Université de Montréal

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

par
Gilles Raîche

Option mesure et évaluation en éducation
Faculté des sciences de l'éducation

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de
Philosophiae Doctor (Ph.D.)

Octobre, 2000

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

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

présentée par :

Gilles Raîche

a été évaluée par un jury composé des personnes suivantes :

Michel Laurier, Université de Montréal, président du jury
Jean-Guy Blais, Université de Montréal, directeur de recherche
Djavid Ajar, Université de Montréal, membre du jury
Richard Bertrand, Université Laval, examinateur externe

Thèse acceptée le :

Sommaire

Cette recherche s'intéresse à l'application des modélisations issues de la théorie de la réponse à l'item au testing adaptatif par ordinateur. Plus spécifiquement, l'impact de la variation des critères retenus pour la règle d'arrêt sur la distribution de probabilité de certaines statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif est étudié. Les règles d'arrêt considérées sont de deux types : selon l'erreur type de l'estimateur du niveau d'habileté et selon le nombre d'items administrés.

Pour 2000 valeurs du niveau d'habileté obtenues de façon aléatoire, des tests adaptatifs sont simulés pour chacune des valeurs retenues pour les règles d'arrêt : soit de 0,20 à 0,85 par sauts de 0,05 pour la règle d'arrêt selon l'erreur type et de 1 à 60 pour la règle d'arrêt selon le nombre d'items administrés. La modélisation de la réponse à l'item est effectuée à partir du modèle logistique à un paramètre et la stratégie d'estimation du niveau d'habileté employée est la méthode de l'espérance a posteriori.

Les résultats des simulations démontrent l'impact de la variation des critères retenus pour la règle d'arrêt sur la distribution de probabilité des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Par conséquent, la distribution d'échantillonnage de l'estimateur du niveau d'habileté est affectée par la variation des

valeurs retenues pour les règles d'arrêt.

Compte tenu des résultats obtenus, il est recommandé d'utiliser la correction de l'estimateur du niveau d'habileté proposée par Bock et Mislevy quelle que soit la règle d'arrêt employée. La règle d'arrêt selon l'erreur type ne doit être utilisée que si l'erreur type retenue est au plus de 0,40. En autant que le niveau d'habileté n'excède pas $\pm 3,00$, l'emploi d'une erreur type retenue pour la règle d'arrêt égale ou inférieure à 0,20 permet toutefois de s'assurer, à toutes fins utiles, que la précision de l'estimateur du niveau d'habileté est constante sur toute l'étendue du niveau d'habileté et que le biais de l'estimateur du niveau d'habileté est nul. Quand à la règle d'arrêt selon le nombre d'items administrés, celle-ci ne doit être appliquée que lorsqu'un minimum de 13 items sont administrés. Cependant l'administration de 40 items ou plus garantie que le biais de l'estimateur du niveau d'habileté soit pratiquement nul quand la correction de Bock et Mislevy est appliquée et que le niveau d'habileté se situe entre -3,00 et 3,00 ; cela sans trop affecter les distributions de probabilité et d'échantillonnage de l'estimateur du niveau d'habileté.

Table des matières

Liste des tableaux

Liste des figures

Liste des équations

1. Introduction

Selon les habitudes développées au 20e siècle pour évaluer les apprentissages réalisés par un étudiant, faire un diagnostic de ses problèmes d'apprentissage, ou encore le classer à l'intérieur d'un groupe pour qu'il puisse recevoir un enseignement approprié, un test papier crayon (Dechef et Laveault, 1999, p. 152), est très souvent administré. 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.

Nous jugeons important de porter à l'attention du lecteur les chevauchements de sens éventuels entre les termes "estimateur", "estimé" et "estimation" dans la pratique de la langue française ; nous les évitons en nous abstenant d'utiliser le terme "estimé" en tant que nom et en adoptant tout de suite une définition univoque pour "estimateur" et "estimation". Ainsi, tout au long de ce texte, l'expression "estimation du niveau d'habileté" est utilisée pour désigner une action qui permet d'obtenir une valeur approchée du niveau d'habileté. Plus spécifiquement, il s'agit d'une méthode d'estimation à partir de laquelle est obtenue la valeur approchée du niveau d'habileté, laquelle est définie comme l'estimateur du niveau d'habileté.

Généralement, le niveau d'habileté d'intérêt est d'ordre cognitif : connaissances en mathématiques, en français, etc. Il peut toutefois être d'ordre affectif ; le niveau d'habileté est alors en lien avec une attitude (Morissette, 1984, p. 303). Il peut aussi être d'ordre psychomoteur, et le niveau d'habileté vise ainsi un comportement moteur (Nadeau, 1988, p. 265). 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.

Tel que le fait remarquer Wainer (1990, p. 6), au 20^{e} siècle un changement majeur apparaît 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 (Hambleton,

Swaminathan et Rogers, 1991, p. 2-5 ; Laurier, 1993b, p. 49-56 ; Wainer, 1990, p. 11) 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 par ordinateur. Nous tenons à signaler que les auteurs cités ont plutôt tendance à présenter les avantages du testing adaptatif par ordinateur en sous-entendant seulement les problèmes et limites associés au test papier crayon fixe et invariable. Nous avons préféré aborder spécifiquement ces problèmes et ces limites avant de présenter la solution proposée à ceux-ci. Nous évoquons ensuite des interrogations auxquelles cette recherche tente de répondre concernant la distribution de probabilité de diverses statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté obtenu à partir de cette solution.

1.1 Problèmes de précision de l'estimateur du niveau d'habileté dans les tests papier crayon fixes et invariables

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 (Laurier, 1993b, p. 41 ; Wainer, 1990, p. 9). 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. Comme le souligne Laurier (1993b, p. 165), 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, soit lorsque les items sont trop difficiles, l'étudiant peut se décourager au point de ne pas compléter le test dans certains cas. Que les items soient trop faciles ou qu'ils soient 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 (Laveault et Grégoire, 1997, p. 304).

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 (Weiss, 1985, p. 775-776). 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.

1.2 Limites d'administration des tests papier crayon

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 (Hambleton, Swaminathan et Rogers, 1991, p. 147). Ainsi, pour les tests à fonction diagnostique ou formative, qui nécessitent le plus souvent une rapide rétroaction, 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 (Wainer, 1990, p. 11). 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, comme nous l'indiquent Green (1990, p. 37), Sands et Waters (1997, p. 9) et Wainer (1990, p. 11), 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 (Laurier, 1993b, p. 40).

1.3 Testing adaptatif par ordinateur

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 par ordinateur (TAO). Le testing adaptatif par ordinateur est une forme de testing sur mesure (*tailored testing*) (Laveault et Grégoire, 1997, p. 304) spécifiquement adaptée à la personne à qui on administre le test. Selon l'avis de plusieurs auteurs (Bock et Mislevy, 1982, p. 431 ; Bunderson, Inouye et Olsen, 1989, p. 382 ; Dodd, de Ayala et Koch, 1995, p. 5 ; Dodd, Koch et de Ayala, 1993, p. 61 ; Hambleton, Swaminathan et Rogers, 1991, p. 146), l'utilisation en éducation du testing adaptatif par ordinateur a été facilitée par l'introduction de propositions de modélisation de la réponse à l'item différentes de celles proposées dans le contexte de la théorie classique des tests. Il s'agit de propositions issues de la théorie de la réponse à l'item, intitulée au départ théorie du trait latent. Selon ces mêmes auteurs, 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 de la réponse à l'item. Déjà en 1977, Urry croyait que le testing adaptatif par ordinateur serait potentiellement l'une des applications les plus utiles des nouvelles propositions de modélisation.

> *«Le testing adaptatif représente une application remarquablement utile de la théorie du trait latent. ...Toutefois, celui-ci devrait être connu d'un plus grand nombre de personnes de façon à ce que ses avantages potentiels quant à l'amélioration de la mesure puissent être appréciés*

(Urry, 1997, p. 181).»[1]

Plusieurs programmes de testing à grande échelle (*large scale testing*), soit des épreuves communes (Davaud et Cardinet, 1992) à des populations complètes, utilisent des versions adaptatives par ordinateur de leurs tests (Dodd et *al.*, 1995, p. 5 ; Meijer et Nering, 1999, p. 187 ; Zwick, Thayer et Wingerski, 1994, p. 121). 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 Board of Nursing*). D'autres organismes emboîtent le pas : la *Psychological Corporation*, le *College Board*, l'*American College Testing*, la Société américaine des pathologistes, l'*American Board of Internal Medicine*, le Département de la défense américaine, 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

---

1 *«Tailored testing represents a remarkably effective application of latent trait theory. ... But a wider audience of test practitionners must become familiar with it, so that its considerable potential for improved measurement in many settings may be realized (Urry, 1997, p. 181).»*

d'application tandis que les travaux d'Auger (1989 ; Auger et Séguin, 1992) sur le testing adaptatif de maîtrise en éducation économique au secondaire et de Laurier en révision de texte (1996), en sont des exemples de travail de recherche.

Le testing adaptatif offre plusieurs avantages par rapport aux tests papier crayon fixes et invariables. Selon Dodd et *al.* (1995, p. 5) et McBride (1997, p. 35), 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 Ayala, 1992b, p. 327), de façon à ce que le niveau de difficulté des items à ces tests ne soit ni trop difficile, ni trop facile (Weiss, 1983, p. 5). Hambleton, Swaminathan et Rogers (1991, p. 145) soulignent que 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. Selon Lord (1980b, p. 201), 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 plus approprié d'estimer le niveau d'habileté indépendamment du choix particulier des items d'une version du test (Hambleton et Swaminathan, 1987, p. 296).

Des propositions de modélisation de la réponse à l'item, 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.

Plusieurs caractéristiques des tests adaptatifs ont reçu une attention particulière et ont été conséquemment l'objet d'études. Ainsi, certains auteurs ont effectué des comparaisons entre l'estimateur du niveau d'habileté obtenu à partir de différentes propositions de modélisation de la réponse à l'item (de Ayala, 1989, 1992b ; Dodd et *al.*, 1995) et de

différentes méthodes d'estimation (Chen, Hou et Dodd, 1998 ; van der Linden, 1999). D'autres ont étudié l'influence de la dimensionnalité de la banque d'items (de Ayala, 1992a), la conformité au postulat d'indépendance locale (Mislevy et Chang, 1998, 2000) ou des caractéristiques de la banque d'items et des différentes règles d'arrêt sur l'estimateur du niveau d'habileté (Dodd et *al.*, 1993). La comparaison de certaines règles de sélection des items a été effectuée (Chang et Ying, 1996, 1999 ; Eggen, T.J.H.M., 1999 ; Schnipke et Green, 1995 ; Veerkamp et Berger, 1997, 1999) comme celle des méthodes pour évaluer le fonctionnement différentiel des items (*differential item functionning*, DIF) (Zwick, 1997 ; Zwick et *al.*, 1994) ou les indices d'ajustement de l'estimateur du niveau d'habileté (*person fit*) (Nering, 1997 ; van Krimpen-Stoop et Meijer, 1999). Cependant plusieurs aspects du testing adaptatif par ordinateur restent à étudier.

1.4 Objectif de la recherche

Parmi les aspects du testing adaptatif par ordinateur qui restent à étudier, les caractéristiques de la distribution de probabilité des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté sont particulièrement importantes puisque leur connaissance et leur compréhension sont nécessaires à l'interprétation de l'estimateur du niveau d'habileté. Ces caractéristiques permettent de se prononcer sur le sens à donner à l'estimateur du niveau d'habileté obtenu en testing adaptatif.

Puisque nous traitons à la fois de distribution de probabilité et de distibution d'échantillonnage, il nous semble utile de rappeler quelques précisions sur la nature, d'un côté, d'une distribution de probabilité et, de l'autre, d'une distribution d'échantillonnage. Une distribution de probabilité associe à chaque valeur possible d'une variable, dite aléatoire, une probabilité de réalisation. Selon la distribution de probabilité utilisée, diverses statistiques peuvent être considérées. Une distribution normale $N(\mu, \sigma)$, par exemple, est caractérisée par sa moyenne et son écart type. De plus, les coefficients d'asymétrie et de kurtose (aplatissement) associés à une distribution normale sont tous deux nuls.

Une distribution d'échantillonnage, pour sa part, correspond à une distribution de probabilité bien spécifique, soit celle d'une statistique obtenue à partir d'un échantillon. Elle permet de connaître la distribution de probabilité d'une statistique obtenue à partir de plusieurs échantillons présentant les mêmes caractéristiques. La distribution d'échantillonnage associe ainsi une probabilité de réalisation à chaque valeur de cette statistique. Cette distribution d'échantillonnage peut aussi correspondre à une distribution normale $N(\mu_\mu, \sigma_\mu)$. À ce moment, la moyenne de cette distribution devient la moyenne de toutes les moyennes tandis que l'écart type associé devient l'erreur type de la moyenne.

Pour les fins de cette recherche, une seule distribution d'échantillonnage est étudiée : il s'agit de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing

adaptatif. Toutefois, la distribution de probabilité de plusieurs statistiques associées à cette distribution d'échantillonnage est également analysée.

Il est aussi important de connaître les caractéristiques de la distribution de probabilité des diverses statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté : moyenne, erreur type, biais, asymétrie, kurtose, proportion de bonnes réponses et nombre d'items administrés. L'étude de ces caractéristiques nous permet, entre autres, de connaître l'étendue de ces valeurs. La première de ces statistiques est la moyenne de la distribution d'échantillonnage de l'estimateur du niveau d'habileté que nous nommons simplement estimateur du niveau d'habileté, en conformité avec la pratique, étant entendu qu'il s'agit en fait de sa moyenne. Il est important de vérifier les caractéristiques de la distribution de probabilité de l'estimateur du niveau d'habileté pour déterminer si cette distribution reproduit exactement celle du niveau d'habileté. Si ce n'était pas le cas, nous pourrions nous interroger sur la pertinence de la méthode d'estimation du niveau d'habileté utilisée.

Si la distribution d'échantillonnage de l'estimateur du niveau d'habileté se distribue selon une loi de probabilité normale, la précision de cet estimateur, lorsqu'elle est mesurée par son erreur type, permet de déterminer un intervalle de confiance autour de l'estimateur du niveau d'habileté. La détermination de cet intervalle de confiance n'est valide que lorsque la distribution d'échantillonnage de l'estimateur du niveau d'habileté est symétrique et qu'elle est mésokurtique, soit ni surélevée ni aplatie.

De plus, on s'attend à ce que l'estimateur du niveau d'habileté obtenu tende vers le niveau d'habileté réel. Dans les faits, est-ce vrai et ne tend-il pas à être toujours soit plus grand, soit plus petit que le niveau d'habileté réel ? En d'autres termes, est-il biaisé ? Et la distribution d'échantillonnage de l'estimateur du niveau d'habilité est-elle influencée par la composition de la banque d'items, par la méthode d'estimation du niveau d'habileté, ou encore par la règle d'arrêt ? Est-ce que ces caractéristiques sont vraies sur tout le continuum du niveau d'habileté ? Et caetera.

En ce sens, les caractéristiques de la distribution de probabilité des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif ont été traitées de façon partielle, ou non spécifique, par tous les auteurs ci-dessus mentionnés. La plupart du temps, cependant, seul le biais ou l'erreur type de l'estimateur du niveau d'habileté a été vraiment l'objet de l'étude de la distribution d'échantillonnage de l'estimateur du niveau d'habileté (Bock et Mislevy, 1982 ; Chen, Hou, Fitzpatrick et Dodd, 1997 ; de Ayala, 1989, 1992b ; Harwell, 1997 ; Roos, Wise et Plake, 1997 ; Vispoel, Rocklin, Wang et Bleiler, 1999 ; Wang, Hanson et Lau, 1999). Dans d'autres cas, seules des représentations graphiques de celle-ci sont présentées (Bock et Mislevy, 1982). Les caractéristiques de la distribution de probabilité des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif n'ont donc pas été étudiées de façon spécifique et systématique. En ce sens, l'asymétrie et la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, caractéristiques importantes d'une distribution de probabilité, seraient à analyser et

l'adéquation de l'interprétation de l'erreur type de l'estimateur du niveau d'habileté serait aussi à vérifier.

De plus, selon Dodd et *al.* (1993), les caractéristiques de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif sont affectées par l'utilisation de différentes règles d'arrêt. Leur étude se limite toutefois à la comparaison de l'utilisation de deux indices différents pour permettre l'arrêt du test : l'erreur type de l'estimateur du niveau d'habileté et la règle du minimum d'information. Ils ne se sont pas intéressés à faire varier l'erreur type désirée de l'estimateur du niveau d'habileté ni le nombre d'items présentés.

L'objectif de cette recherche est précisément d'étudier l'effet de deux règles d'arrêt d'utilisation fréquente sur les caractéristiques de la distribution de probabilité de diverses statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif. Plus exactement, il s'agit de vérifier l'impact de la variation de la règle d'arrêt selon le nombre d'items administrés et de la règle d'arrêt selon la détermination a priori de l'erreur type de l'estimateur du niveau d'habileté sur la distribution de probabilité des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif.

Cette étude nous permettra d'apporter des prescriptions quant aux critères d'application des règles d'arrêt étudiées. Il nous sera ainsi possible d'identifier à quel moment il est

opportun de terminer l'administration d'un test pour obtenir des valeurs satisfaisantes de diverses statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté, donc pour estimer le niveau d'habileté avec suffisamment de précision.

1.5 Organisation du texte

Au deuxième chapitre nous présentons la description et les étapes d'un test adaptatif tout en abordant les développements qui ont présidé à son évolution en éducation. La transition de la forme papier crayon vers l'administration par ordinateur utilisant des propositions différentes de modélisation de la réponse à l'item issues de la théorie de la réponse à l'item est tout spécialement soulignée.

Différentes propositions de modélisation de la réponse à l'item issues de la théorie de la réponse à l'item sont ensuite l'objet du troisième chapitre, tandis que le quatrième chapitre traite des méthodes d'estimation du niveau d'habileté appropriées lorsque ces propositions de modélisation sont utilisées. Les avantages de chacune de ces méthodes d'estimation y sont considérés. Les règles de départ, de suite et d'arrêt sont ensuite abordées au cinquième chapitre ; une place plus importante y est accordée à la description des règles d'arrêt puisqu'elles sont plus spécifiquement l'objet de la recherche.

Le sixième chapitre est consacré à la présentation détaillée des études qui ont apporté une

contribution à la description des caractéristiques de la distribution de probabilité des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction des règles d'arrêt retenues ; ce chapitre se termine par des précisions quant à l'objectif de recherche. La méthodologie proposée pour atteindre l'objectif de recherche est présentée au septième chapitre tandis que les résultats et leur analyse sont traités au chapitre suivant. Enfin, une discussion des résultats et une conclusion terminent la présentation.

2. Description du testing adaptatif

Dans le but de décrire ce qu'est un test adaptatif et de quelle façon il se déroule, nous présentons plusieurs types de tests adaptatifs en utilisant, comme fil conducteur, les développements en éducation qui ont présidé à l'évolution du testing adaptatif. Nous abordons cette évolution en mettant en évidence la transition des tests papier crayon vers l'administration de tests adaptatifs informatisés. Ces derniers reposent toutefois sur une modélisation de la réponse à l'item issue de la théorie de la réponse à l'item qui exige de plus amples explications que celles fournies à l'intérieur de ce chapitre. Nous nous limitons donc ici à introduire les tests adaptatifs informatisés et les propositions de modélisation de la réponse à l'item autour desquelles ils sont construits pour en faire ensuite le sujet spécifique des trois prochains chapitres. Pour les fins de l'illustration du déroulement de chacun de ces tests adaptatifs, nous recourons à une démarche algorithmique proposée par Thissen et Mislevy (1990, p. 103).

2.1 Déroulement d'un test adaptatif

Thissen et Mislevy (1990, p. 103) soulignent que tous les tests, qu'ils soient des tests papier crayon fixes et invariables ou des 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 2.1.

Tableau 2.1
Algorithme décrivant le déroulement normal d'un test papier crayon fixe et invariable

RÈGLE	ACTION
1. Règle de départ	Répondre à une première question, généralement la question #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 lequel 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 avec la première question et répondent

séquentiellement 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 est 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 (Wainer, 1990, p. 10) 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 2.1 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 (Hambleton, Zaal et Pieters, 1991, p. 348). 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.

