envers l'un ou l'autre groupe.

Nous avons voulu, d'entrée de jeu, proposer une définition de biais d'item qui s'éclaircira au cours de ce chapitre et qui nous permettra de la distinguer d'autres notions voisines, mais distinctes. Parcourant les divers textes portant sur les méthodes de détection de biais d'item, et il y en a de très nombreux, nous nous sommes rendu compte qu'ils ne proposaient que très rarement une définition claire et précise de cette notion. Voici donc notre proposition :

Un **item** i sera dit **biaisé** envers un groupe (que ce soit le groupe de référence ou le groupe focal) si les deux critères suivants sont respectés :

1. deux sujets d'habileté égale mais appartenant à des groupes distincts ont une probabilité différente de réussir l'item i (ou, en d'autres mots, en présence d'un FDI) **et**
2. la raison de cette différence de probabilité de réussite n'a rien à voir avec l'interprétation usuelle que l'on fait des scores au test (ou, en d'autres mots, en l'absence de validité).

Cette définition montre que, si la notion de biais peut être basée sur une statistique en rapport avec une différence de probabilité de réussite, elle n'est pas pour autant une notion statistique en elle-même. Le biais d'item ne se résume pas à une valeur ; il procède plutôt d'un jugement basé sur une ou des valeurs. Cette définition montre aussi que nous devrons définir précisément

ce que l'on entend par l'expression « probabilité différente ». En d'autres termes, jusqu'à quel point la différence de probabilité de réussite doit-elle être élevée avant que le critère 1 soit satisfait ? Nous verrons qu'encore aujourd'hui, cette question est matière à débat.

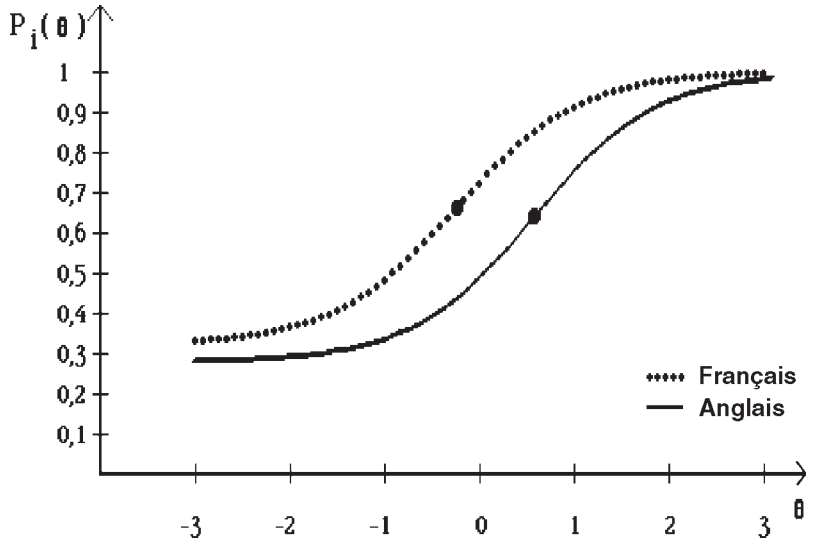
Voyons maintenant comment des notions voisines ont pu être confondues avec celle de biais d'item.

Nous définirons l'impact IM d'un item, noté i, comme la différence entre les valeurs de l'indice classique de difficulté du groupe focal (F) et celles du groupe de référence (R) : $IM_i = p_{Fi} - p_{Ri}$. Est-ce qu'une valeur élevée d'impact relative à un item est nécessairement un signal que l'item est biaisé ? Non ! En effet, même si la valeur de l'impact d'un item est très élevée, cela ne veut pas dire que, à habileté égale, la différence entre les valeurs de l'indice de difficulté sera aussi élevée. Un item dont la valeur de l'impact est élevée signifie encore moins que la raison de la différence est sans rapport avec les interprétations que l'on peut faire des scores. Il peut donc être tout à fait légitime que la valeur de l'impact d'un item soit élevée sans que cet item soit biaisé. Il est connu, par exemple, que les élèves québécois francophones réussissent systématiquement mieux en mathématique que les élèves ontariens anglophones du même âge (CMEC[1], 1993 ; CMEC, 1997 ; CMEC, 2001) pour des raisons qui pourraient être d'ordre historique ou autre. Il serait donc normal que la valeur de l'impact de la plupart des items soit élevée sans nécessairement correspondre à un biais. Imaginons, par ailleurs, que le programme d'études des étudiants anglophones ait été axé beaucoup plus sur la géométrie que le programme d'études des étudiants francophones. Il serait alors bien normal que la valeur de l'impact des items de géométrie soit élevée, voire très élevée. Après tout, plusieurs enquêtes à grande échelle ont justement comme objectif de comparer les programmes d'études de juridictions distinctes et de vérifier les conséquences sur le rendement scolaire des élèves en administrant le même test à des groupes linguistiques ou culturels différents.

La notion statistique la plus communément confondue avec un biais d'item est celle de FDI ou fonctionnement différentiel d'item. La meilleure façon (mais pas la seule) d'exposer cette notion de FDI consiste à examiner deux courbes caractéristiques d'un item (CCI), un peu comme à la figure 8.1.

1. Nous faisons référence ici au Programme des indicateurs du rendement scolaire (PIRS) administré sur une base cyclique à près de 50 000 élèves de 13 et de 16 ans par le Conseil des ministres de l'éducation du Canada (CMEC).

FIGURE 8.1
Courbes caractéristiques de l'item 6 de l'enquête de mathématique du PIRS de 1997. La courbe en pointillés concerne les francophones alors que la courbe en trait plein se rapporte aux anglophones.



Chacune de ces courbes représente la probabilité de réussir un item en fonction de l'habileté des individus provenant de l'un ou l'autre de deux groupes de sujets : un groupe d'anglophones et un groupe de francophones. Attention : ces deux courbes sont associées au même item. L'une renvoie à la probabilité de réussite de cet item pour un groupe d'anglophones alors que l'autre renvoie à la probabilité de réussite du même item pour un groupe de francophones. On peut définir le FDI comme la différence entre la probabilité de réussir cet item pour des élèves d'habileté égale, mais appartenant à des groupes distincts. En examinant la figure 8.1 par exemple, nous voyons bien qu'un sujet anglophone d'habileté moyenne ($\theta = 0$) n'a pas la même chance de réussir cet item qu'un sujet francophone de même habileté. La probabilité de réussite pour le sujet francophone est d'un peu plus de $P_{Fi}(\theta) = 0{,}7$ alors qu'elle est d'un peu moins de $P_{Ai}(\theta) = 0{,}5$ pour le sujet anglophone. Cette façon de définir le FDI revient à considérer l'aire entre les deux CCI. Dans ce contexte, une absence de FDI résulterait en la superposition des deux CCI.

Notons que, définie comme cela, la notion de FDI doit être considérée d'une façon relative. Il serait inexact de ne parler que de présence ou d'absence de FDI, car il y a toujours une certaine différence de probabilité, même très faible, entre les deux CCI : autrement dit, l'aire entre les deux CCI n'est jamais tout à fait nulle. C'est pourquoi nous ne parlerons de FDI que

lorsque seront observées des valeurs anormalement élevées de l'aire entre les CCI. Il faut indiquer également que, comme nous allons le voir bientôt, on pourrait définir le FDI sans avoir recours aux CCI.

Ainsi, le FDI est bien une notion statistique, une certaine valeur relative à une différence de probabilités. Or, il faut le noter, même si le FDI est une notion importante qui fait partie de la définition d'un biais d'item, un item qui comporte un FDI n'est pas nécessairement biaisé. En d'autres mots, un FDI ne mène pas nécessairement à un biais d'item. On pourrait dire qu'un item présentant un FDI satisfait le critère 1 de la définition de biais d'item. Il n'est pas du tout certain cependant que le critère 2 soit éventuellement satisfait. Or ce dernier critère consiste à porter un jugement sur le contenu de l'item. Une fois observé le FDI, il faut savoir si l'écart de probabilité de réussite entre les deux groupes est signifiant ou non, c'est-à-dire en rapport avec l'interprétation que l'on fait habituellement des scores au test. Par exemple, dans le cas du test de mathématique dont il est question plus haut, si les items de résolution de problèmes comportent pour la plupart un FDI, il faut voir si une traduction fautive ne serait pas à l'origine de la différence de probabilité de réussite entre les deux groupes. Notons, par ailleurs, que cette différence pourrait jouer tout autant en faveur du groupe de référence à qui on a administré la version originale de l'instrument qu'en faveur du groupe focal qui a passé la version traduite. En effet, la version traduite pourrait être plus difficile si certains mots ou expressions étaient plus complexes en français qu'en anglais ; au contraire, la version traduite pourrait tout aussi bien contenir des indices favorisant une bonne réponse, indices qui étaient absents dans la version originale anglaise. Il est cependant tout à fait possible, comme nous l'avons déjà exprimé, qu'une valeur élevée de FDI ne mène pas à considérer que l'item est biaisé. Ce serait le cas si plusieurs des items de géométrie favorisaient par exemple le groupe de référence uniquement parce que les élèves de ce groupe avaient été soumis à des cours de mathématique mettant plus l'accent sur la géométrie que les élèves du groupe focal. Il existe aussi des situations qui comportent des jugements beaucoup plus subtils. Imaginons que plusieurs items de résolution de problèmes comportent un FDI, disons en faveur des élèves du groupe de référence, et que l'on se rende compte, après analyse par un comité d'experts, que les items ont été suffisamment bien traduits. Sommes-nous en présence d'un biais d'item ? Peut-être bien, peut-être pas. Cela dépend vraiment de la façon dont les scores seront interprétés. Supposons que les items de résolution de problèmes comportent beaucoup plus de mots en français qu'en anglais (situation fréquente lorsqu'on traduit de l'anglais, langue plus synthétique, au français) et que les élèves de 13 ans n'aient pas encore tous acquis une compétence élevée en compréhension en lecture. Si, d'aventure, les scores sont interprétés comme signifiant une habileté générale en mathématique, sans plus, il pourrait être légitime alors de considérer ces items comme étant biaisés. Si, par contre, le comité d'experts juge que la compréhension en lecture fait partie des habiletés (secondaires) légitimes visées par

(quelques items de) ce test (arguant que quelqu'un qui veut réussir en mathématique, ce qui inclut la résolution de problèmes, doit aussi savoir bien poser des problèmes et donc bien lire, etc.), ces items ne seraient pas considérés comme étant biaisés même si les valeurs observées des FDI étaient jugées élevées.

8.1.1. Approche libérale ou approche conservatrice

De façon générale, on peut distinguer deux approches lorsqu'il est question de détection de biais d'item : l'approche libérale et l'approche conservatrice. Tel qu'exposé par Camilli et Sheppard (1994, p. 149), l'approche libérale consiste à éliminer un item dès qu'un FDI a été identifié. Cette approche permet de minimiser l'erreur de type 2, qui serait d'accepter l'hypothèse qu'il n'y a pas de biais alors qu'en réalité, il y a bel et bien un biais. L'approche conservatrice, au contraire, stipule qu'un FDI peut mener à un biais d'item, mais il faut d'abord montrer que ce FDI est la conséquence d'une autre dimension (autre que le thêta) mesurée par l'item et non pertinente au test. On voit bien que l'approche conservatrice minimise l'erreur de type 1, qui serait de rejeter l'hypothèse qu'il n'y a pas de biais alors qu'en réalité il n'y a pas de biais.

Ces deux approches comportent un lot d'avantages et d'inconvénients. L'approche libérale, par son caractère automatique, convient bien dans un contexte de production à grande échelle où une décision doit être prise rapidement sur l'opportunité de considérer des items comme étant biaisés ou non. Cette approche peut cependant mener à des décisions erronées. Ce serait le cas, par exemple, si le FDI n'était pas vraiment un biais d'item (au sens où nous l'entendons), mais, comme on l'a vu, une particularité de l'item qui, au-delà de l'habileté générale mesurée par le test (p. ex., l'habileté à résoudre des problèmes mathématiques écrits), sollicite une habileté secondaire (p. ex., la compréhension en lecture) bien légitime dans le contexte du test. L'élimination de ce genre d'item constituerait donc une erreur dans le processus de décision.

L'approche conservatrice, pour sa part, convient mieux à un contexte de recherche ou à l'analyse en profondeur d'un test. Elle peut par contre mener également à une mauvaise décision dans le cas où, par exemple, nous ne trouverions pas la raison (qui existe pourtant) d'éliminer un item, donc de le considérer comme étant biaisé. En d'autres termes, cette approche pourrait tromper l'utilisateur qui ne trouverait pas de raison d'éliminer un item ayant un FDI.

Au-delà de ces approches, il est toutefois possible de déboucher sur des décisions erronées, en éliminant par exemple un seul item qui défavorise grandement le groupe de référence alors que la plupart des autres items défavorisent le groupe focal, mais ne sont pas éliminés parce que pas considérés comme des FDI, si bien qu'au total c'est le groupe focal qui est vraiment défavorisé. Dans ce cas-ci, l'élimination de l'item biaisé accentue donc encore plus le caractère injuste du test envers le groupe focal.

8.2. FLORILÈGE DES MÉTHODES EMPIRIQUES DE DÉTECTION DES BIAIS D'ITEM NON FONDÉES SUR LA TRI

La plupart des méthodes dont il sera question dans cette section seront étudiées avec un objectif purement pédagogique, à savoir examiner tous les contours du concept de biais d'item et des méthodes proposées au cours des dernières décennies pour le détecter. Nous nous concentrerons donc sur les méthodes qui ont fait leur marque par le passé, qui ont apporté quelque chose, soit à la définition du concept, soit aux méthodes pour le détecter.

Avouons tout d'abord qu'il n'y a pas de recette miracle quand vient le temps de détecter un biais d'item. Encore aujourd'hui, plusieurs méthodes sont à l'étude et aucune n'a encore reçu l'assentiment général de la communauté des chercheurs en psychométrie ou en édumétrie, que la méthode s'appuie sur un modèle TRI ou non. Tentons de voir, à partir de notre définition de biais d'item, quelles sont les principales caractéristiques des méthodes proposées par le passé, méthodes qui, ici, ne s'appuient pas sur un modèle de la TRI.

Une idée de base qui a présidé à quelques-unes des premières méthodes de détection de biais est celle de la difficulté différentielle (Camilli et Shepard, 1994), qui peut se décrire comme suit. Soit un test qui a été administré à deux groupes, disons le groupe de référence et le groupe focal : selon cette conception, un item est considéré biaisé envers un groupe si la différence de difficulté de cet item entre les groupes est supérieure à la différence moyenne de difficulté de tous les items du test entre ces deux groupes. En d'autres termes, dans la mesure où on observe une différence particulièrement grande entre les indices de difficulté d'un item i, si grande qu'elle dépasse largement celle relative aux autres items, l'item i pourrait être considéré biaisé. Plusieurs méthodes de détection de biais se sont inspirées de cette idée, notamment la méthode *delta plot* d'Angoff (1982), la méthode de Shepard *et al.* (1984) et l'analyse de la variance à mesures répétées (Cleary et Hilton, 1968).

8.2.1. Méthode basée sur l'analyse de la variance

Voyons un peu comment on peut en arriver à détecter un « biais d'item[2] » à partir d'une analyse de la variance. Il faut tout d'abord définir deux facteurs, celui relatif aux groupes (G) et celui relatif aux items (I) : le facteur G possède deux niveaux, le groupe de référence et le groupe focal. Les sujets (S) sont nichés dans le facteur G et le facteur I est croisé avec le facteur G. Le devis d'observation (voir le chapitre 3) s'écrit donc (S:G)×I. Si l'interaction entre le facteur G et le facteur I est significative, c'est qu'il y a un ou plusieurs items

2. L'expression est entre guillemets, car l'interaction significative entre le facteur groupes et le facteur items ne répond nullement à nos critères de biais d'item.

potentiellement biaisés. C'est en procédant à des contrastes concernant ces interactions qu'on pourra identifier l'item ou les items fautifs (ceux qui contribuent le plus à l'interaction).

Un exemple servira à décrire cette méthode présentée ici pour des raisons historiques, car, tel que le mentionnent Camilli et Shepard (1994, p. 34), cette méthode ne peut être recommandée aujourd'hui pour détecter des biais d'item. Le tableau 8.1 montre les indices de difficulté de 4 items pour le groupe de référence et le groupe focal. Si les items 1, 2 et 3 ont été mieux réussis par le groupe focal, il n'en est pas de même pour l'item 4. Cette situation engendre une interaction statistiquement significative au seuil de 5 % entre le facteur G et le facteur I, comme on peut le constater au tableau 8.2. En procédant à des analyses de contrastes, il est possible de détecter quel item est responsable de cette interaction significative. Le tableau 8.3, en effet, montre que, lorsqu'on compare à chacun des trois autres items, l'item 4 induit un verdict statistiquement significatif. La même analyse montre qu'aucune comparaison entre les trois autres items, n'induit de verdict statistiquement significatif.

TABLEAU 8.1
Indices de difficulté de quatre items pour deux groupes

	Groupe focal (n = 15)	Groupe de référence (n = 15)
Item 1	0,7333	0,6667
Item 2	0,6000	0,5333
Item 3	0,6000	0,5333
Item 4	0,2000	0,7333

TABLEAU 8.2
Analyse de la variance à mesures répétées (correction de Greenhouse-Geisser) : le facteur de répétition concerne les items

	Somme des carrés	Degrés de liberté	Carrés moyens	F	Sig.
Items	0,825	2,668	0,309	1,255	0,295
Items * groupes	**2,025**	**2,668**	**0,759**	**3,082**	**0,038**
Erreur	18,400	74,690	0,246		

TABLEAU 8.3
Contrastes entre les quatre items montrant l'interaction significative entre le facteur items et le facteur groupes relative au niveau 4 (l'item 4)

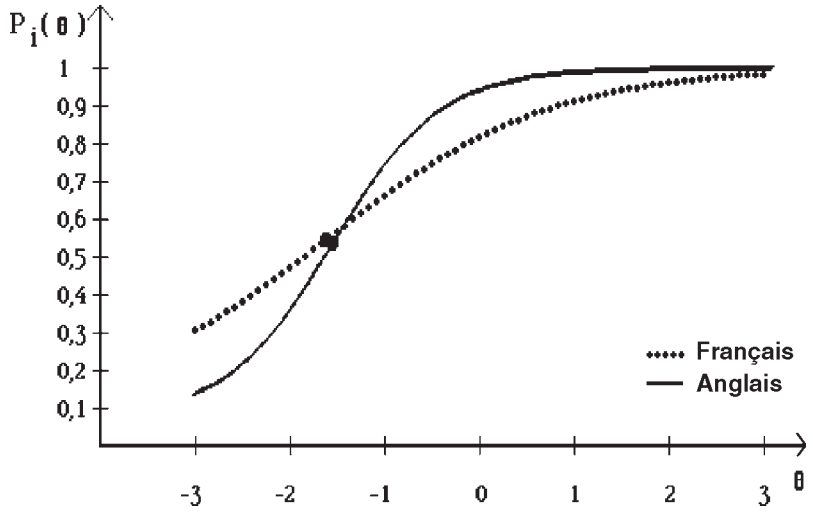
	Items	Somme des carrés	Degrés de liberté	Carrés moyens	F	Sig.
Items	Niveau 1 vs Niveau 4	1,633	1	1,633	5,277	0,029
	Niveau 2 vs Niveau 4	0,300	1	0,300	0,525	0,475
	Niveau 3 vs Niveau 4	0,300	1	0,300	0,840	0,367
Items * groupes	**Niveau 1 vs Niveau 4**	**2,700**	**1**	**2,700**	**8.723**	**0,006**
	Niveau 2 vs Niveau 4	**2,700**	**1**	**2,700**	**4.725**	**0,038**
	Niveau 3 vs Niveau 4	**2,700**	**1**	**2,700**	**7.560**	**0,010**
Erreur	Niveau 1 vs Niveau 4	8,667	28	0,310		
	Niveau 2 vs Niveau 4	16,000	28	0,571		
	Niveau 3 vs Niveau 4	10,000	28	0,357		

8.2.2. Méthode basée sur la régression logistique

La régression logistique a également permis de proposer une méthode de détection de biais d'item qui est de plus en plus en vogue (Clauser et Mazor, 1998). La variable dépendante dans un tel modèle de régression logistique est dichotomique : il s'agit de la réussite ou de l'échec à l'item analysé. Chaque item doit donc être analysé de façon indépendante. Le modèle de régression retenu peut se concevoir en deux ou trois étapes, mais dans chaque cas, la première étape consiste à entrer dans l'équation de régression la variable de contrôle, généralement le score total au test ou encore l'estimé d'habileté TRI. Dans un modèle à deux étapes, la deuxième étape consiste à entrer dans l'équation de régression un bloc de deux variables, le groupe (linguistique, culturel, etc.) et l'interaction groupe*score. Il faut alors tester si l'ajout de ce bloc de deux variables mène à un verdict statistiquement significatif. Si oui, il y a FDI. La méthode en trois étapes est plus élaborée : la deuxième étape consiste à entrer un bloc d'une seule variable, le groupe. La troisième étape consiste à entrer le bloc constitué de l'interaction groupe*score. Dans la mesure où l'ajout de cette variable d'interaction est significatif, le FDI comporterait une composante non uniforme significative, c'est-à-dire que la différence entre les deux groupes ne serait pas uniforme d'un score à l'autre, un peu comme à la figure 8.2. Par exemple, il pourrait y avoir, pour un item donné, une différence importante en faveur des sujets faibles du groupe focal (ici les francophones) et une différence importante en faveur des sujets forts du groupe de référence (ici les anglophones). La logique de cette méthode est relativement simple. La première étape de l'analyse de régression logistique implique une seule variable indépendante, le score total au test. À la deuxième étape, une autre variable est considérée, le groupe. Si l'ajout de cette variable aboutit à un verdict statistiquement significatif, l'interprétation doit être la suivante : une fois considéré le score total au test, le fait d'appartenir à un des deux

groupes (de référence ou focal) explique de façon significative le fait de réussir ou pas à l'item. C'est en ce sens que nous devons parler de FDI. Si, en plus, lors de la troisième étape, l'interaction groupe*score aboutit à un verdict statistiquement significatif c'est que, en plus du score total au test et du fait d'appartenir à un des deux groupes, le fait d'avoir un score total fort ou faible combiné au fait d'appartenir à un des deux groupes explique de façon statistiquement significative la réussite à l'item. C'est seulement lorsque cette interaction mène à un verdict statistiquement significatif que nous disons que le FDI comporte une composante non uniforme significative.

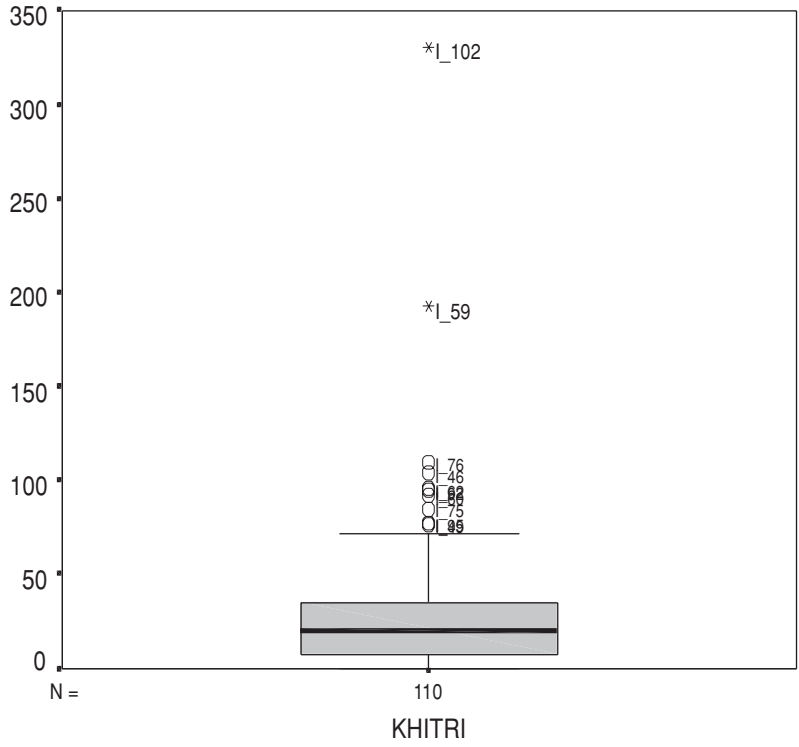
FIGURE 8.2
FDI non uniforme pour un item de l'enquête de mathématique du projet PIRS (1997)



Puisque cette méthode est sensible au nombre de sujets dans chacun des groupes (c'est un test du khi-carré qui détermine si l'apport est significatif ou non), nous proposons une autre façon d'interpréter les FDI. Il s'agit de procéder au déploiement du diagramme en boîte et moustaches (Bertrand et Valiquette, 1986) des valeurs du khi-carré pour tous les items du test. Si la valeur du khi-carré pour un item donné est en même temps statistiquement significative et considérée comme une valeur extrême (une valeur est dite extrême si elle est située à plus de 3 fois l'étendue interquartile du troisième quartile) en regard du diagramme, nous dirons que l'item présente un FDI de catégorie C (FDI sévère). Si la valeur du khi-carré est statistiquement significative et considérée aberrante mais non extrême (une valeur est dite aberrante mais non extrême si elle est située à plus de 1,5 fois mais à moins de 3 fois

l'étendue interquartile du troisième quartile) en regard du diagramme, nous dirons que l'item présente un FDI de catégorie B (FDI modéré). Dans tous les autres cas, nous dirons que l'item présente un FDI de catégorie A (FDI négligeable). La figure 8.3 montre comment il est possible d'isoler les items de catégorie B et de catégorie C. Cette classification permet au comité d'experts chargés de déterminer si les items où on détecte un FDI sont véritablement biaisés d'établir un ordre de priorité dans l'analyse du contenu des items en fonction des ressources à leur disposition.

FIGURE 8.3
Diagramme en boîte et moustaches qui laisse voir deux valeurs extrêmes (catégorie C) représentées par le symbole (*) et plusieurs valeurs aberrantes (catégorie B) représentées par le symbole (O).



Notons que Gierl, Rogers et Klinger (1999) proposent une autre façon de classifier les FDI à partir d'une régression logistique et ce pour tenir compte du trop grand nombre d'items déclarés FDI (faux positifs) si l'interprétation ne tient compte que du test du khi-carré. Supposons que l'on adopte la méthode

en deux étapes développée plus haut. Un item comportera un FDI sévère (catégorie C) si le test du khi-carré est statistiquement significatif et si la différence du R^2 (entre l'étape 1 et l'étape 2) est supérieure à 0,070. Le FDI sera modéré (catégorie B) si le test du khi-carré est statistiquement significatif et si la différence de R^2 (entre l'étape 1 et l'étape 2) se situe entre 0,035 et 0,070. Dans tous les autres cas, le FDI sera considéré négligeable. Si la méthode en trois étapes est adoptée, nous suggérons de tester l'ampleur d'un FDI en recourant à la différence de R^2 entre l'étape 1 et l'étape 3 (et non l'étape 2). Par la suite, on pourra tester l'ampleur de la part de non-uniformité présente dans le FDI en recourant à la différence de R^2 entre l'étape 2 et l'étape 3.