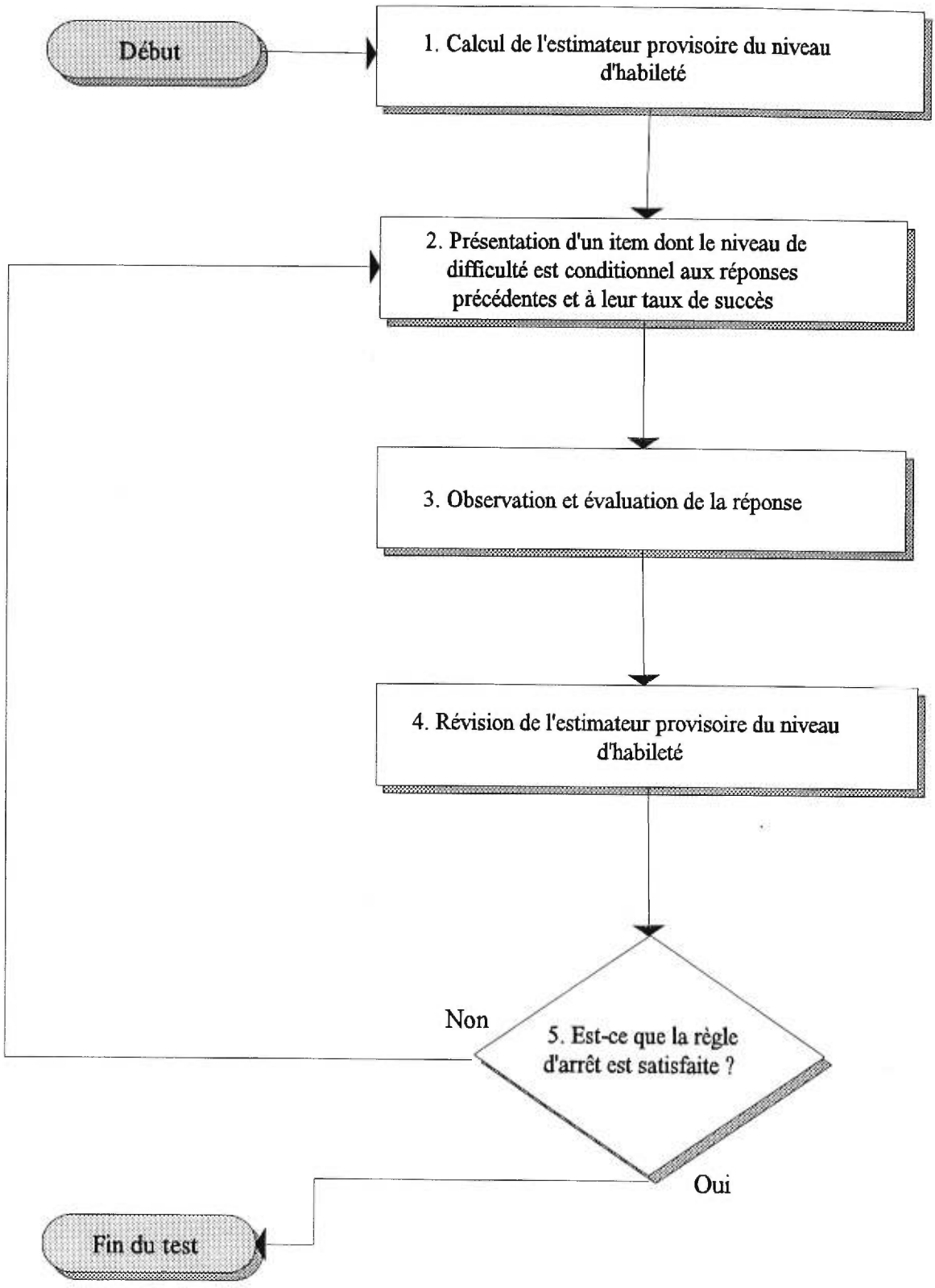

Figure 2.1 Déroulement d'un test adaptatif

2.2 Transformations successives du testing adaptatif

Les grandes étapes du développement du testing adaptatif et de ses transformations successives sont décrites avec pour point de départ le test d'intelligence de Binet auquel Weiss (1982, p. 474) attribue l'application des premiers principes du testing adaptatif en éducation. Selon Weiss (1985, p. 778), il faut toutefois attendre 1950 pour voir apparaître de nouveaux tests adaptatifs, et encore ne sont-ils que des tests expérimentaux utilisés surtout dans le but d'étudier différentes stratégies de testing adaptatif. Ce sont évidemment toujours des tests papier crayon et ils comprennent, pour la plupart, un nombre fixe d'items à administrer.

Dans les pages qui suivent, nous présenterons les types de tests papier crayon généralement cités dans la littérature. Ce sont les tests à deux étapes, à niveaux flexibles et pyramidaux, dont le nombre d'items administrés est fixe. Nous présentons aussi un test dont le nombre d'items administrés est variable, soit le test stratifié. La présentation générale de l'application au testing adaptatif des différentes propositions de modélisation de la réponse à l'item termine ce chapitre, alors que sa description plus détaillée est traitée à l'intérieur des chapitres subséquents.

2.2.1 Test de Binet

Généralement, les premiers principes d'application du testing adaptatif en éducation sont attribués à Binet et à ses collègues, au début des années 1900, lors du développement des tests d'intelligence (Weiss, 1982, p. 474). Le test de Binet, toutefois, n'est pas sous la forme papier crayon, l'administrateur du test prenant en charge la notation des réponses ainsi que la lecture des consignes et des questions. Selon les catégories proposées par Morrisette (1984, p. 229), le test est composé d'items d'examen oral. De plus, les réponses aux items dont est constitué ce test sont construites plutôt que choisies.

Le test d'intelligence de Binet exige qu'au départ l'administrateur détermine, à partir d'informations préalables, l'âge mental de l'enfant. Cette évaluation peut être basée sur des informations contextuelles, tel le groupe d'appartenance, et sur des informations développementales, tel l'âge chronologique. Ainsi, les résultats d'un enfant de huit ans qui présente des problèmes d'apprentissage en classe pourraient laisser croire que son âge mental est inférieur à son âge chronologique. Lors de l'administration du test de Binet, l'administrateur pourrait donc décider de commencer la session en présentant des items dont le niveau de difficulté est approprié à un enfant qui possède un âge mental inférieur à huit ans.

Dans le test de Binet, des items dont le niveau de difficulté est approprié à l'âge mental déterminé au départ sont administrés. Si tous les items dont le niveau de difficulté est

approprié à un âge mental ne sont pas réussis, des items plus faciles, approprié à un âge mental inférieur, sont administrés. Lorsque tous les items d'un même niveau de difficulté sont réussis, il est présumé que l'âge mental de base est identifié. À ce moment, l'administrateur présente des items d'un niveau de difficulté supérieur. Cette stratégie se poursuit jusqu'au moment où le répondant échoue tous les items d'un niveau de difficulté donné. Un estimateur du plafond de l'âge mental est ainsi obtenu et il est présumé que tous les items d'un niveau de difficulté supérieur seraient échoués par le répondant. Le test se termine lorsque l'âge mental de base et le plafond de l'âge mental ont été obtenus. Le calcul de l'âge mental est réalisé en ajoutant un nombre spécifique de mois à l'âge mental de base selon le nombre d'items réussis au-dessus du niveau d'âge mental de base. La démarche algorithmique du test de Binet est présentée au tableau 2.2.

Tableau 2.2
Algorithme décrivant le déroulement du test de Binet

RÈGLE	ACTION
1. Règle de départ	Administrer des items dont le niveau de difficulté correspond à l'âge mental évalué au départ
2. Règle de suite	Administrer des items d'un niveau de difficulté supérieur jusqu'à ce que tous les items d'un niveau aient été échoués. Ensuite, procéder de façon inverse jusqu'à ce tous les items d'un niveau de difficulté inférieur aient été réussis
3. Règle d'arrêt	Terminer l'administration du test lorsqu'un âge mental de base et un plafond de l'âge mental ont été déterminés

Un test d'intelligence comme celui proposé par Binet peut donc être d'une longueur différente pour chaque enfant et les items administrés peuvent être différents selon les réponses obtenues de leur part. Le test tend vers la détermination d'un âge mental, qui se situe entre l'âge mental de base et le plafond de l'âge mental, où chaque enfant réussit environ 50 % des items présentés (Weiss, 1982, p. 474).

Cependant, ce type de test nécessite l'intervention d'un administrateur pour chaque enfant. L'opération est alors longue et onéreuse puisqu'elle ne peut s'effectuer en groupe. De plus, l'administrateur doit posséder des compétences relativement avancées pour pouvoir utiliser et analyser ce type de test. Son utilisation dans un contexte scolaire où des groupes importants d'étudiants sont en cause est donc difficile ; il faut toutefois rappeler que le test d'intelligence de Binet n'était pas prévu pour être utilisé dans de telles conditions.

2.2.2 Test à deux étapes

Un test à deux étapes (*two-stages test*) débute par l'administration d'un test directeur (*routing test*), le même pour tous, qui permet d'obtenir un premier estimateur du niveau d'habileté. Selon le résultat obtenu à ce test directeur, un second test, le test de seconde étape (*second-stage test*), est choisi parmi plusieurs possibilités et est administré. Le test directeur et le test de seconde étape sont tous deux construits selon les principes de la

théorie classique des tests (Lord, 1980a, p. 128). L'objectif du test à deux étapes est d'obtenir un estimateur initial du niveau d'habileté pour ensuite effectuer le choix d'un second test qui permet d'accentuer la précision de l'estimateur final du niveau d'habileté autour d'un niveau de difficulté (Auger, 1989, p. 55). Reckase (1989, p. 12) souligne l'existence de travaux effectués sur cette forme de testing adaptatif dès 1958 par Angoff et Huddleston. Le tableau 2.3 en décrit les règles.

Tableau 2.3
Algorithme décrivant le déroulement d'un test à deux étapes

RÈGLE	ACTION
1. Règle de départ	Administrer un même test directeur à tous
2. Règle de suite	Choisir un test de seconde étape selon le résultat obtenu au test directeur
3. Règle d'arrêt	Terminer le test lorsque le dernier item du test de seconde étape a reçu une réponse

Ce test adaptatif permet d'améliorer, par rapport à un test conventionnel, la précision de l'estimateur du niveau d'habileté tout en utilisant un médium papier crayon (Auger, 1989, p. 55-56). Cependant, l'intervention individualisée d'une personne qui doit calculer le résultat obtenu au test directeur avant de pouvoir administrer le test de seconde étape en fait un test adaptatif relativement lourd à utiliser. Lord (1971, p. 227) et de Gruijter (1980, p. 231) soulignent toutefois qu'il est possible de permettre au répondant de

calculer lui-même son résultat au test directeur. De plus, l'amélioration de la précision de l'estimateur du niveau d'habileté pourrait ne pas être appréciable si les commentaires de Laurier (1993b, p. 42) et d'Auger (1989, p. 56) sont considérés. Selon eux, la précision de l'estimateur du niveau d'habileté au test de seconde étape est fortement influencée par la précision de l'estimateur du niveau d'habileté au test directeur. Auger souligne aussi que, pour obtenir une précision suffisante de l'estimateur provisoire du niveau d'habileté, un grand nombre d'items doit être administré au test directeur : cela constitue, selon lui, une limite à l'utilisation du test à deux étapes.

2.2.3 Test à niveaux flexibles

Lord (1971 : voir Laurier, 1993b, p. 42) propose un test adaptatif papier crayon relativement simple à administrer. Il s'agit du test à niveaux flexibles (*flexilevel test*). La figure 2.2 en illustre un exemple contenant 21 items. La procédure générale consiste à placer en rang, par ordre de niveau de difficulté, les items du test. L'item de niveau de difficulté moyen est donc présenté au début du test. C'est l'item de départ. Selon une bonne ou une mauvaise réponse à cet item, le prochain item administré est choisi à l'intérieur d'une des deux colonnes. La colonne de gauche contient des items dont le niveau de difficulté est inférieur à celui de l'item de départ. À l'opposé, la colonne de droite contient des items dont le niveau de difficulté est supérieur à celui de l'item de départ. Les numéros des items de la colonne de gauche sont imprimés en rouge tandis

que ceux de la colonne de droite sont imprimés en bleu. Pour chacun des items, de la colonne de gauche et de droite, une marque bleue apparaît lorsque le répondant donne une bonne réponse. Lorsqu'il donne une mauvaise réponse, une marque rouge apparaît. Le mécanisme utilisé dans les loteries d'état, qui consiste à gratter certaines plages bien circonscrites d'une feuille, peut servir pour faire apparaître les marques bleues et rouges (Thissen et Mislevy, 1990, p. 104). Lorsque le choix de réponse à un item tourne au bleu, soit une bonne réponse, le prochain item administré est l'item suivant de la colonne de droite (item en bleu). Si le dernier choix de réponse avait fait apparaître une marque rouge, une mauvaise réponse, le prochain item administré aurait été l'item suivant de la colonne de gauche (item en rouge). La procédure se poursuit jusqu'à ce qu'un item de chaque rangée ait été administré.

Item de départ d'un niveau de difficulté moyen (11e item par ordre de niveau de difficulté)			
1.*	Item plus facile (10e item par ordre de niveau de difficulté)	1.+	Item plus difficile (12e item par ordre de niveau de difficulté)
2.*	Item plus facile (9e item par ordre de niveau de difficulté)	2.+	Item plus difficile (13e item par ordre de niveau de difficulté)
.		.	
.		.	
.		.	
10.*	Item plus facile (1er item par ordre de niveau de difficulté)	10.+	Item plus difficile (21e item par ordre de niveau de difficulté)

* Numéros imprimés en rouge

\+ Numéros imprimés en bleu

Figure 2.2 Modèle d'un test adaptatif à niveaux flexibles comportant 21 items (adapté de Lord (1980a, p. 115) et de Thissen et Mislevy (1990, p. 105))

Dans un test à niveaux flexibles, il y a $(N + 1)/2$ items administrés, où N est le nombre total d'items du test incluant l'item de départ. Un test à niveaux flexibles pourrait éventuellement permettre l'administration de 2^N tests différents. Ainsi, dans l'exemple de la figure 2.2, 11 items seront administrés, soit $(21+1)/2$. L'estimateur du niveau d'habileté correspond au nombre de bonnes réponses. Un test à niveaux flexibles suit les règles décrites au tableau 2.4.

Tableau 2.4
Algorithme décrivant le déroulement d'un test à niveaux flexibles

RÈGLE	ACTION
1. Règle de départ	Administrer le même item de départ de niveau de difficulté moyen à tous
2. Règle de suite	Si le choix de réponse tourne au bleu, le prochain item administré est le prochain item non répondu de la colonne de droite. Si le choix de réponse tourne au rouge, le prochain item à administrer est le prochain item non répondu de la colonne de gauche
3. Règle d'arrêt	Terminer le test lorsqu'un item de chaque rangée a été administré, ce qui est équivalent au fait de terminer le test lorsque $(N+1)/2$ items ont été administrés

Ce test adaptatif offre l'avantage de diminuer de moitié le nombre d'items à administrer. De plus, il permet d'éviter des procédures complexes et coûteuses d'appariement (*equating*) pour permettre que tous les répondants, à qui sont administrés des items différents, soient comparés sur la même échelle de mesure (Thissen et Mislevy, 1990,

p. 106).

Les tests à niveaux flexibles semblent toutefois présenter des inconvénients importants (Thissen et Mislevy, 1990, p. 107). Le plus important renvoie à la complexité de la tâche à effectuer. Selon Thissen et Mislevy, certains sujets pourraient avoir plus de difficulté avec le cheminement à travers le test qu'avec la difficulté des items eux-mêmes. De plus, que faire lorsque les directives ne sont pas suivies à la lettre ? Comment estimer le niveau d'habileté lorsque des réponses à des items sont manquantes ? Enfin, Auger (1989, p. 57) indique que les tests à niveaux flexibles présentent l'inconvénient de ne permettre l'administration que d'un seul item par niveau de difficulté.

2.2.4 Test pyramidal

Le test pyramidal est ainsi nommé à cause de la hiérarchisation arborescente des items à l'intérieur du test (Laurier, 1993b, p. 43). Un test adaptatif pyramidal est composé d'un ensemble d'items ordonnés en fonction de leur niveau de difficulté dans une structure similaire à une pyramide (Weiss, 1985, p. 779). Au sommet de la pyramide est placé un item d'un niveau de difficulté moyen. À l'étage suivant, deux items sont disponibles : l'un est d'un niveau de difficulté légèrement supérieur au précédent, l'autre, d'un niveau de difficulté légèrement inférieur. À chaque étage subséquent, deux items supplémentaires, plus difficiles et plus faciles, sont ajoutés. Quatre items sont donc

disponibles au troisième étage à partir du sommet, six au quatrième, et ainsi de suite. Reckase (1989, p. 12) cite des travaux relatifs à cette forme de testing adaptatif par Krathwohl et Hayser dès 1956.

Plusieurs variantes des tests pyramidaux existent (Auger, 1989, p. 56). La version présentée par Laurier (1993b, p. 44) suggère de débuter le test par la présentation d'un item d'un niveau de difficulté moyen. Une bonne réponse à l'item entraîne l'administration d'un item d'un niveau de difficulté supérieur. À l'inverse, une mauvaise réponse entraîne l'administration d'un item plus facile. Le test se termine lorsqu'un nombre fixe d'items a été atteint et l'estimateur du niveau d'habileté correspond au niveau de difficulté du dernier item proposé. Les règles suivantes, illustrées au tableau 2.5, décrivent le déroulement d'un test de ce type.

Tableau 2.5
Algorithme décrivant le déroulement d'un test pyramidal

RÈGLE	ACTION
1. Règle de départ	Administrer le même item de départ de niveau de difficulté moyen à tous les étudiants
2. Règle de suite	Si la réponse est bonne, le prochain item de niveau de difficulté supérieur est administré, sinon le prochain item de niveau de difficulté inférieur est administré
3. Règle d'arrêt	Terminer le test lorsqu'un nombre fixe d'items a été présenté

Laurier (1993b, p. 43) questionne l'efficacité d'une telle stratégie et il souligne qu'il serait possible d'améliorer la précision de l'estimateur du niveau d'habileté en substituant à chaque item, un groupe d'items. Ainsi, pour être dirigé vers un groupe d'items d'un niveau de difficulté légèrement supérieur, la réussite de trois items sur cinq, par exemple, serait nécessaire. Sinon, le prochain groupe d'items administrés serait composé d'items d'un niveau de difficulté légèrement inférieur. Le test pyramidal, contrairement au test à niveaux flexibles, offre donc l'avantage de permettre la construction de tests possédant plusieurs items à chaque niveau de difficulté.

2.2.5 Test stratifié

Le test stratifié (*stratified adaptive* ou *stradaptive*) est un test adaptatif à entrée variable et avec un nombre variable d'items administrés (Auger, 1989, p. 58 ; Hambleton et Swaminathan, 1987, p. 300). Comme le test de Binet, ce test utilise des ensembles d'items organisés en strates. Plusieurs items, le même nombre, sont disponibles à chaque strate. Chaque strate représente un test convergent sur des items d'un niveau de difficulté moyen et les strates sont ordonnées par niveaux croissants de difficulté des items (Auger, 1989, p. 58). De plus, comme le test de Binet, le test stratifié permet de commencer le test par l'administration d'items d'un niveau de difficulté qui varie en fonction d'informations préalables. Contrairement au test de Binet, cependant, le branchement est effectué par suite de la réponse à chaque item. Le test de Binet, compte tenu de cette

dernière remarque, ne peut pas être considéré comme un type de test stratifié. Une bonne réponse conduit à l'administration du prochain item disponible à la strate supérieure alors qu'une mauvaise réponse est associée au prochain item disponible à la strate inférieure. La procédure se continue jusqu'au moment où aucun item d'une strate n'est réussi ; un niveau limite supérieur est alors atteint, comme l'est le plafond de l'âge mental dans le test de Binet. Les règles du test stratifié sont décrites au tableau 2.6.

Tableau 2.6
Algorithme décrivant le déroulement d'un test stratifié

RÈGLE	ACTION
1. Règle de départ	Administrer un premier item en fonction d'informations préalables
2. Règle de suite	Si la réponse est bonne, administrer le prochain item disponible de la strate supérieure ; si la réponse est mauvaise, administrer le prochain item disponible de la strate inférieure
3. Règle d'arrêt	Terminer le test lorsque toutes les réponses aux items d'une strate sont mauvaises

Ce test adaptatif offre l'avantage de réduire considérablement le nombre d'items à administrer. Weiss (1985, p. 781) souligne d'ailleurs que la plupart des versions de tests stratifiés permettent de diminuer de moitié le nombre d'items si elles sont comparées à un test tel que celui de Binet. Toutefois, le test stratifié, comme les tests à niveaux flexibles et à deux étapes, ne permet pas de minimiser l'erreur de mesure associée à

l'estimateur du niveau d'habileté (Auger, 1989, p. 60). Tout comme eux, il ne permet pas non plus le développement de nouvelles versions des tests ni l'ajout de nouveaux items à la banque en cours de route. Les tests adaptatifs fondés sur des propositions de modélisation de la réponse à l'item issues de la théorie de la réponse à l'item permettent de pallier ces inconvénients.

2.2.6 Tests fondés sur des propositions de modélisation de la réponse à l'item issues de la théorie de la réponse à l'item

Ce n'est qu'avec l'introduction de la théorie de la réponse à l'item (*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 de la réponse à l'item.

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. Avantage corollaire : 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 sur mesure.

De plus, un test adaptatif fondé sur des propositions de modélisation de la réponse à l'item issues de la théorie de la réponse à l'item 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 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.

Ces propriétés des tests adaptatifs construits autour de la théorie de la réponse à l'item relèvent toutefois d'un concept théorique qui sera abordé plus loin dans le texte, soit le concept d'invariance. Ce concept d'invariance constitue une propriété importante du modèle de régression sous-jacent aux modélisations de la théorie de la réponse à l'item. Il est nécessaire de réunir des conditions à sa réalisation, ce que nous discutons aussi plus loin.

Différentes règles de début, de suite et d'arrêt peuvent être utilisées à l'intérieur d'un test adaptatif fondé sur des modélisations de la réponse à l'item issues de la théorie de la réponse à l'item. Devant cette diversité, il est difficile de réaliser une synthèse en un seul tableau. C'est pourquoi les prochains chapitres, soient les chapitres 3, 4 et 5, sont dédiés spécifiquement à la présentation de la théorie de la réponse à l'item, à des propositions de modélisation issues de cette théorie, aux méthodes d'estimation de l'estimateur du niveau d'habileté et au déroulement d'un test adaptatif fondé sur ces propositions de modélisation.

3. Théorie de la réponse à l'item

La théorie classique des tests (Lord et Novick, 1968 ; Novick, 1966) met l'accent surtout sur les caractéristiques globales du test, principalement sa fidélité et sa validité, ainsi que sur le score obtenu à ce test. Elle s'intéresse peu à la réponse aux items et encore moins à l'interaction entre la réponse à chacun des items d'un test et le niveau d'habileté du répondant. Tout au plus, permet-elle de connaître le niveau de difficulté d'un item, ainsi que son niveau de discrimination, sans toutefois tenir compte du fait que la réponse à l'item est affectée par le niveau d'habileté du répondant. De plus, dans la théorie classique des tests, il est présumé que la précision du score obtenu, telle qu'elle est mesurée par la fidélité du test, est la même quel que soit le niveau d'habileté. Il s'agit donc d'une théorie au service de l'analyse des tests, plutôt que d'une théorie au service de l'analyse de la réponse à chacun des items d'un test. D'ailleurs la théorie de la réponse à l'item, comme son nom le suggère, a été avancée pour répondre à cette lacune de la théorie classique des tests.

La théorie de la réponse à l'item est issue principalement des travaux de Lawley (1944), de Rasch (1960) et de Lord (1952). Sous le nom de théorie du trait latent (*latent trait theory*) (Birnbaum, 1968), de théorie de la réponse à l'item (*item response theory*) (Baker, 1992 ; Bock, 1997a ; Hambleton, Swaminathan et Rogers, 1991) ou de théorie de la courbe caractéristique de l'item (*item characteristic curve theory*) (Blais, 1987, p. 2), elle s'est développée de façon importante à partir des années soixante-dix (Blais,

1987, p. 3).

La théorie de la réponse à l'item a été proposée pour remédier à certaines limites de la théorie classique des tests (Hambleton et Swaminathan, 1987, p. 1-4 ; Hambleton, Swaminathan et Rogers, 1990, p. 5 ; Weiss et Yoes, 1990, p. 70-71). Ces limites font en sorte que, dans la théorie classique des tests, les paramètres d'items sont différents selon le niveau d'habileté du groupe de sujets dont les résultats contribuent à la réalisation de l'estimation de ceux-ci, que l'estimateur du niveau d'habileté varie en fonction de la valeur des paramètres des items administrés et que l'erreur type de l'estimateur du niveau d'habileté est considérée comme homogène quel que soit le niveau d'habileté.

Ainsi, la théorie de la réponse à l'item permet, contrairement à la théorie classique des tests (Lord et Novick, 1968), d'utiliser des paramètres associés aux items dont les estimateurs sont indépendants du niveau d'habileté de l'échantillon particulier de sujets à l'intérieur duquel ils ont été obtenus (Blais et Ajar, 1992, p. 10 ; Weiss et Yoes, 1990, p. 70). Les paramètres d'items sont alors dits invariants par rapport au groupe. La proportion de bonnes réponses à un item, soit l'indice de difficulté d'un item utilisé en théorie classique des tests, n'est pas invariante par rapport au groupe où elle est obtenue. Ainsi, pour deux groupes de sujets dont les niveaux d'habileté moyens diffèrent largement, l'estimateur de la difficulté d'un item sera différent pour chacun des groupes.

Pour la même raison, Weiss et Yoes (1990, p. 70) soulignent que le paramètre de

discrimination des items utilisé dans la théorie classique des tests, soit le coefficient de corrélation bisérial ou point-bisérial, n'est pas invariant par rapport au groupe de sujets où il a été obtenu.

De plus, dans la théorie de la réponse à l'item, contrairement à la théorie classique des tests, l'estimateur du niveau d'habileté n'est pas indépendant des items utilisés pour effectuer la mesure (Weiss et Yoes, 1990, p. 71). Lorsque la théorie classique des tests est appliquée, l'estimateur du niveau d'habileté est égal au score total obtenu au test, lui-même fonction du niveau de difficulté des items administrés. Par conséquent, le score total obtenu à un autre test, composé d'items dont le niveau de difficulté moyen est plus faible ou plus élevé, risque de ne pas être le même. Par contre, dans la théorie de la réponse à l'item, lorsque l'échantillon de sujets qui sert à la calibration des paramètres des items est homogène, n'importe quel sous-ensemble d'items qui satisfait aux exigences de la modélisation de la réponse à l'item peut servir pour obtenir l'estimateur du niveau d'habileté. Il est alors question d'invariance de l'estimateur du niveau d'habileté par rapport aux items. L'échantillon de sujets est considéré homogène lorsque les sujets qui le composent sont tirés d'une population de personnes qui peuvent toutes être caractérisées par les mêmes paramètres de groupe (Baker, 1992, p. 133), soit le niveau d'habileté dans le cas présent. À ce moment, tous les sujets qui possèdent le même niveau d'habileté affichent tous la même probabilité d'obtenir une bonne réponse à chacun des items (Hambleton, Swaminathan et Rogers, 1991, p. 8). Cette invariance de l'estimateur du niveau d'habileté par rapport aux items est une caractéristique intéressante

d'un test adaptatif où, justement, les items administrés sont différents d'une personne à une autre. Les tests peuvent ainsi être vraiment sur mesure. C'est pourquoi les propositions de modélisation de la réponse à l'item issues de la théorie de la réponse à l'item trouvent une application tout à fait appropriée en testing adaptatif. Il faut toutefois insister sur la nécessité de satisfaire aux exigences spécifiques de chacune de ces propositions de modélisation.

Enfin, comme l'explique Thissen (1990, p. 163), dans la théorie classique des tests, la précision de l'estimateur du niveau d'habileté est dérivée du calcul du coefficient de fidélité du test et, de ce fait, est considérée comme étant la même à tous les niveaux d'habileté. Il y aurait ainsi homogénéité de l'erreur type de l'estimateur du niveau d'habileté. Par contre, la théorie de la réponse à l'item tient compte du fait que la précision de l'estimateur du niveau d'habileté n'est pas nécessairement la même à tous les niveaux d'habileté. Ainsi, lorsqu'un test est surtout composé d'items faciles, la précision de l'estimateur du niveau d'habileté devrait être plus élevée pour une personne dont le niveau d'habileté est faible que pour celle dont le niveau d'habileté est élevé. Cette dernière caractéristique de la théorie de la réponse à l'item fait en sorte qu'il est possible qu'en testing adaptatif la règle d'arrêt soit conditionnelle à l'atteinte d'un niveau de précision prédéterminé de l'estimateur du niveau d'habileté (Thissen et Mislevy, 1990, p. 114).

3.1 Modélisation de la réponse à l'item

La théorie de la réponse à l'item postule qu'il est possible de spécifier une fonction mathématique reliant la probabilité d'une réponse à un item au niveau d'habileté du répondant (Goldstein, 1994a, p. 366 ; Goldstein, 1994b, p. 109 ; Laveault et Grégoire, 1997, p. 291 ; van der Linden et Hambleton, 1997, p. v). Certaines des modélisations proposées dépendent du type de réponses aux items : réponses dichotomiques (Lord, 1952 ; Lord et Novick , 1968, p. 365), réponses polytomiques (Baker, 1992, p. 251-288 ; Bock, 1997b ; Roberts, Donoghue et Laughlin, 2000 ; Thissen, 1988), réponses ordonnées (Samejima, 1997b), réponses polytomiques partiellement ordonnées (Wilson, 1992) et réponses continues (Samejima, 1973b). D'autres modélisations dépendent de l'échelle de mesure postulée pour le niveau d'habileté : catégorielle, classes latentes (Gitomer et Rock, 1993) ; continue, trait latent (Samejima, 1973b, 1974) ; ou hybride, classes latentes et trait latent (Yamamoto et Gitomer, 1993). Dans d'autres cas, des habiletés différentes contribuent à produire la réponse à l'item : les modèles sont alors multidimensionnels (Ackerman, 1994 ; Goldstein et Wood, 1989, p. 160-162 ; Luecht, 1996 ; McDonald, 1982, p. 381-384, 1997, 2000 ; Reckase, 1985, 1997). Des modèles non paramétriques existent aussi (Junker et Sijtsma, 2000 ; Mokken, 1982 ; Mokken et Lewis, 1997 ; Molenaar, 1997 ; Ramsay, 1991, 1993a, 1993b, 1997 ; Stout, 1990).

Dans presque tous les tests adaptatifs utilisés actuellement, le type de réponses aux items est dichotomique (bonne ou mauvaise réponse) et l'unidimensionnalité de l'habileté sur

une échelle continue est supposée. En fonction de ces caractéristiques, trois modèles ont été privilégiés en testing adaptatif (Hambleton, Swaminathan et Rogers, 1991, p. 12-18 ; Wainer et Mislevy, 1990, p. 68-72). Ils ne diffèrent que par le nombre de paramètres impliqués dans la fonction modélisant la probabilité d'obtenir une bonne réponse à un item selon le niveau d'habileté. Ce sont les modèles à un, à deux et à trois paramètres. Ces trois modèles sont présentés ainsi que celui, moins fréquent, à quatre paramètres.

Dans le modèle à un paramètre, seul le niveau de difficulté de l'item, b, est considéré. Le modèle à deux paramètres ajoute un paramètre de discrimination, a, qui correspond à la pente maximale de la fonction. Le modèle à trois paramètres intègre de plus un paramètre de pseudo-chance, c. Dans ce modèle, il est postulé que la probabilité d'une bonne réponse à un item n'est pas nécessairement nulle lorsque le niveau d'habileté est très faible. De façon similaire, le modèle à quatre paramètres incorpore un paramètre, γ, qui correspond à la valeur asymptotique supérieure de la probabilité d'une bonne réponse à l'item. Il y est postulé que, même si le niveau d'habileté est très élevé, la probabilité d'une bonne réponse à l'item n'est pas nécessairement égale à 1, donc certaine. Ce dernier modèle, pourtant intéressant au plan théorique, ne semble cependant pas utilisé dans la pratique. Barton et Lord ont été incapables de lui trouver des avantages significatifs (1981 : voir Hambleton et Swaminathan, 1987, p. 49).

3.2 Modèle logistique à un paramètre

Comme le souligne Blais (1987, p. 24), le modèle logistique à un paramètre a été développé par Georges Rasch (1960) indépendamment des travaux de Lord (1952) et de Birnbaum (1968), et presque simultanément à ces travaux. Il est souvent nommé le modèle de Rasch (Rasch 1960 : voir Molenaar, 1995, p. 15), quoique Rasch ait formulé ce modèle de manière différente, mais mathématiquement équivalente. On reconnaît le modèle à un paramètre par l'expression de la probabilité d'une bonne réponse à un item, $P(r = 1|\theta)$, à partir de la fonction logistique suivante :

$$P(r = 1|\theta) = \frac{1}{1 + e^{-(\theta - b)}} \qquad \textbf{(3.1)}$$

où r est la réponse à l'item, θ est le niveau d'habileté et b le niveau de difficulté de l'item. La valeur de r est égale à 1 pour une bonne réponse et à 0 pour une mauvaise réponse tandis que θ et b peuvent prendre des valeurs comprises dans l'intervalle $[-\infty, \infty]$.

Il semble important de souligner qu'une modélisation à partir d'une loi normale, dans l'esprit de la formulation originale du modèle à deux paramètres par Lord (1952, p. 5) et Lord et Novick (1968, p. 366), a aussi été proposée (Baker, 1992, p. 17 ; Hambleton et Swaminathan, 1987, p. 50) :

$$P(r = 1|\theta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-(z^2)/2} dz \qquad \textbf{(3.2)}$$

où z est égal à θ - b.

C'est d'ailleurs pourquoi, dans toutes les fonctions logistiques utilisées, autant à un paramètre qu'à deux, trois ou quatre, une constante D est fréquemment utilisée. La fonction logistique (équation 3.1) prend alors la forme suivante :

$$P(r = 1|\theta) = \frac{1}{1 + e^{-D(\theta - b)}} \qquad \textbf{(3.3)}$$

La constante D, proposée par Haley (1952 : voir Baker, 1992, p. 16 ; Camilli, 1994 ; Wright et Stone, 1979, p. 21), ramène la variance de la fonction logistique à 1. Les probabilités calculées à partir de la fonction logistique et de la fonction normale N(0,1) (équation 3.2) sont à ce moment presque identiques. Selon Haley, l'erreur maximale d'approximation de l'ogive normale par l'ogive logistique lorsque la constante D est égale à 1,70 ne dépasse pas 0,01 sur tout le continuum d'intégration [$-\infty$, ∞].

Les probabilités obtenues à partir des deux fonctions, logistique et normale, sont présentées à l'annexe I. Comme prévu, l'erreur d'approximation de l'ogive normale par l'ogive logistique y est d'au plus 0,01 en valeur absolue. L'utilisation de la constante D pour estimer les paramètres des fonctions logistiques n'a pour but que de rendre comparable les estimateurs obtenus à partir des modèles basés sur la loi normale. Certains auteurs l'utilisent, d'autres, non. Il convient toujours de vérifier si les

paramètres ont été estimés en utilisant cette constante. De plus, il semble opportun de spécifier qu'en 1968, Birnbaum suggérait d'utiliser des fonctions logistiques plutôt que des fonctions basées sur une ogive normale. En fait, les fonctions logistiques sont plus simples d'utilisation car, tout comme leurs dérivées, elles permettent un calcul sans exiger le recours à des méthodes d'approximation numérique. Les fonctions basées sur la loi normale n'offrent pas cet avantage. Conséquemment, seuls les modèles logistiques, qu'ils soient à un, deux, trois ou quatre paramètres, sont traités dans ce présent chapitre.

La figure 3.1 présente trois courbes du modèle à un paramètre qui se distinguent par la valeur du paramètre b. Plus la valeur de b est élevée, plus le niveau de difficulté de l'item correspondant est élevé. Ces courbes sont les courbes caractéristiques des items (CCI). Elles permettent d'observer la probabilité d'obtenir une bonne réponse à un item en fonction du niveau d'habileté. Selon le modèle à un paramètre, la probabilité d'obtenir une bonne réponse est nulle lorsque le niveau d'habileté est très faible. Elle est certaine lorsque le niveau d'habileté est élevé.

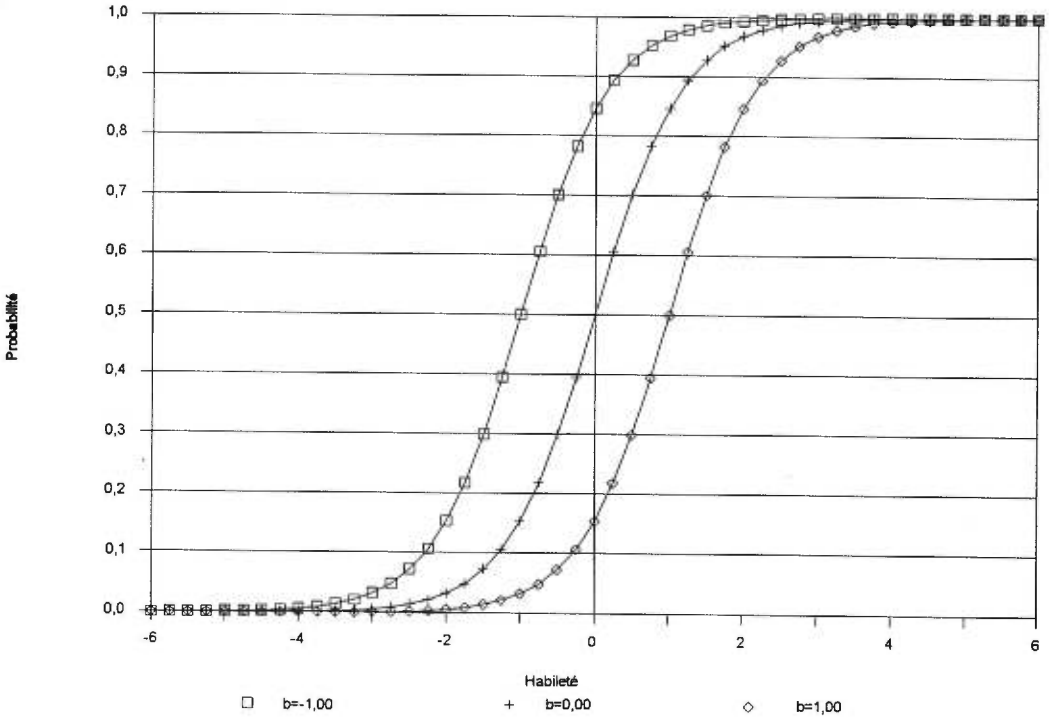

Figure 3.1 Courbe caractéristique d'item (CCI) du modèle logistique à un paramètre selon trois niveaux de difficulté de l'item (D = 1,70)

De plus, il semble important de noter que certains auteurs (Fischer, 1974 : voir Fischer, 1995, p. 37 ; Hambleton et Swaminathan, 1987, p. 47) ont attribué au modèle de Rasch la particularité de permettre de placer les niveaux d'habileté et de difficulté de l'item sur une échelle de rapport. Il serait ainsi possible de dire qu'une personne est deux fois plus habile qu'une autre ou qu'un item est deux fois plus difficile qu'un autre, lorsque le rapport entre deux niveaux d'habileté ou deux niveaux de difficulté est égal à 2. Fischer (1995, p. 37-38) rejettera plus tard cette interprétation et attribuera au modèle de Rasch tout au plus la possibilité de situer les niveaux d'habileté et de difficulté de l'item sur une échelle d'intervalle. Une échelle d'intervalle permet d'obtenir une égalité des intervalles entre les mesures à tous les points de l'échelle de mesure (Stevens, 1946, p. 5). Comme le souligne Stevens, il est alors possible de calculer des moyennes et des écarts types ainsi que d'effectuer des transformations linéaires. Il est important de pouvoir situer les niveaux d'habileté sur une échelle d'intervalle puisqu'à ce moment la distance entre les niveaux d'habileté est comparable sur toute l'étendue de l'échelle. Cela permet, comme le notent Wright (1977) ainsi que Wright et Stone (1979, p. 9), de réaliser des opérations arithmétiques sur les niveaux d'habiletés ; la comparaison de groupes et l'application de procédures statistiques telles que la régression linéaire, l'analyse de variance ou l'analyse factorielle sont alors facilitées.

3.3 Modèle logistique à deux paramètres

Le modèle logistique à deux paramètres propose d'ajouter au niveau de difficulté, b, un second paramètre, la discrimination. Le paramètre de discrimination, a, peut prendre des valeurs qui varient entre $-\infty$ et ∞. Il est toutefois inhabituel d'obtenir des valeurs supérieures à 2 et, lorsque la valeur est négative, l'item devrait être rejeté (Hambleton, Swaminathan et Rogers, 1991, p. 15). Selon Blais (1987, p. 24), l'intervalle raisonnable du paramètre de discrimination varie entre 0,50 et 2,00. Dans ce modèle, les items ne partagent pas tous le même pouvoir de discrimination lorsque le niveau d'habileté est égal au niveau de difficulté. L'indice de discrimination est proportionnel à la pente de la courbe caractéristique de l'item lorsque le niveau d'habileté est égal au niveau de difficulté de l'item. Un item dont le paramètre de discrimination affiche une valeur négative indique que la probabilité d'une bonne réponse à l'item diminue avec le niveau d'habileté. Si le modèle est monotone croissant, alors la valeur du paramètre de discrimination est fixée par cette condition et alors $a > 0$. L'équation de la fonction logistique à deux paramètres prend la forme suivante :

$$P(r = 1|\theta) = 1 + \frac{1}{1 + e^{-Da(\theta - b)}} \qquad \textbf{(3.4)}$$

La figure 3.2 présente deux courbes caractéristiques d'item lorsque le niveau de difficulté b est maintenu à 0,00 alors que l'indice de discrimination, a, varie de 0,50 à 2,00. La

figure 3.2 montre clairement que les courbes caractéristiques des items n'affichent pas la même pente lorsque le niveau d'habileté est égal à 0,00.