Le tableau 8.4 montre le résultat obtenu de la régression logistique réalisée sur les quatre items dont les indices de difficulté se retrouvent au tableau 8.1. Nous avons priviligié ici la méthode en deux étapes. C'est pourquoi il y a deux degrés de liberté pour le test du khi-carré : un pour le facteur groupe et un autre pour l'interaction groupe*score. Notons que seul l'item 4 en arrive à un verdict statistiquement significatif (au seuil 0,05) et doit donc être considéré comme FDI. Compte tenu du très petit nombre d'items, nous n'avons pas eu recours au diagramme en boîte et moustaches. En utilisant la classification de Gierl, Rogers et Klinger (1999), il appert que l'item 4 présente un FDI sévère (la différence de R^2 entre les étapes 1 et 2 est de 0,593 – 0,340 = 0,253).

Tableau 8.4
Test du khi-carré pour le bloc de deux variables groupe et groupe*score dans le cas où la variable dépendante est le score (0 ou 1) à l'item. Seul l'item 4 mène à un verdict statistiquement significatif.

Item 1		**Khi-carré**	**dl**	**Sig.**
	Bloc	1,989	2	0,370
	Modèle	22,592	3	0,000
Item 2		**Khi-carré**	**dl**	**Sig.**
	Bloc	0,591	2	0,744
	Modèle	3,757	3	0,289
Item 3		**Khi-carré**	**dl**	**Sig.**
	Bloc	2,489	2	0,288
	Modèle	19,007	3	0,000
Item 4		**Khi-carré**	**dl**	**Sig.**
	Bloc	8,798	2	0,012
	Modèle	17,622	3	0,001

8.2.3. Méthode de Mantel-Haenszel

C'est sur la méthode de Mantel-Haenszel que nous nous attarderons. C'est certainement, parmi les méthodes non fondées sur la TRI, celle qui a reçu le plus d'attention au fil des ans. Nous allons décrire sommairement le principe de cette méthode dans cette section, puis y revenir avec une application beaucoup plus poussée dans une section ultérieure. Les détails de cette méthode doivent être attribués à Holland et Thayer (1986).

Supposons donc que l'on veuille voir si un item d'un test de n items administré à deux groupes, le groupe de référence et le groupe focal, peut être considéré comme présentant un FDI. La première étape de cette méthode consiste à construire un tableau à double entrée comme celui présenté au tableau 8.5, pour chaque score observé X_k (k = 1 à n) au test.

TABLEAU 8.5
Fréquences de sujets de chaque groupe, dont le score au test est X_k, qui ont réussi ou échoué l'item i.

Score = X_k		
	Réussite à l'item (1)	**Échec à l'item (0)**
Groupe de référence	A	B
Groupe focal	C	D

La lettre A dans le tableau renvoie au nombre de sujets du groupe de référence qui ont réussi cet item, parmi ceux dont le score total au test est X_k. Les autres lettres du tableau, B, C et D, s'interprètent de façon similaire. Dans ce cas la statistique α_{MH} de Mantel-Haenszel est donnée par

$$\alpha_{MH} = \frac{\sum AD/T}{\sum BC/T}$$

où T = A + B + C + D et où la somme est prise sur tous les scores observés, c'est-à-dire sur tous les tableaux comme celui présenté plus haut. Notons qu'il y aura une valeur de α_{MH} pour chaque item. L'origine de cette formule est instructive. Considérons la statistique suivante, appelée le rapport de chances (*odds ratio*).

$$OR = \frac{p_R / q_R}{p_F / q_F} = \frac{AD}{BC}$$

où

p_R désigne la proportion de sujets du groupe de référence qui ont réussi l'item,

q_R désigne la proportion de sujets du groupe de référence qui ont échoué l'item,

p_F désigne la proportion des sujets du groupe focal qui ont réussi l'item,

q_F désigne la proportion des sujets du groupe focal qui ont échoué l'item.

Le rapport OR peut alors être interprété comme suit : si ce rapport est supérieur à 1, c'est que le rapport entre la proportion de réussite et la proportion d'échec est supérieur pour les sujets du groupe de référence. Ceux-ci ont donc plus de chances de réussir l'item. Mais le rapport OR est égal à AD/BC. Ainsi, la statistique α_{MH} est donc fonction du rapport entre les chances de réussir du groupe de référence et les chances de réussir du groupe focal.

En général, la statistique qui est testée est le logarithme népérien de α_{MH}, soit

$$\beta_{MH} = ln\ (\alpha_{MH})$$

L'interprétation de β_{MH} a une valeur heuristique puisque si le FDI favorise le groupe de référence, c'est que OR, donc α_{MH} aussi, est supérieur à 1 ; en conséquence $\beta_{MH} > 0$. Inversement, si le FDI favorise le groupe focal, alors $\beta_{MH} < 0$.

Appliquée aux données du tableau 8.1, la méthode MH en arrive au même verdict que celui précédemment trouvé avec les autres méthodes : c'est l'item 4 qui présente un FDI.

Le tableau 8.6 donne les fréquences de base qui ont permi de calculer la statistique α_{MH} en regard de l'item 4. Puisque le test comprend 4 items et que les items sont codés 1 pour une réussite et 0 pour un échec, les scores totaux possibles au test sont 0, 1, 2, 3 et 4. Les fréquences de réussite et d'échec pour les sujets des deux groupes sont donc présentées en cinq mini-tableaux, soit un pour chacun des cinq scores totaux.

Dans ce cas, $\alpha_{MH} = \dfrac{\sum AD/T}{\sum BC/T} = \dfrac{0/2+0/3+30/13+9/8+0/4}{0/2+0/3+1/13+1/8+0/4} = 17$, et $\beta_{MH} = ln\ (17) = 2{,}833$.

Puisque le niveau de probabilité observée[3] relatif à l'item 4 est, selon le tableau 8.7, de 0,01, c'est donc qu'il y a un FDI. Comme la valeur de β_{MH} est positive, ce FDI favorise le groupe de référence. Un regard rapide au tableau 8.1 nous rassure sur le sens de ce FDI puisque plus de 73 % des sujets du groupe de référence ont réussi cet item alors que seulement 20 % des sujets du groupe focal faisaient de même.

3. Bien qu'il soit possible de calculer une statistique de test associée à α_{MH} suivant une loi du khi-carré avec 1 degré de liberté (Holland et Thayer, 1986), nous utiliserons à la section 8.5 une autre façon de juger de l'ampleur de α_{MH}.

TABLEAU 8.6
Fréquences de sujets de chaque groupe, dont le score au test est X_k, qui ont réussi ou échoué l'item 4 du tableau 8.1.

Score = 0	Réussite à l'item (1)	Échec à l'item (0)
Groupe de référence	0	1
Groupe focal	0	1
Score = 1		
Groupe de référence	0	1
Groupe focal	0	2
Score = 2		
Groupe de référence	5	1
Groupe focal	1	6
Score = 3		
Groupe de référence	3	1
Groupe focal	1	3
Score = 4		
Groupe de référence	3	0
Groupe focal	1	0

TABLEAU 8.7
Valeurs des statistiques α_{MH}, β_{MH} et niveau de signification observée. Seul l'item 4 donne un verdict statistiquement significatif.

Item 1	
α_{MH}	0,167
β_{MH}	–1,792
Erreur-type de β_{MH}	1,354
Sig.	**0,186**
Item 2	
α_{MH}	0,517
β_{MH}	–0,659
Erreur-type de β_{MH}	0,915
Sig.	**0,471**
Item 3	
α_{MH}	0,267
β_{MH}	–1,322
Erreur-type de β_{MH}	1,089
Sig.	**0,225**
Item 4	
α_{MH}	17,000
β_{MH}	2,833
Erreur-type de β_{MH}	1,105
Sig.	**0,010**

8.3. FLORILÈGE DES MÉTHODES EMPIRIQUES DE DÉTECTION DES BIAIS D'ITEM FONDÉES SUR LA TRI

Comme nous allons le voir, toutes les méthodes émanant des modèles de la TRI sont basées soit sur une fonction de l'aire entre les deux CCI, soit sur un test de signification en rapport avec les paramètres d'item. Ces méthodes reviennent toutes plus ou moins à quantifier une différence entre les CCI. Elles requièrent toutes, cependant, des tailles d'échantillon relativement élevées surtout lorsque c'est la modélisation à trois paramètres qui convient.

Plusieurs méthodes basées sur la TRI ont déjà été proposées pour identifier un FDI. On n'a qu'à penser à celles suggérées par Berk (1982), Hulin *et al.* (1983), Camilli et Shepard (1994) et Raju *et al.* (1995). Nous ne voulons pas toutes les expliciter, mais nous désirons tout de même en présenter quelques-unes, l'objectif étant de montrer l'évolution qu'elles ont connue au cours des dernières années. La section 8.5 permettra de présenter en détail les méthodes sur lesquelles nous nous attarderons le plus.

Méthode de Wright, Mead et Draba (1976)

Cette méthode s'applique spécifiquement au modèle de Rasch (un paramètre) et teste la différence entre les indices de difficulté (les b_i) des deux groupes R et F par la formule

$$z_i = (b_{iR} - b_{iF}) \,/\, (SE(b_{iR})^2 + SE(b_{iF})^2)^{0,5}$$

où z_i suit une loi normale centrée et réduite et SE signifie l'erreur-type.

Ainsi, l'indice de FDI est la différence $b_{iR} - b_{iF}$, mais le test de signification est fait sur la statistique z_i.

Méthode de l'aire signée et de l'aire non signée (Rudner, 1977)

Il s'agit de calculer l'aire entre les deux CCI en utilisant l'intégration tout le long de l'axe du paramètre d'habileté θ.

$$SA = \int \left[P_{iR}(\theta) - P_{iF}(\theta) \right] d\theta$$

$$UA = \sqrt{\int \left[P_{iR}(\theta) - P_{iF}(\theta) \right]^2 \, d\theta}$$

Selon Camilli et Shepard (1994), cette méthode présente par contre deux problèmes. Les valeurs de SA (*signed area*) ou de UA (*unsigned area*) peuvent être infinies dans le cas du modèle à trois paramètres ; en plus, elles ne prennent pas en compte la distribution des sujets : il se peut par exemple que les sujets soient surtout concentrés dans un intervalle donné (p. ex., [–1, +1]) alors que la méthode suppose une distribution uniforme tout le long de l'axe du paramètre d'habileté.

Méthode RMSD de Linn, Levine, Hastings et Wardrop (1981)

Plutôt que d'intégrer tout le long de l'axe d'habileté, ces auteurs proposent d'approximer l'intégration employée par Rudner (1977) par une statistique basée sur 600 points de cet axe. La racine carrée de la moyenne des différences au carré ou *root mean square difference* (RMSD) constitue la statistique de choix.

$$RMSD_i = \{1/600\ \Sigma_j[P_{iR}(\theta_j) - P_{iF}(\theta_j)]^2\}^{0,5}$$

L'aire entre les CCI est alors approximée en divisant l'échelle θ entre –3 et +3 en 600 mini-intervalles égaux. C'est une méthode facile à expliquer qui, en plus, possède un support visuel non négligeable. Cette méthode règle un des problèmes rencontrés en employant la méthode précédente proposée par Rudner (1977), parce que les valeurs de $RMSD_i$ ne peuvent être infinies. Mais elle suppose toujours une distribution uniforme des théta. Nous verrons que le recours à la méthode suivante permettra de lever ce problème de distribution uniforme. En effectuant un diagramme en boîte et moustaches des valeurs $RMSD_i$ il est possible de cibler les items dont l'approximation de l'aire est la plus grande : ce seront les valeurs $RMSD_i$ considérées comme aberrantes ou extrêmes (les *outliers*) sur le diagramme.

Méthode de l'aire entre les CCI de Shepard, Camilli et Williams (1984) reprise par Camilli et Shepard (1994)

La méthode de Shepard *et al.* (1984) permettra de régler le problème généré par le recours à une distribution uniforme forcée des théta. Suivant cette méthode, les différences de probabilités ne seront comptabilisées que pour les n_F sujets du groupe focal. Des deux indices proposés par ces auteurs, l'un (SPD-θ[4]) est signé, c'est-à-dire que les valeurs sont tantôt positives tantôt négatives, et l'autre (UPD-θ) est non signé, c'est-à-dire que les valeurs sont toujours positives.

$$SPD\text{-}\theta = \Sigma_j[P_{iR}(\theta_j) - P_{iF}(\theta_j)] / n_F \qquad \text{où } j = 1, 2, \ldots, n_F.$$

$$UPD\text{-}\theta = (\Sigma_j[P_{iR}(\theta_j) - P_{iF}(\theta_j)]^2 / n_F)^{0,5} \qquad \text{où } j = 1, 2, \ldots, n_F.$$

Il faut noter que les deux jeux de paramètres, (a_R, b_R, c_R) pour le groupe de référence et (a_F, b_F, c_F) pour le groupe focal, doivent être équilibrés (*equated*) avant de calculer ces indices. Une bonne façon d'arriver à cet équilibre est de calibrer les deux jeux de paramètres ensemble en utilisant l'option *not reached* de BILOG-3 pour identifier les sujets qui n'ont pas atteint les items concernés. Ainsi, pour chaque item i du test, il faut générer deux autres

4. Selon Camilli et Shepard (1994, p. 67), il faut lire *signed probability difference controlling for* θ.

items, l'un, i_R, pour les sujets du groupe de référence et l'autre, i_F, pour les sujets du groupe focal. Il faut cependant supposer un certain nombre d'items communs au groupe de référence et au groupe focal.

Les auteurs ne proposent cependant pas de test de signification pour l'un ou l'autre de ces indices. C'est à Raju *et al.* (1995) que nous devons d'avoir développé un test de signification pour des indices similaires.

Méthode non compensatoire de Raju* et al. *(1995)

Cette méthode pousse encore plus loin les méthodes précédentes, car elle permet d'obtenir des indices à partir desquels il existe un test de signification. Loin de supposer une distribution uniforme des thêta, elle considère la distribution observée des thêta des n_F sujets du groupe focal (suivant la recommandation de Shepard *et al.*, 1984 ; voir aussi Camilli et Shepard, 1994, p. 67). Ainsi, pour chaque item i, l'indice $NCDIF_i$ (*non compensatory DIF*) est donné par

$$NCDIF_i = \varepsilon_j(d_{ij}^2) = \sigma_{d_{ij}}^2 + \overline{d_{ij}^2} \text{ où } d_{ij} = [P_{iR}(\theta_j) - P_{iF}(\theta_j)] \text{ et où } j = 1, 2, \ldots, n_F.$$

En somme, pour obtenir $NCDIF_i$, il s'agit de calculer les différences bien connues d_{ij} pour les seuls thêta des n_F sujets du groupe focal, de calculer la variance de ces différences et la moyenne des carrés de ces différences.

Le test du khi-carré (avec n_F degrés de liberté) relatif à cette statistique est le

$$\chi^2 = \frac{n_F \times NCDIF_i}{\sigma_{d_{ij}}^2}$$

Il faut noter que l'indice $NCDIF_i$ est non compensatoire, donc non signé, c'est-à-dire que les valeurs de cet indice sont toujours positives. Nous allons décrire plus abondamment cette méthode à la section 8.5.

Méthode DFT de Raju* et al. *(1995)

Mis à part l'indice non compensatoire $NCDIF_i$ proposé par Raju, celui-ci a suggéré une tout autre façon de concevoir le FDI, soit en définissant le fonctionnement différentiel de test (FDT), c'est-à-dire la somme des FDI. Il faut d'abord calculer, pour chacune des n_F valeurs de thêta du groupe focal, le FDT, soit en gros la différence entre les deux courbes caractéristiques de test (CCT), la première CCT constituée à partir des estimés de paramètre d'item du groupe focal et l'autre CCT à partir des estimés de paramètre d'item du groupe de référence. Il faut se souvenir qu'à chaque valeur θ_j, un thêta du

groupe focal, on peut faire correspondre un score vrai. Si les deux CCT sont juxtaposées, l'aire entre les deux CCT sera nulle et il n'y aura pas besoin d'investiguer le FDI de chacun des items. Sinon, il faudra éliminer les items qui contribuent le plus au FDT jusqu'à ce que les CCT coïncident[5].

La logique de cette méthode peut s'exprimer de la façon suivante :

1. Obtenir les scores vrais $V_F(\theta_j)$ et $V_R(\theta_j)$ pour les individus du groupe focal : c'est comme si, pour chaque individu du groupe focal, on obtenait deux scores vrais, le premier, $V_F(\theta_j)$, calculé à partir de la calibration du groupe focal (un premier jeu de paramètres d'items (a_F, b_F, c_F)) et le second, $V_R(\theta_j)$, calculé à partir du second jeu de paramètres d'items associé au groupe de référence (a_R , b_R, c_R). Ainsi, $V_F(\theta_j) = \Sigma_i P_{iF}(\theta_j)$ et $V_R(\theta_j) = \Sigma_i P_{iR}(\theta_j)$, où la somme est prise sur tous les items du test.
2. Obtenir l'indice de fonctionnement différentiel de test FDT

   $$FDT = \varepsilon_j(D_j^2) = \sigma^2_{D_j} + \overline{D_j^2}$$

   où $D_j = V_R(\theta_j) - V_F(\theta_j)$.
3. Obtenir la contribution de chaque item au FDT : l'indice compensatoire $CDIF_i$: Raju a montré que la valeur du FDT était la somme des valeurs de l'indice compensatoire $CDIF_i$ pour chaque item.
   $FDT = \Sigma_i CDIF_i$
   où

   $$CDIF_i = \varepsilon_j(d_{ij}D_j) = \sigma_{d_{ij}D_j} + \overline{d_{ij}\,D_j}$$

   Notons que l'indice $CDIF_i$ étant compensatoire, donc signé, ses valeurs peuvent être positives ou négatives. Pour obtenir une valeur de l'indice $CDIF_i$, il s'agit d'additionner deux termes : le premier terme est la covariance entre d_{ij} et D_j prise sur les n_F sujets qui font partie du groupe focal et le second terme est la moyenne des produits des $d_{ij} = P_{iR}(\theta_j) - P_{iF}(\theta_j)$ et des $D_j = V_R(\theta_j) - V_F(\theta_j)$.
4. Élimination des items dont le $CDIF_i$ est le plus élevé tout en étant > 0.

Si le test du khi-carré du FDT est statistiquement significatif au seuil 0,01 et si la valeur de FDT dépasse une valeur critique C préalablement définie (Raju *et al.*, 1995 proposent C = 0,006), on peut commencer par éliminer, un à un, les items dont l'indice $CDIF_i$ est le plus élevé tout en étant supérieur à 0 puisque la somme des $CDIF_i$ est égale à FDT. La méthode s'arrête lorsque, après avoir enlevé un item, la valeur de FDT < 0,006 ou lorsque le test du khi-carré devient non statistiquement significatif au seuil 0,01.

5. Puisque chaque individu passe un test différent, cette méthode ne peut être opérationnelle dans le cas du testing adaptatif (voir le chapitre 9).

Méthode de la différence de modèles de Thissen* et al. *(1993) telle que décrite par Camilli et Shepard (1994, p. 74-96)

Nous présentons cette méthode, car elle est originale et se démarque des méthodes précédemment proposées dans cette section. Nous sommes cependant d'avis qu'elle risque d'être particulièrement laborieuse dans la plupart des situations qui impliquent la production rapide des résultats d'une étude de biais d'item.

Supposons que nous ayons un test de n items pour lequel nous voulons investiguer la présence d'items comportant un FDI. Disons que c'est l'item 5 que nous voulons étudier. C'est le modèle à trois paramètres qui est choisi. Nous allons tester le FDI de l'item 5 en estimant les paramètres d'items de deux façons, puis retenir, à chaque fois, la dernière valeur observée de la statistique – 2*Loglikelihood*, une indication de l'ajustement analytique du modèle.

La première façon implique l'estimation des paramètres des n items à l'aide, par exemple, de BILOG-3 en utilisant tous les sujets des deux groupes, le groupe de référence (R) et le groupe focal (F). Il s'agit alors de conserver la dernière valeur du –2*Loglikelihood* présente dans le fichier de sortie relatif à la phase 2 de la sortie de BILOG-3. Cette valeur est une indication de l'ajustement analytique du modèle à trois paramètres pour les n items et tous les sujets.

La deuxième façon d'obtenir la valeur de la statistique –2*Loglikelihood* nécessite de recoder cet item et d'en faire deux autres (pseudo-)items :

item 5_R : c'est l'item 5, mais considéré atteint par les sujets du groupe R **et** non atteint par ceux du groupe F ;

item 5_F : c'est encore l'item 5, mais considéré atteint par les sujets du groupe F **et** non atteint par ceux du groupe R.

En utilisant les items 5_F et 5_R plutôt que l'item 5 en version originale, le test a maintenant n + 1 items dont les paramètres doivent être estimés une seconde fois par BILOG-3. La dernière valeur de la statistique –2*Loglikelihood* du fichier de sortie relatif à la phase 2 doit être conservée. Cette valeur est une indication de l'ajustement analytique du modèle à trois paramètres pour les n + 1 items : c'est-à-dire que cette nouvelle valeur est une indication de l'ajustement analytique dans le cas où on suppose, pour l'item 5, un jeu distinct de paramètres pour chaque groupe. Rappelons que la valeur du –2*Loglikelihood* obtenue de l'analyse des n items renvoyait à l'ajustement analytique du modèle qui suppose un seul jeu de paramètres pour l'item 5, les deux groupes étant combinés.

La différence entre ces deux valeurs du –2*Loglikelihood* est distribuée selon une loi du khi-carré avec 3 degrés de liberté (en effet, la deuxième façon d'estimer les paramètres comporte un item de plus, donc trois paramètres de plus à estimer).

L'hypothèse nulle postule que l'ajustement analytique n'est pas différent dans les deux cas. Ainsi, selon cette hypothèse, que l'on estime les paramètres de l'item évalué (item 5) avec les deux groupes ou d'une façon séparée (une pour le groupe R et une autre pour le groupe F) ne change rien à l'ajustement analytique : les deux modèles s'ajustent aussi bien aux données. Si le test du khi-carré est significatif, cela nous amène à rejeter l'hypothèse nulle, donc à considérer que le modèle comportant n + 1 items s'ajuste mieux aux données de façon statistiquement significative, c'est-à-dire que nous devons postuler la présence d'un FDI pour cet item.

La mauvaise nouvelle, c'est qu'il faut refaire la même procédure pour chacun des items du test, une entreprise qui, pour peu que le test soit long, peut devenir extrêmement laborieuse. Par ailleurs, tel que le font remarquer Camilli et Shepard (1994, p. 88), cette méthode de détection de biais permet du même coup d'obtenir des paramètres d'items équilibrés (*equated*), donc de pouvoir calculer l'aire entre les CCI des deux groupes à l'étude à l'aide par exemple de l'indice signé SPD-$\theta = \Sigma_j[P_{iR}(\theta_j) - P_{iF}(\theta_j)] / n_F$. Si le signe de cet indice est positif, c'est que $P_{iR}(\theta_j)$ possède une valeur supérieure à $P_{iF}(\theta_j)$, c'est à dire que les sujets du groupe de référence sont favorisés par rapport à ceux du groupe focal. Si le signe est négatif, les sujets du groupe focal sont les plus favorisés.

8.4. APPLICATION DES MÉTHODES NON BASÉES SUR LA TRI

Plusieurs méthodes d'identification d'un FDI ne s'appuyant pas sur un modèle de la TRI ont été proposées au cours des dernières décennies. Nous en avons décrit un certain nombre à la section 8.2. L'objectif de la présente section est d'élaborer davantage, à partir de la description d'un exemple détaillé, l'interprétation associée aux FDI pour deux de ces méthodes parmi les plus prometteuses, à savoir la méthode de Mantel-Haenszel et la régression logistique.

En 1997, le programme des indicateurs du rendement scolaire (PIRS) administré par le Conseil des ministres de l'Éducation du Canada (CMEC) lançait une enquête visant à évaluer les connaissances mathématiques des élèves canadiens de 13 ans et de 16 ans. Un test de 125 items a été élaboré à cette fin et administré à plus de 25 000 élèves. La stratégie d'administration de ces items est connue sous le nom de testing en deux étapes (*two-staged testing*) : à la première étape, un test de classement formé des 15 premiers items (de difficulté moyenne) est d'abord administré à tous les élèves. Puis, à la deuxième étape, selon le résultat obtenu à ce premier test, les élèves doivent faire les items restants en suivant ces règles : les élèves qui ont réussi 10 items ou moins commencent par l'item le plus facile (I_16), les élèves qui ont réussi entre 11 et 13 items n'ont pas à répondre aux 25 items les plus faciles et commencent à l'item I_41 alors que les élèves qui ont réussi 14 ou 15 items n'ont pas à répondre aux 50 items les plus faciles et débutent à l'item I_66.

Les résultats que nous allons présenter viennent d'une étude que nous avons menée (Bertrand et Laroche, 1999) afin d'identifier les items biaisés en faveur de l'un ou l'autre des deux groupes linguistiques canadiens ; ils ne concernent que les 110 items de ce test administrés à la deuxième étape de cette méthode, soit les items I_16 à I_125.

8.4.1. Méthode de Mantel-Haenszel

C'est Holland et Thayer (1986) qui ont proposé l'emploi d'une statistique, d'abord proposée par Mantel et Haenszel (1959) pour développer une méthode de détection de FDI. Les caractéristiques de base de cette méthode ont été présentées à la section 8.2. Nous avons montré comment obtenir la statistique α_{MH} . Il existe plusieurs façons de déterminer si un item présente un FDI à partir de cette statistique. Par exemple, il est possible de développer une statistique qui se distribue selon une loi du khi-carré avec 1 degré de liberté. Mais il est généralement admis que la méthode que nous allons maintenant décrire et qui est devenu la norme de l'industrie (Roussos *et al.*, 1999) permet une interprétation plus nuancée des items présentant un FDI. Il s'agit de calculer la valeur $\Delta_{MH} = -2{,}35\ ln(\alpha_{MH})$. Cette transformation de la statistique α_{MH} permet d'obtenir une échelle de valeurs centrée à 0 et qui reflète les différences de difficulté des items (Holland et Thayer, 1985). De plus, les valeurs de Δ_{MH} qui sont négatives correspondent aux items qui favorisent le groupe de référence et les valeurs positives de Δ_{MH} correspondent aux items qui favorisent le groupe focal. Si la valeur absolue de Δ_{MH} est supérieure à 1,5 **et** significativement supérieure à 1 (au seuil de signification $\alpha = 0{,}05$), l'item est classé de catégorie C (FDI sévère). Si la valeur absolue de Δ_{MH} est inférieure à 1 **ou** non significativement supérieure à 0 (au seuil de signification $\alpha = 0{,}05$), l'item est classé de catégorie A (FDI négligeable). Dans tous les autres cas de figure, l'item est classé de catégorie B (FDI modéré).