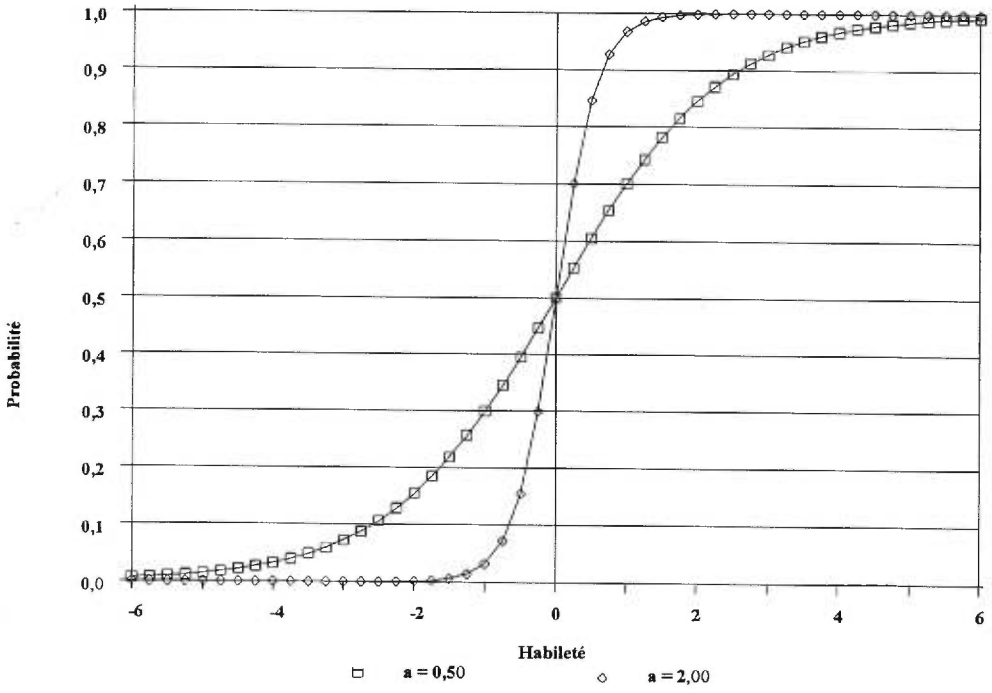

Figure 3.2 Courbe caractéristique d'item (CCI) du modèle logistique à deux paramètres selon deux valeurs du paramètre de discrimination (D = 1,70, b = 0,00)

3.4 Modèle logistique à trois paramètres

Le modèle logistique à trois paramètres ajoute un troisième élément aux deux modèles précédents : l'indice de pseudo-chance, c. Selon ce modèle, la probabilité d'une bonne réponse à un item n'est pas nécessairement nulle lorsque le niveau d'habileté est très faible. Des facteurs externes au niveau d'habileté peuvent affecter la probabilité d'une bonne réponse. Par exemple, dans un choix de réponses *vrai ou faux,* la réponse *vrai* pourrait être naturellement préférée par les individus dont le niveau d'habileté est très faible. Quoique le paramètre de pseudo-chance puisse prendre des valeurs comprises entre 0,00 et 1,00, sa valeur ne devrait pas dépasser celle d'une réponse au hasard et serait proportionnelle au nombre de choix de réponses (Blais, 1987, p. 24 ; Hambleton, Swaminathan et Rogers, 1991, p. 17). Selon Laveault et Grégoire (1997, p. 294), la valeur du paramètre de pseudo-chance est généralement inférieure à celle qui correspondrait à un choix de réponses complètement au hasard. La figure 3.3 montre de tels types d'items. L'item dont le paramètre de pseudo-chance c est égal à 0,50 est réussi une fois sur deux lorsque le niveau d'habileté est très faible. Selon Wainer et Mislevy (1990, p. 72), ce modèle est largement utilisé dans les applications de testing à grande échelle. L'équation du modèle logistique à trois paramètres est la suivante :

$$P(r = 1|\theta) = c + \frac{1 - c}{1 + e^{-Da(\theta - b)}} \qquad \textbf{(3.5)}$$

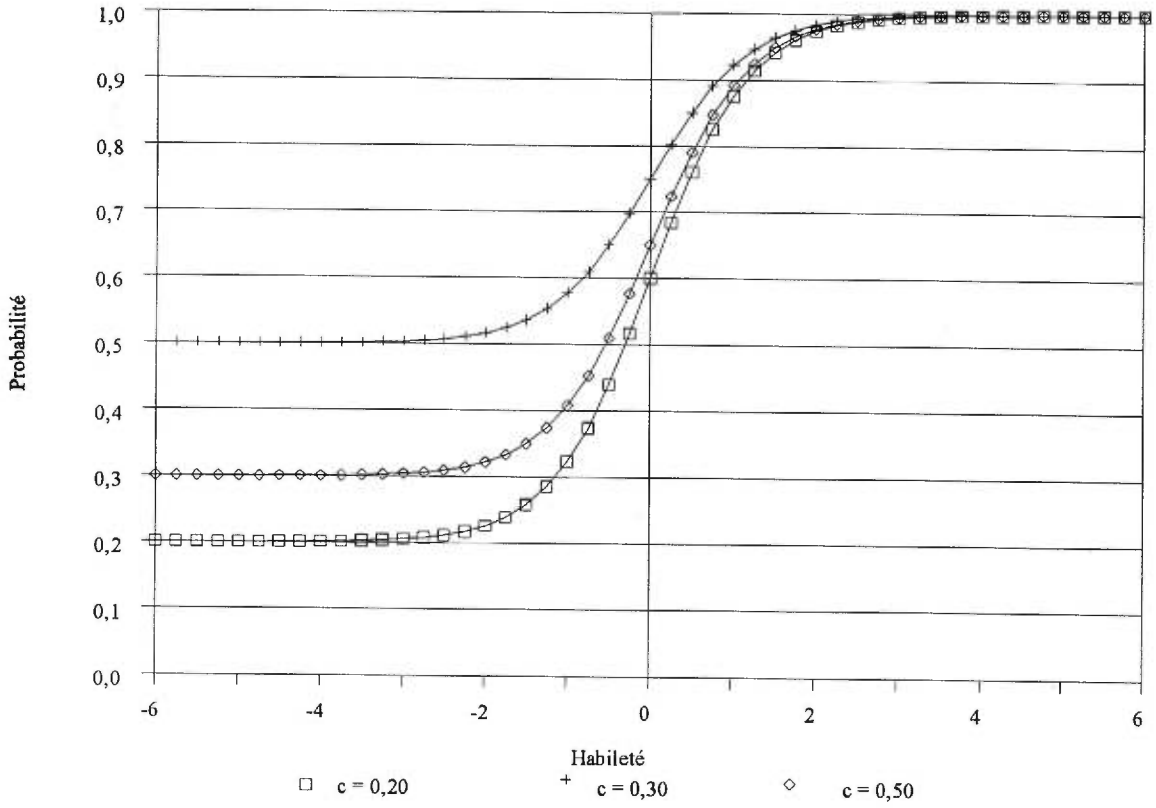

Figure 3.3 Courbe caractéristique d'item (CCI) du modèle logistique à trois paramètres selon trois valeurs du paramètre de pseudo-chance (D = 1,70, a = 1,00, b = 0,00)

3.5 Modèle logistique à quatre paramètres

Dans le modèle logistique à quatre paramètres, on suppose que la probabilité d'une bonne réponse à un item n'est pas nécessairement égale à 1,00 lorsque le niveau d'habileté est très élevé. On incorpore alors le paramètre γ à la fonction logistique pour en tenir compte, et celui-ci varie à l'intérieur de l'intervalle compris entre 0,00 et 1,00. La fonction logistique à quatre paramètres semble toutefois n'être jamais utilisée, puisque des problèmes d'estimation numérique lui sont associés lorsque les méthodes d'estimation par vraisemblance maximale sont appliquées. La fonction est la suivante (Hambleton et Swaminathan, 1987, p. 49) :

$$P(r = 1|\theta) = c + \frac{(\gamma - c)}{1 + e^{-Da(\theta - b)}} \qquad \textbf{(3.6)}$$

La figure 3.4 illustre la courbe caractéristique d'item de ce modèle lorsque le paramètre γ prend respectivement les valeurs 0,50, 0,80 et 0,90. Un tel modèle a été proposé par McDonald (1967, p. 67) et par Burton et Lord (1981 : voir Hambleton et Swaminathan, 1987, p. 50) dans le contexte d'une modélisation basée sur la loi normale. Ils l'ont cependant abandonné puisqu'ils ne lui ont pas trouvé d'avantages appréciables dans le gain de précision des mesures.

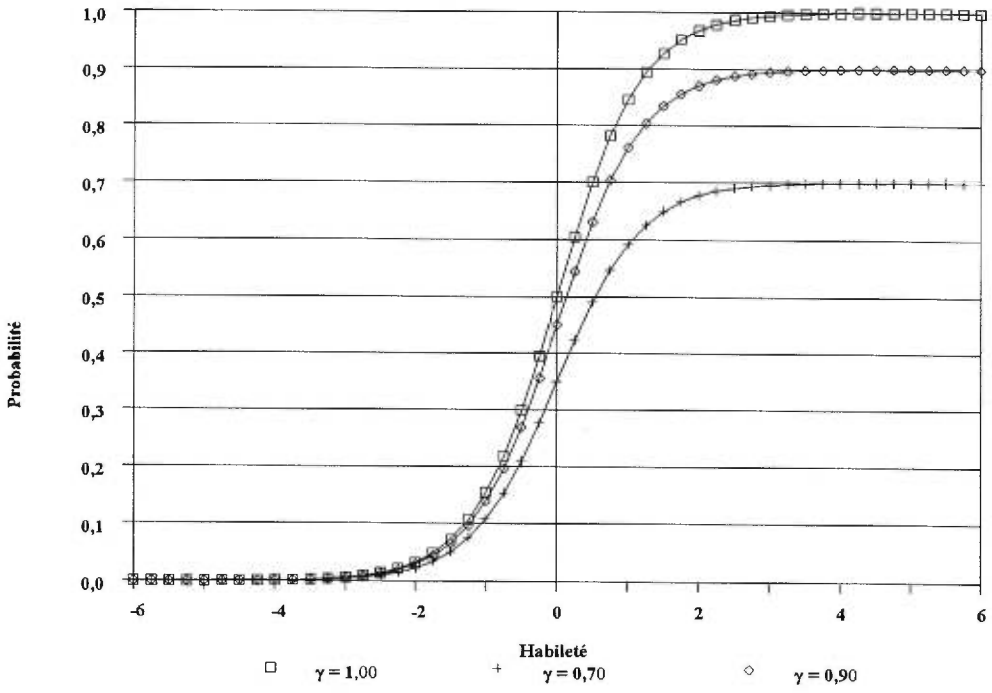

Figure 3.4 Courbe caractéristique d'item (CCI) du modèle logistique à quatre paramètres selon trois valeurs du paramètre γ (D = 1,70, a = 1,00, b = 0,00, c = 0,00)

4. Méthodes d'estimation du niveau d'habileté dans la théorie de la réponse à l'item

Il existe des stratégies qui permettent d'estimer simultanément les paramètres associés aux items et ceux associés aux personnes. Il s'agit surtout des méthodes de vraisemblance maximale conjointe (Baker, 1992, p. 85-113) (*joint maximum likelihood*, JML), de maximum de vraisemblance conditionnelle (Molenaar, 1995, p. 44-47) (*conditional maximum likelihood*, CML) ou de vraisemblance maximale marginale (Baker, 1992, p. 171-192) (*marginal maximum likelihood*, MML). En testing adaptatif, cependant, l'estimation des paramètres d'items est généralement réalisée à l'avance de façon à utiliser, à l'intérieur d'une banque, des items dont les valeurs des paramètres sont déjà connues. Par conséquent, seules des stratégies permettant de réaliser l'estimation des paramètres de personnes, dans ce cas-ci le niveau d'habileté, sont abordées.

Considérant que l'estimation du niveau d'habileté est conditionnelle à la connaissance préalable des paramètres d'items (a, b, c et γ), ce chapitre décrit trois méthodes d'estimation : méthode de vraisemblance maximale (ML), méthode bayésienne de maximisation a posteriori (MAP) et méthode de l'espérance a posteriori (EAP). Pour chacune de ces méthodes, nous présentons les fonctions qui permettent le calcul de l'estimateur du niveau d'habileté ainsi que de son erreur type.

4.1 Méthode de vraisemblance maximale (ML)

La méthode de vraisemblance maximale (*maximum likelihood, ML*) consiste à trouver le niveau d'habileté qui maximise la probabilité d'apparition de la séquence de bonnes et de mauvaises réponses. Le mode de la distribution d'échantillonnage de l'estimateur du niveau d'habileté est ainsi utilisé comme estimateur du niveau d'habileté.

C'est donc une fonction qui maximise :

$$P(R|\theta) = \prod_{i=1}^{n} P_i(r_i|\theta)^{r_i} Q_i(r_i|\theta)^{1-r_i} \qquad \textbf{(4.1)}$$

où n est le nombre d'items, R est le vecteur de bonnes et de mauvaises réponses $\{r_1 \ldots r_n\}$, chaque élément de R prenant une valeur de 1, lors d'une bonne réponse, et de 0, lors d'une mauvaise réponse, $P_i(r_i|\theta)$ est une des fonctions logistiques présentées aux équations 3.3, 3.4, 3.5 ou 3.6 et $Q_i(r_i|\theta)$ est égal à $1 - P_i(r_i|\theta)$.

Généralement, on maximise cette fonction en utilisant la procédure de Newton-Raphson (Andersen, 1995, p. 278-280 ; Yakowitz et Szidarovszky, 1986, p. 203-210) sur le logarithme naturel de cette fonction. Le produit se transforme alors en somme et la procédure itérative converge plus facilement. La procédure de Newton-Raphson est une procédure itérative qui maximise une fonction en corrigeant la valeur de l'estimateur

$ML(\theta)$ à chaque itération t par la soustraction du rapport de la dérivée première à la dérivée seconde. Lorsque la valeur de ce rapport est négligeable, la procédure d'estimation s'arrête. Le rapport h correspond d'ailleurs à l'erreur de calcul. Ainsi,

$$ML_{t+1}(\theta) = ML_t(\theta) - h_t \qquad \textbf{(4.2)}$$

où

$$h_t = \frac{P_t'(R|\theta)}{P_t''(R|\theta)} \qquad \textbf{(4.3)}$$

La méthode de vraisemblance maximale n'est cependant pas applicable lorsque le vecteur de réponses R est composé uniquement de bonnes ou de mauvaises réponses, puisque que le maximum de la fonction correspond alors à l'infini. Il faut aussi se méfier des résultats obtenus par cette méthode lorsque le nombre d'items est petit et que la fonction logistique utilisée est celle à trois ou à quatre paramètres, puisque la fonction peut afficher plusieurs solutions. Plusieurs maximums locaux peuvent alors être présents (Samejima, 1973a, p. 223-226). Lord (1980a, p. 59) indique que, dans le cas du modèle à trois paramètres, la procédure donne de bons résultats lorsque le nombre d'items est supérieur à 20, situation où il n'existe habituellement qu'un seul maximum local. La procédure de Newton-Raphson est un algorithme qui permet de trouver un maximum local et, lorsque plusieurs maximums locaux existent, l'algorithme ne permet pas de choisir celui pour lequel la probabilité de la séquence de bonnes et de mauvaises réponses est maximale.

À cause de l'existence potentielle de plusieurs maximums locaux, la méthode de vraisemblance maximale n'est pas la méthode de choix en testing adaptatif où l'estimateur du niveau d'habileté peut être estimé avec un nombre d'items assez faible et où un niveau de fidélité de 0,90 est souvent obtenu avec un nombre d'items inférieur à 20.

Il faut toutefois souligner que certains algorithmes qui permettraient de déterminer le maximum absolu d'une fonction, plutôt qu'un maximum local, sont actuellement à l'étude. Il s'agit d'une famille de stratégies, les chaînes de Markov de type Monte Carlo (*Monte Carlo Markov Chain*, MCMC) dont certaines applications à la théorie de la réponse à l'item seraient possibles (Almond et Mislevy, 1999 ; Baker, 1998 ; Patz et Junker, 1997a, 1997b, 1999 ; Tanner, p. 151-152). Toutefois, pour le moment du moins, le comportement des chaînes de Markov de type Monte Carlo pose plusieurs problèmes d'application. Ainsi, la convergence des calculs n'est pas toujours assurée, la vitesse de celle-ci est généralement très lente et le calcul des estimateurs peut être affecté par l'autocorrélation entre les données simulées (Cowles et Carlin, 1996 ; Fismen, 1997, p. 8 ; Kass, Carlin, Gelman et Neal, 1997, p. 6-8 ; Patz et Junker, 1997a, p. 15-16, 1997b, p. 16-17, 1999, p. 168).

Il existe un concept important relié à l'erreur type de l'estimateur du niveau d'habileté obtenue par la méthode de vraisemblance maximale. Il s'agit du concept d'information au sens de Fisher (1922), qui est associé à la mesure de la précision de l'estimateur du niveau d'habileté (Baker, 1992, p. 72-73). Le concept d'information est tout de même

occasionnellement appliqué lorsque l'estimateur du niveau d'habileté est obtenu par une des méthodes, abordées plus loin, qui utilisent la moyenne de la distribution d'échantillonnage comme estimateur du niveau d'habileté. À ce moment, l'erreur type de l'estimateur du niveau d'habileté est toutefois plus appropriée que l'information. Dans la théorie classique des tests, la fidélité constitue une mesure globale de la précision d'un test et la précision de l'estimateur du niveau d'habileté est la même peu importe le niveau d'habileté. Dans la théorie de la réponse à l'item, la précision de l'estimateur du niveau d'habileté n'est pas la même à tous les niveaux d'habileté. 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 de vraisemblance maximale (Baker, 1992, p. 79, 81). Il est ainsi possible de déterminer le niveau d'habileté où un test est le plus efficace au sens où l'information est maximale (Laurier, 1993b, p. 64). L'information fournie par chacun des items d'un test est égale à :

$$I_i(\theta) = \frac{[P'_i(r_i|\theta)]^2}{P_i(r_i|\theta)Q_i(r_i|\theta)} \qquad \textbf{(4.4)}$$

$P'_i(r_i|\theta)$ étant la dérivée première de la fonction de probabilité logistique, $P_i(r_i|\theta)$.

L'information offre, de plus, l'avantage d'être additive : l'information totale d'un test à un niveau d'habileté donné est tout simplement la somme de l'information fournie par chacun des n items du test. Ainsi,

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \qquad \textbf{(4.5)}$$

où $I(\theta)$ est l'information totale du test.

L'erreur type de l'estimateur du niveau d'habileté obtenue par la méthode de vraisemblance maximale s'obtient en calculant l'inverse de la somme des fonctions d'information aux items (Séguin et Auger, 1986, p. 38-39) :

$$S_{ML(\theta)} = \frac{1}{\sqrt{I(\theta)}} = \frac{1}{\sqrt{\sum_{i=1}^{n} I_i(\theta)}} \qquad \textbf{(4.6)}$$

Il est à noter que cette relation entre l'erreur type et la fonction d'information ne tient que lorsque l'estimateur du niveau d'habileté est non biaisé (Lord, 1980a, p. 71). Un estimateur est dit non biaisé lorsque la moyenne de la distribution d'échantillonnage d'un paramètre, ici le niveau d'habileté, est égale à la vraie valeur du paramètre (Bertrand, 1986, p. 178 ; Freund et Walpole, 1980, p. 311 ; Moore et McCabe, 1993, p. 262). La valeur du biais de l'estimateur du niveau d'habileté (Weiss, 1982, p. 482) obtenue par la méthode de vraisemblance maximale à un niveau d'habileté donné est égale à :

$$BIAIS_{ML(\theta)} = \frac{\Sigma(ML(\theta) - \theta)}{N} \qquad \textbf{(4.7)}$$

où N est le nombre de valeurs différentes de l'estimateur du niveau d'habileté.

Lorsque l'estimateur du niveau d'habileté est biaisé, sa valeur est soit supérieure, soit inférieure, à la vraie valeur du niveau d'habileté. À ce moment, la relation entre l'erreur type et la fonction d'information ne peut pas être appliquée.

4.2 Méthode bayésienne de maximisation a posteriori (MAP)

Pour contourner certains problèmes inhérents à la méthode de vraisemblance maximale, tel que l'impossibilité d'estimer le niveau d'habileté lorsque le vecteur de réponses est constant, il a été suggéré (Weiss et Yoes, 1990, p. 85) d'appliquer une approche bayésienne. Dans cette méthode, on utilise de l'information a priori dans le but d'estimer a posteriori le niveau d'habileté, en incorporant cette information a priori dans la fonction de maximisation. Le mode de la distribution a posteriori est alors utilisé comme estimateur du niveau d'habileté. Cette méthode consiste à maximiser la fonction de probabilité suivante :

$$P(R|\theta,\ f(\theta)) = f(\theta) \prod_{i=1}^{n} P_i(r_i|\theta)^{r_i} Q_i(r_i|\theta)^{1-r_i} \qquad \textbf{(4.8)}$$

où $f(\theta)$ est une fonction de probabilité a priori du niveau d'habileté. Fréquemment, les

fonctions de probabilité normale ou logistique sont utilisées (Hambleton et Swaminathan, 1987, p. 92). La maximisation de cette fonction est effectuée à partir de la méthode de Newton-Raphson comme pour la méthode de vraisemblance maximale. La méthode de vraisemblance maximale peut être considérée comme un cas particulier de la méthode bayésienne de maximisation a posteriori (MAP) où la fonction de probabilité a priori est une loi uniforme U($-\infty$, ∞).

Comme le souligne Samejima (1972, p. 7, 1997a, p. 474-475), le critère de l'existence d'un maximum unique doit aussi être satisfait lorsque la méthode bayésienne de maximisation a posteriori est utilisée. Cela implique que la méthode bayésienne de maximisation a posteriori devrait nécessiter, comme la méthode de vraisemblance maximale, un nombre important d'items afin d'assurer une solution unique à l'estimation du niveau d'habileté lorsque que le modèle à trois paramères est utilisé. C'est qu'elle repose sur le même procédé de calcul, soit la maximisation d'une fonction qui, lorsque le nombre d'items est faible, peut ne pas conduire à une solution unique. Selon nous, il est fort probable qu'un nombre minimal de 20 items soit aussi nécessaire. Pour cette même raison, la méthode bayésienne de maximisation a posteriori ne semble pas la plus appropriée dans un contexte de testing adaptatif lorsque le modèle logistique à trois paramètres est utilisé et qu'il n'y a pas de contraintes quant à un minimum de 20 items à administrer.

4.3 Méthode de l'espérance a posteriori (EAP)

Dans les deux méthodes précédentes on effectue l'estimation en utilisant le mode de la distribution d'échantillonnage comme estimateur du niveau d'habileté. La méthode de l'espérance a posteriori (*expected a posteriori*, EAP) utilise le premier moment, donc la moyenne, comme estimateur. Cette méthode a l'avantage de toujours donner une solution unique à l'estimateur du niveau d'habileté, contrairement aux méthodes basées sur le mode. Le modèle à quatre paramètres pourrait ainsi devenir plus intéressant.

Bock et Mislevy (1982 ; Mislevy et Stocking, 1987, p. 35) proposent d'utiliser l'espérance mathématique a posteriori comme estimateur du niveau d'habileté. Comme la méthode bayésienne de maximisation a posteriori, la méthode de l'espérance a posteriori utilise de l'information provenant de la distribution de probabilité a priori du niveau d'habileté. Lorsque la distribution de probabilité a priori suit une loi uniforme, la méthode de l'espérance a posteriori devient tout simplement une méthode qui repose uniquement sur l'espérance mathématique. Veerkamp et Berger (1997, p. 210) présentent une application de la méthode de l'espérance mathématique (E). Dans les faits, cette méthode semble être peu utilisée ; sa variante bayésienne, soit celle de l'espérance a posteriori, lui est préférée.

Dans la méthode de l'espérance a posteriori l'estimateur du niveau d'habileté est égal à :

$$EAP(\theta) = \frac{\int_{-\infty}^{\infty} \theta f(\theta) \prod_{i=1}^{n} P(r_i|\theta)^{r_i} Q(r_i|\theta)^{1-r_i}}{\int_{-\infty}^{\infty} f(\theta) \prod_{i=1}^{n} P(r_i|\theta)^{r_i} Q(r_i|\theta)^{1-r_i}} \qquad \textbf{(4.9)}$$

tandis que l'erreur type de l'estimateur du niveau d'habileté correspond à :

$$S_{EAP(\theta)} = \left[\frac{\int_{-\infty}^{\infty} [\theta - E(\theta)]^2 f(\theta) \prod_{i=1}^{n} P(r_i|\theta)^{r_i} Q(r_i|\theta)^{1-r_i}}{\int_{-\infty}^{\infty} f(\theta) \prod_{i=1}^{n} P(r_i|\theta)^{r_i} Q(r_i|\theta)^{1-r_i}} \right]^{1/2} \qquad \textbf{(4.10)}$$

où $f(\theta)$ est une fonction de probabilité a priori du niveau d'habileté et où r_i prend une valeur de 1, lors d'une bonne réponse, et de 0, lors d'une mauvaise réponse.

La méthode de l'espérance a posteriori est recommandée par Thissen et Mislevy (1990, p. 113) dans le contexte du testing adaptatif, à cause de la simplicité des calculs impliqués et de la précision obtenue de l'estimateur du niveau d'habileté. Ainsi, les calculs ne sont pas itératifs, ne requièrent pas l'évaluation des dérivées de la fonction utilisée pour modéliser la réponse à l'item et une solution unique existe toujours, même lorsque les réponses sont toutes bonnes ou toutes mauvaises. Quant à la précision de l'estimateur du niveau d'habileté, selon les mêmes auteurs, aucun autre estimateur que celui obtenu par

la méthode de l'espérance a posteriori ne permet d'obtenir un plus petit carré moyen de l'erreur (*root mean square error*). Le carré moyen de l'erreur est une mesure qui, dans son calcul, tient compte à la fois de l'erreur type et du biais de l'estimateur du niveau d'habileté. Le carré moyen de l'erreur de l'estimateur du niveau d'habileté (de Ayala, Schafer et Sava-Bolesta, 1995, p. 390 ; Weiss, 1982, p. 482), lorsque la méthode de l'espérance a posteriori est utilisée, est égal à :

$$\text{CARRÉ MOYEN}_{EAP(\theta)} = \sqrt{\frac{\Sigma(EAP(\theta) - \theta)^2}{N}} = \sqrt{S^2_{EAP(\theta)} + \text{BIAIS}^2_{EAP(\theta} \quad \textbf{(4.11)}$$

où N est le nombre de valeurs différentes de l'estimateur du niveau d'habileté.

4.4 Comparaison de l'erreur type de l'estimateur du niveau d'habileté selon la méthode d'estimation utilisée

Le tableau 4.1 compare l'estimateur du niveau d'habileté obtenu selon les méthodes d'estimation précédentes ainsi que son erreur type. Les estimations sont effectuées à partir d'un exemple proposé par Wainer (1983, p. 70) basé sur une modélisation logistique à un paramètre de la réponse à l'item où la constante D n'est pas utilisée. Parmi ces méthodes d'estimation, Wainer ne présente que celles qui reposent sur le calcul de la vraisemblance maximale ; nous y avons ajouté les méthodes basées sur l'espérance

mathématique.

L'erreur type de l'estimateur du niveau d'habileté est plus faible lorsque les méthodes de maximisation a posteriori (MAP) et de l'espérance a posteriori (EAP) sont utilisées. La méthode de l'espérance mathématique (E) vient en troisième lieu, suivie de la méthode de vraisemblance maximale (ML). Ce sont donc les méthodes de maximisation a posteriori et de l'espérance a posteriori qui permettent d'obtenir la plus grande précision de l'estimateur du niveau d'habileté. L'impact de la distribution de probabilité a priori du niveau d'habileté sur l'estimation du niveau d'habileté est aussi à noter. Ainsi, l'estimateur du niveau d'habileté obtenu par les méthodes utilisant une distribution de probabilité a priori du niveau d'habileté, MAP(θ) et EAP(θ), se rapproche plus de la moyenne de cette distribution a priori (dans l'exemple utilisé, de forme normale avec une moyenne de 0 et un écart type de 1 et notée N(0,1)) que lorsque les méthodes non bayésiennes de vraisemblance maximale et de l'espérance mathématique sont appliquées.

Tableau 4.1
Estimateur du niveau d'habileté et erreur type de l'estimateur du niveau d'habileté selon quatre méthodes d'estimation

VECTEUR DE RÉPONSES	ML(θ)	$S_{ML(\theta)}$	MAP(θ)	$S_{MAP(\theta)}$	E(θ)	$S_{E(\theta)}$	EAP(θ)	$S_{EAP(\theta)}$
1100000000	-1,84	0,89	-1,10	0,62	-1,81	0,75	-1,12	0,62
1110000000	-1,55	0,79	-0,72	0,61	-1,20	0,79	-0,74	0,61
1111000000	-0,56	0,76	-0,36	0,60	-0,59	0,77	-0,36	0,61
1111100000	0,00	0,74	0,00	0,60	0,00	0,76	0,00	0,60
1111110000	0,56	0,76	0,36	0,60	0,59	0,77	0,36	0,61
1111111000	1,55	0,79	0,72	0,61	1,20	0,79	0,74	0,61
1111111100	1,84	0,89	1,10	0,62	1,81	0,75	1,12	0,62

b = -2,00, -1,60, -1,10, -0,70, -0,20, 0,20, 0,70, 1,10, 1,60, 2,00 (Wainer, 1983, p. 70)
Distribution a priori N(0,1)

5. Déroulement d'un test adaptatif basé sur la théorie de la réponse à l'item

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 5.1 et le tableau 5.1 décrivent le déroulement d'un tel test. Dans le présent chapitre, nous présentons pour chacune des règles considérées des stratégies proposées par la littérature.

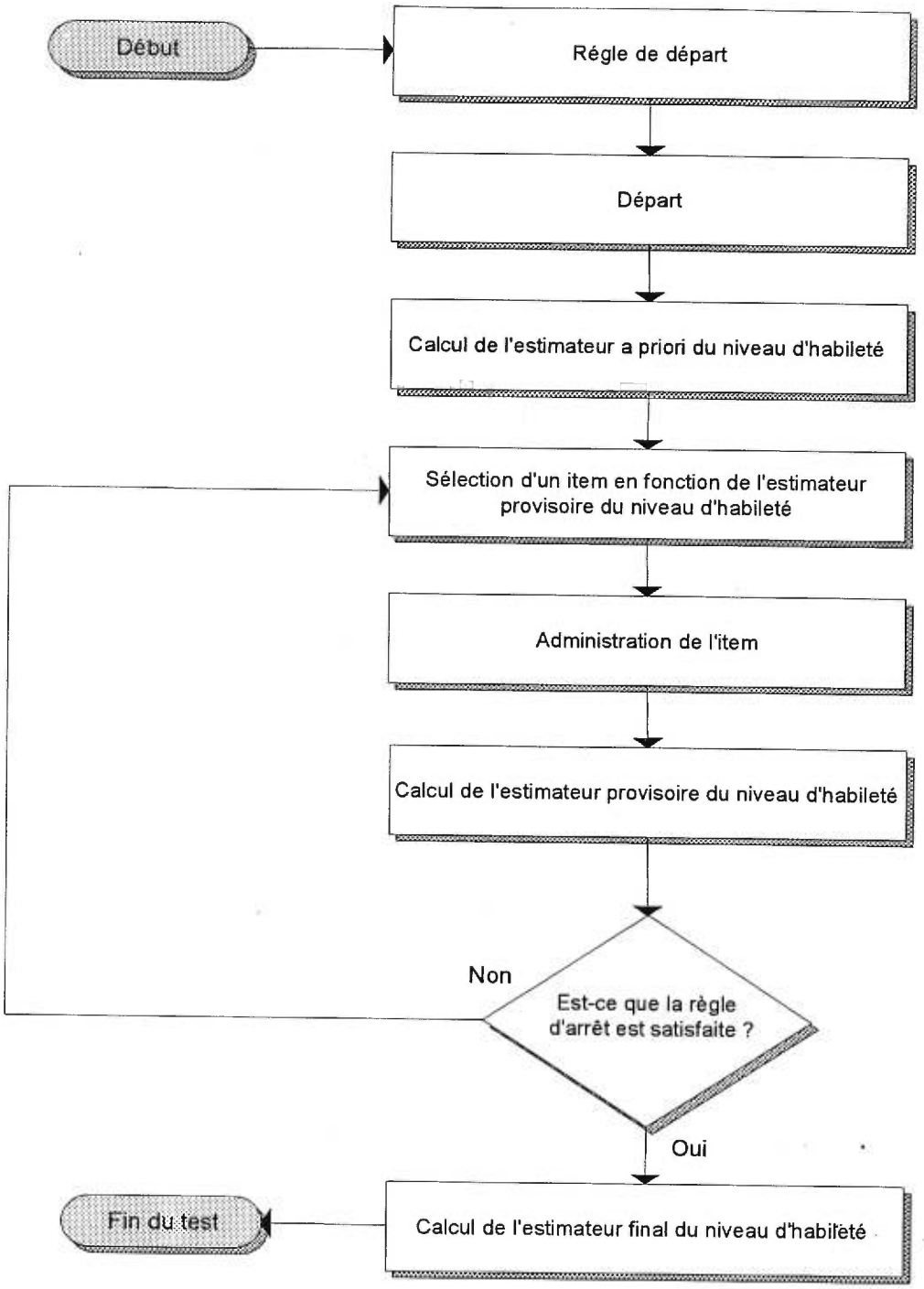

Figure 5.1 Structure d'un test adaptatif

Tableau 5.1
Algorithme décrivant le déroulement d'un test adaptatif

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'item qui puisse fournir une quantité d'information minimale au niveau d'habileté estimé

5.1 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. Une règle de départ basée sur l'information a priori disponible à propos du niveau d'habileté doit être adoptée. Selon Thissen et Mislevy (1990, p. 109), la moyenne de la population d'où provient l'individu en situation de testing, estimée préalablement selon une modélisation issue de la théorie de la réponse à l'item, 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.

5.2 Règle de suite et estimation provisoire du niveau d'habileté

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 et 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. Enfin, 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*).

5.2.1 Stratégie 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 de Urry.

La méthode de sélection par information maximale sans contrainte (*unconstrained maximum information selection*) permet de choisir un item pour lequel l'information évaluée au niveau d'habileté estimé provisoirement après l'administration de l'item j est maximale (Lord, 1980b, p. 199), l'information à la prochaine estimation du niveau d'habileté étant évaluée en conformité avec les fonctions 4.4 et 4.5.

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 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 alternative 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.

5.2.2 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 bayésienne 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 bayésienne 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 bayésienne d'Owen pour le calcul de l'estimateur du niveau d'habileté et de l'erreur type de l'estimateur du niveau d'habileté sont présentées à l'annexe II. 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. Ils soulignent toutefois un inconvénient important dans l'application de la méthode bayésienne 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 bayésienne 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, a, 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 l'utilisation de minitests est suggérée par Wainer et Kiely (1987).

5.2.3 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 influencent la réponse aux items suivants.

Prometteuse, selon Wainer et *al.* (1990, p. 253-254), cette stratégie exige toutefois des modélisations de la réponse à l'item plus sophistiquées telles que celles qui sont utilisées

dans les modèles à réponses nominales ou ordonnées. Thissen (1993) propose d'ailleurs certains modèles spécifiques à une démarche de testing adaptatif par minitests. L'étude des minitests ne nous intéresse pas à l'intérieur de cette recherche à cause précisément de cette nécessité d'utiliser des modélisations de la réponse à l'item qui lui sont spécifiques.

5.2.4 Estimateur provisoire du niveau d'habileté

Selon Thissen et Mislevy (1990, p. 113), les méthodes d'estimation provisoire du niveau d'habileté après l'administration de j items les plus utilisées sont celles basées sur les fonctions de vraisemblance telle que la méthode de vraisemblance maximale (section 4.1). Les méthodes bayésiennes d'estimation du niveau d'habileté sont aussi utilisées, soient la méthode bayésienne de maximisation a posteriori (section 4.2) et la méthode de l'espérance a posteriori (section 4.3). 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 bayésienne 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.

5.3 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 (1990, p. 115), 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 par vraisemblance maximale ou une méthode d'estimation bayésienne. En fait, dans les méthodes d'estimation bayésienne, 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 de vraisemblance maximale, et par Chang et Stout (1993) pour les méthodes bayésiennes. Selon eux, l'estimateur du niveau d'habileté obtenu par la méthode de vraisemblance maximale, ainsi que par les méthodes bayésiennes, tend vers la vraie valeur du niveau d'habileté lorsque le nombre d'items administrés tend vers l'infini.

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, un nombre maximal d'items à administrer doit de plus être fixé 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 de la théorie classique des tests. Comme dans le cas de la règle d'arrêt basée sur le nombre d'items administrés, aucun critère absolu n'a été arrêté quant à l'erreur type de l'estimateur du niveau d'habileté prédéterminée à retenir.

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 puisse 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*), un temps d'administration prédéterminé peut être fixé (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.

5.4 Estimateur final du niveau d'habileté

Toutes les méthodes précédentes, utilisées 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 bayésienne d'Owen, alors que l'estimateur final du niveau d'habileté est calculé par une des méthodes de vraisemblance maximale, de maximisation a posteriori ou d'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 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 conséquemment 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.

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.

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.

6. Caractéristiques de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction des règles d'arrêt

Quelques recherches ont eu, directement ou indirectement, pour objet d'étude les règles d'arrêt en testing adaptatif. Par exemple, les travaux de Dodd, Koch et de Ayala (1993) s'intéressent spécifiquement à la comparaison des caractéristiques de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt. Dodd, Koch et de Ayala se limitent cependant à une simple prescription sur l'utilisation d'une règle plutôt qu'une autre. Ils ne s'intéressent pas à l'impact de la variation du critère dans les règles d'arrêt sur les caractéristiques de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif. Vispoel, Wang et Bleiler (1997), McBride et Martin (1983) ainsi que Bock et Mislevy (1982) apportent quelques informations à ce sujet, sans que ce soit le principal objet de leurs travaux.

Ces auteurs utilisent toutefois des méthodes d'estimation du niveau d'habileté et de son erreur type qui sont adéquates lorsque le nombre d'items administrés tend vers l'infini. Les caractéristiques de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, dont celles de l'erreur type de l'estimateur du niveau d'habileté, lorsque le nombre d'items administrés est petit, comme en testing adaptatif, sont abordées par Hoijtink et Boomsma (1995, 1996), Samejima (1994) ainsi que par Warm (1989).

6.1 Dodd, Koch et de Ayala (1993)

L'étude réalisée par Dodd, Koch et de Ayala vise un examen de l'effet de certaines caractéristiques d'une banque d'items et de l'utilisation de deux règles d'arrêt sur le comportement de l'estimateur du niveau d'habileté et de l'erreur type de l'estimateur du niveau d'habileté en testing adaptatif. Ils utilisent le modèle de crédit partiel (*partial credit model*) de Masters (1982) comme modélisation de la réponse à l'item et les deux règles d'arrêt spécifiquement étudiées sont la règle d'obtention d'une erreur type prédéterminée de l'estimateur du niveau d'habileté (section 5.3) et une règle basée sur l'information minimale de l'item (section 5.3).

La méthodologie est agencée en quatre opérations : création de quatre banques d'items, simulation de données en testing non adaptatif, estimation des paramètres d'items et simulation d'un test adaptatif. La première banque d'items est composée d'items de difficulté moyenne dont la fonction d'information est légèrement surélevée (*peaked*). Plus spécifiquement, il s'agit de la fonction d'information d'un test qui contiendrait tous les items de cette banque. La seconde banque est composée d'items faciles dont la fonction d'information est surélevée. La troisième banque est composée d'items difficiles dont la fonction d'information est surélevée. Enfin, la quatrième banque est constituée autant d'items faciles que d'items difficiles dont la fonction d'information est bimodale. Il est à noter que les auteurs ne justifient pas l'utilisation des distributions de probabilité du niveau de difficulté des items et ne fournissent pas d'explications quant à leur

interprétation d'un item facile ou d'un item difficile. Ils n'en fournissent pas plus en ce qui concerne leur interprétation d'une fonction d'information surélevée, d'autant plus qu'ils appliquent ce concept, comme celui de la bimodalité, d'ailleurs, à la fonction d'information d'un test qui contiendrait tous les items d'une banque d'items plutôt qu'à une fonction de probabilité. La lecture d'une figure qui décrit ces banques d'items permet tout de même d'approximer certaines de ces valeurs : une banque contient des items faciles lorsque la fonction d'information d'un test, qui contiendrait tous les items d'une banque d'items, est maximale si le niveau d'habileté est égal à environ -1,50, et des items difficiles, si le niveau d'habileté est égal à environ 1,50. Quatre autres banques d'items constitués de 60 items sont aussi créées en doublant tout simplement les quatre banques de 30 items.