Le tableau 8.8 montre le résultat de l'analyse du fonctionnement différentiel des 110 items du test de mathématique du PIRS. Deux items seulement sont considérés comme présentant un FDI sévère (de catégorie C), soit l'item I_102 et l'item I_25. Puisque, dans les deux cas, la valeur de Δ_{MH} est négative, ces deux items favorisent le groupe de référence, soit celui constitué des élèves canadiens anglophones. Un de ces deux items, I_25, est classé comme un des items les plus faciles du test alors que l'autre, I_102, est classé comme un des items les plus difficiles. Suivant la classification exposée plus haut et au regard du tableau 8.8, nous voyons que six items ont été classés comme présentant un FDI modéré (catégorie B), soit les items I_121, I_56, I_100, I_75, I_66 et I_113 alors que tous les autres items présentaient un FDI négligeable (catégorie A).

TABLEAU 8.8
Valeurs des statistiques α_{MH}, Δ_{MH}, erreur-type associée à la statistique Δ_{MH}, valeurs inférieure et supérieure de l'intervalle de confiance à 95 % et catégorie du FDI pour l'enquête du PIRS de 1997. Deux items présentent un FDI sévère, I_102 et I_25. Six autres items présentent un FDI modéré.

Item	α_{MH}	Δ_{MH}	Erreur-type	Inférieure	Supérieure	Catégorie
I_102	2,205	–1,858	0,090	1,681	2,034	C
I_25	1,946	–1,565	0,081	1,406	1,724	C
I_121	1,721	–1,276	0,191	0,902	1,649	B
I_56	1,719	–1,274	0,083	1,110	1,437	B
I_100	0,585	1,262	0,141	0,986	1,538	B
I_75	0,613	1,151	0,073	1,008	1,294	B
I_66	0,624	1,110	0,094	0,927	1,293	B
I_113	1,589	–1,089	0,108	0,877	1,301	B
Tous les autres items sont classés de catégorie A.						

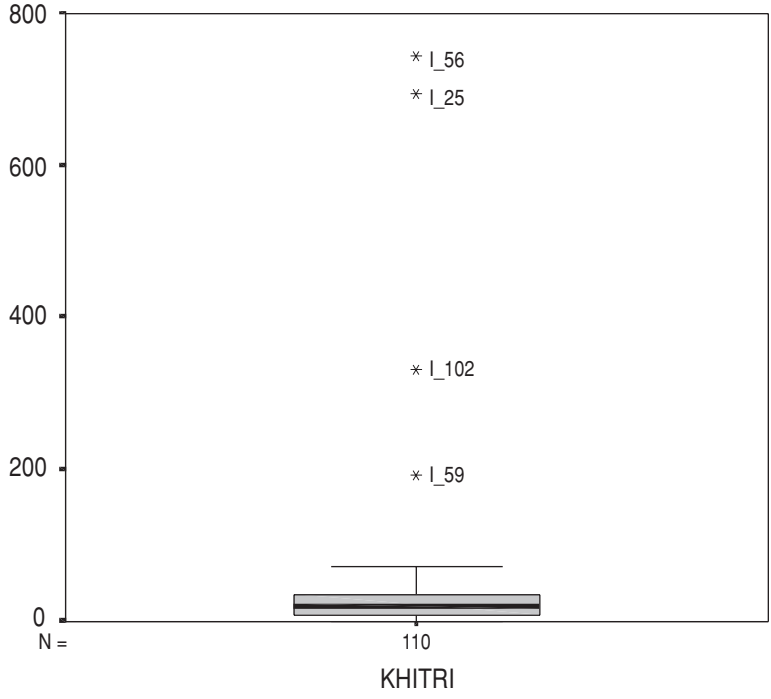
8.4.2. Méthode basée sur la régression logistique

Appliquée aux 110 items de l'enquête de mathématique du PIRS, la méthode de détection du FDI fondée sur la régression logistique décrite à la section 8.2 a permis de produire les résultats présentés à la figure 8.4 où la statistique KHITRI (la valeur du khi-carré) est mise en évidence. Quatre items possèdent un FDI sévère[6] selon cette méthode, soit les items I_56, I_25, I_102 et I_59. On se souviendra que seuls les items I_102 et I_25 avaient été reconnus FDI sévères selon la méthode de Mantel-Haenszel qui, rappelons-le, ne permet pas de détecter une composante non uniforme de FDI aussi bien que le fait la régression logistique : or, comme on peut le constater à la figure 8.2, l'item I_56, détecté FDI par la régression logistique et non par la méthode de Mantel-Haenszel, comporte une forte composante non uniforme.

Selon la figure 8.5, ce sont les items I_76, I_46, I_62, I_92, I_60, I_75, I_95 et I_49 qui présentent un FDI modéré selon la méthode fondée sur la régression logistique. On se rappellera que les items de FDI modéré selon la méthode de Mantel-Haenszel étaient I_121, I_56, I_100, I_75, I_66 et I_113. Il faut tout de même noter un certain écart entre les résultats obtenus selon ces deux méthodes.

6. Utilisant la classification (plutôt conservatrice) de Gierl *et al.* (1999), seulement 2 items sont considérés FDI, l'item I_25 un FDI sévère et l'item I_56 un FDI modéré.

FIGURE 8.4
Diagramme en boîte et moustaches indiquant que quatre des 110 items de l'enquête PIRS de 1997, soit les items I_56, I_25, I_102 et I_59, représentés par le symbole (*), sont considérés comme présentant un FDI sévère.

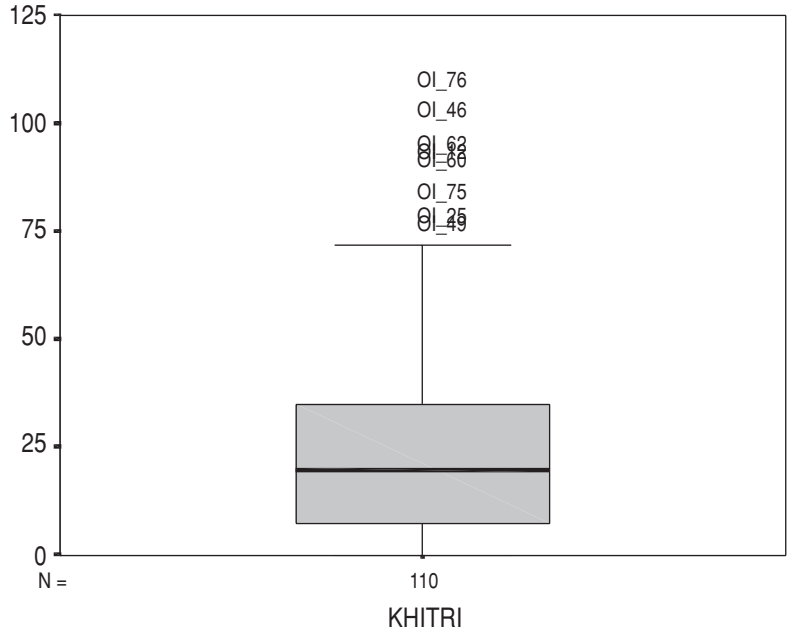


8.5. APPLICATION DES MÉTHODES TRI DE DÉTECTION DE FDI

8.5.1. La méthode non compensatoire NCDIF$_i$ de Raju

Rappelons que la méthode non compensatoire de Raju *et al.* (1995) vise à trouver les items pour lesquels l'aire entre les CCI des deux groupes linguistiques est exagérément grande. Selon cette méthode, un item i sera jugé comme présentant un FDI si la valeur de NCDIF$_i$ est supérieure à 0,006 **et** si le test du khi-carré qui y est associé donne un verdict statistiquement significatif (au seuil critique de 0,01). Suivant cette méthode, sept des 110 items de l'enquête du projet PIRS de 1997 ont été reconnus comme présentant un FDI. Il s'agit des items I_102, I_25, I_56, I_91, I_59, I_121 et I_76. Les six premiers items doivent être considérés comme présentant un FDI sévère alors que le septième est modéré. La figure 8.6 montre quels items ont obtenu les plus grandes valeurs en regard de la statistique NCDIF$_i$.

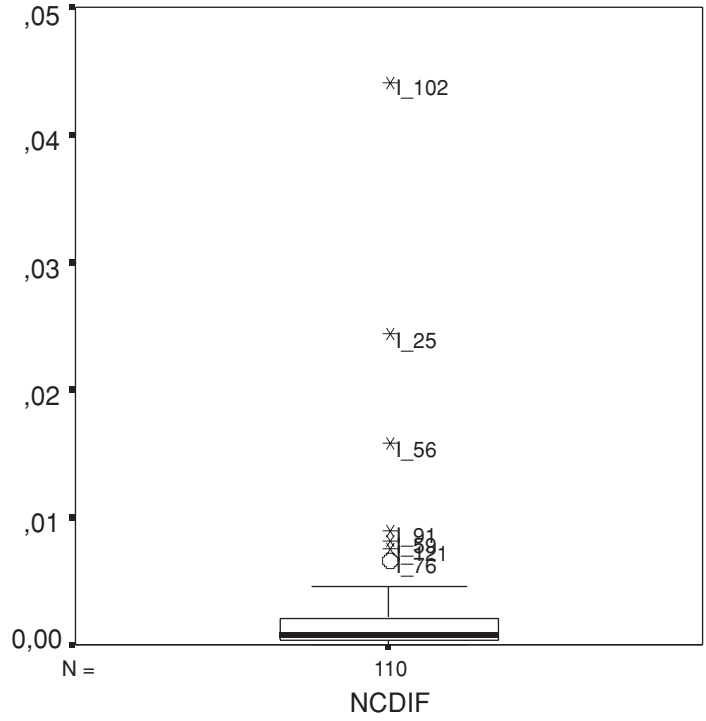
FIGURE 8.5
Diagramme en boîte et moustaches indiquant que huit[7] des 110 items de l'enquête PIRS de 1997, soit les items I_76, I_46, I_62, I_92, I_60, I_75, I_95 et I_49 repérés par le symbole (o), sont considérés comme présentant un FDI modéré.



Quatre de ces sept items, à savoir I_102, I_25, I_56 et I_121, avaient été repérés comme présentant un FDI sévère ou modéré à l'aide de la méthode de Mantel-Haenszel. Par contre, cinq des sept items considérés FDI par la méthode non compensatoire de Raju ont été reconnus comme étant FDI sévère ou modéré par la méthode s'appuyant sur la régression logistique : seuls les items I_91 et I_121 n'ont pu être détectés par cette dernière méthode. Il est tout de même remarquable que la méthode de Raju, en apparence si distincte de celle s'appuyant sur la régression logistique, donne des résultats somme toute assez convergents.

7. Puisqu'il n'est pas possible de voir clairement les numéros des items présentant un FDI à partir de la figure elle-même, ces huit valeurs aberrantes ont été identifiées en utilisant la commande *Explore* de SPSS associée à la production du diagramme en boîte et moustaches.

FIGURE 8.6
Diagramme en boîte et moustaches indiquant les valeurs de l'indice compensatoire $NCDIF_i$ de Raju pour les 110 items de l'enquête du PIRS en 1997



8.5.2. La méthode des différences de modèles de Thissen

La méthode de Thissen, dont nous avons développé les éléments de base à la section 8.4, implique la comparaison de deux modèles : le premier suppose l'estimation des paramètres de l'item étudié en utilisant tous les sujets des deux groupes : le groupe des élèves anglophones et le groupe des élèves francophones. Le second suppose que les paramètres de l'item étudié sont mieux estimés en considérant une estimation différente pour chaque groupe d'élèves. La statistique de test utilisée, que nous appelons LOGDIF, soit la différence des deux –2*Loglikelihood*, suit une loi du khi-carré avec 3 degrés de liberté pour les items s'ajustant au modèle de réponse aux items à trois paramètres. La procédure originale implique de reconnaître un item comme présentant un FDI si la valeur de la statistique de test mène à un verdict statistiquement significatif (au niveau de signification $\alpha = 0{,}01$). Notons qu'ici, la valeur du

khi-carré critique (au seuil $\alpha = 0{,}01$) avec 3 degrés de liberté est de 11,341. Si on l'appliquait à notre cas, compte tenu des très grands effectifs, environ 80 % des items présenteraient un FDI !

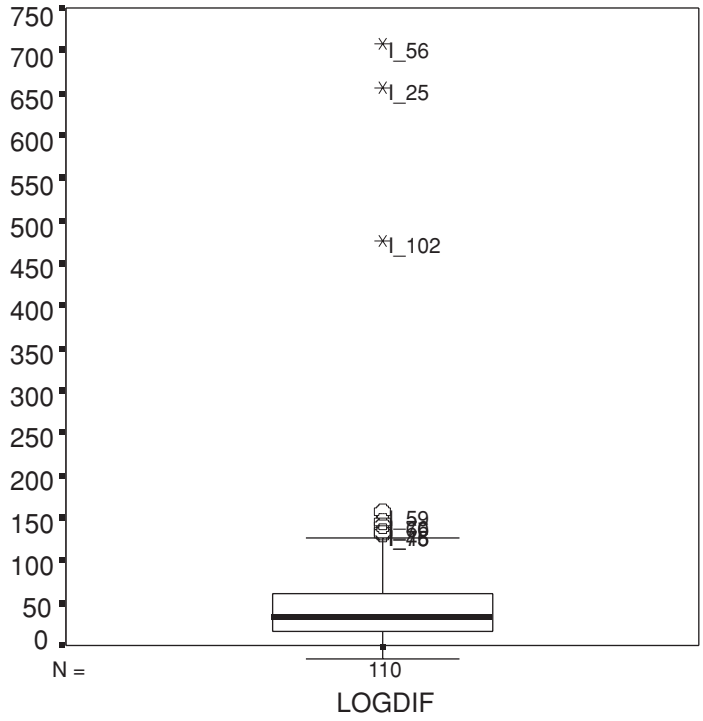
Il appert que cette méthode va trop souvent mener à des verdicts statistiquement significatifs puisque la statistique du khi-carré est très sensible à la taille de l'échantillon. En conséquence, à l'instar de la méthode basée sur la régression logistique, nous proposons de constituer trois catégories de FDI après avoir interprété les valeurs exagérément élevées du diagramme en boîte et moustaches. Nous dirons donc qu'un FDI est sévère ou de catégorie C pour les items qui mènent à une valeur extrême d'après le diagramme en boîte et moustaches ; le FDI sera modéré ou de catégorie B pour les items qui mènent à une valeur aberrante qui n'est pas extrême. Dans tous les autres cas, le FDI sera considéré négligeable ou de catégorie A. Dans notre cas, selon la figure 8.7, trois items révèlent un FDI sévère, soit les items I_56, I_25 et I_102. Ces trois items sont aussi reconnus comme présentant un FDI sévère par la méthode NCDIF de Raju et la méthode s'appuyant sur la régression logistique. On se souviendra que les items I_25 et I_102 étaient reconnus comme FDI sévères par la méthode de Mantel-Haenszel. Selon la méthode de Thissen, cinq items présentent un FDI modéré, soit les items I_59, I_76, I_60, I_75 et I_46. Ces cinq items sont considérés comme présentant un FDI modéré ou sévère selon la méthode de la régression logistique. Les items I_59 et I_76 présentent également un FDI selon la méthode NCDIF. Parmi ces cinq items, seul l'item I_75 est considéré FDI modéré par la méthode de Mantel-Haenszel.

8.5.3. La méthode de Shepard, Camilli et Williams (1984)

Tel qu'indiqué à la section 8.2, la méthode de Shepard, Camilli et Williams (1984), basée sur le calcul de l'aire entre les CCI, tient compte de la distribution des sujets du groupe focal pour calculer les indices SPD-θ et UPD-θ. Nous utiliserons l'indice SPD-θ pour interpréter le sens du FDI puisqu'il s'agit d'un indice signé. Nous aurons recours à l'indice non signé UPD-θ pour quantifier l'ampleur du FDI. À défaut de test de signification, nous proposons d'interpréter l'indice UPD-θ de la façon suivante : seront considérés FDI les items dont la valeur de UPD-θ sera supérieure à 0,10, ce qui correspond à une différence moyenne de probabilité de réussite entre les deux groupes (à habileté égale) de 0,10. Si la valeur de l'indice UPD-θ est supérieure à 0,10 et qu'elle est reconnue extrême sur un diagramme en boîte et moustaches, nous dirons que le FDI est sévère ou de catégorie C. Le FDI sera considéré modéré ou de catégorie B si la valeur de l'indice UPD-θ est supérieure à 0,10 ou si elle est reconnue aberrante sans être extrême. Dans tous les autres cas, le FDI sera considéré négligeable et de catégorie A. Cette méthode de détection de FDI est intéressante dans la mesure où elle comporte un support visuel non négligeable en plus de mener à une interprétation qui correspond à une certaine

intuition : la valeur de ces indices constitue en effet une différence moyenne de probabilité de réussite d'un item (calculée à chaque niveau d'habileté) entre les sujets du groupe de référence et les sujets du groupe focal.

FIGURE 8.7
Diagramme en boîte et moustaches indiquant les valeurs aberrantes (O) et extrêmes (*) de l'indice LOGDIF de Thissen pour les 110 items de l'enquête du PIRS en 1997.

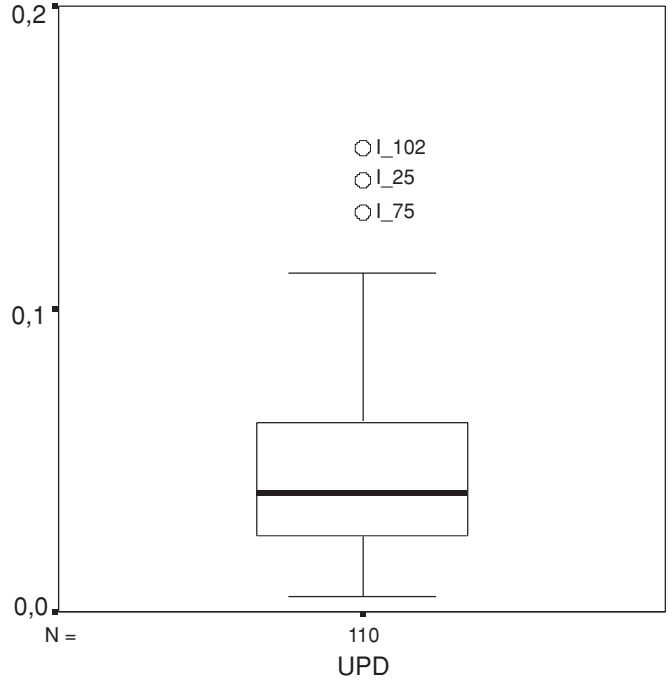


En appliquant ces règles, nous avons reconnu dix items présentant un FDI, nommément les items I_102, I_25, I_56, I_75, I_72, I_60, I_49, I_62, I_35 et I_63. Tous ces FDI sont considérés modérés puisque, comme on peut le constater au tableau 8.9, les valeurs de l'indice UPD sont supérieures à 0,10, mais, comme on le voit à la figure 8.8, aucune valeur de l'indice UPD n'est considérée extrême. Seulement trois valeurs sont aberrantes. La valeur de l'indice SPD étant négative pour la majorité des items, il appert que ces items, pour la plupart, favorisent les sujets du groupe focal, les francophones.

TABLEAU 8.9
Valeurs de l'indice SPD et de l'indice UPD pour les items du PIRS

Numéro d'item	SPD	UPD
I_25	0,11	0,14
I_35	–0,09	0,11
I_49	–0,10	0,10
I_56	0,08	0,10
I_60	–0,10	0,11
I_62	–0,10	0,11
I_63	–0,09	0,11
I_72	–0,10	0,11
I_75	–0,11	0,13
I_102	0,13	0,15

FIGURE 8.8
Diagramme en boîte et moustaches indiquant les valeurs aberrantes (O) de l'indice UPD de Shepard, Camilli et Williams (1984) pour les 110 items de l'enquête du PIRS en 1997.



La figure 8.9 montre les CCI se référant à l'item I_62 (favorisant le groupe focal, les francophones) alors que la figure 8.10 se rapporte à l'item I_102, qui favorise le groupe de référence, les anglophones.

FIGURE 8.9
Courbes caractéristiques de l'item I_62 pour les anglophones et les francophones. La valeur de SPD est négative, soit –0,10, signe que cet item favorise les sujets du groupe focal, à savoir les francophones.

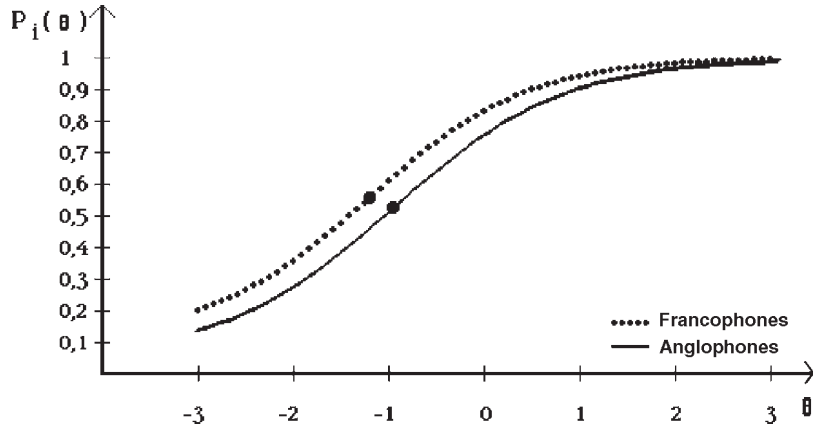
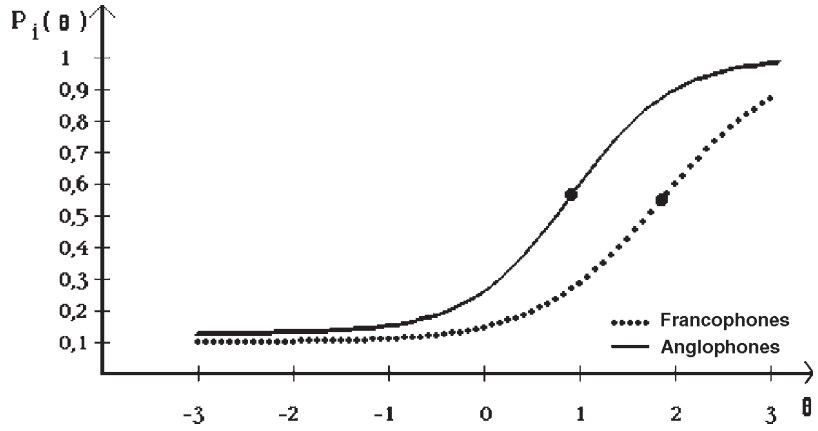


FIGURE 8.10
Courbes caractéristiques de l'item I_102 pour les anglophones et les francophones. La valeur de SPD est positive, soit 0,13, signe que cet item favorise les sujets du groupe de référence, à savoir les anglophones.



8.6. SYNTHÈSE DES RÉSULTATS

Le tableau 8.10 présente la synthèse des résultats que nous avons produits à la suite de l'utilisation des cinq méthodes de détection de FDI, soit la méthode de Mantel-Haenszel, la régression logistique, la méthode non compensatoire (NCDIF) de Raju, la méthode de différence de modèles de Thissen et la méthode de l'aire (SPD, UPD) de Shepard, Camilli et Williams. Comme le révèle ce tableau, 20 des 110 items ont été reconnus comme présentant un FDI modéré ou sévère par au moins une des cinq méthodes. Sur ces 20 items, trois seulement ont fait l'unanimité et ont donc été reconnus FDI modérés ou sévères par toutes les méthodes : il s'agit des items I_102 et I_25 et I_56. Ces trois items favorisent les anglophones puisque la valeur de l'indice SPD est positive. Les items I_102 et I_25, en effet, ont été reconnus comme présentant un FDI sévère par quatre méthodes. L'item I_56, pour sa part, a été reconnu FDI sévère par trois méthodes et FDI modéré par les deux autres méthodes. Le FDI de l'item I_75 a été identifié par quatre des cinq méthodes. Seule la méthode de l'indice non compensatoire de Raju n'a pu détecter l'item I_75. Trois autres items, I_59, I_60 et I_76 ont été reconnus FDI par trois des cinq méthodes. Quatre autres items, I_46, I_49, I_62 et I_121 ont été reconnus FDI par seulement deux des cinq méthodes. Tous les autres items de ce tableau ont été reconnus FDI par une seule des cinq méthodes. Notons que tous les items détectés FDI par la méthode de différence de modèles de Thissen l'ont aussi été par au moins une autre et parfois par plusieurs autres méthodes.

Il est instructif de noter que parmi les items détectés FDI par une seule méthode, c'est la méthode de Mantel-Haenszel (MH) qui remporte la palme avec trois items alors que la régression logistique et la méthode de Shepard *et al.* (UPD) en révèlent deux chacune, la méthode non compensatoire (NCDIF) un seul item et, enfin, aucun item dans le cas de la méthode de différence de modèles de Thissen. Cette observation est d'autant plus étonnante que la méthode de Mantel-Haenszel est souvent perçue comme la norme de l'industrie.

Le tableau 8.11 fait état des taux d'entente entre les méthodes prises deux à deux. Chaque valeur du tableau indique le rapport entre le nombre d'items reconnus FDI conjointement par deux méthodes et le nombre d'items reconnus FDI par l'une ou l'autre de ces deux méthodes. Par exemple, comme on peut le constater au tableau 8.10, quatre items ont été identifiés comme présentant un FDI conjointement par la méthode de Mantel-Haenszel et la méthode non compensatoire (NCDIF) de Raju. Or, puisque onze items ont été identifiés FDI par l'une ou l'autre de ces deux méthodes, la valeur du taux d'entente entre ces deux méthodes est de 4/11 ou 0,3636. La dernière ligne de ce tableau indique la moyenne des taux d'entente associés à l'une ou l'autre des méthodes. Ce taux moyen peut être considéré ici comme une indication globale de la capacité, d'une méthode donnée à détecter des FDI que les autres méthodes ont aussi détectés.

TABLEAU 8.10
Items reconnus FDI sévères (S) ou modérés (M) par l'une ou l'autre de cinq méthodes : Mantel-Haenszel (MH), régression logistique (Reg.log), méthode non compensatoire de Raju (NCDIF), méthode de différence de modèles de Thissen (Diff.mod) et méthode de l'aire (UPD) de Shepard, Camilli et Williams.