Par la suite, des patrons de réponse aux items des banques sont générés aléatoirement pour 3000 répondants tirés d'une population où le niveau d'habileté se distribue selon une loi normale N(0,1). Les patrons de réponse obtenus permettent ensuite d'estimer les paramètres d'items qui seront utilisés dans la simulation de 16 tests adaptatifs différents : selon deux tailles de la banque d'items, 30 ou 60 items ; selon les quatre caractéristiques de la banque d'items décrites au paragraphe précédent et selon deux règles d'arrêt qui sont exposées dans les lignes qui suivent. Les 16 tests adaptatifs sont simulés en se basant sur un échantillon aléatoire de 200 répondants, le même pour les 16 tests, tiré parmi les 3000 obtenus précédemment. Les tests simulés débutent tous par l'administration d'un item d'un niveau de difficulté égal à 0,00 et la sélection des items se fait selon une

stratégie de maximisation par la méthode de sélection par information maximale sans contrainte. Les tests se terminent lorsqu'au maximum 20 items ont été administrés ou que le critère établi pour l'une des deux règles d'arrêt étudiées est atteint : une erreur type de 0,30 pour la stratégie basée sur l'erreur type de l'estimateur du niveau d'habileté ($S_{ML(\theta)} = 0,30$) et une valeur de l'information de 0,45 pour la règle basée sur l'information minimale de l'item ($I(\theta) = 0,45$). Le tableau 6.1 présente l'algorithme qui décrit le déroulement des tests adaptatifs.

Dodd et *al.* présentent ensuite des statistiques descriptives pour chacun des 16 tests adaptatifs : estimateur du niveau d'habileté, erreur type associée à l'estimateur du niveau d'habileté et nombre d'items administrés. Le tableau 6.2, ainsi que les figures 6.1 et 6.2, présentent un sommaire des résultats de Dodd et *al.*

Tableau 6.1
Algorithme décrivant le déroulement des tests adaptatifs dans l'étude réalisée par Dodd, Koch et de Ayala (1993)

RÈGLE	ACTION
1. Règle de départ	Administrer un item dont le niveau de difficulté est égal à 0,00
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, ML(θ), selon une stratégie de maximisation de l'information par la méthode de maximisation sans contrainte
3. Règle d'arrêt	Terminer le test en fonction de deux règles d'arrêt : lorsque l'erreur type de l'estimateur du niveau d'habileté est égale à 0,30 ou lorsqu'il n'y a plus d'item qui puisse fournir une quantité d'information supérieure à 0,45 au niveau d'habileté estimé ; dans les deux cas, un maximum de 20 items est administré
	L'estimateur final du niveau d'habileté est obtenu selon la méthode de vraisemblance maximale

Selon ces résultats, le nombre moyen d'items administrés est plus petit dans la méthode basée sur l'erreur type seulement lorsque la banque d'items est composée d'items de niveau de difficulté moyen : respectivement 14,88 items contre 18,12 et 13,61 items contre 19,51 selon que la banque contient 30 et 60 items. De plus, Dodd et *al.* remarquent que, lorsque la banque est composée d'items d'un niveau de difficulté moyen et que la règle d'arrêt est basée sur l'information minimale de l'item, le nombre maximal de 20 items est fréquemment utilisé. Cela serait dû, selon eux, au nombre plus important d'items qui fournissent une plus grande quantité d'information à un niveau d'habileté

moyen. Dans tous les autres cas, quand le niveau de difficulté est facile, difficile ou mixte, le nombre moyen d'items administrés est supérieur. Lorsque toutes les banques d'items sont considérées, la règle d'arrêt basée sur l'erreur type de l'estimateur du niveau d'habileté nécessite en moyenne une administration de trois items de plus : 16,65 items contre 13,52.

Tableau 6.2
Moyenne de l'erreur type de l'estimateur du niveau d'habileté, $S_{ML(\theta)}$, et nombre moyen d'items administrés (n) en fonction des caractéristiques de la banque d'items et de la règle d'arrêt utilisée : $S_{ML(\theta)} = 0,30$ ou $I(\theta) = 0,45$ (adapté de Dodd, Koch et de Ayala (1993, p. 71))

BANQUE D'ITEMS	NOMBRE D'ITEMS	CARACTÉRISTIQUES		RÈGLE D'ARRÊT	$S_{ML(\theta)}$ (n)
		NIVEAU DE DIFFICULTÉ	FONCTION D'INFORMATION		
1	30	Moyen (b=0,00)	Légement surélevée	$S_{ML(\theta)}=0,30$	0,30 (14,88)
2	30	Moyen (b=0,00)	Légement surélevée	$I(\theta)=0,45$	0,29 (18,12)
3	30	Facile (b=-1,50)	Fortement surélevée	$S_{ML(\theta)}=0,30$	0,38 (17,45)
4	30	Facile (b=-1,50)	Fortement surélevée	$I(\theta)=0,45$	0,58 (11,00)
5	30	Élevé (b=1,50)	Fortement surélevée	$S_{ML(\theta)}=0,30$	0,40 (18,23)
6	30	Élevé (b=1,50)	Fortement surélevée	$I(\theta)=0,45$	0,66 (9,64)
7	30	Mixte	Bimodale	$S_{ML(\theta)}=0,30$	0,32 (17,78)
8	30	Mixte	Bimodale	$I(\theta)=0,45$	0,45 (11,26)
9	60	Moyen (b=0,00)	Légement surélevée	$S_{ML(\theta)}=0,30$	0,29 (13,61)
10	60	Moyen (b=0,00)	Légement surélevée	$I(\theta)=0,45$	0,25 (19,51)
11	60	Facile (b=-1,50)	Fortement surélevée	$S_{ML(\theta)}=0,30$	0,35 (17,07)
12	60	Facile (b=-1,50)	Fortement surélevée	$I(\theta)=0,45$	0,56 (12,02)
13	60	Élevé (b=1,50)	Fortement surélevée	$S_{ML(\theta)}=0,30$	0,37 (17,69)
14	60	Élevé (b=1,50)	Fortement surélevée	$I(\theta)=0,45$	0,62 (11,04)
15	60	Mixte	Bimodale	$S_{ML(\theta)}=0,30$	0,31 (16,45)
16	60	Mixte	Bimodale	$I(\theta)=0,45$	0,37 (15,59)
Toutes les banques				$S_{ML(\theta)}=0,30$	0,34 (16,65)
				$I(\theta)=0,45$	0,47 (13,52)

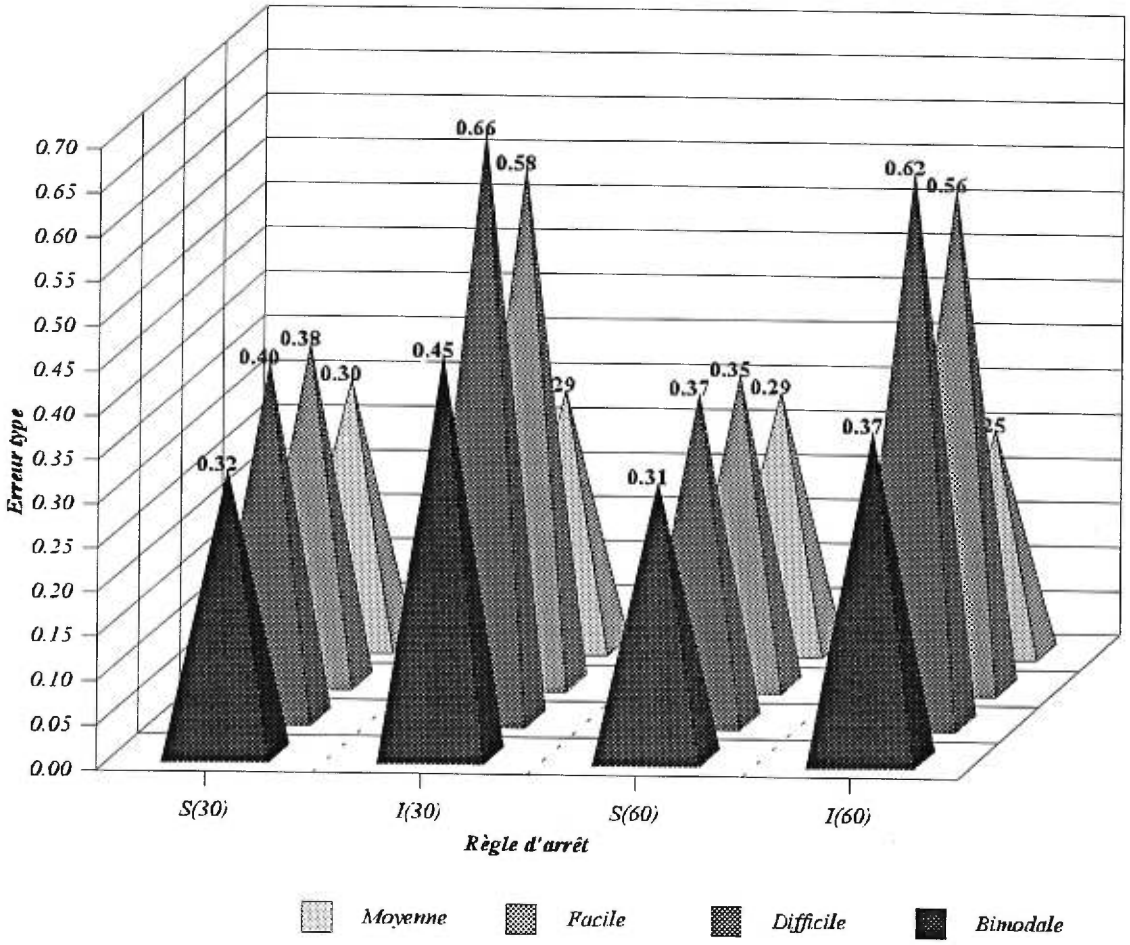

Figure 6.1 Moyenne de l'erreur type de l'estimateur du niveau d'habileté, $S_{ML(\theta)}$, en fonction des caractéristiques de la banque d'items et de la règle d'arrêt utilisée : $S_{ML(\theta)} = 0{,}30$ ou $I(\theta) = 0{,}45$ (adapté de Dodd, Koch et de Ayala, (1993, p. 71))

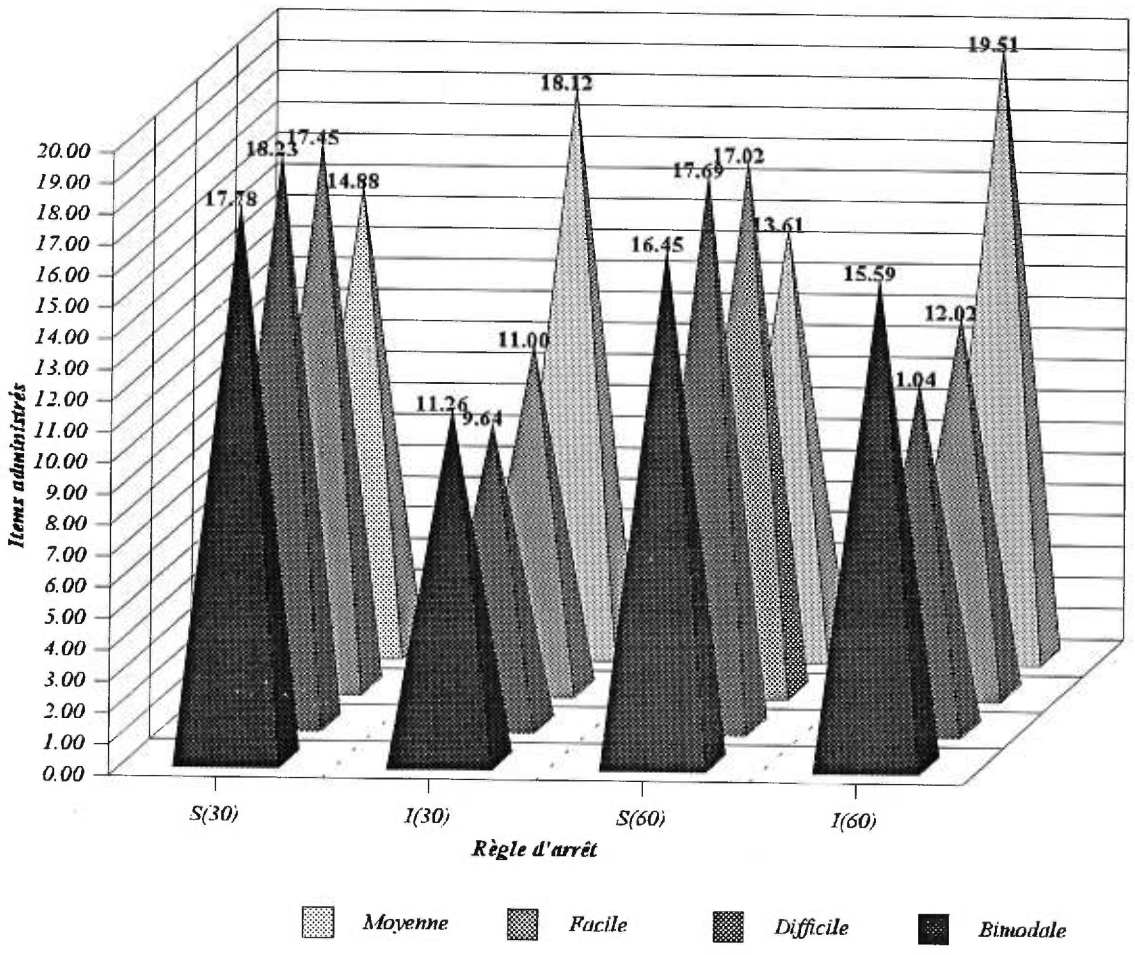

Figure 6.2 Nombre moyen d'items administrés en fonction des caractéristiques de la banque d'items et de la règle d'arrêt utilisée : $S_{ML(\theta)} = 0{,}30$ ou $I(\theta) = 0{,}45$ (adapté de Dodd, Koch et de Ayala, (1993, p. 71))

En ce qui concerne l'erreur type de l'estimateur du niveau d'habileté, la situation inverse se produit. La moyenne de l'erreur type de l'estimateur du niveau d'habileté est plus importante lorsque la règle d'arrêt basée sur l'erreur type est utilisée avec une banque d'items composée d'items d'un niveau de difficulté moyen : respectivement 0,30 contre 0,29 et 0,29 contre 0,25 selon que la banque contient 30 et 60 items. Lorsqu'on tient compte globalement de toutes les banques d'items, la moyenne de l'erreur type de l'estimateur du niveau d'habileté est toutefois plus petite lorsque la règle d'arrêt basée sur l'erreur type est utilisée : 0,34 contre 0,47.

De plus, Dodd et *al.* remarquent que, lorsque la règle d'arrêt basée sur l'information minimale de l'item est utilisée, les tests adaptatifs se terminent fréquemment de façon prématurée quand les premiers estimateurs du niveau d'habileté affichent des valeurs extrêmes, aucun item dont l'information au niveau d'habileté estimé est égale ou supérieure à 0,45 n'étant disponible. Cette situation peut aussi expliquer, selon eux, qu'une corrélation se situant entre 0,77 et 0,78 est obtenue entre l'estimateur du niveau d'habileté, $ML(\theta)$, et le niveau d'habileté simulé, θ, lorsque la règle d'arrêt basée sur l'information minimale de l'item est utilisée et que les banques d'items sont composées d'items faciles ou difficiles. Ces corrélations sont de 0,89 à 0,93 lorsque la règle d'arrêt basée sur l'erreur type est utilisée. Pour toutes ces raisons, les auteurs concluent à la supériorité de la règle d'arrêt basée sur l'erreur type.

L'étude de Dodd et *al.* permet seulement de conclure que la règle d'arrêt basée sur

l'erreur type de l'estimateur du niveau d'habileté est généralement plus efficace que celle basée sur l'information minimale de l'item. Puisqu'ils ont fixé les valeurs dans chacune des règles d'arrêt, $I(\theta) = 0,45$ et $S_{ML(\theta)} = 0,30$, il n'est pas possible de connaître l'effet de la variation de ces valeurs sur l'estimateur du niveau d'habileté en testing adaptatif.

6.2 Vispoel, Wang et Bleiler (1997)

Vispoel et *al.* comparent des versions adaptatives et non adaptatives de tests qui visent l'estimation du niveau d'habileté de rappel des mélodies. Ils effectuent une étude qui permet de comparer :

1) une version adaptative (TAO1) avec des versions non adaptatives et non informatisées, soient le *Seashore measures of musical talents* (SEA), le *Wing standardized tests of musical intelligence* (WING) et un test construit à partir des items de la version adaptative (FIXE), en utilisant des données simulées ;

2) une version adaptative (TAO2) avec des versions non adaptatives et non informatisées, soient le *Seashore measures of musical talents*, le *Wing standardized tests of musical intelligence*, le *Drake musical aptitude tests* et l'*Advanced measures of music audition* en utilisant des données réelles ;

3) une version adaptative (TAO3) avec les versions non adaptatives informatisées citées au point précédent, mais de façon informatisée cette fois-ci, en utilisant des données réelles.

La fidélité des tests; le nombre d'items administrés et différentes mesures de validité sont utilisés pour comparer les versions. Dans les versions adaptatives et non adaptatives, une modélisation logistique à trois paramètres de la réponse à l'item est utilisée. Le test débute par l'administration d'un item de difficulté moyenne et les items suivants sont sélectionnés selon une stratégie, non identifiée, de maximisation de l'information. Les auteurs n'indiquent pas la méthode utilisée pour calculer l'estimateur provisoire du niveau d'habileté. L'estimateur final du niveau d'habileté est obtenu à partir de la méthode de l'espérance a posteriori. Le test adaptatif se termine après l'administration de 30 items, soit la règle d'arrêt. L'algorithme décrivant le déroulement de ces tests adaptatifs est présenté au tableau 6.3.

Tableau 6.3
Algorithme décrivant le déroulement des tests adaptatifs dans l'étude réalisée par Vispoel, Wang et Bleiler (1997)

RÈGLE	ACTION
1. Règle de départ	Administrer un item d'un niveau de difficulté moyen
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 La méthode de calcul de l'estimateur provisoire du niveau d'habileté n'est pas indiquée : il est probable que ce soit la même méthode que celle qui est utilisée pour obtenir l'estimateur final du niveau d'habileté
3. Règle d'arrêt	Terminer le test après l'administration de 30 d'items La méthode de l'espérance a posteriori est utilisée pour obtenir l'estimateur final du niveau d'habileté

Cette étude donne des résultats qui permettent d'étudier la relation entre le nombre d'items administrés et la fidélité du test. Le tableau 6.4 et la figure 6.3 illustrent cette relation ; ils ont été construits à partir des résultats publiés à l'intérieur de leur texte.

Vispoel et *al.* (p. 48) estiment la fidélité des versions TAO1, FIXE, WING et SAE en calculant la corrélation, $r_{\theta EAP(\theta)}$, entre le niveau d'habileté simulé, θ, et l'estimateur du niveau d'habileté obtenu dans les simulations, EAP(θ). Ils présentent les résultats relatifs à la fidélité de ces versions uniquement à l'intérieur d'une figure sans en offrir les valeurs exactes (p. 49). Les valeurs présentées au tableau 6.4 et à la figure 6.3 sont donc des

approximations ; elles reproduisent les valeurs de la fidélité des versions en fonction du nombre d'items administrés obtenues par Vispoel et *al.* En ce qui concerne les versions TAO2 (p. 53) et TAO3 (p. 56), ils obtiennent une approximation du coefficient de fidélité calculée à partir de l'erreur type de l'estimateur du niveau d'habileté (p. 52). Dans ce cas, ils fournissent les valeurs exactes que nous reproduisons aussi au tableau 6.4 et à la figure 6.3.

Le tableau 6.4 et la figure 6.3 permettent de constater que la fidélité des tests obtenue à partir des versions adaptatives (TAO1, TAO2 et TAO3) est toujours supérieure à celle obtenue à partir des versions non adaptatives (WING et SAE) et qu'elle augmente aussi plus rapidement en fonction du nombre d'items administrés. Dans les versions adaptatives, déjà au 4^{e} item, la fidélité est de plus de 0,80 et d'au moins 0,90 à partir du 8^{e} item. À partir de ce nombre d'items administrés, la fidélité n'augmente que très peu ; dans le cas du WING et du SAE, la fidélité augmente lentement et continuellement jusqu'au 30^{e} item administré sans même atteindre les valeurs affichées par les versions adaptatives. La version non adaptative FIXE donne toutefois, quant à la fidélité, des résultats tout aussi satisfaisants que les versions adaptatives. Il est à noter que cette version est composée des mêmes items que la version TAO1, celle qui affiche les plus hautes valeurs de la fidélité.

Même s'ils fournissent des résultats à chaque valeur du nombre d'items administrés, Vispoel et *al.* utilisent au maximum 30 items. Il aurait été intéressant qu'ils continuent

leur exploration au-delà de ce nombre d'items administrés. De plus, ils apportent peu d'information en ce qui concerne d'autres statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté. En ce sens, par exemple le biais de l'estimateur du niveau d'habileté aurait pu davantage retenir leur attention.

De plus, comme souligné précédemment, Vispoel et *al.* ne fournissent pas toutes les informations quant au déroulement des versions adaptatives, principalement en ce qui a trait au calcul de l'estimateur provisoire et à la stratégie de maximisation de l'information.

Tableau 6.4
Fidélité des tests en fonction du nombre d'items administrés (adapté de Vispoel, Wang et Bleiler (1997, p. 49, 53 et 56))

	VERSIONS ADAPTATIVES			VERSIONS NON ADAPTATIVES		
ITEMS ADMINISTRÉS	TAO1	TAO2	TAO3	FIXE	WING	SEA
1	0,67	0,37	0,35	0,54	0,30	0,21
2	0,81	0,61	0,60	0,67	0,33	0,33
3	0,87	0,74	0,74	0,77	0,42	0,36
4	0,91	0,81	0,81	0,82	0,47	0,42
5	0,93	0,84	0,84	0,85	0,50	0,45
6	0,94	0,87	0,87	0,87	0,54	0,47
7	0,95	0,89	0,89	0,90	0,66	0,50
8	0,95	0,90	0,90	0,90	0,66	0,51
9	0,96	0,91	0,91	0,91	0,67	0,53
10	0,96	0,92	0,92	0,91	0,67	0,56
11	0,96	nd.*	0,93	0,92	0,70	0,57
12	0,96	nd.	0,93	0,92	0,76	0,61
13	0,96	nd.	0,94	0,92	0,80	0,67
14	0,96	nd.	0,94	0,93	0,81	0,70
15	0,97	0,94	0,95	0,94	0,82	0,72
16	0,97	nd.	nd.	0,94	0,85	0,75
17	0,97	nd.	nd.	0,95	0,86	0,75
18	0,97	nd.	nd.	0,95	0,86	0,76
19	0,97	nd.	nd.	0,95	0,86	0,80
20	0,97	0,95	0,96	0,95	0,87	0,81
21	0,97	nd.	nd.	0,95	0,87	0,81
22	0,97	nd.	nd.	0,96	0,87	0,82
23	0,97	nd.	nd.	0,96	0,88	0,85
24	0,97	nd.	nd.	0,96	0,89	0,86
25	0,97	0,96	0,96	0,96	0,90	0,86
26	0,97	nd.	nd.	0,96	0,90	0,87
27	0,97	nd.	nd.	0,96	0,90	0,88
28	0,97	nd.	nd.	0,96	0,90	0,88
29	0,97	nd.	nd.	0,96	0,91	0,89
30	0,97	0,97	0,96	0,96	0,91	0,90

* non disponible

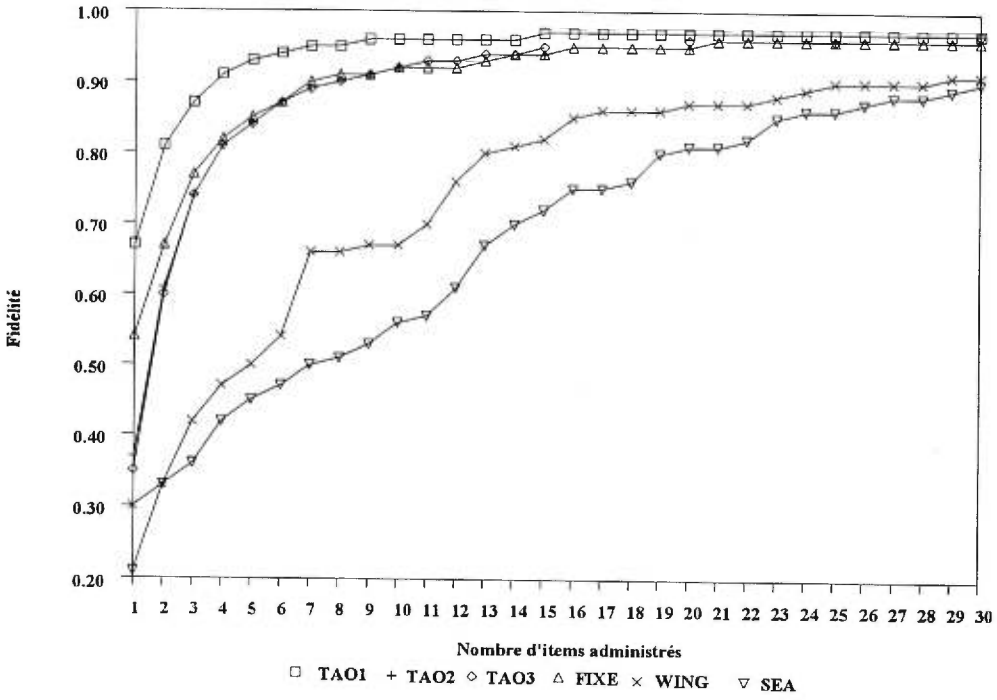

Figure 6.3 Fidélité des tests en fonction du nombre d'items administrés (adapté de Vispoel, Wang et Bleiler (1997, p. 49, 53 et 56))

6.3 McBride et Martin (1983)

McBride et Martin (1983) vérifient si, à longueur égale, un test adaptatif est plus fidèle et valide qu'un test conventionnel de longueur fixe. Un groupe reçoit deux versions équivalentes d'un test adaptatif tandis qu'un autre groupe reçoit deux versions équivalentes d'un test conventionnel de longueur fixe administré par ordinateur. Toutes ces versions comportent 30 items. Dans le but de contrôler les caractéristiques des items administrés dans chacune des versions, les deux types de tests, adaptatif et conventionnel, sont construits à partir d'items tirés de la même source, soit une banque commune de 150 items mesurant le niveau d'habileté verbale. Un modèle à trois paramètres est utilisé. Le niveau de difficulté, b, des items varie entre -2 et 2, la valeur moyenne du paramètre de discrimination, a, est de 1,24 et aucun item n'affiche une valeur du paramètre de pseudo-chance, c, supérieure à 0,30. Ces items ont été préalablement calibrés à l'intérieur d'un échantillon composé de recrues de la marine américaine selon la méthode auxiliaire (*ancillary method*) proposée par Urry (1975 : voir Urry, 1977, p. 187, 195). La méthode auxiliaire recourt aux indices de difficulté et de discrimination de la théorie classique des tests pour obtenir une approximation des paramètres de difficulté et de discrimination propres aux modélisations de la théorie de la réponse à l'item (Lord, 1980a, p. 33-34). Les sujets, au nombre de 530, sont assignés au hasard au groupe A, versions adaptatives, et au groupe C, versions conventionnelles.

Les versions adaptatives du test débutent en prenant pour point de départ une distribution

a priori N(0,1) du niveau d'habileté et les items sont sélectionnés en utilisant la valeur de l'estimateur provisoire du niveau d'habileté calculée par la méthode bayésienne d'Owen, OWEN(θ). Les versions du test se terminent lorsque 30 items ont été administrés. La méthode d'estimation finale du niveau d'habileté n'est pas précisée, mais il est fort probable que la même méthode utilisée pour l'estimation provisoire soit appliquée. L'erreur type est calculée après l'administration de chaque item selon la méthode bayésienne d'Owen, $S_{OWEN(\theta)}$. La fidélité, autant pour les versions adaptatives que pour les versions conventionnelles, est estimée par la corrélation entre les formes alternatives des tests. Le déroulement des tests adaptatifs est décrit au tableau 6.5.

Le tableau 6.6 et la figure 6.4, pour leur part, présentent les résultats obtenus par McBride et Martin en ce qui concerne la fidélité des versions adaptatives (TAO) et conventionnelles (FIXE). Ces résultats permettent de tirer des conclusions similaires à celles de l'étude précédente de Vispoel et *al.* À longueur égale, les versions adaptatives affichent des valeurs de la fidélité qui sont toujours supérieures à celles calculées dans les versions conventionnelles. Déjà, à partir du 9^{e} item, la fidélité est de 0,80. Dans les versions conventionnelles, 17 items doivent être administrés pour obtenir une telle valeur. Pour parvenir à la même fidélité, la version adaptative nécessite donc l'administration de presque deux fois moins d'items. Les valeurs de la fidélité obtenues par McBride et Martin croissent toutefois moins rapidement avec le nombre d'items administrés que dans le cas de l'étude Vispoel et *al.*

Tableau 6.5
Algorithme décrivant le déroulement des tests adaptatifs dans l'étude réalisée par McBride et Martin (1983)

RÈGLE	ACTION
1. Règle de départ	Administrer un item dont le niveau de difficulté est près de la moyenne d'une distribution N(0,1)
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, OWEN(θ), selon une stratégie de maximisation de l'information par la méthode de maximisation sans contrainte
3. Règle d'arrêt	Terminer le test lorsque 30 items ont été administrés
	L'estimation finale du niveau d'habileté est réalisée fort probablement selon la méthode d'Owen

Tableau 6.6
Fidélité (formes alternatives) des tests en fonction du nombre d'items administrés (adapté de McBride et Martin (1983, p. 231))

ITEMS ADMINISTRÉS	TAO	FIXE
1	0,45	0,16
2	0,51	0,13
3	0,57	0,27
4	0,67	0,43
5	0,73	0,46
6	0,75	0,49
7	0,78	0,61
8	0,79	0,66
9	0,80	0,68
10	0,81	0,69
11	0,82	0,71
12	0,83	0,76
13	0,85	0,75
14	0,85	0,76
15	0,86	0,77
16	0,86	0,79
17	0,86	0,80
18	0,87	0,82
19	0,87	0,82
20	0,88	0,83
21	0,88	0,85
22	0,88	0,85
23	0,89	0,85
24	0,89	0,86
25	0,89	0,86
26	0,89	0,86
27	0,89	0,87
28	0,89	0,88
29	0,89	0,88
30	0,89	0,89

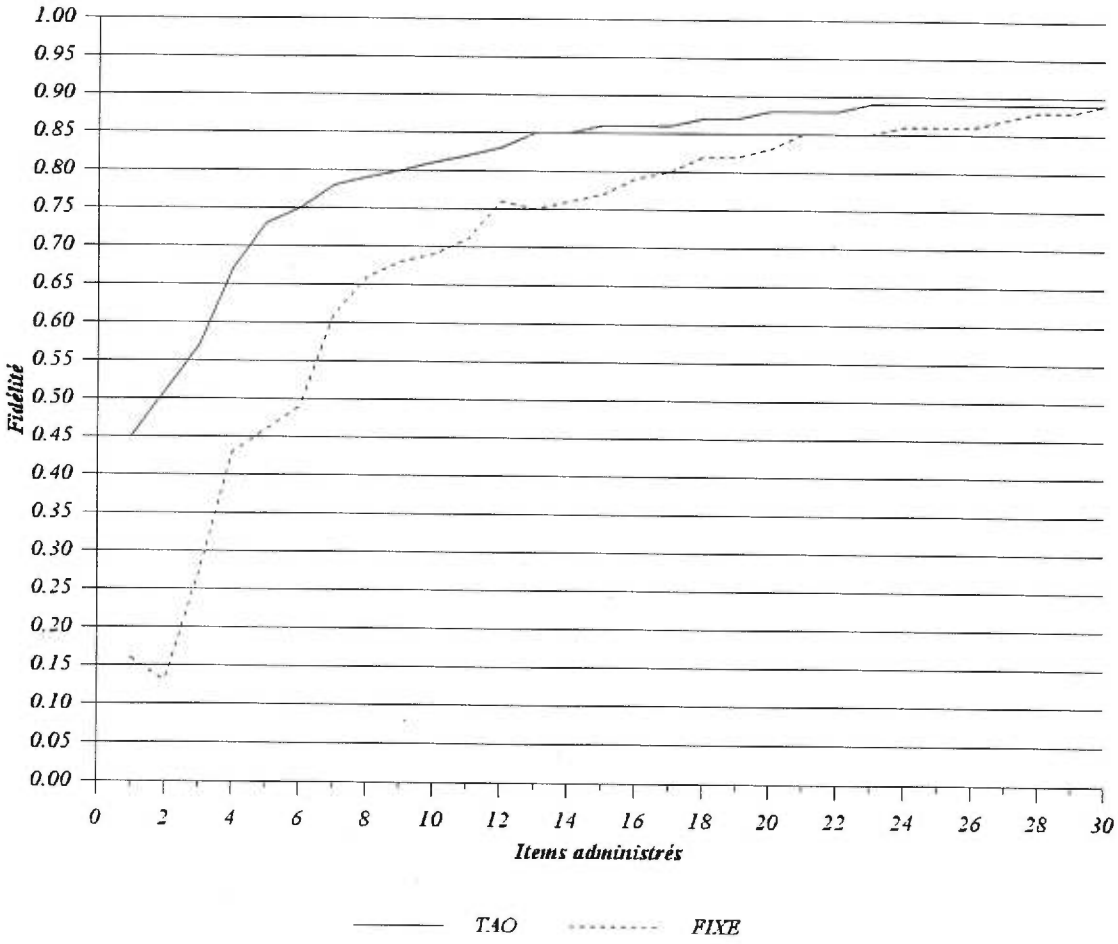

Figure 6.4 Fidélité des tests (formes alternatives) en fonction du nombre d'items administrés (adapté de McBride et Martin (1983, p. 231))

Rappelons que McBride et Martin utilisent une méthode d'estimation du niveau d'habileté et de l'erreur type de l'estimateur du niveau d'habileté qui est affectée par l'ordre de présentation des items, soit la méthode bayésienne d'Owen. Cette caractéristique non désirée de la méthode bayésienne d'Owen a été discutée à la section 5.2.4 traitant des méthodes d'estimation provisoire du niveau d'habileté. L'utilisation d'une méthode d'estimation indépendante de l'ordre de présentation des items, telles que les méthodes de vraisemblance maximale (ML), de maximisation a posteriori (MAP) ou de l'espérance a posteriori (EAP), serait plus appropriée. De plus, McBride et Martin, comme Vispoel et *al.*, ne présentent pas les valeurs de l'erreur type de l'estimateur du niveau d'habileté, $S_{OWEN(\theta)}$, ou de l'information obtenue au niveau d'habileté estimé après l'administration de chaque item. Ce renseignement nous aurait permis d'analyser la relation entre le nombre d'items administrés et l'erreur type de l'estimateur du niveau d'habileté telle qu'elle est calculée, de façon usuelle, en testing adaptatif.

6.4 Bock et Mislevy (1982)

L'étude de Bock et Mislevy (1982) nous offre l'occasion d'analyser les avantages d'une méthode d'estimation du niveau d'habileté en testing adaptatif qui ne soit pas affectée par l'ordre de présentation des items. Plus spécifiquement, ces auteurs analysent les avantages de la méthode de l'espérance a posteriori (EAP).

Au hasard, 500 patrons de réponses sont générés à partir d'une banque d'items précalibrés. On fait débuter les tests adaptatifs en assumant une distribution a priori N(0,1) du niveau d'habileté et on utilise les modélisations logistiques à deux et trois paramètres de la réponse à l'item. La règle de sélection du prochain item n'est pas spécifiée et les estimateurs provisoires et finaux du niveau d'habileté sont obtenus par la méthode de l'espérance a posteriori. Les tests adaptatifs se terminent lorsque des valeurs prédéterminées de l'erreur type de l'estimateur du niveau d'habileté sont atteintes ($S_{EAP(\theta)}$ = 0,40, 0,30 et 0,20). Le tableau 6.7 décrit le déroulement des tests adaptatifs réalisés à l'intérieur de l'étude de Bock et Mislevy. La banque d'items est composée d'items dont les paramètres de discrimination et de pseudo-chance sont fixés respectivement à 1,00 et 0,20.

Bock et Mislevy remarquent que le nombre moyen d'items administrés est moindre lorsque le niveau d'habileté est estimé à partir d'une modélisation logistique à deux paramètres plutôt que par une modélisation logistique à trois paramètres. Ainsi, pour obtenir une valeur de la fidélité de 0,91 ($S_{EAP(\theta)}$ = 0,30), 17 items doivent être administrés en moyenne lorsqu'une modélisation à deux paramètres est utilisée, contre 25 dans le cas d'une modélisation à trois paramètres. Dans le cas de la modélisation à deux paramètres, les auteurs nous indiquent que la valeur de la fidélité est déjà de 0,84 dès le 9^{e} item administré.

Tableau 6.7
Algorithme décrivant le déroulement des tests adaptatifs dans l'étude réalisée par Bock et Mislevy (1982)

RÈGLE	ACTION
1. Règle de départ	Administrer un item dont le niveau de difficulté est égal à 0,00
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, EAP(θ), selon une stratégie non spécifiée de maximisation de l'information
3. Règle d'arrêt	Terminer le test selon trois valeurs de la règle d'arrêt basée sur l'erreur type de l'estimateur du niveau d'habileté : 0,20, 0,30 ou 0,40 L'estimation finale du niveau d'habileté est réalisée selon la méthode de l'espérance à posteriori

Plus pertinente au présent projet est la présentation par les auteurs d'un profil spécifique de réponses aux items lorsque le niveau d'habileté est de -0,50. Le tableau 6.8 et la figure 6.5 illustrent le déroulement du test adaptatif ainsi simulé et présentent la séquence des valeurs de l'erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$. Ces données permettent d'analyser de plus près la relation entre le nombre d'items administrés et la valeur de l'erreur type de l'estimateur du niveau d'habileté. Ainsi, après l'administration de 20 items, la valeur de l'erreur type de l'estimateur du niveau d'habileté observée est de 0,25. Elle est déjà de 0,42 après l'administration de 10 items.

Tableau 6.8
Profil spécifique de l'erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, dans un test adaptatif en fonction du nombre d'items administrés (adapté de Bock et Mislevy (1982, p. 434))

ITEMS ADMINISTRÉS	$S_{EAP(\theta)}$
1	0,89
2	0,83
3	0,74
4	0,70
5	0,56
6	0,57
7	0,50
8	0,48
9	0,48
10	0,42
11	0,36
12	0,36
13	0,36
14	0,33
15	0,31
16	0,30
17	0,29
18	0,27
19	0,26
20	0,25

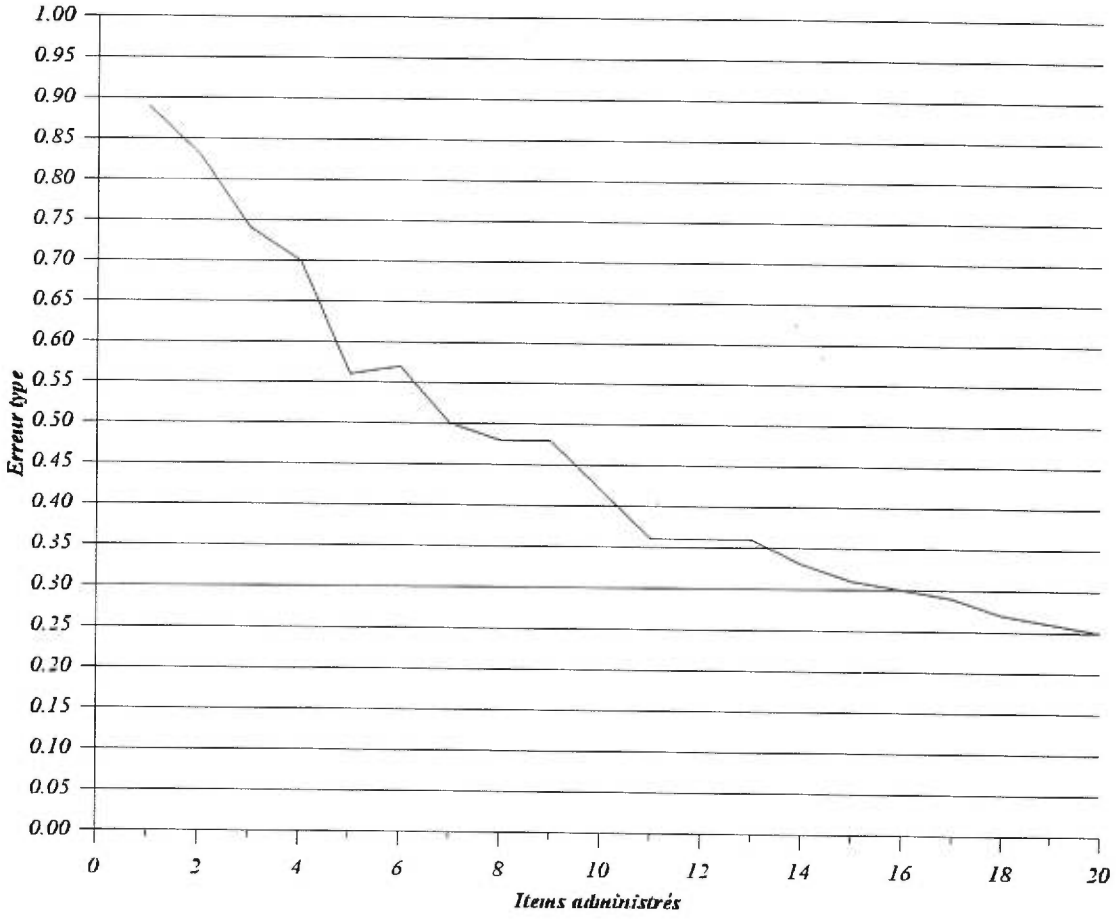

Figure 6.5 Erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, en fonction du nombre d'items administrés (adapté de Bock et Mislevy (1982, p. 434))

L'étude de Bock et Mislevy, comme toutes celles commentées dans ce chapitre, utilise cependant des estimateurs dont le comportement est théoriquement valide seulement lorsque le nombre d'items tend vers l'infini. Un comportement asymptotique est alors présumé.