Item	MH	Reg.log	NCDIF	Diff.mod	UPD
I_102	S	S	S	S	M
I_25	S	S	S	S	M
I_121	M	–	S	–	–
I_56	M	S	S	S	M
I_100	M	–	–	–	–
I_75	M	M	–	M	M
I_66	M	–	–	–	–
I_113	M	–	–	–	–
I_72	–	–	–	–	M
I_59	–	S	S	M	–
I_76	–	M	M	M	–
I_46	–	M	–	M	–
I_92	–	M	–	–	–
I_60	–	M	–	M	M
I_95	–	M	–	–	–
I_49	–	M	–	–	M
I_91	–	–	S	–	–
I_62	–	M	–	–	M
I_35	–	–	–	–	M
I_63	–	–	–	–	M

L'observation de ces valeurs nous suggère les constatations suivantes. Malgré le fait que l'indice non compensatoire de Raju (NCDIF) et l'indice UPD de Shepard *et al.* soient basés sur une idée similaire, à savoir quantifier la part d'aire entre la CCI du groupe de référence et la CCI du groupe focal, ils constituent les deux méthodes qui s'entendent le moins avec, selon le tableau 8.11, un taux d'entente de seulement 0,2142. Les deux méthodes qui ont généré le taux d'entente le plus élevé sont la régression logistique et la différence de modèles avec, selon le tableau 8.11, une valeur de 0,6667. Pourtant, les bases théoriques de ces deux méthodes sont relativement différentes : en outre, contrairement à la méthode de différence de modèles, la méthode de la régression logistique n'est pas fondée sur les modèles de réponses aux items[8]. Toujours suivant le tableau 8.11, la méthode qui obtient un taux moyen d'entente le plus faible est la méthode de Mantel-Haenszel (0,3082). Par contre, la méthode de la différence de modèles, avec 0,4712, donne le taux moyen d'entente le plus élevé.

8. Il faut toutefois noter que la variable de contrôle utilisée dans le cas de la régression logistique est le score thêta.

Tableau 8.11
Taux d'entente entre chaque paire de méthodes détection de biais.

	MH	Reg.log	NCDIF	Diff.mod	UPD
MH	–	0,2500	0,3636	0,3333	0,2857
Reg.log	0,2500	–	0,3571	0,6667	0,4667
NCDIF	0,3636	0,3571	–	0,5	0,2142
Diff.mod	0,3333	0,6667	0,5	–	0,3846
UPD	0,2857	0,4667	0,2142	0,3846	–
Moyenne	**0,3082**	**0,4351**	**0,3587**	**0,4712**	**0,3378**

8.7. CONSTATS, REMARQUES ET LIMITES DES MÉTHODES PROPOSÉES

Bien que les méthodes de détection de biais puissent théoriquement se diviser en deux catégories, celles basées sur un critère externe et celles basées sur un critère interne, nous nous en sommes tenus aux méthodes qui renvoient à un critère interne (généralement le score total au test ou l'estimé d'habileté thêta), de loin celles qui ont reçu le plus d'attention au cours des dernières années. Les méthodes qui requièrent un critère externe exigent que ce critère soit aussi exempt de biais et somme toute valide. Mais comme on l'a vu, nous ne pouvons garantir que les interprétations faites à partir des scores à un test soient toujours valides. En outre, il faut administrer, corriger et interpréter ce critère, augmentant d'autant les ressources nécessaires et le temps requis pour le testing. Les principaux désavantages des méthodes axées sur un critère interne touchent la contamination du critère par les items biaisés et l'impossibilité de détecter un biais présent dans tous les items du test. Concernant la contamination du critère, il est cependant possible, comme l'ont montré Holland et Thayer (1988) par le passé, d'épurer le critère des items biaisés en adoptant une méthode itérative, bien que celle-ci soit beaucoup plus dispendieuse et difficile à gérer. Selon l'approche conservatrice, en effet, il faut réunir un comité pour étudier les FDI avant de les éliminer, ajoutant ainsi une étape fâcheuse et parfois trop longue à une méthode déjà ardue. Quant à la possibilité de détecter un biais présent dans tous les items, il s'agit là d'une limite bien connue des méthodes à critère interne. Encore que la procédure menant à la construction du test, pour peu qu'elle ait été rigoureuse, aurait dû permettre d'éviter ce genre de biais. À moins que ce ne soient les procédures permettant la détection de biais de méthode ou de biais de concept qui aient déjà détecté ce genre d'anomalie qui, présente dans tout le test, devrait tout de même être un peu visible. Supposons par exemple qu'un test ne contienne que des items de pêche et de chasse, défavorisant les filles en général. Les méthodes de détection de biais d'item ne permettraient probablement pas de mettre ce biais en relief. Cependant, si une méthode de biais de concept a été mise en place et

si, par conséquent, une analyse factorielle est effectuée, il y a bien des chances que les facteurs ne soient pas tout à fait les mêmes d'un groupe (les gars) à l'autre (les filles) : il pourrait même y avoir un facteur (significatif) de plus pour les filles. D'un autre côté, supposons qu'un test formé d'items à choix de réponses soit administré à deux groupes, les Québécois et les Maliens. Si d'aventure les Maliens ne sont pas habitués à ce genre d'item, un biais de méthode est prévisible, mais il ne sera pas détecté par les procédures de biais d'item puisqu'il affecte tous les items, donc aussi le test, le critère interne. Il faut juste supposer que, dans ce cas, la procédure menant à la détection de biais de méthode aura pu identifier et corriger ce genre de situation fâcheuse.

Comme autre limite, mentionnons que la méthode de Mantel-Haenszel est mieux adaptée à la détection de FDI uniformes que de FDI non uniformes. De plus, dans le cas de la détection de FDI uniformes pour des items s'ajustant à un modèle à trois paramètres, plus l'item analysé est difficile, plus la statistique de Mantel-Haenszel a tendance à diminuer artificiellement, causant ainsi un biais d'estimation non désiré et un manque de puissance à détecter un FDI (Roussos, Schnipke et Pashley, 1999).

Ces limites montrent pourquoi il est possible de trouver des différences entre les méthodes présentées ici et qu'il vaut mieux combiner au moins deux méthodes (Gierl *et al.*, 1999) si les ressources sont disponibles. De plus, en présence d'items difficiles, il vaut mieux utiliser une autre méthode que celle de Mantel-Haenszel, pourtant la norme de l'industrie. C'est une remarque qui devrait toucher plus particulièrement les personnes aux prises avec le développement de tests de sélection, farcis d'items difficiles, et les tests adaptatifs qui doivent eux aussi contenir plusieurs items difficiles.

La régression logistique est une méthode intéressante, mais elle ne permet de découvrir le sens du biais que si on emploie la procédure en trois étapes. Dans cette éventualité, c'est le signe du coefficient B de l'étape 2 associé au groupe qui permet de déterminer le sens du FDI.

Nous concluons ce chapitre par les constats suivants :

1. Toutes les méthodes employées ici ont détecté les trois items présentant les FDI les plus manifestes : I_102, I_25 et I_56.
2. Une procédure de détection de biais d'item ne s'appuyant que sur un test de signification est difficilement défendable dans la mesure où les items détectés FDI dépendront, en partie du moins, de la taille des échantillons du groupe de référence et du groupe focal. Il nous semble donc impératif de jumeler à ce genre de méthode, une valeur critique de type « grandeur de l'effet » (*effect size*), au-delà de laquelle un item sera déclaré FDI.
3. Le FDI est un concept relatif : il n'y a pas et il n'y aura jamais d'interprétation univoque et absolue d'un FDI. En ce sens, déclarer qu'un item présente un FDI comporte une certaine part d'arbitraire, de risque, lorsqu'il est question de fixer un seuil par exemple. Il faut

donc voir toute entreprise de détection de FDI de manière plutôt relative. L'idée est plutôt d'identifier les items qui comportent le plus de FDI de manière à ce qu'ils soient examinés par un comité d'experts qui jugeront s'il y a lieu de considérer ces items comme étant biaisés ou non.

4. Un item peut être considéré comme présentant un FDI dans un contexte donné, mais pas dans un autre : un item de chasse, par exemple, portant sur la distance parcourue par une flèche et présentant un FDI contre les filles risque de ne pas être considéré biaisé s'il fait partie d'un test visant à mesurer les connaissances sportives, mais serait, selon toute éventualité, considéré biaisé s'il faisait partie d'un test de résolution de problèmes mathématiques. Tout dépend de l'interprétation que l'on fait des scores ou des résultats émanant d'un test : nous touchons ici à la validité.
5. Il faut cibler les groupes sur lesquels une étude de FDI doit être effectuée puisque, à la limite, tous les items pourraient être considérés comme présentant un FDI contre l'un ou l'autre des très multiples sous-groupes possibles.
6. La procédure FDT de Raju décrite à la section 8.3, qui repose sur la différence entre les courbes caractéristiques de test, donne des résultats diamétralement opposés aux autres méthodes et notamment à la procédure non compensatoire de Raju. Si les items I_100, I_93, I_85, I_94 et I_70 sont considérés comme présentant un FDI selon la méthode FDT, un seul de ces items, I_100, a été déclaré FDI dans le tableau 8.10 et, en plus, une seule méthode, celle de Mantel-Haenszel, l'a détecté.

CHAPITRE 9

Le testing adaptatif

Gilles Raîche, professeur
Université du Québec à Montréal

Selon les habitudes développées au XX^{e} siècle pour évaluer les apprentissages réalisés par un étudiant, faire un diagnostic de ses problèmes d'apprentissage ou le classer à l'intérieur d'un groupe pour qu'il puisse recevoir un enseignement approprié, on administre très souvent un test papier-crayon. Il s'agit d'un test où l'étudiant inscrit ses réponses, choisies ou construites, sur une feuille de papier à l'aide d'un crayon. Le test vise principalement à estimer le niveau d'habileté de celui-ci dans un domaine de connaissances spécifique pour permettre, par la suite, de porter un jugement sur ses apprentissages ou connaissances et de prendre une décision quant à une sanction, un classement ou un diagnostic.

Généralement, le niveau d'habileté d'intérêt est d'ordre cognitif ; connaissances en mathématique, en français, etc. Il peut toutefois être d'ordre affectif ; le niveau d'habileté est alors en lien avec une attitude. Il peut aussi être d'ordre psychomoteur et le niveau d'habileté vise ainsi un comportement moteur. Dans tous ces cas, le test ne permet d'obtenir qu'un estimateur de ce niveau d'habileté : il n'est qu'une occasion pour l'étudiant de manifester son habileté.

Au XX^e siècle apparaît un changement majeur dans l'utilisation des tests, changement qui s'intensifie après la Deuxième Guerre mondiale ; leur administration, surtout individuelle au départ, devient de plus en plus appliquée à de grands groupes (*mass administration*). Conséquemment, pour accélérer et faciliter la correction, les réponses à ces tests sont habituellement choisies plutôt que construites et, d'un étudiant à un autre, le même nombre de questions et les mêmes questions sont administrées. De plus, le temps maximal qui est imparti pour répondre au test est le même pour tous. Ce type de test est alors dit fixe et invariable. Il faut tout de même noter que ce ne sont pas seulement les tests composés d'items à réponses choisies qui peuvent être fixes et invariables ; les tests à réponses construites peuvent aussi l'être. Toutefois, nous ne nous intéressons ici qu'à un seul type de test, soit celui composé d'items à réponses choisies.

Plusieurs problèmes de précision de l'estimateur du niveau d'habileté et plusieurs limites à l'administration d'un tel test papier-crayon fixe et invariable existent cependant. Nous décrivons ici ces problèmes ainsi que ces limites pour ensuite présenter une proposition de solution à ceux-ci, soit le testing adaptatif.

9.1. PROBLÈMES DE PRÉCISION ET LIMITES À L'ADMINISTRATION DES TESTS PAPIER-CRAYON

Dans un test papier-crayon fixe et invariable, le niveau de difficulté des items auxquels doit répondre l'étudiant ne correspond pas toujours au niveau d'habileté de ce dernier. L'étudiant peut faire face à certains items trop faciles ou trop difficiles pour lui. Dans le premier cas, aucun défi n'est relevé, et l'étudiant peut avoir l'impression de perdre son temps. Cela peut alors se traduire par des réponses erronées de la part de l'étudiant parce que celui-ci ne se concentre pas sur la tâche, qui lui semble sans intérêt. Dans le second cas, lorsque les items sont trop difficiles, l'étudiant peut se décourager au point de ne pas terminer le test. Que les items soient trop faciles ou trop difficiles, un manque de motivation de la part de l'étudiant peut alors se produire avec un impact potentiel sur la précision de l'estimateur du niveau d'habileté obtenu.

De plus, pour permettre l'administration d'un test papier-crayon fixe et invariable à des étudiants dont le niveau d'habileté varie beaucoup, ce test doit être constitué d'items dont le niveau de difficulté est très varié. Des items faciles ne sont donc pas nécessairement administrés à des étudiants dont le niveau d'habileté est faible, tandis que des items difficiles ne sont pas forcément administrés aux élèves dont le niveau d'habileté est plus élevé. Pour cette raison surtout, les tests papier-crayon fixes et invariables ne permettent généralement pas d'obtenir un estimateur précis du niveau d'habileté, dans les points extrêmes de l'échelle d'habileté, où les niveaux d'habileté sont très faibles ou très élevés. Weiss (1982, p. 474) souligne ainsi que plus ce type de

test permet d'estimer une large étendue de niveaux d'habileté, donc plus il est constitué d'items dont le niveau de difficulté varie de très facile à très difficile, moins la précision du test est élevée. À l'inverse, lorsque le test est composé d'items dont le niveau de difficulté varie peu, donc lorsqu'ils sont destinés à estimer un niveau d'habileté spécifique, une plus grande précision de l'estimateur du niveau d'habileté est obtenue lorsque les items administrés ne sont ni trop faciles, ni trop difficiles pour l'étudiant. C'est ce que souligne Weiss (1982, p. 474) lorsqu'il met en relief le dilemme entre la largeur de bande et la fiabilité du test.

Lors de l'administration d'un test papier-crayon fixe et invariable, il est à noter que l'étudiant ne peut recevoir immédiatement son résultat au test ; il doit attendre que celui-ci soit corrigé. Ainsi, pour les tests à fonction diagnostique ou formative, qui nécessitent le plus souvent une rétroaction rapide, les délais de correction constituent une limite importante à leur utilisation.

Une autre limite à l'administration d'un test papier-crayon fixe et invariable est que la correction n'est pas totalement automatisée ; il y a nécessité d'une intervention humaine dans la correction du test, soit par une correction manuelle, soit par la manipulation de feuilles réponses destinées à être traitées par un lecteur optique. Il serait possible de corriger le test plus rapidement en éliminant complètement cette étape ; il y aurait ainsi une diminution des coûts de correction et une réduction potentielle du nombre d'erreurs de correction. Avec ce type de test, lorsque la correction est manuelle, Laurier (1993b, p. 228) a d'ailleurs remarqué jusqu'à 10 % d'erreurs dans le calcul de l'estimateur du niveau d'habileté.

De plus, un test papier-crayon fixe et invariable ne peut être adapté à l'étudiant auquel il est administré puisque tous les étudiants reçoivent la même version du test. Il est ainsi impossible de modifier le nombre d'items administrés, ou les items eux-mêmes, en fonction du niveau d'habileté de l'étudiant et de la précision obtenue de l'estimateur de son niveau d'habileté. Le test n'est donc pas personnalisé.

Le format des items est habituellement assez limité. Ainsi, les séquences vidéo et les éléments auditifs sont peu employés et, lorsque c'est le cas, dans des conditions souvent inadéquates. Par exemple, les tests de classement en langue seconde comportent souvent une section visant à estimer le niveau d'habileté en compréhension orale. À cette fin, l'étudiant doit écouter un texte enregistré sur cassette et par la suite répondre à des items destinés à estimer son niveau d'habileté en compréhension auditive.

Enfin, des problèmes de sécurité peuvent se poser lors de l'administration d'un test. Ainsi, il peut y avoir plagiat au moment même de l'administration du test. Ou encore, la confidentialité des réponses peut être affectée par la circulation d'une copie du test, de la feuille réponse ou de la grille de correction.

9.2. DÉROULEMENT D'UN TEST ADAPTATIF

Pour remédier à ces problèmes de précision de l'estimateur du niveau d'habileté et à ces limites d'administration, des chercheurs ont proposé l'utilisation du testing adaptatif (TA). Le testing adaptatif est une forme de testing sur mesure (*tailored testing*) spécifiquement adaptée à la personne à qui on administre le test. Le testing adaptatif a connu de multiples transformations depuis son introduction ; test à deux étapes, tests à niveaux flexibles, test pyramidal ou test stratifié. Ces diverses formes de test adaptatif étant abordées ailleurs par Auger (1989, p. 51-71), Laurier (1993b, p. 37-46) et Raîche (2000, p. 18-36), nous jugeons plus approprié de ne traiter que de la forme de test adaptatif la plus prometteuse, soit celle qui se base sur les propositions modernes de modélisation des réponses aux items. En fait, l'utilisation du testing adaptatif a été facilitée principalement depuis l'introduction de propositions de modélisation des réponses aux items différentes de celles proposées dans le contexte de la théorie classique des tests. Il s'agit de propositions issues de la théorie des réponses aux items. L'accessibilité à des micro-ordinateurs de plus en plus puissants et offerts à des prix abordables a permis l'application de ces nouvelles propositions de modélisation des réponses aux items.

Plusieurs programmes de testing à grande échelle (*large-scale testing*) utilisent des versions adaptatives par ordinateur de leurs tests. C'est le cas, notamment, de plusieurs tests développés par l'Educational Testing Service (ETS) tels que le SAT (*Scholastic Asssessment Test*), le GRE (*Graduate Record Examination*), le PRAXIS (successeur du NTE pour l'évaluation des enseignants) et le NCLEX (examen du National Council of State Boards of Nursing). D'autres organismes emboîtent le pas : la Psychological Corporation, le College Board, l'American College of Testing, la Société américaine des pathologistes, l'American Board of Internal Medicine, le ministère de la Défense des États-Unis, etc. Même le concepteur de logiciels Microsoft utilise maintenant des versions adaptatives de ses tests de certification (Microsoft, 2000).

Au Québec, toutefois, peu de versions adaptatives de tests ont été élaborées et, dans plusieurs cas, il s'agit de travaux de recherche plutôt que d'applications à un programme de testing à grande échelle. Le programme CAPT (*Computerized Adaptive Placement Test*), développé par Laurier (1993a, 1993b, 1993c, 1998, 1999a, 1999b) et visant le classement en français langue seconde au niveau post-secondaire, est un exemple d'application, tandis que les travaux d'Auger (1989 ; Auger et Séguin, 1992) sur le testing adaptatif de maîtrise en éducation économique au secondaire, de Laurier en révision de texte (1996) et de Raîche (1994, 2000, 2001a, 2001b) et Raîche et Blais (2002a, 2002b, 2002c) sur la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing sont des exemples de travail de recherche.

Le testing adaptatif, principalement lorsqu'il est administré par ordinateur, offre plusieurs avantages par rapport aux tests papier-crayon fixes et invariables. L'une des caractéristiques les plus importantes du testing adaptatif

est de permettre l'administration d'items dont le niveau de difficulté correspond au niveau d'habileté de la personne passant le test. À l'opposé des tests papier-crayon fixes et invariables, où tous les items du test sont administrés sans égard pour le niveau d'habileté de la personne, le testing adaptatif permet l'administration de tests sur mesure, de façon à ce que le niveau de difficulté des items à ces tests ne soit ni trop difficile, ni trop facile. Le nombre d'items administrés, tout comme la durée de l'administration, sont ainsi réduits par rapport à une version papier-crayon du test, sans que la précision de l'estimateur du niveau d'habileté diminue pour autant. Le testing adaptatif devrait d'ailleurs permettre d'obtenir un estimateur plus précis du niveau d'habileté, plus spécifiquement lorsque le niveau d'habileté est faible ou élevé.

En testing adaptatif, chaque personne peut recevoir une version du test dont les items ont un niveau de difficulté adapté à son niveau d'habileté et dont la séquence des items peut varier d'une personne à une autre. Toutefois, cette caractéristique du testing adaptatif fait en sorte que le nombre de bonnes réponses au test ne permet plus de comparer les personnes entre elles puisqu'elles obtiennent toutes, selon certains auteurs (Weiss, 1985, p. 776), environ le même pourcentage de bonnes réponses aux items. Il serait alors plus approprié d'estimer le niveau d'habileté indépendamment du choix particulier des items d'une version du test.

Des propositions de modélisation des réponses aux items, telles que celles décrites par Goldstein et Wood (1989) ou par Thissen et Steinberg (1986), ont facilité l'utilisation du testing adaptatif en permettant justement d'estimer le niveau d'habileté indépendamment du choix particulier des items d'une version du test. Toutefois, les calculs exigés par les différentes modélisations mathématiques proposées ne permettaient pas, jusqu'à tout récemment, l'application du testing adaptatif à des situations réalistes, pendant des opérations d'inscription scolaire, par exemple. L'accessibilité à un ordinateur central ou à un mini-ordinateur n'était pas toujours possible en raison à la fois des coûts d'utilisation et de la disponibilité physique des appareils. Les micro-ordinateurs offrent maintenant une puissance de calcul suffisante pour supporter ces propositions de modélisations, et ce à un coût abordable.

Tous les tests, qu'il s'agisse de tests papier-crayon fixes et invariables ou de tests adaptatifs administrés par ordinateur, peuvent être décrits par un ensemble de règles, un algorithme, composé de trois éléments. Le premier de ces éléments concerne la façon de déterminer quelle sera la première question présentée. Le second élément concerne la façon de déterminer quelle sera la question qui suivra une question donnée. Enfin, le dernier élément consiste à déterminer le moment à partir duquel l'administration des questions doit cesser.

Ainsi, les tests varient selon les éléments de l'algorithme qui définissent les règles de départ, de suite et d'arrêt. Un test papier-crayon fixe et invariable dont le nombre de questions est fixe peut, par exemple, être caractérisé par un algorithme relativement simple, comme celui illustré au tableau 9.1.

Tableau 9.1
Algorithme décrivant le déroulement normal d'un test papier-crayon fixe et invariable (d'après Raîche, 2000, p. 19)

RÈGLE	ACTION
1. Règle de départ	Répondre à une première question, généralement la question nº 1
2. Règle de suite	Répondre à une prochaine question, généralement la suivante
3. Règle d'arrêt	Terminer le test lorsqu'une réponse a été donnée à la dernière question

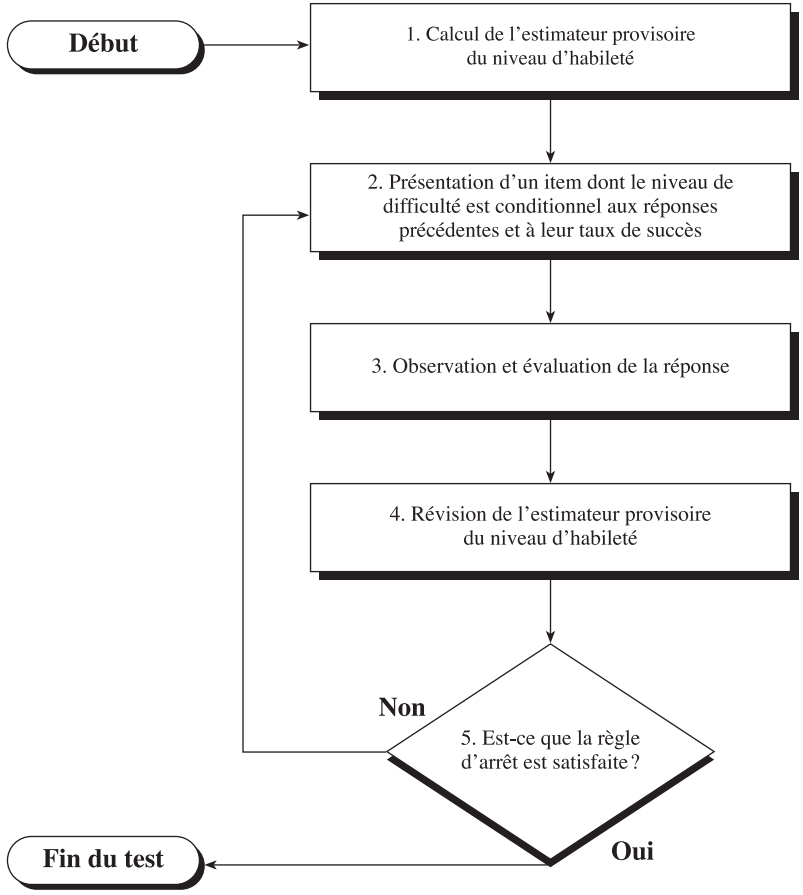
Dans cette démarche, invariablement et quelle que soit la personne, les mêmes questions sont présentées dans le même ordre à tous et à toutes. Toutefois, la personne peut, à sa guise, commencer avec n'importe quel item. Les questions sont présentées à tous dans le même ordre, mais le point de départ est laissé à la discrétion du répondant. Dans les faits, même si presque tous débutent par la première question et répondent de manière séquentielle aux questions suivantes, la suite n'est pas nécessairement la même pour tous.

Dans un test adaptatif, à l'opposé, la première question proposée, les questions subséquentes, l'ordre de ces questions ainsi que la fin du test peuvent varier d'une personne à une autre selon des règles préétablies. Les règles de départ, de suite et d'arrêt permettent de présenter une première question selon des caractéristiques préalables du répondant, de déterminer quelle sera la prochaine question à administrer en fonction de la réponse à la question précédente ou, encore, de mettre fin au test lorsque des conditions qui dépendent des réponses du répondant ont été satisfaites. En ce sens, le test est sur mesure, individualisé, selon les caractéristiques préalables et les réponses de chaque répondant. En fait, dans un test adaptatif, l'objectif est de reproduire le comportement qu'aurait un examinateur expérimenté qui prendrait des décisions sur les questions à administrer au répondant, donc sur les informations à obtenir pour permettre d'estimer le plus précisément possible son niveau d'habileté. Ainsi, lorsqu'un examinateur pose une question trop difficile, il peut ajuster à la baisse le niveau de difficulté de la prochaine question. En effet, l'examinateur apprendrait peu sur le répondant en persistant à ne lui proposer que des questions trop difficiles ou trop faciles, questions auxquelles il n'obtiendrait que de mauvaises ou de bonnes réponses. Au contraire, pour lui permettre d'estimer le niveau d'habileté du répondant le plus précisément possible, l'examinateur devrait tenter d'ajuster le niveau de difficulté des questions au niveau d'habileté du répondant.

La figure 9.1 illustre, de manière générale, le déroulement d'un test adaptatif. Au départ, un estimateur provisoire du niveau d'habileté du répondant est déterminé. Cet estimateur peut être obtenu en se basant sur des caractéristiques du répondant telles que son âge, des résultats antérieurs à d'autres tests ou, tout simplement, un estimateur fourni par le répondant lui-même. En l'absence d'informations préalables sur les caractéristiques du répondant, le niveau de difficulté de la première question est fréquemment

fixé à un niveau moyen. À la suite de la réponse choisie par le répondant, un nouvel estimateur provisoire de son niveau d'habileté est alors calculé et une nouvelle question est administrée. Tant que la règle d'arrêt n'est pas satisfaite, de nouvelles questions, dont le niveau de difficulté est conditionnel aux réponses précédentes et à leur taux de succès, sont présentées. Cette règle d'arrêt peut être aussi simple que de cesser le test lorsqu'un nombre fixe de questions a été présenté, comme elle peut être aussi complexe que de mettre fin à l'administration du test lorsqu'un niveau prédéterminé de précision de l'estimateur du niveau d'habileté est atteint.

FIGURE 9.1
Déroulement général d'un test adaptatif (d'après Raîche, 2000, p. 22)



9.3. LE TESTING ADAPTATIF : UNE APPLICATION FORT PERTINENTE DE LA THÉORIE DES RÉPONSES AUX ITEMS

Ce n'est qu'avec l'introduction de la théorie des réponses aux items (*item response theory*) par Lord (1952) que les applications et le développement des tests adaptatifs peuvent prendre réellement leur envol. Weiss (1982, p. 475-476) souligne quatre avantages importants des tests adaptatifs construits autour de la théorie des réponses aux items.

Premièrement, l'obtention d'un estimateur du niveau d'habileté qui se situe sur la même échelle de mesure que le niveau de difficulté des items devient possible. Les tests adaptatifs précédents ne permettaient pas de répondre à cette correspondance métrique parce qu'ils étaient construits autour de la théorie classique des tests. En second lieu, il y a un avantage corollaire à ceci : le niveau d'habileté peut être estimé à partir de n'importe quel sous-ensemble d'items administrés. Cette caractéristique est très utile en testing adaptatif puisqu'elle permet d'administrer des items différents à des personnes différentes, tout en permettant d'obtenir des scores sur une même échelle. Les tests peuvent donc être réellement considérés sur mesure.

Troisièmement, un test adaptatif fondé sur des propositions de modélisation des réponses aux items issues de la théorie des réponses aux items peut être conçu de façon telle que les branchements soient conditionnels à des caractéristiques supplémentaires au seul niveau de difficulté des items. Ainsi, le pouvoir de discrimination et la probabilité de réussir un item sans pour autant connaître la réponse, la pseudo-chance (*pseudo-guessing*), peuvent être pris en considération.

Enfin, un dernier avantage souligné par Weiss (1982) est que la règle d'arrêt peut être basée sur la précision de l'estimateur du niveau d'habileté après chaque réponse. La règle d'arrêt peut ainsi être conditionnelle à l'atteinte d'un niveau de précision prédéterminé de l'estimateur du niveau d'habileté.

Dans un test adaptatif, où sont présentés des items dont le niveau de difficulté se rapproche le plus possible du niveau d'habileté, des décisions doivent être prises en ce qui concerne les caractéristiques du ou des premiers items administrés ; autrement dit, une règle de départ doit être établie. Par suite de la performance à un premier item ou aux premiers items, d'autres items dont le niveau de difficulté est de plus en plus près du niveau d'habileté sont proposés ; il est alors question de la règle de suite. Enfin, un ou des critères ayant pour but de décider de mettre fin à la situation de mesure doivent être adoptés ; il s'agit de la règle d'arrêt.

La figure 9.2 et le tableau 9.2 décrivent le déroulement d'un tel test. Nous présentons, pour chacune des règles considérées, soient celles de départ, de suite et d'arrêt, des stratégies proposées par la littérature.

FIGURE 9.2
Structure d'un test adaptatif basé sur la théorie des réponses aux items (d'après Raîche, 2000, p. 71)

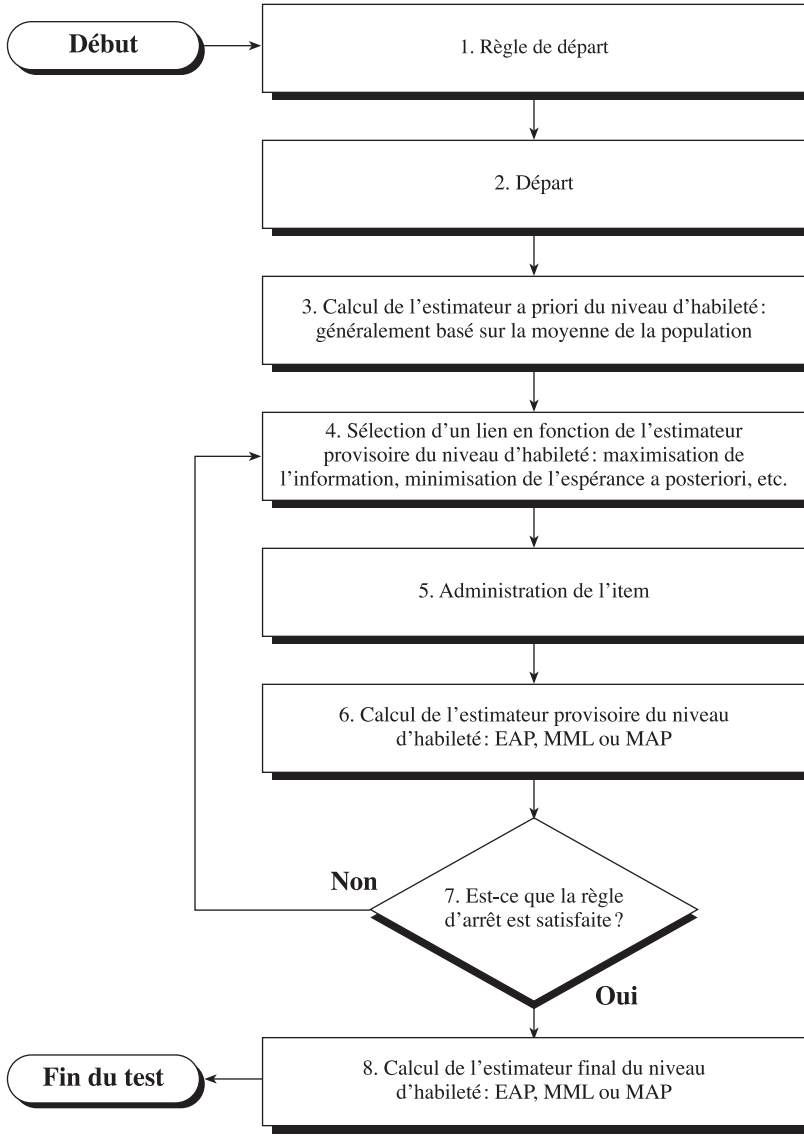


TABLEAU 9.2
Algorithme décrivant le déroulement d'un test adaptatif basé sur la théorie des réponses aux items (d'après Raîche, 2000, p. 72)

RÈGLE	ACTION
1. Règle de départ	Administrer un item dont le niveau de difficulté est conditionnel à certaines caractéristiques du candidat
2. Règle de suite	Administrer un item dont le niveau de difficulté se rapproche de la valeur de l'estimateur provisoire du niveau d'habileté
3. Règle d'arrêt	Terminer le test après l'administration d'un nombre prédéterminé d'items, lorsqu'une erreur-type prédéterminée de l'estimateur du niveau d'habileté est obtenue ou lorsqu'il n'y a plus d'items qui puissent fournir une quantité d'information minimale au niveau d'habileté estimé

9.3.1. Les stratégies quant à la règle de départ

Un test adaptatif débute généralement par l'administration d'un item dont le niveau de difficulté est conditionnel à l'information disponible a priori ; moyenne de groupe, âge ou même appréciation subjective de la part de l'étudiant évalué. Il faut donc adopter une règle de départ basée sur l'information a priori disponible à propos du niveau d'habileté. Selon Thissen et Mislevy (2000, p. 107), la moyenne de la population d'où provient l'individu en situation de testing, estimée préalablement selon une modélisation issue de la théorie des réponses aux items, est un estimateur provisoire de départ raisonnable du niveau d'habileté. Un estimateur préalable du niveau d'habileté moyen peut être obtenu, par exemple, comme le souligne van der Linden (1999, p. 22), à partir des administrations précédentes du test adaptatif à d'autres étudiants. Le niveau de difficulté du premier item administré est ainsi égal au niveau d'habileté moyen de la population. Le premier item présenté est alors un item dont les paramètres permettent une discrimination optimale lorsque le niveau d'habileté est égal à la moyenne du niveau d'habileté de la population.

Laurier (1993b, p. 146-148), quant à lui, utilise des informations qu'il obtient directement auprès du répondant en cours d'administration d'un test de classement en langue seconde. Ainsi, avant d'administrer les items du test, le répondant doit répondre à quelques questions qui permettent d'obtenir des renseignements sur son habileté perçue et sur son expérience antérieure dans la langue seconde. Des questions du type : *À quelle année remonte le dernier cours suivi dans la langue seconde ?* ou encore, *Quel est ton degré d'aisance dans la langue seconde ?* Le niveau de difficulté du premier item administré est tributaire du niveau d'habileté établi en fonction des réponses à ces questions préalables.

On pourrait aussi imaginer une stratégie très simple pour obtenir de l'information a priori sur le niveau d'habileté d'un étudiant lors de l'administration d'un test de classement en anglais langue seconde, un test, non adaptatif pour le moment, qui est d'ailleurs utilisé actuellement dans la plupart des

collèges et cégeps du Québec, soit le TCALS II (Laurier, Froio, Paero et Fournier, 1999). Il s'agirait d'utiliser les résultats obtenus en anglais en secondaire V et IV. Selon des résultats non publiés obtenus chez les étudiants inscrits au Collège de l'Outaouais (n = 1715), le coefficient de détermination entre les notes en secondaire V et le résultat au TCALS II est d'ailleurs de 0,64, une valeur assez importante pour justifier l'utilisation du résultat en secondaire V comme estimateur du niveau d'habileté a priori dans une éventuelle version adaptative du TCALS II.

Il faut aussi signaler que la détermination de l'estimateur a priori du niveau d'habileté, $\hat{\theta}_{a\ priori}$, peut affecter l'estimateur final du niveau d'habileté, $\hat{\theta}$, lorsque trop peu d'items sont administrés. C'est pourquoi les auteurs (Thissen et Mislevy, 1989, p. 110) suggèrent d'utiliser la même valeur de l'estimateur a priori du niveau d'habileté pour toutes les personnes à qui est administré un test adaptatif. Raîche (2000, p. 188-189) a réalisé une modélisation de l'impact de la détermination de l'estimateur a priori sur la valeur obtenue du biais de l'estimateur du niveau d'habileté en fonction de quatre valeurs courantes de la règle d'arrêt selon l'erreur-type ; 0,40, 0,35, 0,30 et 0,20. À titre de rappel, le biais de l'estimateur du niveau d'habileté correspond à la valeur moyenne de la différence entre l'estimateur du niveau d'habileté et le niveau d'habileté, soit $\overline{\hat{\theta} - \theta}$. Les équations de régression cubique (équations 9.1 à 9.4) permettent de calculer le biais de l'estimateur du niveau d'habileté lorsque l'estimateur a priori du niveau d'habileté a été fixé à 0,00. Dans ces équations $Biais_{0,40}$, $Biais_{0,35}$, $Biais_{0,30}$ et $Biais_{0,20}$ représentent le biais de l'estimateur du niveau d'habileté selon les quatre valeurs retenues de la règle d'arrêt selon l'erreur-type.

$$Biais_{0,40} = 0,00206 - 0,15132\,\theta + 0,00040\,\theta^2 - 0,00078\,\theta^3 \qquad (9.1)$$

$$Biais_{0,35} = 0,00419 - 0,11620\,\theta + 0,00127\,\theta^2 - 0,00088\,\theta^3 \qquad (9.2)$$

$$Biais_{0,30} = 0,00962 - 0,08577\,\theta + 0,00593\,\theta^2 - 0,00007\,\theta^3 \qquad (9.3)$$

$$Biais_{0,20} = -0,01069 - 0,04019\,\theta + 0,00096\,\theta^2 - 0,00041\,\theta^3 \qquad (9.4)$$

Le tableau 9.3, ainsi que la figure 9.3, présentent les valeurs du biais de l'estimateur du niveau d'habileté prédites par ces fonctions. Il y est très clair que plus l'erreur-type retenue pour la règle d'arrêt est élevée, situation où moins d'items sont administrés, plus le biais de l'estimateur du niveau d'habileté est important. On remarque d'ailleurs que la valeur du biais de l'estimateur du niveau d'habileté peut s'approcher de la valeur de l'erreur-type retenue, voire la surpasser, lorsque le niveau d'habileté s'éloigne considérablement de l'estimateur a priori et que l'erreur-type retenue pour la règle d'arrêt est inférieure à 0,20. À titre d'exemple, lorsque l'erreur-type retenue pour la règle

d'arrêt est égale à 0,40 et que le niveau d'habileté est égal à –3,00, le biais de l'estimateur du niveau d'habileté atteint 0,48. Le tableau 9.3 et la figure 9.3 nous permettent aussi de constater que le biais de l'estimateur du niveau d'habileté est peu important, quelle que soit la valeur du niveau d'habileté, lorsque le niveau d'habileté ne dépasse pas 1,00 en valeur absolue ; dans ces conditions, il est au plus de | 0,15 |. Ces résultats ne sont pas surprenants ; ils concordent avec ceux obtenus par la communauté scientifique. Nous devons donc faire preuve de prudence lorsque l'estimateur du niveau d'habileté s'éloigne considérablement de la valeur de l'estimateur a priori et que l'erreur-type retenue pour la règle d'arrêt est élevée. Ces considérations nous amènent à recommander, contrairement à ce que suggèrent Thissen et Mislevy, l'utilisation de valeurs adaptées au sujet comme estimateur a priori du niveau d'habileté en testing adaptatif. Une règle de départ qui permet l'utilisation de valeurs adaptées au sujet comme estimateur a priori du niveau d'habileté offre aussi l'avantage de minimiser l'exposition des mêmes items à différents sujets puisque, pour des raisons de sécurité, il faut s'assurer que le premier item administré puisse varier d'un répondant à un autre. Sinon, il serait risqué que les répondants se transmettent l'information et soient ainsi informés à l'avance du contenu du premier item administré.

TABLEAU 9.3
Biais de l'estimateur du niveau d'habileté selon la distance entre l'estimateur a priori et le niveau d'habileté en fonction de quatre valeurs de l'erreur-type retenue pour la règle d'arrêt ($S_{\hat{\theta}}$) lorsque l'estimateur a priori du niveau d'habileté est fixé à 0,00

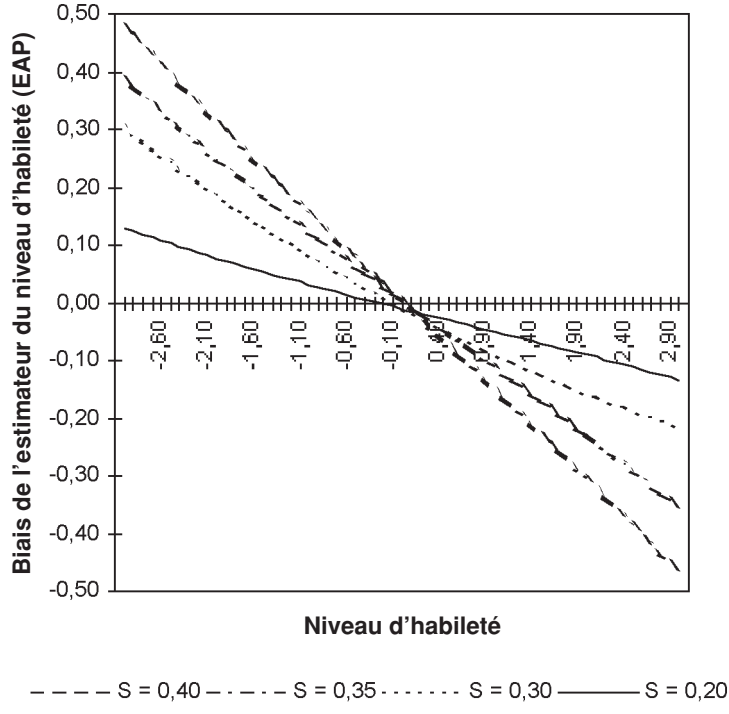
	$S_{\hat{\theta}}$			
Niveau d'habileté $\hat{\theta}$	0,40	0,35	0,30	0,20
– 3,00	0,48	0,39	0,30	0,13
– 2,00	0,31	0,25	0,19	0,08
– 1,00	0,15	0,12	0,08	0,03
0,00	0,00	0,00	– 0,01	– 0,01
1,00	– 0,15	– 0,11	– 0,09	– 0,05
2,00	– 0,31	– 0,23	– 0,16	– 0,09
3,00	– 0,47	– 0,36	– 0,22	– 0,13

9.3.2. Les stratégies quant à la règle de suite

Selon la performance au premier item ou aux items précédents, un item optimal, dont le niveau de difficulté se rapproche de l'estimateur provisoire du niveau d'habileté, doit être sélectionné, puis administré et ainsi de suite, jusqu'à ce que la règle d'arrêt soit satisfaite. Deux stratégies sont couramment utilisées pour sélectionner le prochain item à administrer lorsqu'un estimateur provisoire du niveau d'habileté, basé sur les réponses précédentes et des

informations auxiliaires, est disponible (Thissen et Mislevy, 1990, p. 111). Il s'agit des stratégies de maximisation de l'information (*maximum information*) et de minimisation de l'espérance de l'erreur-type a posteriori (*minimum expected posterior standard deviation*) de l'estimateur du niveau d'habileté. Ces deux stratégies, selon certains auteurs (Thissen et Mislevy, 1990, p. 112-113 ; Wainer et Kiely, 1987, p. 188), peuvent toutefois provoquer un déséquilibre du contenu des items lorsque différentes valeurs du paramètre de discrimination sont reliées à des domaines de contenu différents. Wainer et Kiely (1987) proposent une stratégie de sélection des items permettant d'exercer un meilleur contrôle sur l'équilibre du contenu des items, celle des minitests (*testlets*). Enfin, de nouvelles stratégies ont été récemment proposées pour tenir compte de contraintes spécialisées dans la sélection des items (Hetter et Sympson, 1997 ; van der Linden et Pashley, 2000). Nous présentons maintenant ces diverses stratégies quant à la règle de suite.

FIGURE 9.3
Biais de l'estimateur du niveau d'habileté en testing adaptatif selon quatre valeurs de l'erreur-type (S) retenue pour la règle d'arrêt lorsque l'estimateur a priori du niveau d'habileté est fixé à 0,00



Stratégies de maximisation de l'information

La première de ces stratégies de sélection du prochain item à administrer consiste à choisir l'item pour lequel l'information est maximale. Plusieurs méthodes peuvent être utilisées pour maximiser l'information ; par information maximale sans contrainte, par une table des valeurs de l'information pour chaque item ou par la méthode d'Urry.

La méthode de sélection par information maximale sans contrainte (*unconstrained maximum information selection*) permet de choisir un item pour lequel l'information, au sens de Fisher (1922), évaluée au niveau d'habileté estimé provisoirement après l'administration de l'item i est maximale (Lord, 1980, p. 199). C'est le concept d'information qui permet d'obtenir une mesure de la précision de l'estimateur du niveau d'habileté lorsque celui-ci est obtenu par la méthode du maximum de vraisemblance (Baker, 1992, p. 79-81). Tel que souligné au chapitre 4, l'information fournie par l'item i au niveau d'habileté θ est évaluée en conformité avec l'équation 9.5.

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \tag{9.5}$$

où $P_i(\theta)$ correspond à la probabilité d'obtenir une bonne réponse à l'item i calculée selon une des modélisations issues de la théorie des réponses aux items sachant que le niveau d'habileté est égal à θ. $Q_i(\theta)$ correspond à la probabilité d'obtenir une mauvaise réponse à l'item i et $P_i'(\theta)$ est égale à la dérivée première de la fonction de probabilité. L'information offre l'avantage d'être additive, de sorte que l'information totale à un test à un niveau d'habileté fixé est égale à la somme de l'information fournie par chacun des items administrés.

Il est cependant possible d'obtenir, avec une précision satisfaisante, une approximation de l'information fournie au prochain item en recourant à une table de valeurs où l'information (calculée au préalable) apportée par chacun des items d'une banque d'items disponibles est indiquée pour différentes valeurs du niveau d'habileté. La procédure de sélection consiste alors à choisir l'item qui fournit le plus d'information à une valeur rapprochée du niveau d'habileté. Selon Thissen et Mislevy (1990, p. 111), cette méthode a l'avantage d'être moins exigeante en temps de calcul tout en permettant d'obtenir une approximation généralement satisfaisante.

Urry (1970, p. 82) propose une méthode de remplacement et relativement simple qui consiste à choisir le prochain item de façon telle que le niveau de difficulté, b, de cet item soit le plus près possible de l'estimateur provisoire du niveau d'habileté. Cette méthode est équivalente à la méthode d'information maximale sans contrainte et à la méthode de la table des valeurs

lorsque le modèle logistique à un paramètre est utilisé puisque, dans ce modèle, l'information est maximale lorsque le niveau de difficulté de l'item est égal à l'estimateur du niveau d'habileté.

Stratégie de minimisation de l'espérance de l'erreur-type a posteriori

La seconde stratégie de sélection du prochain item à administrer consiste à choisir un item qui minimise l'espérance de l'erreur-type a posteriori de l'estimateur du niveau d'habileté. Owen (Jensema, 1974, 1977 ; Owen, 1975) propose une méthode bayesienne basée sur une stratégie de mise à jour récursive de l'estimateur de l'habileté. Cette fonction utilise un modèle à deux paramètres basé sur la loi normale. Owen (1975), ainsi que Thissen et Mislevy (1990, p. 112), soulignent que, dans la méthode bayesienne d'Owen, une approximation de la loi normale par une loi logistique est fréquemment appliquée.

À cause de la complexité de leur représentation, les équations utilisées dans la méthode bayesienne d'Owen pour le calcul de l'estimateur du niveau d'habileté et de l'erreur-type de l'estimateur du niveau d'habileté ne sont pas présentées ici. Selon Thissen et Mislevy (1990, p. 112), ces équations, quoique complexes, permettent de diminuer le temps de calcul de façon significative puisqu'elles ne reposent pas sur des calculs itératifs, comme c'est le cas dans la méthode de sélection par information maximale sans contrainte. Ces auteurs soulignent toutefois un inconvénient important dans l'application de la méthode bayesienne d'Owen ; l'estimateur du niveau d'habileté et l'erreur-type de celui-ci varient avec l'ordre de présentation des items. C'est une propriété indésirable en testing adaptatif, où les valeurs obtenues de l'estimateur du niveau d'habileté et de son erreur-type devraient être indépendantes de l'ordre de présentation des items. Pour cette raison, selon Thissen et Mislevy, l'utilisation de la méthode bayesienne d'Owen, tenant compte de l'amélioration de la puissance de calcul des ordinateurs, est de moins en moins de mise en testing adaptatif.

Thissen et Mislevy (1990, p. 113), ainsi que Wainer, Dorans, Green, Mislevy, Steinberg et Thissen (1990, p. 240), soulignent aussi que les stratégies de maximisation de l'information et de minimisation de l'espérance de l'erreur-type a posteriori de l'estimateur du niveau d'habileté peuvent provoquer des séquences problématiques de présentation des items. Ces stratégies font en sorte que les items dont le paramètre de discrimination est élevé sont sélectionnés plus fréquemment. Selon eux, cette situation peut mener à un déséquilibre du contenu des items lorsque différentes valeurs du paramètre de discrimination sont reliées à des domaines de contenu différents. C'est pour pallier ce problème que Wainer et Kiely (1987) suggèrent l'utilisation de minitests.

Minitests

Wainer et Kiely (1987) proposent une stratégie qui pourrait permettre d'exercer un meilleur contrôle sur l'équilibre du contenu des items. Ils suggèrent de sélectionner des groupes d'items (*item clusters*) plutôt que des items isolés. Ainsi, selon la performance à un premier minitest, un minitest optimal est sélectionné puis administré. Selon Wainer et Kiely, cette stratégie permettrait d'exercer un contrôle sur plusieurs aspects reliés au contexte d'un test adaptatif. Il serait ainsi possible d'annuler l'effet indésirable de l'ordre de présentation d'un item dans un test, qui peut varier d'une administration du test à une autre. Il serait aussi possible de mieux contrôler les effets croisés (*cross-information*) qui se produisent lorsque l'administration d'un item fournit des informations qui influent sur la réponse aux items suivants.

Un exemple de minitest appliqué au domaine de la statistique est offert à la figure 9.4. On peut remarquer que les réponses aux items 2, 3 et 4 ne sont pas indépendantes de la réponse fournie à l'item 1 puisque la valeur de la moyenne est nécessaire à la réussite de ces items. De plus, la réussite des items 3 et 4 nécessite la connaissance de la valeur de l'écart-type calculé à l'item 2. Il est alors possible d'attribuer, soit un score de succès ou d'échec au minitest, soit un score variant entre 0 et 4 représentant un échec, un succès partiel ou un succès total au minitest.

FIGURE 9.4
Exemple d'un minitest

Soit la série de valeurs suivante : 10, 25, 5, 15, 30, 2, 24, 18 et 30

1.	Quelle est la valeur de la moyenne arithmétique ?	
	a) 17,67 b) 21,32 c) 25,01 d) 10,00	1 point
2.	Quelle est la valeur de l'écart-type ?	
	a) 2,32 b) 45,10 c) 5,51 d) 10,43	1 point
3.	Quelle est la valeur du coefficient d'asymétrie ?	
	a) 4,55 b) –0,28 c) –8,76 d) 0,12	1 point
4.	Quelle est la valeur du coefficient de kurtose ?	
	a) –1,39 b) 56,34 c) –8,02 d) 1,39	1 point
	TOTAL	4 points

Prometteuse selon Wainer *et al.* (1990, p. 253-254), cette stratégie exige toutefois des modélisations des réponses aux items plus sophistiquées que celles qui sont utilisées dans les modèles habituels. Les modèles à réponses nominales ou ordonnées se montrent alors intéressants. Dans cette veine, Thissen (1993) propose d'ailleurs certains modèles spécifiques à une démarche de testing adaptatif par minitests, principalement la modélisation des réponses aux items par crédit partiel (*partial credit model*) de Masters (1982). La recherche sur l'utilisation des minitests en testing adaptatif est très active actuellement. Plus récemment, des auteurs tels que Wainer, Bradlow et Du (2000), Glas, Wainer et Bradlow (2000) ainsi que Vos et Glas (2000) proposent

diverses stratégies pour soutenir la mise en œuvre des minitests lors de tests adaptatifs. Wainer, Bradlow et Wu (2000) explorent l'utilisation d'une généralisation multidimensionnelle de la modélisation logistique à trois paramètres. Glas, Wainer et Bradlow (2000) explorent l'utilisation des méthodes d'estimation basées sur les chaînes de Markov Monte Carlo (*Monte Carlo Markov Chain*) tandis que Vos et Glas appliquent la stratégie des minitests aux tests de maîtrise.

Nouvelles stratégies de sélection des items

Enfin, de nouvelles stratégies ont été récemment proposées pour tenir compte de contraintes spécialisées dans la sélection des items. Ces stratégies reçoivent actuellement beaucoup d'attention de la part des chercheurs et il est fort probablement trop tôt pour vraiment juger de leur supériorité sur les stratégies présentées plus haut.

Dans la plupart des cas, il s'agit de contraintes qui visent à minimiser la probabilité d'exposition de chacun des items qui composent la banque d'items disponibles. Par exemple, Hetter et Sympson (1997) ont développé une stratégie de sélection du prochain item visant à réduire les séquences d'items prédictibles et, ainsi, la surexposition éventuelle des items qui fournissent le plus d'information au sens de Fisher. Van der Linden (2000), ainsi que van der Linden et Pashley (2000), pour leur part, suggèrent l'utilisation de tests fantômes (*shadow tests*), soit, plus ou moins des super-minitests satisfaisant à des contraintes plus complexes que celle qui vise le simple contrôle de la surexposition des items. Les tests fantômes ainsi construits sont ceux qui fournissent le plus d'information tout en répondant aux contraintes spécifiées. Ces contraintes peuvent s'adresser aux caractéristiques des items, telles que le nombre de mots, le nombre de choix de réponse. Elles peuvent aussi dicter l'administration d'items différents selon les personnes à qui sont administrés les tests ; par exemple, selon la langue, le sexe ou la culture.

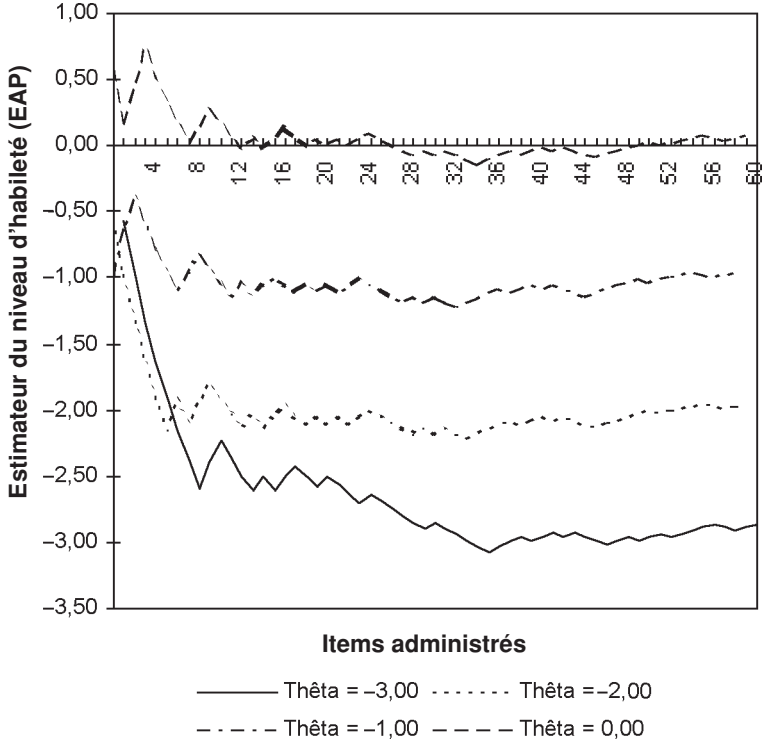
9.3.3. Stratégies d'estimation provisoire du niveau d'habileté

Selon Thissen et Mislevy (1990, p. 113), les méthodes d'estimation provisoire du niveau d'habileté les plus utilisées après l'administration de J items sont celles basées sur les fonctions de vraisemblance telle que la méthode du maximum de vraisemblance (*maximum likelihood, ML*) (section 6.1). On utilise aussi les méthodes bayesiennes d'estimation du niveau d'habileté, soit la méthode bayesienne de maximisation a posteriori (*maximization a posteriori, MAP*) (section 6.2) et la méthode de l'espérance a posteriori (*expected a posteriori, EAP*) (section 6.3). Wainer et Thissen (1987, p. 353) ont comparé différentes méthodes d'estimation du niveau d'habileté et en arrivent à la conclusion que les estimateurs du niveau d'habileté obtenus par la méthode de l'espérance a posteriori sont ceux dont l'erreur-type est généralement la plus petite.

Selon Thissen et Mislevy (1990, p. 113), la méthode bayesienne d'Owen est quelquefois utilisée puisque la précision de l'estimateur provisoire du niveau d'habileté est moins importante à cette étape que la rapidité des calculs.

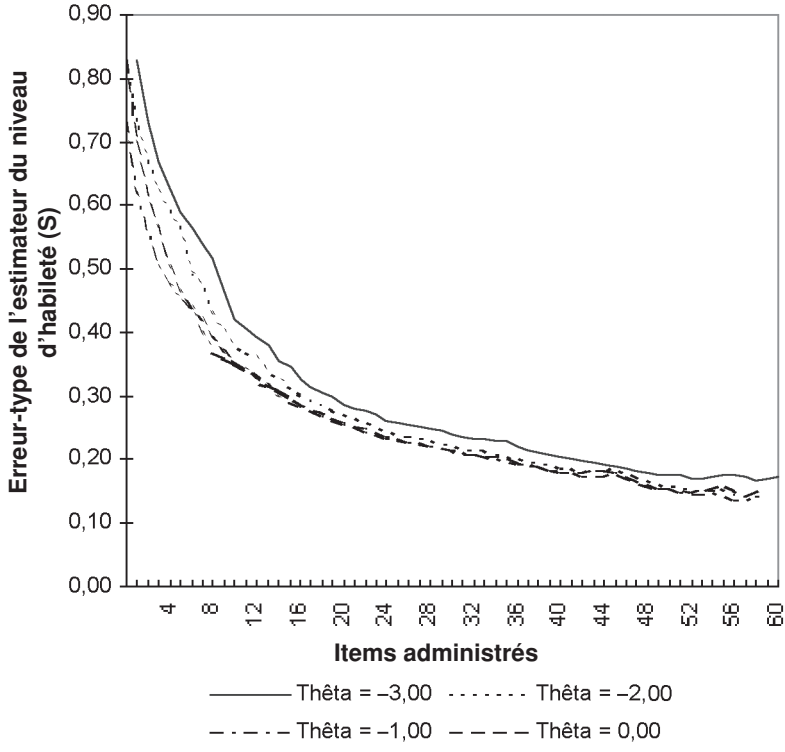
Les figures 9.5 et 9.6 représentent respectivement les valeurs de l'estimateur provisoire du niveau d'habileté et de l'erreur-type associés à chacun des 60 items, valeurs obtenues par une simulation de quatre tests adaptatifs. La modélisation logistique à un paramètre et la méthode de l'estimateur a posteriori (EAP) sont utilisées. Les quatre tests adaptatifs diffèrent par la valeur du niveau d'habileté simulé pour quatre sujets ; –3,00, –2,00, –1,00 et 0,00. On peut remarquer à la figure 9.5 que la convergence de l'estimateur du niveau d'habileté est plus rapide quand le niveau d'habileté est fixé à 0,00. En fait, lorsque peu d'items sont administrés et que le niveau d'habileté s'éloigne substantiellement de 0,00, le biais de l'estimateur du niveau d'habileté est assez important.

FIGURE 9.5
Estimateur du niveau d'habileté en testing adaptatif en fonction du nombre d'items administrés selon quatre valeurs du niveau d'habileté



Selon la figure 9.6, plus le niveau d'habileté s'éloigne de 0,00, plus l'erreur-type de l'estimateur du niveau d'habileté est importante. Ce phénomène devient toutefois de moins en moins important avec l'augmentation du nombre d'items administrés. À partir du 12e item administré, l'erreur-type est d'environ 0,40, tandis qu'elle n'est que de 0,20 autour du 40e item.

FIGURE 9.6
Erreur-type de l'estimateur du niveau d'habileté en testing adaptatif en fonction du nombre d'items administrés selon quatre valeurs du niveau d'habileté



Raîche et Blais (2002b) expérimentent actuellement des méthodes d'estimation qui sont elles-mêmes adaptatives, soit l'estimation par intervalle d'intégration adaptatif, l'estimation par estimateur a priori adaptatif et l'estimation avec correction adaptative pour biais. Toutes ces stratégies visent à centrer l'estimation provisoire θ_j du niveau d'habileté autour de l'estimateur provisoire précédent θ_{j-1} du niveau d'habileté.

La méthode d'estimation par intervalle d'intégration adaptatif est utilisée pour permettre d'ajuster l'intervalle d'intégration de la méthode de l'espérance a posteriori (EAP) à la valeur de l'estimateur du niveau d'habilité obtenue au cycle d'estimation précédent. Dans la méthode de l'estimation de l'espérance a posteriori, les limites d'intégration sont généralement fixées au plus à ±4,00 autour d'un point milieu égal à 0,00. La méthode d'estimation par intervalle adaptatif fait en sorte que les limites d'intégrations varient de ± 4,00 autour de la valeur de l'estimateur provisoire précédent θ_{j-1}.

La méthode d'estimation par estimateur a priori adaptatif peut aussi bien s'appliquer à la méthode de maximisation a posteriori qu'à la méthode de l'espérance a posteriori où l'estimateur a priori du niveau d'habileté est fixe tout au long de l'administration du test adaptatif. Dans la méthode d'estimation par estimateur a priori adaptatif, l'estimateur a priori varie donc en fonction de la valeur de l'estimateur provisoire θ_{j-1} obtenu après l'administration de l'item précédent.

Enfin, la méthode d'estimation avec correction adaptative pour biais applique l'ajustement proposé par Bock et Mislevy (1982, p. 439-442) pour diminuer l'importance du biais de l'estimateur du niveau d'habileté. Bock et Mislevy effectuent cette correction en divisant l'estimateur du niveau d'habileté par une approximation du coefficient de fidélité, r_{tt} :

$$r_{tt} = 1 - S^2_{\hat{\theta}} \tag{9.6}$$

où $S_{\hat{\theta}}$ correspond à l'erreur-type associée à l'estimateur du niveau d'habileté. L'estimateur corrigé du niveau d'habileté $\hat{\theta}_c$ devient alors égal à

$$\hat{\theta}_c = \frac{\hat{\theta}}{1 - S^2_{\hat{\theta}}} \tag{9.7}$$

Généralement, la correction de Bock et Mislevy n'est appliquée qu'à l'estimateur final du niveau d'habileté. Dans la méthode d'estimation avec correction adaptative pour biais, elle est effectuée à chaque estimation provisoire du niveau d'habileté.

Le tableau 9.4 illustre la simulation des résultats obtenus après l'administration de chaque item à quatre tests adaptatifs qui se terminent au 15^{e} item chez une personne dont le niveau d'habileté est égal à –3,00. Le premier test adaptatif utilise la méthode de l'espérance a posteriori (EAP), tandis que des stratégies d'estimation adaptative sont utilisées dans les trois autres tests adaptatifs. La méthode usuelle de l'espérance a posteriori et la méthode avec intervalle d'intégration adaptatif sont celles qui présentent la valeur la plus importante du biais de l'estimateur du niveau d'habileté, soit 0,40

(–2,60 – (–3,00) = 0,40). Toutefois, la méthode avec intervalle d'intégration adaptative permet d'obtenir une meilleure précision de l'estimateur du niveau d'habileté puisque l'erreur-type (0,31) est inférieure à celle obtenue par la méthode de l'espérance a posteriori (0,35). Cependant, c'est la méthode avec estimateur a priori adaptatif qui semble la plus intéressante, car elle permet d'obtenir l'estimateur du niveau d'habileté le moins biaisé, soit seulement –0,07, et que l'erreur-type associée est parmi les plus petites.

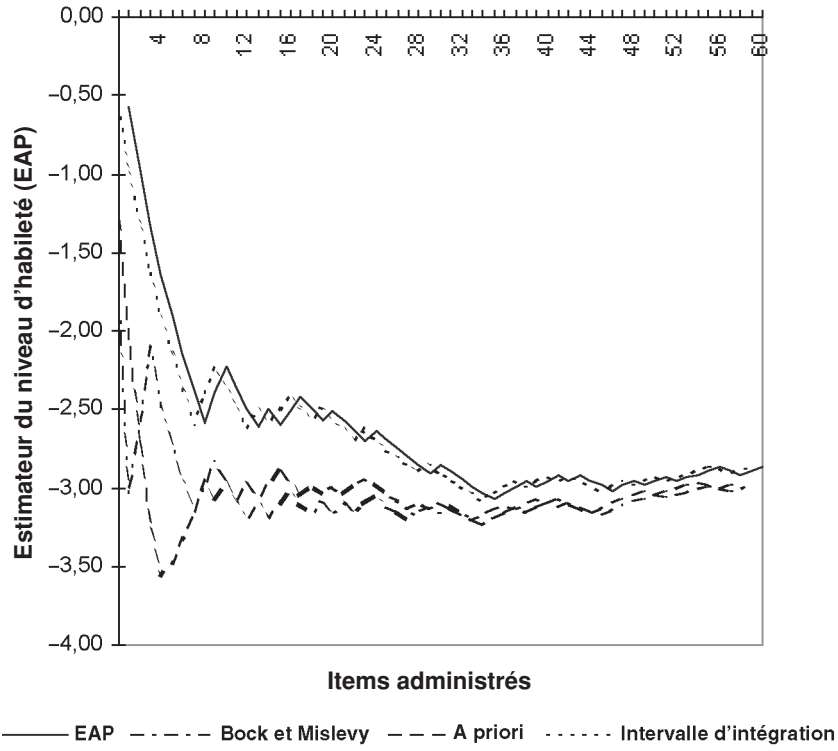
TABLEAU 9.4
Estimateur du niveau d'habileté, erreur-type de celui-ci et réponse à l'item en testing adaptatif en fonction du nombre d'items administrés selon quatre méthodes d'estimation lorsque le niveau d'habileté est égal à –3,00

	EAP			Correction de Bock et Mislevy adaptative			Estimateur a priori adaptatif			Intervalle d'intégration adaptatif		
Item	r*	$\hat{\theta}$	$S_{\hat{\theta}}$	r	$\hat{\theta}$	$S_{\hat{\theta}}$	r	$\hat{\theta}$	$S_{\hat{\theta}}$	r	$\hat{\theta}$	$S_{\hat{\theta}}$
1	0	–0,57	0,83	0	–1,81	0,83	0	–0,57	0,83	0	–0,57	0,83
2	0	–0,99	0,73	0	–3,05	0,73	0	–1,30	0,73	0	–0,99	0,69
3	0	–1,34	0,67	1	–2,56	0,67	0	–2,05	0,62	0	–1,34	0,61
4	0	–1,64	0,62	1	–2,11	0,62	0	–2,74	0,56	0	–1,64	0,56
5	0	–1,91	0,59	0	–2,48	0,59	0	–3,25	0,50	0	–1,91	0,50
6	0	–2,15	0,56	0	–2,74	0,56	0	–3,56	0,47	0	–2,16	0,46
7	0	–2,38	0,54	0	–2,96	0,49	1	–3,49	0,45	0	–2,38	0,44
8	0	–2,59	0,51	0	–3,16	0,47	1	–3,34	0,43	0	–2,60	0,42
9	1	–2,39	0,46	1	–2,99	0,43	1	–3,16	0,40	1	–2,40	0,39
10	1	–2,24	0,42	1	–2,84	0,40	1	–2,97	0,37	1	–2,24	0,37
11	0	–2,37	0,41	0	–2,97	0,38	0	–3,09	0,36	0	–2,37	0,35
12	0	–2,50	0,39	0	–3,09	0,37	1	–2,98	0,35	0	–2,50	0,34
13	0	–2,62	0,38	0	–3,21	0,36	0	–3,08	0,34	0	–2,62	0,33
14	1	–2,50	0,36	1	–3,10	0,34	1	–2,99	0,32	1	–2,51	0,32
15	0	–2,60	0,35	0	–3,20	0,33	0	–3,07	0,31	0	–2,60	0,31

* r est égal à 1 lors d'une bonne réponse à l'item et à 0 lors d'une mauvaise réponse.

La figure 9.7 permet d'obtenir une représentation visuelle des valeurs disponibles au tableau 9.4. La plus rapide convergence vers la valeur du niveau d'habileté des méthodes de l'estimateur a priori adaptatif et de la correction pour biais de Bock et Mislevy est très nette.

FIGURE 9.7
Estimateur du niveau d'habileté en testing adaptatif en fonction du nombre d'items administrés selon quatre méthodes d'estimation provisoire du niveau d'habileté lorsque le niveau d'habileté est égal à –3,00



9.3.4. Stratégie quant à la règle d'arrêt

Deux règles sont généralement utilisées dans le but de mettre fin au test. La première consiste à arrêter le test après l'administration d'un nombre fixe et prédéterminé d'items. Aucun critère absolu n'a été arrêté quant à ce nombre d'items. Selon Thissen et Mislevy (2000, p. 113), l'administration d'un nombre minimal de 20 items permet d'obtenir un estimateur du niveau d'habileté presque identique, que l'on utilise la méthode d'estimation du maximum de vraisemblance ou une méthode d'estimation bayesienne. En fait, dans les méthodes d'estimation bayesienne, plus le nombre d'items administrés est élevé, moins la fonction de probabilité a priori a d'impact sur l'estimateur obtenu (Chen, Hou, Fitzpatrick et Dodd, 1997, p. 425). De plus, à partir de leur étude de différents estimateurs du niveau d'habileté, Hoijtink et Boomsma

(1995, p. 68) recommandent d'utiliser au moins 10 items pour permettre d'obtenir un estimateur du niveau d'habileté dont le biais et la variance ne sont pas trop importants. Selon eux, les méthodes usuelles d'estimation du niveau d'habileté sont valides lorsque le nombre d'items tend vers l'infini. Le comportement asymptotique des estimateurs a été discuté par Warm (1989), pour la méthode du maximum de vraisemblance, et par Chang et Stout (1993), pour les méthodes bayesiennes. Selon eux, l'estimateur du niveau d'habileté obtenu par la méthode du maximum de vraisemblance, ainsi que par les méthodes bayesiennes, tend vers la vraie valeur du niveau d'habileté lorsque le nombre d'items administrés tend vers l'infini. Raîche (2000), ainsi que Raîche et Blais (2002a) et Blais et Raîche (2000), arrivent à la conclusion que l'administration d'aussi peu que 13 items est suffisante lorsque la situation n'exige pas que l'erreur-type de l'estimateur du niveau d'habileté soit très petite. Ainsi, lorsqu'on utilise la modélisation logistique à un paramètre et la méthode de l'espérance a posteriori, l'erreur-type de l'estimateur du niveau d'habileté est égale à 0,40 avec seulement 12 items administrés. Toutefois, en dehors de l'intervalle [–1,50, 1,50], le biais est important. Quand la précision exigée est plus importante, l'administration d'au moins 40 items est nécessaire. L'erreur-type est alors égale à 0,20 ou moins.

Une seconde règle d'arrêt consiste à terminer l'administration du test lorsqu'une erreur-type prédéterminée de l'estimateur du niveau d'habileté est obtenue. En pratique, il faut aussi fixer un nombre maximal d'items à administrer au cas où l'erreur-type de l'estimateur du niveau d'habileté serait impossible à calculer ou trop longue à obtenir. Cette règle d'arrêt permet, d'après Thissen et Mislevy (1990, p. 114), d'obtenir la même erreur-type à tous les niveaux d'habileté estimés. C'est ce qui explique qu'un test adaptatif utilisant cette règle d'arrêt se conforme au postulat d'homogénéité de la variance de l'estimateur du niveau d'habileté en théorie classique des tests. Selon Raîche (2000), Raîche et Blais (2002a) et Blais et Raîche (2000), pour que le biais de l'estimateur du niveau d'habileté ne soit pas trop important aux valeurs extrêmes du niveau d'habileté, l'erreur-type de l'estimateur du niveau d'habileté doit être d'au plus 0,40 lorsqu'on utilise la modélisation logistique à un paramètre et que la méthode de l'espérance a posteriori. Toutefois, si on désire obtenir l'homogénéité de l'erreur-type de l'estimateur du niveau d'habileté sur un intervalle important du niveau d'habileté, l'erreur-type retenue pour la règle d'arrêt doit être d'au plus 0,20. L'homogénéité de l'erreur-type de l'estimateur du niveau d'habileté est importante quand vient le moment d'appliquer certaines procédures de comparaison de moyennes telles que des tests t de Student ou des analyses de la variance ; ces procédures reposent d'ailleurs sur le postulat de l'homogénéité de la variance.

Dans certaines situations spécifiques, d'autres règles d'arrêt peuvent être utilisées. Ainsi, Dodd (1990) et Dodd, Koch et de Ayala (1993), à l'intérieur d'études de certaines règles d'arrêt, ont utilisé une règle basée sur l'information minimale de l'item (*minimum item information*). Selon cette

stratégie, l'administration du test se termine lorsqu'il n'y a plus d'items dans la banque d'items qui puissent fournir une quantité d'information minimale prédéterminée au niveau d'habileté estimé. Dans une autre situation, lorsque certains tests sont destinés à mesurer l'exactitude des réponses dans un test où le fait de répondre avec rapidité est important (*accuracy at speed*), on peut fixer un temps d'administration prédéterminé (Thissen et Mislevy, 1990, p. 115). Ces auteurs ne recommandent pas d'utiliser cette règle d'arrêt pour les tests de puissance. D'autre part, Hambleton, Zaal et Pieters (1990, p. 351), Kingsbury et Weiss (1983) ainsi que Davey, Godwin et Mittelholtz (1997) suggèrent une stratégie d'arrêt adaptée aux tests critériés *(criterion-referenced testing)* ; le test se termine lorsque la probabilité d'assignation à un niveau de maîtrise ciblé dépasse une valeur prédéterminée.

9.3.5. Estimateur final du niveau d'habileté

Toutes les méthodes utilisées précédemment pour calculer l'estimateur provisoire du niveau d'habileté peuvent servir au calcul de l'estimateur final du niveau d'habileté. Il n'est cependant pas nécessaire que l'estimateur final soit calculé de la même façon que l'estimateur provisoire. Ainsi, selon Thissen et Mislevy (1990, p. 113), il est fréquent que l'estimateur provisoire du niveau d'habileté soit calculé par la méthode bayesienne d'Owen, alors que l'estimateur final du niveau d'habileté est calculé par la méthode du maximum de vraisemblance, de la maximisation a posteriori ou de l'espérance a posteriori. En fait, selon eux, une grande précision de l'estimateur du niveau d'habileté n'est pas nécessaire en cours de testing.

Thissen et Mislevy (1990, p. 115) soulignent que, lorsque les méthodes de la maximisation a posteriori et de l'espérance a posteriori sont utilisées pour calculer l'estimateur final du niveau d'habileté, l'influence de la distribution a priori diminue avec l'augmentation du nombre d'items. Selon eux, il peut donc être plus sûr d'utiliser la même distribution a priori pour tous, question de justice (*test fairness*), surtout lorsque le nombre d'items administrés est petit. Toutefois, ni les travaux de Raîche et Blais (2002b) sur les méthodes d'estimation provisoires adaptatives, ni les commentaires formulés à la section traitant des stratégies quant à la règle de départ ne nous permettent d'endosser le point de vue de Thissen et Mislevy. Leur position serait acceptable seulement dans des situations irréalistes, où trop peu d'items seraient administrés.

Bock et Mislevy (1982) suggèrent d'utiliser la méthode de l'espérance a posteriori pour calculer l'estimateur final du niveau d'habileté. Leurs études indiquent que l'estimateur final du niveau d'habileté obtenu par cette méthode affiche, en général, une valeur plus petite de son erreur-type. Comme nous l'avons signalé plus haut, à la section qui traite de l'estimateur provisoire du niveau d'habileté, ils suggèrent aussi d'utiliser une correction du biais en divisant l'estimateur final du niveau d'habileté par une approximation du coefficient

de fidélité. Selon Raîche (2000, p. 191-194), la correction proposée par Bock et Mislevy est fortement recommandée lorsque le nombre d'items administrés est inférieur à 40 ou que l'erreur-type de l'estimateur du niveau d'habileté retenue pour la règle d'arrêt est supérieure à 0,20.

Certains auteurs ont proposé d'utiliser des méthodes d'estimation du niveau d'habileté qui seraient moins affectées par des patrons de réponses atypiques ou par l'effet potentiel du petit nombre d'items sur le comportement des estimateurs. Selon eux, l'utilisation d'estimateurs robustes (Hoijtink et Boomsma, 1995, p. 54 ; Mislevy et Bock, 1982 ; Thissen et Mislevy, 1990, p. 115 ; Wainer, 1983, p. 71) pourrait être plus appropriée. En ce sens, Mislevy et Bock (1982) suggèrent l'utilisation d'une méthode d'estimation à double pondération (*biweight*) tandis que Wainer et Thissen (1987, p. 344-345) ainsi que Wainer et Wright (1980), explorent une méthode reposant sur une technique de rééchantillonnage sans remise (*jackknife)*, soit la méthode AMJACK. Dans la pratique, toutefois, ces méthodes semblent peu employées, car les situations pour lesquelles elles ont été proposées au départ sont extrêmes et laissent peu de crédibilité quant à leur capacité d'estimer le niveau d'habileté.

9.4. CONSIDÉRATIONS DIVERSES

9.4.1. Une formule de prophétie adaptée aux tests adaptatifs

Dans la théorie classique des tests, la formule de prophétie de Spearman-Brown (Laveault et Grégoire, 1997, p. 154 ; Wainer et Thissen, 2001, p. 31) permet de prédire le nombre d'items qu'il est nécessaire d'administrer $n_{prédit}$ pour obtenir un niveau de fidélité désiré connaissant le niveau de fidélité r_{tt} observé à partir d'un test de longueur n. Par extension, la formule de prophétie permet aussi de prédire le niveau de fidélité qu'afficherait un test n fois plus long ou n fois plus court que le test qui a servi à calculer la fidélité.

Raîche et Blais (2002a) ont élaboré des formules de prophétie spécifiques à un test adaptatif construit autour d'une modélisation logistique à un paramètre. Ainsi, l'équation 9.8 permet de prédire l'erreur-type de l'estimateur du niveau d'habileté $S_{prédite}$ lorsque n fois plus d'items sont administrés. Selon Raîche et Blais, une différence d'au plus 0,04 est obtenue quant à l'erreur-type prédite de l'estimateur du niveau d'habileté.

$$S_{prédite} = \sqrt{1 - \frac{n \times (1 - S^2_{\hat{\theta}})}{1 + (n-1) \times (1 - S^2_{\hat{\theta}})}} \qquad (9.8)$$

Par exemple, à partir des données disponibles au tableau 9.4, si nous tentons de prédire l'erreur-type obtenue à la suite de l'administration du quatrième item à partir de l'erreur-type obtenue après l'administration du premier item (0,83), nous obtiendrons :

$$S_{\text{prédite}} = 0,60 = \sqrt{1 - \frac{4 \times (1 - 0,83^2)}{1 + (4-1) \times (1 - 0,83^2)}}$$

La différence entre l'erreur-type de l'estimateur du niveau d'habileté obtenue au quatrième item, selon la méthode EAP, et l'erreur-type prédite de l'estimateur du niveau d'habileté (0,62 – 0,60) n'est que de 0,02.

À partir de l'équation 9.9, nous pouvons aussi prédire le nombre d'items qu'il est nécessaire d'administrer $n_{\text{prédit}}$ pour obtenir une valeur prédéterminée de l'erreur-type de l'estimateur du niveau d'habileté.

$$n_{\text{prédit}} = n \times \left(\frac{(1 - S^2_{\text{prédite}}) \times \left[1 - (1 - S^2_{\hat{\theta}})\right]}{(1 - S^2_{\hat{\theta}}) \times \left[1 - (1 - S^2_{\text{prédite}})\right]} \right) \tag{9.9}$$

où n correspond au nombre d'items administrés au moment où nous obtenons une erreur-type de l'estimateur du niveau d'habileté.

Toujours en nous basant sur les données fournies au tableau 9.4, connaissant l'erreur-type obtenue à la suite de l'administration du second item (0,73) selon la méthode EAP, nous désirons prédire le nombre d'items nécessaires pour obtenir une erreur-type égale à environ 0,39. En insérant les valeurs requises à l'intérieur de l'équation 9.9, nous obtenons :

$$n_{\text{prédit}} = 12,72 = 2 \times \left(\frac{(1 - 0,39^2) \times \left[1 - (1 - 0,73^2)\right]}{(1 - 0,73^2) \times \left[1 - (1 - 0,39^2)\right]} \right)$$

Au tableau 9.4, un nombre minimal de 12 items était nécessaire pour obtenir une erreur-type de l'estimateur du niveau d'habileté égale à 0,39. La différence entre la valeur prédite et la valeur obtenue n'est alors que de 0,72 (12,00 – 12,72).

Il est intéressant de constater la précision des valeurs obtenues à partir de ces équations avec si peu d'items administrés.

9.4.2. Logiciels disponibles

Les premiers tests adaptatifs basés sur les modélisations issues de la théorie des réponses aux items ont été développés à partir du début des années 70 pour les besoins de la Marine américaine en sélection de personnel, principalement autour des travaux de McBride, Urry, et Weiss (voir McBride et Martin, 1983, p. 223-236). Au début, ces tests, difficilement disponibles au grand public, nécessitaient l'utilisation d'ordinateurs centraux puissants. Avec l'arrivée des micro-ordinateurs, il a été possible de développer des logiciels plus accessibles à coût plus abordable. L'un de ces premiers logiciels de testing adaptatif a été MICROCAT (Assessment Systems Corporation, 1984). MICROCAT a été conçu pour être utilisé sous le système d'exploitation DOS et accepte des fichiers graphiques au plus au format CGA. Toujours sur le marché actuellement, on lui préfère généralement une version plus contemporaine, MICROTEST, fonctionnant sous Windows et qui permet la gestion des fichiers multimédias accessibles par ce système d'exploitation.

Au Canada, Laurier a développé des versions de tests adaptatifs spécifiquement destinés au classement en langue seconde (Laurier, 1999a, 1999b). L'un de ces logiciels, FrenchCapt (*French Computerized Adaptive Placement Test*), est actuellement utilisé dans certaines universités québécoises. Raîche (2000, p. xxx-xxxii) propose aussi un simulateur de tests adaptatifs, SIMCAT, utilisant une modélisation logistique à un paramètre et développé en langage SAS.

Actuellement, les logiciels développés au Canada ou au Québec ne permettent pas, à notre connaissance, à la fois de modifier aisément la banque d'items et de présenter les items sous divers formats : audio, vidéo ou autres. Seuls MICROCAT et MICROTEST permettent ces opérations. Toutefois, leur coût élevé et la non-disponibilité de banques d'items préalablement calibrées dans une langue autre que l'anglais en limitent l'usage par la communauté francophone. C'est pourquoi, à notre avis, le développement de tels logiciels et la mise en marché de banques d'items spécialisées correspond actuellement à un besoin important dans la communauté francophone.

Pour ceux et celles qui désireraient se lancer dans cette aventure, Linacre (2000) rend disponible sur Internet le code source en langage Microsoft BASIC d'un test adaptatif, UCAT. Ce logiciel rend relativement facile la gestion de la banque d'items au domaine de notre choix et permet de plus, caractéristique non négligeable, la recalibration en ligne des items qui composent la banque d'items.

9.5. DÉFIS ET ENJEUX DU TESTING ADAPTATIF

D'un point de vue technologique, nous sommes maintenant prêts à mettre en application des instruments de mesure sous la forme de tests adaptatifs. Des ordinateurs suffisamment puissants sont disponibles à coût abordable. Les

algorithmes de calcul sont relativement bien développés. Les expériences d'administration de tests adaptatifs à grande échelle sont chose courante et une industrie de l'administration des tests adaptatifs est en émergence aux États-Unis. Cependant, comme le soulignent Wainer et Eignor (2000), les tests adaptatifs ne sont pas près de remplacer les tests papier-crayon. Les coûts associés à l'administration des tests adaptatifs sont encore, pour le moment, plus élevés que ceux des tests papier-crayon. Par exemple, Wainer et Eignor (2000, p. 285) considèrent que les frais d'administration du TOEFL aux États-Unis dans sa version papier-crayon seraient encore de seulement 35 $ à 40 $ par individu, plutôt qu'actuellement d'environ 100 $ pour la version adaptative par ordinateur. Le rapport coûts-avantages n'est donc pas encore justifié ; ce qui n'enlève rien à la nécessité de nous préparer pour l'avenir, un avenir qui, d'ailleurs, pourrait s'avérer assez rapproché.

Outre ces considérations financières, les tests adaptatifs continuent de poser de nouveaux défis à la modélisation des réponses aux items. La plupart de ces défis sont apparus en dehors du contexte des tests adaptatifs. En fait, ils étaient déjà l'objet de recherches sur l'application des modélisations des réponses aux items aux tests papier-crayon. Toutefois, les problèmes associés à l'administration des tests adaptatifs ont fait ressortir de façon plus aiguë l'importance de ces enjeux. Sans rechercher l'exhaustivité, voici quelques-uns de ces enjeux et défis.

Selon nous, le défi le plus important des prochaines années sera de proposer des algorithmes de sélection du prochain item qui permettent d'exercer un contrôle quant à l'exposition des items de la banque d'items. Actuellement, il s'agit d'un domaine de recherche très actif. Les algorithmes traditionnels de sélection du prochain item, qui maximisent l'information ou qui minimisent l'erreur-type de l'estimateur du niveau d'habileté, favorisent l'administration des mêmes items aux individus de même niveau d'habileté. Les auteurs étudient actuellement diverses stratégies dont le but est d'éviter que les mêmes items soient administrés trop fréquemment aux personnes qui affichent un niveau d'habileté similaire. La sécurité des tests adaptatifs et, par conséquent, la crédibilité de ces tests est en jeu.

Dans les tests adaptatifs, comme dans les tests papier-crayon, le format de réponse aux items est presque toujours de type réponse à choix multiples. Les modélisations des réponses aux items utilisées pour ce type d'items reposent sur un postulat selon lequel la réponse à un item est indépendante de celle fournie à l'item précédent. Cependant, en éducation, les tests sont fréquemment composés d'items dont la réponse est affectée par la réponse à l'item précédent sous la forme de minitests. De nouvelles modélisations des réponses aux items, ici les minitests, sont alors nécessaires.

Il faut disposer d'un nombre important d'items pour constituer des banques d'items qui justifient l'utilisation d'un test adaptatif. Les paramètres des items qui composent ces banques doivent de plus être estimés au préalable

auprès d'échantillons de taille suffisante. En dehors des évaluations à grande échelle, les efforts et les coûts associés à ces opérations sont souvent trop importants pour qu'elles soient réellement pratiques. Par exemple, il est actuellement difficile d'administrer en classe un test adaptatif pour soutenir soit l'évaluation formative, soit l'évaluation diagnostique. Il ne serait pas justifiable d'investir des efforts humains et financiers considérables dans la création d'une banque d'items pour les besoins d'un seul groupe cours. Les auteurs (Bejar, 1993 ; Embretson, 1999) s'intéressent actuellement à des solutions à ce problème. Principalement autour de la modélisation du test logistique linéaire (*linear logistic test model*) proposée par Fischer (1995), ces auteurs proposent des solutions pour déterminer le paramètre de difficulté d'un item à partir de ses caractéristiques (*model-based item generation*). Ces solutions devraient permettre de produire des items en cours d'administration d'un test adaptatif en fonction d'attributs divers sans être astreint à calibrer préalablement les paramètres des items d'une banque. Il est à noter que cette solution à la génération d'items en ligne présente aussi une solution alternative aux problèmes de surexposition des items.

Les tests de personnalité, de valeurs ou d'intérêts devraient être tout désignés pour profiter des avantages de la technologie du testing adaptatif. Ces tests comportent généralement un nombre important de questions et le temps d'administration est assez long. Ils gagneraient à être élaborés sous forme adaptative pour permettre de diminuer de manière substantielle leur longueur et leur temps d'administration. Puisque le construit mesuré n'est plus unique, mais plutôt multidimensionnel, ces tests nécessitent toutefois des modélisations des réponses aux items plus complexes que celles qui ont été traditionnellement abordées en testing adaptatif. Les règles d'arrêt, de suite et de fin usuelles d'un test adaptatif doivent être aussi repensées pour soutenir ces modélisations multidimensionnelles des réponses aux items. Pour le moment, il reste encore beaucoup de travail à faire quant à la modélisation multidimensionnelle des réponses aux items ; par conséquent, peu de versions de tests adaptatifs permettent de mesurer des construits multidimensionnels.

Nous pourrions aborder plus en détail bien d'autres enjeux et défis associés aux tests adaptatifs, tels que l'analyse des items à réponse construites (Bennett et Ward, 1993), les mesures d'ajustement des patrons de réponses (*person fit*) ou le fonctionnement différentiel d'item (*differential item functioning*). Le terrain de jeu est vaste et nous pouvons prédire, sans que la probabilité de se tromper soit trop importante, que le testing adaptatif sera un sujet d'intérêt en recherche, et pas seulement en éducation, pour plusieurs années encore.

Exercices

1. Décrivez un test papier-crayon conventionnel à partir des trois règles présentées dans ce chapitre.

2. Proposez un algorithme qui permet de décrire le déroulement d'une entrevue de sélection d'emploi à partir des trois règles présentées dans ce chapitre.

3. Suggérez une stratégie pour la règle de début dans un test adaptatif qui assure une protection contre la transmission de l'information quant au contenu du premier item administré.

4. Proposez un minitest qui permet de mesurer l'habileté à développer un test adaptatif.

5. L'erreur-type retenue pour la règle d'arrêt étant égale à 0,20, estimez le biais de l'estimateur du niveau d'habileté lorsque le niveau d'habileté réel est égal à –2,50, –1,76, –0,82, 0,05, 1,13 et 4,00.

6. À partir du tableau 9.4, appliquez la correction de Bock et Mislevy à l'estimateur final du niveau d'habileté après l'administration de chaque item lorsque la méthode de l'espérance a posteriori est utilisée. Comparez les nouvelles valeurs que vous avez calculées avec les valeurs de l'estimateur du niveau d'habileté qui proviennent des méthodes d'estimation adaptatives.

7. Quel impact a la correction pour biais de Bock et Mislevy sur l'erreur-type de l'estimateur du niveau d'habileté ? Plus précisément, l'erreur-type de la valeur corrigée de l'estimateur du niveau d'habileté devrait-elle augmenter, diminuer ou rester stable ? Justifiez votre réponse par des exemples tirés des résultats obtenus à la question 6.

8. À partir des estimateurs du niveau d'habileté du tableau 9.4, calculez la matrice des corrélations entre les diverses méthodes d'estimation du niveau d'habileté. Quelles sont les méthodes qui affichent les liens les plus marquées entre elles ?

9. Prédisez l'erreur-type de l'estimateur du niveau d'habileté après l'administration du 15^{e} item, sachant que l'erreur-type de l'estimateur du niveau d'habileté est égale à 0,54 à la suite de l'administration du septième item.

10. Prédisez le nombre d'items à administrer pour obtenir une erreur-type de l'estimateur du niveau d'habileté égale à 0,20 sachant que l'erreur-type obtenue après l'administration de 10 items est égale à 0,42.

Corrigé des exercices nécessitant des calculs

5. Selon l'équation 9.4.

Réponse :

Niveau d'habileté	Biais estimé
–2,50	0,10
–1,76	0,07
–0,82	0,02
0,05	–0,01
1,13	–0,06
4,00	–0,18

6. Selon l'équation 9.7.

Réponse :

Item	$\hat{\theta}$	$S_{\hat{\theta}}$	Correction de Bock et Mislevy $\hat{\theta}_C$
1	–0,57	0,83	–1,83
2	–0,99	0,73	–2,12
3	–1,34	0,67	–2,43
4	–1,64	0,62	–2,66
5	–1,91	0,59	–2,93
6	–2,15	0,56	–3,14
7	–2,38	0,54	–3,36
8	–2,59	0,51	–3,50
9	–2,39	0,46	–3,04
10	–2,24	0,42	–2,71
11	–2,37	0,41	–2,85
12	–2,50	0,39	–2,94
13	–2,62	0,38	–3,06
14	–2,50	0,36	–2,88
15	–2,60	0,35	–2,97

8. À partir des valeurs de l'estimateur du niveau d'habileté obtenues selon les quatre méthodes d'estimation.

   Réponse :

	EAP	Correction de Bock et Mislevy adaptative	Estimateur a priori adaptatif	Intervalle d'intégration adaptatif
EAP		0,74	0,90	1,00
Correction de Bock et Mislevy adaptative			0,52	0,74
Estimateur a priori adaptatif				0,90
Intervalle d'intégration adaptatif				

9. À partir de l'équation 9.8.

   Réponse : 0,40

10. À partir de l'équation 9.9.

    Réponse : 51,40, soit 52 items.

Bibliographie

Adams, R.J. et Khoo, S.T. (1992). *QUEST : the interactive test analysis system*. Melbourne : Australian Council for Educational Research.

Allen, M.J. et Yen, W.M. (1979). *Introduction to measurement theory*. Monterey : Brooks and Cole.

American Psychological Association. (1985, 1992, 1999). *Standards for educational and psychological testing*. Washington : APA.

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

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

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

Angoff, W.H. (1982). Use of difficulty and discrimination indices for detecting item bias. Dans : R.A. Berk (dir.), *Handbook of methods for detecting test bias.* Baltimore : The Johns Hopkins University Press.

Assessment System Corporation. (1984). *User's manual for the MICROCAT testing system*. St. Paul, MN : Assessment System Corporation.

Auger, R. (1989). Étude de praticabilité du testing adaptatif de maîtrise des apprentissages scolaires au Québec : une expérimentation en éducation économique secondaire 5. Thèse de doctorat non publiée. Montréal : Université du Québec à Montréal.

Auger, R. et Séguin, S.P. (1992). Le testing adaptatif avec interprétation critérielle, une expérience de praticabilité du TAM pour l'évaluation sommative des apprentissages au Québec. *Mesure et évaluation en éducation*, *15*, 1 et 2, 103-145.

Bain, D. et Pini, G. (1996). *La généralisabilité* : *mode d'emploi*. Genève : Centre de recherches psychopédagogiques (SRED).

Baker, F.B. (1985). *The basics of item response theory*. Portsmouth, NH : Heinemann.

Baker, F.B. (1992). *Item response theory : parameter estimation techniques*. New York : Marcel Dekker.

Beaton, A.E. (1987). *Implementing the new design : the NAEP 1983-1984 technical report*. Princeton, NJ : Educational Testing Service.

Beck, A.T., Rush, A., Shaw, B. et Emery, G. (1979). *Cognitive therapy of depression*. New York : Guilford Press.

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

Bejar, I.I. (1993). A generative approach to psychological and educational measurement. Dans : N. Frederiksen, R.J. Mislevy et I.I. Bejar (dir.), *Test theory for a new generation of tests*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Bennett, R.E. et Ward, W.C. (1993). *Construction versus choice in cognitive measurement : issues in constructed responses, performance testing, and portfolio assessment*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Ben-Simon, A. et Cohen, Y. (1990). *Rosenbaum's test of unidimensionality : sensitivity analysis*. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Boston.

Berk, R.A. (dir.) (1982). *Handbook of methods for detecting test bias.* Baltimore : The Johns Hopkins University Press.

Berka, K. (1983). *Measurement : its concepts, theories and problems*. Dordecht, D. Reidel.

Bertrand, R., Boiteau, N., Gauthier, N., Compain, C., Frenette, É., Laprise, A., Léger-Bourgoin, N. et Jeanrie, C. (2001). La gestion des biais de concept, des biais de méthode et des biais d'item dans le contexte des enquêtes du Programme des indicateurs de rendement scolaire (PIRS). Rapport de recherche (247 pages). Sainte-Foy : Université Laval.

Bertrand, R., Dupuis, F.A. et Garneau, M. (1993). Effets des caractéristiques des items sur le rôle des composantes impliquées dans la performance en résolution de problèmes mathématiques écrits : une étude de validité de construit. Document inédit. Québec : Université Laval.

Bertrand, R. et Jeanrie, C. (dir.). (1995). Théories modernes de la mesure : enjeux et perspectives. *Mesure et évaluation en éducation, 17*, 2.

Bertrand, R. et Laroche, L. (1999). *IRT design for the School Achievement Indicators Program (SAIP)*. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Montréal.

Bertrand, R. et Leclerc, M. (1984). La fiabilité des données d'un instrument d'observation des enseignants en classe de mathématique. *Revue des sciences de l'éducation, 10*, 2, 311-329.

Bertrand, R. et Valiquette, C. (1986). *Pratique de l'analyse statistique des données*. Sainte-Foy : Presses de l'Université du Québec.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Dans : F.M. Lord et M.R. Novick (dir.), *Statistical theories of mental test scores*. Reading, Mass : Addison-Wesley.

Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology, 6*, 258-276.

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

Blais, J.G. et Laurier, M.D. (1995). *Methodological considerations in using DIMTEST to assess unidimensionality*. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, San Francisco.

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

Blais, J.G. et Raîche, G. (2002). *Some features of the estimated sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules*. Communication présentée au 11th Biannual International Objective Measurement Workshop. Nouvelle Orléans : IOMW.

Blalock, H.M. (1982). *Conceptualization and measurement in the social sciences*. Beverly Hills : Sage.

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

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

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

Bock, R.D. et Lieberman, M. (1970). Fitting a response model for n dichotomously scores items. *Psychometrika*, *35*, 179-197.

Bock, R.D. et Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 4, 431-444.

Bock, R.D. et Zimowski, M.F. (1995). Multiple group IRT. Dans : W. van der Linden et R. Hambleton (dir.), *Handbook of item response theory*. New York : Springer-Verlag.

Bond, T.G. et Fox, C.M. (2001). *Applying the Rasch model*. Mahwah, NJ : Lawrence Erlbaum Associates.

Brennan, R.L. (1979). Handbook for Gapid : a Fortran IV computer program for generalizability analyses with single facet designs. *ACT technical report No. 34*. Iowa City : The American College Testing Program.

Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City : The American College Testing Program.

Brennan, R.L. (2001). *Generalizability theory*. New York : Springer-Verlag.

Brennan, R.L. et Kane, M. (1977). Signal/noise ratios for domain-referenced tests. *Psychometrika, 42*, 609-625.

Brogden, H.E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology*, *37*, 65-76.

Camilli, G. et Shepard, L.A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA : Sage.

Campbell, N.R. (1920). *Physics : the elements*. Londres : Cambridge University Press.

Campbell, D.T. et Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago : Rand McNally.

Cardinet, J. et Tourneur, Y. (1978). Le calcul des marges d'erreur dans la théorie de la généralisabilité. Service d'étude des méthodes et des moyens d'enseignement. Document 780.410/CT. Mons : Université de l'État.

Cardinet, J. et Tourneur, Y. (1985). *Assurer la mesure*. Berne : Peter Lang.

Cardinet, J., Tourneur, Y. et Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement*, *13*, 2.

Chang, H.H. et Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 1, 37-52.

Chen, S.K., Hou, L., Fitzpatrick, S.J. et Dodd, B.G. (1997). The effect of population distribution and method of theta estimation on computerized adaptive testing (CAT) using the rating scale model. *Educational and Psychological Measurement*, *57*, 3, 422-439.

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

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*, 5-32.

Clauser, B.E. et K.M. Mazor (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement : Issues and Practices*, *printemps*, 31-44.

Cleary, T.A. et Hilton, T.L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28*, 61-75.

Cliff, N. (1977). A theory of consistency or ordering generalizable to tailored testing. *Psychometrika*, *42*, 375-399.

Cliff, N. (1983). Evaluating Guttman scales : some old and new thoughts. Dans : H. Wainer et S. Messick (dir.), *Principles of modern psychological measurement* (p. 283-301). Hillsdale, NJ : Lawrence Erlbaum Associates.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*, 3, 186-190.

Cohen, L. (1979). Approximate expression for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *32*, 113-120.

Cook, T.D. et Campbell, D.T. (1979). Quasi-experimentation : design and analysis issues for field settings. Boston : Houghton Mifflin.

Crick, J.E. et Brennan, R.L. (1982). *GENOVA : a generalized analysis of variance system*. Dorchester : University of Massachusetts at Boston.

Crocker, L. et Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York : Holt, Rinehart et Winston.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L.J., Gleser, G.C., Nanda, H. et Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York : John Wiley.

Cronbach, L.J. et Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*.

Cronbach, L.J., Rajaratnam, N. et Gleser, G.C. (1963). Theory of generalizability : a liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology*, *16*.

Davey, T., Godwin, J. et Mittelholtz, D. (1997). Developing and scoring an innovative computerized writing assessment. *Journal of Educational Measurement*, *34*, 1, 21-41.

De Champlain, A. et Gessaroli, M.E. (1991). *Assessing test dimensionality using an index based on nonlinear factor analysis*. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Chicago.

De Gruijter, D.N.M. et van der Kamp, L.J.Th. (dir.) (1984). *Advances in psychological and educational measurement*. Londres : John Wiley.

Divgi, D.R. (1980). *Dimensionality of binary items : use of a mixed model*. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Boston.

Dodd, B.G. (1990). The effect of item selection procedure and step size on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement*, *14*, 4, 355-366.

Dodd, B.G., Koch, W.R. et de Ayala, R.J. (1993). Computerized adaptive testing using the partial credit model effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement*, *53*, 1, 61-77.

Donlon, T. F. et Fischer, F.E. (1968). An index of individual's agreement with group-determined item difficulties. *Educational and Psychological Measurement*, *28*, 105-113.

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

Dorans, N.J. et Lawrence, I.M. (1987). *The internal construct validity of the SAT*. Princeton, NJ : Educational Testing Service.

Drasgow F., Levine, M.V. et McLaughlin, M.E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement*, *15*, 2, 171-191.

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

Drasgow, F., Levine, M.V. et Williams, E. (1982). Advances in appropriateness measurement. Manuscrit non publié.

Du Toit, M. (dir.) (2003). *IRT from SSI*. Lincolnwood, IL : Scientific Software International.

Embretson, S.E. (1983). Construct validity : construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 1.

Embretson, S.E. (1985). Multicomponent latent trait models for test desing. Dans : S.E. Embretson (dir.), *Test design, developments in psychology and psychometrics.* Orlando, FL : Academic Press.

Embretson, S.E. (1997). Multicomponent response models. Dans : W.J. Van der Linden et R.K. Hambleton (dir.), *Handbook of modern item response theory*. New York : Springer.

Embretson, S.E. (1999). Generating items during testing : psychometric issues and models. *Psychometrika, 64*, 4, 407-433.

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

Ferguson, A., Myers, C.S., Bartlett, R.J., Banister, H., Bartlett, F.C., Brown, W., Campbell, W., Craik, K.J.W., Drever, N.R., Guild, J., Houstoun, R.A., Irwin, J.O., Kaye, G.W.C., Philpott, S.J.F., Richardson, L.F., Shaxby,

J.H., Smith, T., Thouless, R.H. et Tucker, W.S. (1940). Quantitative estimates of sensory events : final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Advancement of Science, 1*, 339-349.

Fischer, G.H. (1995). The linear logistic test model. Dans : G.H. Fischer et I.W. Molenaar (dir.), *Rasch models : foundations, recent developments, and applications*. New York : Springer-Verlag.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London (A), 222*, 309-368.

Fraser, C. (1988). *NOHARM II : a Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, NSW : University of New England, Center for Behavioral Studies.

Gauthier, N. (2003). Étude de l'effet de l'ordre de difficulté des items, de la longueur du test et de la flexibilité de la révision des items sur la structure factorielle et les scores d'un test de mathématique informatisé. Thèse de doctorat inédite. Québec : Université Laval.

Gierl, M.J., Rogers, W.T. et Klinger, D.A. (1999). Using statistical and judgmental reviews to identify and interpret translation differential item functioning. *The Alberta Journal of Educational Research*, *45*, 4, 353-376.

Glas, C.A.W., Wainer, H. et Bradlow, E.T. (2000). MML and EAP estimation in testlet-based adaptive testing. Dans : W.J. van der Linden et C.A.W. Glas (dir.), *Computerized adaptive testing : theory and practice*. Dordrecht : Kluwer.

Gleser, G.C., Cronbach, L.J. et Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, *30*.

Goldstein, H. (1994). Recontextualizing mental measurement. *Educational Measurement : Issues and Practice, 13*, 1, 16-19.

Goldstein, H. et Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, *42*, 139-167.

Greaud, V.A. (1988). *Some effects of applying unidimensional IRT to multidimensional tests*. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Nouvelle-Orléans.

Green, B.F. (1956). A method of scalogram analysis using summary statistics. *Psychometrika*, *21*, 79-88.

Green, S.B., Salkind, N.J. et Akey, T.M. (2000). *Using SPSS for Windows*. Upper Saddle River, NJ : Prentice-Hall.

Green, S.B., Lissitz, R.W. et Mulaik, S.A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, *37*, 827-838.

Gulliksen, H. (1950). *Theory of mental tests*. New York : John Wiley.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.

Guttman, L. (1950). The basis for scalogram analysis. Dans : S.A. Souffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. et Star, J.A. Claussen (dir.), *Measurement and prediction* (p. 60-90). Princeton, NJ : Princeton University Press.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. Dans : R.L. Linn (dir.), *Educational measurement*, 3e éd. (p. 147-200). New York : Macmillan.

Hambleton, R.K. et Murray L. (1983). Some goodness of fit investigations for item response models. Dans : R.K. Hambleton (dir.), *Applications of item response theory* (p. 71-94). Vancouver : Educational Research Institute of British Columbia.

Hambleton, R.K. et Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement*, *10*, 287-302.

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

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

Hambleton, R.K., Zaal, J.N. et Pieters, J.M.P. (1991). Computerized adaptive testing : theory, applications, and standards. Dans : R.K. Hambleton et J.N. Zaal (dir.), *Advances in educational and psychological testing : theory and applications*. Boston : Kluwer.

Harman, H.H. (1976). Modern factor analysis. Chicago : University of Chicago Press.

Harnisch, D.L. et Linn, R.L. (1981). Analysis of item response patterns : Questionable test data and dissimilar curriculum pratices. *Journal of Educational Measurement, 18*, 133-146.

Harrison, D.A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, *11*, 91-115.

Hattie, J.A. (1984). An empirical study of various indices for detecting unidimensionality. *Multivariate Behavioral Research*, *19*, 49-78.

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

Hattie, J.A., Krakowski, K., Rogers, H.J. et Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, *20*, 1-14.

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

Henning, G.T., Hudson, T. et Turner, J. (1985). Item response theory and the assumption of unidimensionality. *Language Testing*, *2*, 141-154.

Hetter, R.D. et Sympson, J.B. (1997). Item exposure control in CAT-ASVAB. Dans : W.A. Sands, B.K. Waters et J.R. McBride (dir.), *Computerized adaptive testing : from inquiry to application*. Washington : American Psychological Association.

Ho, D.Y.F. (1996). Filial piety and its psychological consequences. Dans : M.H. Bond (dir.), *Handbook of Chinese psychology* (p. 155-165). Honk-Kong : Oxford University Press.

Hoijtink, H. et Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. Dans : G.H. Fischer et I.W. Molenaar (dir.), *Rasch models : foundations, recent developments, and applications*. New York : Springer-Verlag.

Holland, P.W. et Rosenbaum, P.R. (1986). Conditional association and unidimensionality assumption in monotone latent variable models. *The Annals of Statistics*, *14*, 1523-1543.

Holland, P.W. et Thayer, D.T. (1985). *An alternative definition of the ETS delta scale of item difficulty* (ETS Research Report No. 85-43). Princeton, NJ : Educational Testing Service.

Holland, P.W. et Thayer, D.T. (1986). *Differential item functionning and the Mantel-Haenszel procedure*. Technical Report. Princeton, NJ : Educational Testing Service.

Holland, P.W. et Thayer, D.T. (1988). *Stability of the MH D-DIF statistics across populations* (PRPC Report). Princeton, NJ : Educational Testing Service.

Hulin, C.L., Drasgow, F. et Parsons, C.K. (1983). *Item response theory : applications to psychological measurement.* Homewood, IL : Dow Jones Irwin.

Hulin, C.L., Lissak, R.I. et Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristics curves : A Monte Carlo study. *Applied Psychological Measurement*, *6*, 249-260.

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

Hutten, L. (1980). *Some empirical evidence for latent trait model selection*. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Boston.

Ip, E.H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, *66*, 109-132.

Janssen, R., Hoskens, M. et DeBoeck, P. (1991). A test of Embretson's multicomponent model on vocabulary items. Dans : R. Steyer et K. Wideman (dir.), *Psychometric Methodology* (p. 187-190). Stuttgart : Springer-Verlag.

Jensema, C.J. (1974). An application of latent trait mental test theory. *British Journal of Mathematical and Statistical Psychology*, *27*, 29-48.

Jensema, C.J. (1976). A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement*, *36*, 705-715.

Jensema, C.J. (1977). Bayesian tailored testing and the influence of item bank characteristics. *Applied Psychological Measurement*, *1*, 1, 111-120.

Joe, G. et Woodward, J. (1976). Some developments in multivariate generalizability. *Psychometrika, 41*.

Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23*, 187-200.

Kaiser, H.F. (1970). A second-generation Little Jiffy. *Psychometrika*, *35*, 401-415.

Karabatsos, G. (1999). *Rasch vs. two- and three-parameter logistic model*. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Montréal.

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

Kim, H.R. et Stout, W.F. (1993). *A robustness study of ability estimation in the presence of latent trait multidimensionality using the Junker/Stout index of dimensionality*. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Atlanta.

Kingsbury, G.G. et Weiss, D.J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. Dans : D.J. Weiss (dir), *New horizons in testing : latent trait test theory and computerized adaptive testing*. New York : Academic Press.

Kingston, N.M. et Dorans, N.J. (1985). The analysis of item-ability regressions : an exploratory IRT model fit tool. *Applied Psychological Measurement*, *8*, 147-154.

Klauer, K.C. (1995). The assessment of person fit. Dans G.H. Fischer et I.W. Molenaar (dir.), *Rasch models, foundations, recent developments and applications*, New York : Springler-Verlag, 97-110.

Kline, P. (1994). *An easy guide to factor analysis*. Londres : Routledge.

Krantz, D.H., Luce, R.D., Suppes, P. et Tversky, A. (1971). *Foundations of measurement : additive and polynomial representations.* San Diego : Academic Press.

Kuder, G.F. et Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

Kuhn, T.S. (1983). *La structure des révolutions scientifiques*. Paris : Flammarion.

Lapointe, A.E., Mead, N.A. et Askew, J.M. (1992). *Learning mathematics.* Princeton, NJ : Educational Testing Service.

Laurier, M. (1993a). *Les tests adaptatifs en langue seconde*. Communication lors de la 16e session d'étude de l'ADMÉÉ à Laval. Montréal : Association pour le développement de la mesure et de l'évaluation en éducation.

Laurier, M. (1993b). *L'informatisation d'un test de classement en langue seconde*. Québec : Université Laval, Faculté des lettres.

Laurier, M. (1993c). Un test adaptatif en langue seconde : la perception des apprenants. Dans : R. Hivon (dir), *L'évaluation des apprentissages*. Sherbrooke : Éditions du CRP.

Laurier, M. (1996). Pour un diagnostic informatisé en révision de texte. *Mesure et évaluation en éducation*, *18*, 3, 85-106.

Laurier, M. (1998). Méthodologie d'évaluation dans des contextes d'apprentissage des langues assistés par des environnements informatiques multimédias. *Études de linguistique appliquée*, *A110*, 247-255.

Laurier, M. (1999a). Testing adaptatif et évaluation des processus cognitifs. Dans. C. Depover et B. Noël (dir.), *L'évaluation des compétences et des processus cognitifs : modèles, pratiques et contextes*. Bruxelles : De Boeck Université.

Laurier, M. (1999b). The development of an adaptive test for placement in French. *Studies in Language Testing*, *10*, 122-135.

Laurier, M., Froio, L., Paero, C. et Fournier, M. (1999). *L'élaboration d'un test provincial pour le classement des étudiants en anglais langue seconde, au collégial*. Québec : Ministère de l'Éducation, Direction générale de l'enseignement collégial.

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

Laveault, D. et Grégoire, J. (2002). *Introduction aux théories des tests en sciences humaines (2e édition)*. Paris : De Boeck.

Lawley, D.N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh*, *61*, 273-287.

Lazarsfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. Dans : S.A. Souffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Staret et J.A. Claussen (dir.), *Measurement and prediction* (p. 362-412). Princeton, NJ : Princeton University Press.

Lazarsfeld, P.F. (1959). Latent structure analysis. Dans : S. Koch (dir.), *Psychology : a study of science (Vol. 3),* New York, McGraw-Hill.

Lazarsfeld, P.F. et Henry, N.W. (1968). *Latent structure analysis*. Boston : Houghton Mifflin.

Leary, L.F. et Dorans, N.J. (1985). Implications for altering the context in which test items appear : a historical perspective on an immediate concern. *Review of Educational Research*, *55*, 387-413.

Leclerc, M., Bertrand, R. et Dufour, N. (1986). Correlations between teaching practices and class achievement in introductory algebra. *Teaching and Teacher Education*, *2*, 4, 355-365.

Levine, M.V. et Drasgow, F. (1982). Appropriateness measurement : review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.

Levine, M.V. et Drasgow F. (1988). Optimal appropriateness measurement. *Psychometrika*, *53*, 2, 161-176.

Levine, M.V. et Rubin, D.F. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

Linacre, J.M. (2000). *Computer-adaptive testing : a methodology whose time has come*. MESA memorandum no 69. Chicago : MESA Psychometric Laboratory, University of Chicago.

Linacre, J.M. et Wright, B.D. (1995). *A user's guide to BIGSTEPS*. Chicago : Mesa Press.

Linn, R.L. (dir.) (1989). *Educational measurement* (3e éd.). New York : Macmillan.

Linn, R.L., Levine, M.V., Hastings, C.N. et Wardrop, J.L. (1981). Item bias in a test of reading comprehension. *Applied Psychological Measurement, 5*, 159-173.

LLabre, M.M. (1980). Estimating variance components with unbalanced designs in generalizability theory. Boston, AERA.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monograph,* 61.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *9*.

Longford, N.T. (1985). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Manuscrit inédit.

Lord, F.M. (1952). A theory of test scores. *Psychometric Monographs*, no 7.

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

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

Luce, R.D., Krantz, D.H., Suppes, P. et Tversky, A. (1990). *Foundations of measurement : representation, axiomatization, and invariance*. San Diego : Academic Press.

Luecht, R.M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, *20*, 389-404.

Lumsden, J. (1957). A factorial approach to unidimensionality. *Australian Journal of Psychology*, *9*, 105-111.

Mantel, N. et Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Mantel, N. et Haenszel, W. (1959). Statistical aspects of the retrospective study of disease. *Journal of the National Cancer Institute*, *11*, 3-31.

Marcoulides, G.A. (1986). Alternative methods for non-negative variance component estimation : Applications to generalizability theory. Manuscrit inédit. Los Angeles : University of California.

Martin, O. (1999). *La mesure de l'esprit*. Paris : L'Harmattan.

Martin-Lof, P. (1974). The notion of redundancy and its use as a quantitative measure of the discrepancy between a statistical hypothesis and a set of observational data. *Scandinavian Journal of Statistics*, *1*, 3-18.

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

Masters, G.N. et Wright, B.D. (1997). The partial credit model. Dans : W.J. van der Linden et R.K. Hambleton, *Handbook of modern item response theory* (p. 101-122). New York : Springer.

McArthur, D.L. (1987). Analysis of patterns : the S-P technique. Dans : D.L. McArthur (dir.), *Alternative approaches to the assessment of achievement*. Boston : Kluwer Academic.

McArthur, D.L. (dir.). (1987). *Alternative approaches to the assessment of achievement*. Boston : Kluwer Academic.

McBride, J.R. et Martin, J.T. (1983). Reliability and validity of adaptive tests in a military setting. Dans : D.J. Weiss (dir) : *New horizons in testing : latent trait test theory and computerized adaptive testing*. New York : Academic Press.

McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric Monograph*, nº 15.

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

McDonald, R.P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, *6*, 379-396.

McDonald, R.P. (1985). Unidimensional and multidimensional models for item response theory. Dans : D. J. Weiss (dir.), *Proceedings of the 1982 Item Response Theory and Computer Adapative Testing Conference*. Minneapolis : University of Minnesota.

McDonald, R.P. (1997). Normal-ogive multidimensional model. Dans : W.J. van der Linden et R.K. Hambleton (dir.), *Handbook of modern item response theory* (p. 258-270). New York : Springer.

McDonald, R.P. (1999). *Test theory*. Mahwah, NJ : Lawrence Erlbaum Associates.

McDonald, R.P. et Ahlawat, K.S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology*, *27*, 82-99.

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

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, *43*, 289-374.

Meijer, R.R, Molenaar, I.W. et Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, *18*, 2, 111-120.

Meredith, W. et Kearns, J. (1973). Empirical Bayes point estimate of latent trait scores without knowledge of the trait distribution. *Psychometrika*, *38*, 533-554.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 9, 74-149.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*.

Messick, S. (1988). The once and future issues of validity : assessing the meaning and consequences of measurement. Dans : H. Wainer et H.I. Braun (dir.), *Test validity*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. Dans : R.L. Linn (dir.) *Educational measurement* (3e éd.). New York : Macmillan.

Michell, J. (1999). Measurement in psychology : critical history of a methodological concept. New York : Cambridge University Press.

Microsoft (2000). Adaptive testing. <www.windowsgalore.com/cert/adaptive_testing>, accessible le 27 août.

Mislevy, R.J. et Bock, R.D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement*, *42*, 2, 725-737.

Mislevy, R.J. et Bock R.D. (1990). *BILOG-3 : item analysis and test scoring with binary logistic models*. Mooresville, *IN*: Scientific Software Inc.

Mokken, R.J. (1997). Nonparametric models for dichotomous responses. Dans : W.J. van der Linden et R.K. Hambleton, *Handbook of modern item response theory*. New York : Springer.

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

Molenaar, I.W., Debets, P., Sijtsma, K. et Hemker, B.T. (1994). Guide de l'usager pour le logiciel *MSP*. Groningen, Pays-Bas : iecProGAMMA.

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

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

Muraki, E. (1997). A generalized partial credit model. Dans : W.J. van der Linden et R.K. Hambleton, *Handbook of modern item response theory*. New York : Springer.

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

Muraki, E. et Carslon, J.E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*, 73-90.

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 551-560.

Muthen, B. (1984). A general structure equation model with dichotomous, ordered category, and latent variable indicators. *Psychometrika*, *49*, 115-132.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses : comparisons of different approaches. *Journal of Educational Measurement*, *31*, 17-35.

Nandakumar, R. et Stout, W. (1993). Refinement of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*, 41-68.

Nering, J. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*, 2, 121-129.

Neyman, J. et Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, *16*, 1-32.

Nunnaly, J.C. (1978). *Psychometric theory*. New York : McGraw-Hill.

Orlando, M. et Thissen, D. (2000). New item fit indices for dichotomus item response theory model. *Applied Psychological Measurement*, *24*, 50-64.

Owen, R. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351-356.

Programme d'indicateurs du rendement scolaire. (1993). *Rapport sur l'évaluation en mathématique I*. Toronto : Conseil des ministres de l'Éducation du Canada.

Programme d'indicateurs du rendement scolaire. (1997). *Rapport sur l'évaluation en mathématique II*. Toronto : Conseil des ministres de l'Éducation du Canada.

Programme d'indicateurs du rendement scolaire. (2001). *Rapport sur l'évaluation en mathématique III*. Toronto : Conseil des ministres de l'Éducation du Canada.

Raîche, G. (1994). *La simulation de modèle sur ordinateur en tant que méthode de recherche : le cas concret de l'étude de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt*. Actes du 6e colloque de l'Association pour la recherche au collégial. Montréal : Association pour la recherche au collégial.

Raîche, G. (2000). *La distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt : selon l'erreur-type et selon le nombre d'items administrés*. Thèse de doctorat inédite. Montréal : Université de Montréal.

Raîche, G. (2001a). *Principes et enjeux du testing adaptatif : de la loi des petits nombres à la loi des grands nombres*. Communication présentée dans le cadre du 69e congrès de l'Association canadienne française pour l'avancement de la science. Sherbrooke : ACFAS.

Raîche, G. (2001b). *Pour une évaluation sur mesure des étudiants : défis et enjeux du testing adaptatif*. Communication présentée dans le cadre de la 23e session d'études de l'Association pour le dévelopement de la mesure et de l'évaluation en éducation. Québec : ADMÉÉ.

Raîche, G. et Blais, J.G. (2002a). Étude de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt dans le contexte de l'application du modèle de Rasch. *Mesure et évaluation en éducation*, 24 (2-3).

Raîche, G. et Blais, J.G. (2002b). *Practical considerations about expected a posteriori estimation in adaptive testing : adaptive* a priori*, adaptive correction for bias, and adaptive integration interval*. Communication présentée au 11th Biannual International Objective Measurement Workshop. Nouvelle Orléans : IOMW.

Raju, N.S., van der Linden, W.J. et Fleer, P.F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19,* 4, 353-368.

Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*, 4.

Ramsay, J.O. (1993). *TESTGRAF. Programme informatique pour l'analyse nonparamétrique des réponses aux items d'un test*. Montréal : Université McGill.

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

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

Reckase, M.D. (1990). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests*. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Boston.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. Dans : W.J. van der Linden et R.K. Hambleton, *Handbook of modern item response theory*. New York : Springer.

Reckase, M.D. (1998). Converting boundaries between National Assessment Governing Board performance categories to points on the National Assessment of Educational Progress score scale : the 1996 science NAEP process. *Applied Measurement in Education, 11,* 9-21.

Reckase, M.D. et McKinley, R.L. (1983). *The definition of difficulty and discrimination for multidimensional item response theory models*. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Montréal.

Rosenbaum, P.R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*, 425-435.

Rosenbaum, P.R. (1985). Comparing distributions of item response for two groups. *British Journal of Mathematical and Statistical Psychology*, *38*, 206-215.

Roussos, L.A., Schnipke, D.L. et Pashley, P.J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24*, 3, 292-322.

Rudner, L.M. (1977). An approach to biased item identification using latent trait measurement theory. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, New York.

Rulon, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9*, 99-103.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometric Monograph No. 17. Iowa City : Psychometric Society.

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

Samejima, F. (1997). Graded response model. Dans : W.J. van der Linden et R.K. Hambleton, *Handbook of modern item response theory*. New York : Springer.

Sato, T. (1975). *The construction and interpretation of S-P tables*. Tokyo : Meiji Tosho.

Schmitt, N., Cortina, J.M. et Whitney, D.J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement, 17*, 2, 143-150.

Shavelson, R.J. et Webb, N.M. (1991). *Generalizability theory : a primer*. Newbury Park, CA : Sage.

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

Sijtsma, K. (1998). Methodology review : nonparametric IRT approaches to the analysis of dichotomous items scores. *Applied Psychological Measurement*, *22*, 3-31.

Sirotnik, K.A. (1987). Toward more sensible achievement measurement : a retrospective. Dans : D.L. McArthur (dir.), *Alternative approaches to the assessment of achievement*. Boston : Kluwer Academic.

Smith, P. (1978). Sampling errors of variance components in small multifacet generalizability studies. *Journal of Educational Statistics*, *3*.

Smith, P. (1980). Some approaches to determining the stability of estimated variance components. Boston, AERA.

Smith, P.C., Kendall, L.M. et Hulin, C.L. (1969). *The measurement of satisfaction in work and retirement*. Skokie, IL : Rand McNally.

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

Stevens, S.S. (1951). *Handbook of experimental psychology*. New York : Wiley.

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

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

Stout, W., Nandakumar, R., Junker, B., Chang, H. et Steidinger, D. (1991). *DIMTEST and TESTSIM, programs for dimensionality testing and test simulation*. University of Illinois at Urbana-Champaign, Département de Statistique.

Suen, H.K. (1990). *Principles of test theories*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Suppes, P., Krantz, D.H., Luce, R.D. et Tversky, A. (1989). *Foundations of measurement : geometrical, threshold, and probability representations.* San Diego : Academic Press.

Swaminathan, H. et Gifford, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, *7*, 175-191.

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

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

Swygert, K.A., McLeod, L.D. et Thissen, D. (2001). Factor analysis for items or testlets scored in more than two categories. Dans : D. Thissen et H. Wainer (dir.), *Test scoring* (p. 217-259). Mahwah, NJ : Lawrence Erlbaum Associates.

Thibault, J. (1992). *L'apport de fidélité intra-individuelle de trois modes de conception distincts estimés selon le modèle logistique à trois paramètres et selon le modèle polytomique de Bock-Samejima utilisés en TRI*. Thèse de doctorat. Sainte-Foy : Université Laval.

Thissen, D.B. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika*, *47*, 175-186.

Thissen, D. (1991). *MULTILOG user's guide : multiple categorical item analysis and test scoring using item response theory*. Chicago : Scientific Software International.

Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. Dans : N. Frederiksen, R.J. Mislevy et I.I. Bejar (dir.), *Test theory for a new generation of tests*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Thissen, D. et Mislevy, R.J. (2000). Testing algorithms. Dans : H. Wainer, D. Eignor, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (dir.), *Computerized adaptive testing : a primer*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Thissen, D. et Orlando, M. (2001). Item response theory for items scored in two categories. Dans : D. Thissen et H. Wainer (dir.), *Test scoring* (p. 73-140). Mahwah, NJ : Lawrence Erlbaum Associates.

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

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

Thissen, D., Steinberg, L. et Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. Dans : H. Wainer et H.I. Braun (dir.), *Test Validity*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L. et Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. Dans : P.W. Holland et H. Wainer (dir.), *Differential item functioning* (p. 67-114). Hillsdale, NJ : Lawrence Erlbaum Associates.

Thissen, D. et Wainer, H. (dir.) (2001). *Test scoring*. Mahwah, NJ : Lawrence Erlbaum Associates.

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

Torgerson, W.S. (1958). *Theory and methods of scaling*. New York : Wiley.

Trabin, T.E. et Weiss, D.J. (1983). The person response curve : fit of individuals to item characteristic curve models. Dans : D.J. Weiss (dir.), *New horizons in testing*. New York : Academic Press.

Traub, R.E. (1994). *Reliability for the social sciences.* Newbury Park, CA : Sage.

Urry, V.W. (1970). *A Monte Carlo investigation of logistic mental models*. Thèse de doctorat inédite. West Lafayette, IN : Purdue University.

Urry, V.W. (1974). Approximation to item parameters of mental test models and their uses. *Educational and Psychological Measurement*, *34*, 253-269.

Van de Vijver, F.J.R. et Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA : Sage.

Van den Wollenberg, A. (1988). Testing a latent trait model. Dans : R. Langeheine et J. Rost, *Latent trait and latent class models*. Londres : Plenum Press.

van der Linden, W.J. (1986). The changing conception of testing in education and psychology. *Applied Psychological Measurement*, *10*, 325-352.

van der Linden, W.J. (1994). Fundamental measurement and the fundamentals of Rasch measurement. Dans : M. Wilson (dir.). *Objective measurement : theory into practice, vol. 2*. Norwood, NJ : Ablex.

van der Linden, W.J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement*, *20*, 373-388.

van der Linden, W.J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied Psychological Measurement*, *23*, 1, 21-29.

van der Linden, W.J. (2000). Constrained adaptive testing with shadow tests. Dans : W.J. van der Linden et C.A.W. Glas (dir.), *Computerized adaptive testing : theory and practice*. Dordrecht : Kluwer.

van der Linden, W.J. et Glas, C.A.W. (dir.) (2000). *Computerized adaptive testing : theory and practice*. Dordrecht : Kluwer.

van der Linden, W.J. et Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York : Springer.

van der Linden, W.J. et Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. Dans : W.J. van der Linden et C.A.W. Glas (dir.), *Computerized adaptive testing : theory and practice*. Dordrecht : Kluwer.

vos, H.J. et Glas, C.A.W. (2000). Testlet-based adaptive mastery testing. Dans : W.J. van der Linden et C.A.W. Glas (dir.), *Computerized adaptive testing : theory and practice*. Dordrecht : Kluwer.

Wainer, H. (1983). Are we correcting for guessing in the wrong direction ? Dans : D.J. Weiss (dir.) : *New horizons in testing : latent trait test theory and computerized adaptive testing*. New York : Academic Press.

Wainer, H., Bradlow, E.T. et Du, Z. (2000). Teslet response theory : an analog for the 3PL model useful in testlet-based adaptive testing. Dans : W.J. van der Linden et C.A.W. Glas (dir.), *Computerized adaptive testing : theory and practice*. Dordrecht : Kluwer.

Wainer, H., Dorans, N.J., Green, B.F., Mislevy, R.J., Steinberg, L. et Thissen, D. (1990). Future challenges. Dans : H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (dir.), *Computerized adaptive testing : a primer*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Wainer, H. et Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. Dans : H. Wainer, D. Eignor, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (dir.), *Computerized adaptive testing : a primer*. Hillsdale, NJ : Lawrence Erlbaum Associates.

Wainer, H. et Kiely, G.L. (1987). Item clusters and computerized adaptive testing : a case for testlets. *Journal of Educational Measurement*, *24*, 3, 185-201.

Wainer, H. et Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*, 4, 339-368.

Wainer, H. et Thissen, D. (2001). True score theory : the traditional method. Dans : D. Thissen. et H. Wainer (dir.), *Test scoring*. Mahwah, NJ : Lawrence Erlbaum Associates.

Wainer, H. et Thissen, D. (2001). *Test scoring*. Mahwah, NJ : Lawrence Erlbaum Associates.

Wainer, H. et Wright, B. (1980). Robust estimation of ability in the Rasch model. *Psychometrika*, *45*, 371-390.

Wang, M. (1988). *Measurement bias in the application of a unidimensional model to multidimensional item-response data*. Texte présenté à l'occasion de la rencontre annuelle de l'American Educational Research Association, Nouvelle-Orléans.

Warm, T.A. (1978). *A primer of item response theory (technical report 941278)*. Oklahoma City : U.S. Coast Guard Institute.

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

Webb, N.M. (1987). Generalizability theory and achievement testing. Dans : D.L. McArthur (dir.), *Alternative approaches to the assessment of achievement*. Boston : Kluwer Academic.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 4, 473-492.

Weiss, D.J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology*, *53*, 6, 774-789.

Whitely, S.E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.

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

Wilson, D., Wood, R. et Gibbons, R.D. (1987). *TESTFACT : test scoring, item statistics and factor analysis.* Mooresville, IN : Scientific Software Inc.

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

Wise, S.L. (1983). Comparisons of order analysis and factor analysis in assessing the dimensionality of binary data. *Applied Psychological Measurement*, *7*, 311-312.

Woodruff, D. (1990). Conditional standard error of measurement in prediction. *Journal of Educational Measurement, 27*, 191-208.

Wright, B.D. (1997). A history of social science measurement. *Educational Measurement : Issues and Practice, 16,* 4, 33-45.

Wright, B.D. et Linacre, J.M. (1991). *Winsteps Rasch measurement computer program.* Chicago : MESA Press.

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

Wright, B.D., Mead, R.J. et Draba, R.E. (1976). Detecting and correcting test item bias with a logistic response model. *Research memorAndum No. 22*, Statistical Laboratory. Chicago : University of Chicago.

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

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

Wu, M.L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalized item response models.* Mémoire de maîtrise inédit. Melbourne : University of Melbourne.

Wu, M.L., Adams, R.J. et Wilson, M.R. (1998). *CONQUEST : Generalised item response modelling software.* Melbourne : Australian Council for Educational Research.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

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

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

Zickar, M.J. et Drasgow F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20,* 1, 71-87.

Zimowski, M.F., Muraki, E., Mislevy, R.J. et Bock, R.D. (1996). *BILOG-MG : Multiple-group IRT analysis and test maintenance for binary items.* Chicago : Scientific Software.

Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24,* 293-308.

Index