6.5 Caractéristiques de la distribution d'échantillonnage de l'estimateur du niveau d'habileté lorsque le nombre d'items administrés est petit

6.5.1 Samejima (1994)

Lorsque le nombre d'items administrés tend vers l'infini, l'estimateur du niveau d'habileté se distribue selon une distribution normale, que soit utilisée la méthode de vraisemblance maximale (Hambleton et Swaminathan, 1987, p. 89 ; Klauer, 1990 : voir Hoijtink et Boomsma, 1995, p. 56) ou la méthode de maximisation a posteriori (Chang et Stout, 1993). En ce qui concerne l'estimation par la méthode de l'espérance a posteriori, au moment où Hoijtink et Boomsma écrivent ces lignes (1995, p. 58), il n'existe aucun résultat théorique quant aux propriétés asymptotiques des estimateurs. Au meilleur de nos connaissances, c'est encore le cas aujourd'hui.

Dans les méthodes d'estimation par vraisemblance maximale et de maximisation a posteriori, il existe un biais entre l'estimateur du niveau d'habileté et le niveau

d'habileté. Lorsque le nombre d'items administrés tend vers l'infini, ce biais devient nul. Lord (1983, p. 237) et Samejima (1994) ont étudié le biais dans la méthode d'estimation par vraisemblance maximale. Ils proposent une fonction permettant de l'évaluer et, de cette fonction, Samejima (1994) dérive des fonctions modifiées qui permettent le calcul de l'information et de l'erreur type de l'estimateur du niveau d'habileté en fonction du biais de l'estimateur du niveau d'habileté. Lorsque la méthode de maximisation a posteriori est utilisée, Warm (1989, p. 429) propose des fonctions du même type dans le but de calculer le biais de l'estimateur du niveau d'habileté ainsi que l'information ajustée en fonction de ce biais.

Samejima (1994, p. 242) souligne d'ailleurs que l'utilisation de ces fonctions modifiées serait plus appropriée dans le contexte du testing adaptatif quand la règle d'arrêt est basée sur l'erreur type de l'estimateur du niveau d'habileté.

Samejima (1994) a plus spécifiquement étudié l'impact de l'ajustement de l'erreur type de l'estimateur du niveau d'habileté en utilisant les fonctions modifiées tenant compte du biais de l'estimateur du niveau d'habileté obtenu par la méthode de vraisemblance maximale. À cette fin, elle compare trois méthodes pour calculer l'erreur type de l'estimateur du niveau d'habileté lorsque cet estimateur est obtenu par la méthode de vraisemblance maximale. Plus précisément, elle applique trois formules pour calculer l'information ($I_{ML(\theta)}$) et, de ces formules, elle dérive l'erreur type de l'estimateur du niveau d'habileté ($S_{ML(\theta)}$). La première de ces méthodes de calcul utilise une

approximation de la limite minimale de la variance de l'estimateur du niveau d'habileté et est symbolisée par $S_{\Upsilon(\theta)}$, tandis que la seconde intègre une approximation de la limite minimale du carré moyen de l'estimateur du niveau d'habileté et est notée $S_{\Xi(\theta)}$. La troisième méthode de calcul de l'erreur type de l'estimateur du niveau d'habileté est tout simplement celle qui est habituellement utilisée pour obtenir cette erreur type quand la méthode de vraisemblance maximale est appliquée, $S_{ML(\theta)}$.

Pour permettre la comparaison de l'erreur type obtenue à partir de ces trois méthodes de calcul, la simulation d'un test conventionnel composé de 30 items équivalents (a = 1, b = 0, c = 0 et D = 1,70) est réalisée. Le choix de ces paramètres d'items est justifié par Samejima (p. 236) du fait que dans un test adaptatif, sauf pour les premiers items administrés, les items sélectionnés devraient être équivalents en ce qui a trait à leur niveau de difficulté. La simulation est effectuée en utilisant six distributions hypothétiques du niveau d'habileté, θ : N(0,00, 1,00), N(-0,80, 1,00), N(0,00, 0,50), N(-0,80, 0,50), N(-1,60, 0,50) et N(-2,40, 0,50). Dans le cas des deux premières distributions, 1998 simulations sont effectuées, tandis que pour les quatre autres, 2004 simulations sont réalisées.

Selon les résultats obtenus et illustrés à la figure 6.6, la différence entre la valeur de l'erreur type de l'estimateur du niveau d'habileté ajustée en fonction du biais de l'estimateur du niveau d'habileté et la valeur de l'erreur type calculée sans ajustement devient très importante lorsque le niveau d'habileté est en dehors de l'intervalle compris

entre -1,00 et 1,00. Par exemple, lorsque $\theta = -1,60$, l'erreur type de l'estimateur du niveau d'habileté calculée sans ajustement du biais, $S_{ML(\theta)}$, est égale à 0,49 alors qu'elle est de 0,73 et de 0,77, selon la fonction modifiée utilisée ($S_{T(\theta)}$ ou $S_{\Xi(\theta)}$), lorsque le biais de l'estimateur du niveau d'habileté est considéré. Quand la moyenne de la distribution d'échantillonnage de l'estimateur du niveau d'habileté est égale à -2,40, l'erreur type ajustée en fonction du biais va jusqu'à atteindre des valeurs de 2,76 et 2,89 : l'erreur type de l'estimateur du niveau d'habileté obtenue sans ajustement est de 0,92, ce qui constitue une valeur déjà très importante.

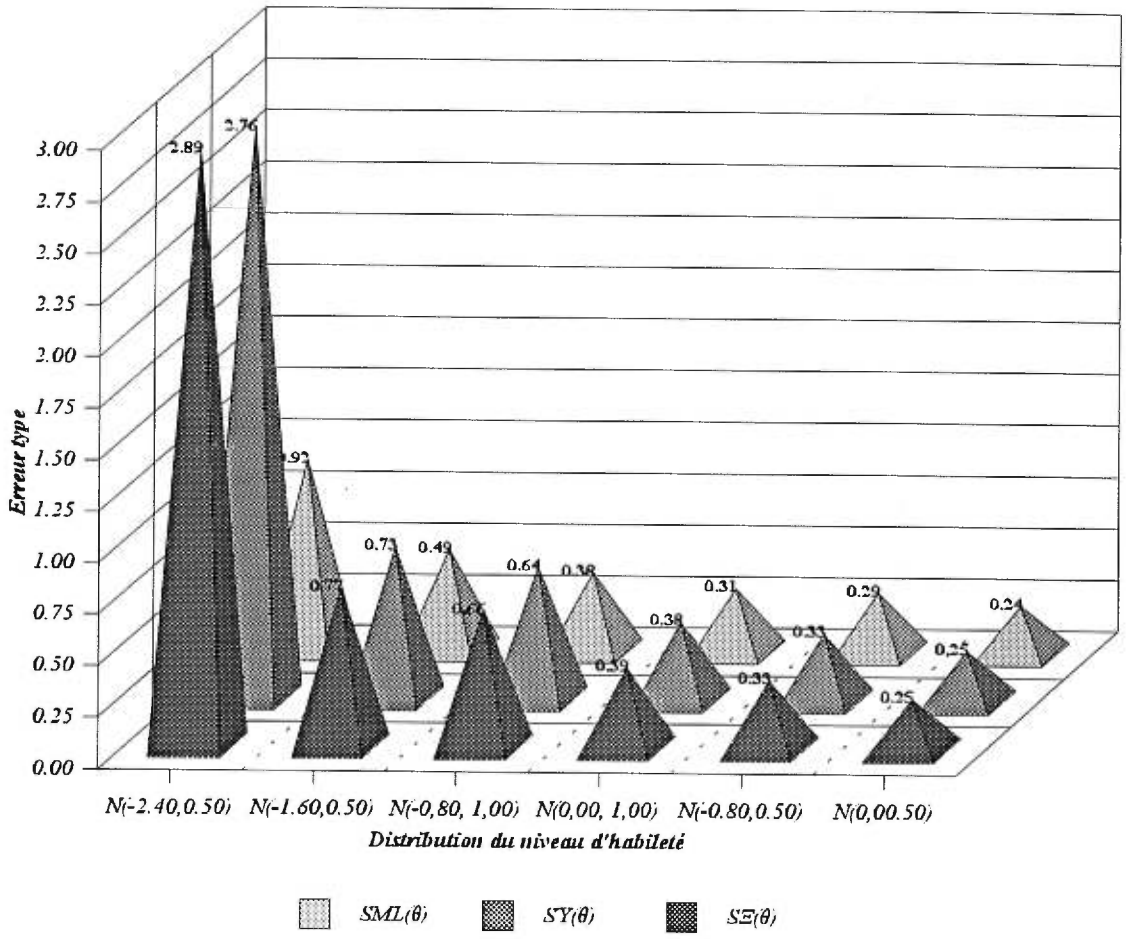

Figure 6.6 Erreur type de l'estimateur du niveau d'habileté en fonction de la distribution de probabilité du niveau d'habileté et de la méthode de calcul de l'erreur type tenant compte, $S_{Y(\theta)}$ et $S_{\Xi(\theta)}$, ou non, $S_{ML(\theta)}$, du biais de l'estimateur du niveau d'habileté (adapté de Samejima, 1994, p. 239)

Les résultats obtenus par Samejima montrent que, même avec l'administration de 30 items, la prise en compte du biais de l'estimateur du niveau d'habileté peut affecter de manière importante la valeur de l'erreur type de l'estimateur du niveau d'habileté. Toutefois, comme Samejima (p. 242) le souligne, ce type de test est peu réaliste dans le contexte d'un test adaptatif où le niveau de difficulté des items se rapproche de plus en plus du niveau d'habileté avec l'augmentation du nombre d'items administrés. Il est aussi à noter que Samejima limite son étude à l'administration de 30 items. Ainsi aucune information n'est disponible en ce qui concerne l'impact de la variation du nombre d'items administrés.

6.5.2 Hoijtink et Boomsma (1995, 1996)

Hoijtink et Boomsma (1995, 1996) comparent les valeurs du biais et de la variance d'erreur de l'estimateur du niveau d'habileté obtenues par les méthodes de vraisemblance maximale ($S^2_{ML(\theta)}$) et de maximisation a posteriori ($S^2_{MAP(\theta)}$).

Ils simulent 1000 valeurs du niveau d'habileté selon 13 distributions normales dont la moyenne varie entre 0,00 et 3,00. La modélisation logistique à un paramètre de la réponse à l'item est appliquée et le nombre d'items administrés est de 5, 15 et 25. Dans la méthode de maximisation a posteriori, la distribution de probabilité a priori du niveau d'habileté correspond à N(0, 1). Dans la méthode de vraisemblance maximale, une

valeur de -5,00 est attribuée arbitrairement à l'estimateur du niveau d'habileté si aucune bonne réponse n'est obtenue. À l'inverse, si tous les items reçoivent une bonne réponse, l'estimateur du niveau d'habileté est fixé à 5,00.

Lorsque le nombre d'items administrés est égal à 5, le niveau de difficulté des items correspond à -1,50, -0,75, 0,00, 0,75 et 1,50. Quand 15 items sont administrés, leur niveau de difficulté varie entre -2,50 et 2,50 tandis qu'il se situe entre -3,00 et 3,00 lorsque 25 items sont administrés.

Pour chacune des valeurs simulées du niveau d'habileté, Hoijtink et Boomsma calculent les valeurs théoriques du biais et de la variance d'erreur, $S^2_{ML(\theta)}$ et $S^2_{MAP(\theta)}$, de l'estimateur du niveau d'habileté et les comparent aux valeurs empiriques calculées à partir des valeurs connues du niveau d'habileté simulé.

Leurs résultats indiquent que, lorsque le nombre d'items administrés est égal à 25, le biais de l'estimateur du niveau d'habileté se rapproche plus de la valeur empirique que lorsque le nombre d'items administrés est égal à 5 ou à 15. Ils tirent les mêmes conclusions en ce qui concerne la variance d'erreur de l'estimateur du niveau d'habileté. Selon eux (1996), lorsque la méthode de vraisemblance maximale est utilisée, le nombre d'items administrés doit être au minimum de 15 pour que la distance entre les valeurs théoriques et empiriques soit suffisamment petite. La méthode de maximisation a posteriori, quant à elle, nécessiterait l'administration de 25 items pour que la distance

entre les valeurs théoriques et empiriques soit suffisamment petite. Hoijtink et Boomsma n'indiquent pas, toutefois, ce qu'ils considèrent comme étant une distance suffisamment petite.

Dans le but de faciliter l'interprétation des résultats obtenus par ces auteurs en 1996, nous avons réalisé une approximation des valeurs du biais, de la variance et de l'erreur type, lorsque le niveau d'habileté prend une valeur extrême, soit $\theta = 3{,}00$, à partir des figures 1 à 6 présentées dans leur étude. Nous la présentons au tableau 6.9.

Au tableau 6.9, nous remarquons que le biais de l'estimateur du niveau d'habileté est toujours supérieur lorsque la méthode de maximisation a posteriori est utilisée. La distribution a priori N(0,1) semble donc affecter sensiblement les valeurs obtenues de l'estimateur du niveau d'habileté. Les valeurs théoriques du biais de l'estimateur du niveau d'habileté sont aussi toujours supérieures aux valeurs empiriques, sauf lorsque la méthode de vraisemblance maximale est appliquée et que 25 items sont administrés. Il y a ainsi surestimation des valeurs réelles du biais de l'estimateur du niveau d'habileté. Même avec 25 items administrés, la valeur théorique du biais de l'estimateur du niveau d'habileté calculée à partir de la méthode de maximisation a posteriori, égale à -1,20, est différente de 0,30 unités de la valeur empirique, égale à -0,90. Il est toutefois à noter que les limites des niveaux de difficulté des tests administrés à partir de 5, 15 et 25 items sont respectivement de $\pm 1{,}50$, $\pm 2{,}50$ et $\pm 3{,}00$, les valeurs théoriques et empiriques du biais de l'estimateur du niveau d'habileté ont sûrement été affectées par ces limites

différentes selon le nombre d'items administrés. Il aurait été préférable, selon nous, d'utiliser les mêmes limites quel que soit le nombre d'items administrés, cela aurait permis de rendre comparables les trois types d'administration selon le nombre d'items utilisés.

Quant à la variance d'erreur de l'estimateur du niveau d'habileté, elle est toujours plus importante lorsque la méthode de vraisemblance maximale, $S^2_{ML(\theta)}$, est utilisée plutôt que la méthode de maximisation a posteriori, $S^2_{MAP(\theta)}$. De plus, les valeurs théoriques de la variance d'erreur de l'estimateur du niveau d'habileté sont, dans tous les cas, supérieures aux valeurs empiriques. Ce n'est qu'avec l'administration de 25 items que les valeurs théoriques et empiriques de la variance se rapprochent, et ce pour les deux méthodes d'estimation du niveau d'habileté.

Tableau 6.9
Approximation, lorsque $\theta = 3,00$, des valeurs du biais, de la variance d'erreur et de l'erreur type obtenues à partir des figures 1 à 6 présentées dans l'étude réalisée par Hoijtink et Boomsma (1996, p. 317-322)

	5 ITEMS (b = -1,50 à 1,50)		15 ITEMS (b = -2,50 à 2,50)		25 ITEMS (b = -3,00 à 3,00)	
	MLE	MAP	MLE	MAP	MLE	MAP
Valeurs empiriques du biais	0,90	-2,00	0,25	-1,25	0,15	-0,90
Valeurs théoriques du biais	6,00	-10,00	0,50	-2,50	0,15	-1,20
Valeurs empiriques de S^2	2,50	0,25	1,00	0,15	0,58	0,15
Valeurs théoriques de S^2	12,50	0,50	1,75	0,45	0,60	0,22
Valeurs empiriques de S	1,58	0,50	0,55	0,40	0,76	0,39
Valeurs théoriques de S	3,54	0,71	1,32	0,67	0,77	0,45

Les résultats obtenus par Hoijtink et Boomsma (1995, 1996), tout comme ceux de Samejima (1994), ont toutefois été obtenus dans un contexte de testing non adaptatif et, de plus, le nombre d'items administrés n'a pas été l'objet de variations permettant d'analyser en détail les propriétés de l'estimateur du niveau d'habileté lorsque le nombre d'items administrés est petit. Il y aurait aussi intérêt à effectuer des études en ce qui concerne l'estimation du niveau d'habileté par la méthode de l'espérance a posteriori.

6.6 Précisions sur l'objectif de recherche

Dans cette recherche, nous abordons spécifiquement l'étude de l'impact de la variation des critères dans deux règles d'arrêt, selon le nombre d'items administrés et selon l'erreur type de l'estimateur du niveau d'habileté sur la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Cette recherche réalisée dans un contexte de testing adaptatif tient compte de l'estimation du niveau d'habileté lorsqu'un petit nombre d'items est administré et lorsque la méthode de l'espérance a posteriori est appliquée pour calculer l'estimateur du niveau d'habileté.

Plusieurs valeurs du niveau d'habileté sont générées au hasard et, pour chacune de ces valeurs, nous calculons l'estimateur du niveau d'habileté en fonction des critères retenus dans les deux règles d'arrêt. Par la même occasion, nous calculons différentes statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté telles que l'erreur type, l'asymétrie, la kurtose, le biais, la proportion de bonnes réponses et le nombre d'items administrés. Les valeurs de l'asymétrie et de la kurtose que nous obtenons seront utiles pour déterminer l'allure de la distribution d'échantillonnage de l'estimateur du niveau d'habileté et pour illustrer jusqu'à quel point cette dernière s'éloigne d'une distribution normale.

Les résultats obtenus nous permettront de porter un jugement sur le nombre d'items à administrer ainsi que sur la valeur à retenir de l'erreur type de l'estimateur du niveau

d'habileté lorsque la règle d'arrêt basée sur une valeur prédéterminée de l'erreur type de l'estimateur du niveau d'habileté est utilisée. Nous porterons aussi un jugement sur le nombre d'items à administrer lorsque la règle d'arrêt basée sur le nombre d'items à administrer est utilisée.

7. Méthodologie

Dans ce chapitre, nous présentons la méthodologie retenue pour étudier différentes statistiques associées la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de la variation des critères dans les règles d'arrêt. Dans le but d'exercer un contrôle strict de la situation de testing et qu'un nombre important d'observations soit disponible, nous proposons une simulation informatisée. Selon Harwell (1997, p. 266) et Harwell, Stone, Hsu et Kirisci (1996, p.103), l'utilisation d'une simulation dans les études sur la théorie de la réponse à l'item est appropriée lorsqu'on étudie la distribution d'échantillonnage des estimateurs ou que l'on compare plusieurs méthodes visant le même objectif, sans qu'il soit possible d'obtenir une solution analytique exacte. Selon les mêmes auteurs, l'utilisation d'une simulation permet aussi plus facilement la manipulation de différents facteurs à la fois, ce qui n'est pas toujours possible à réaliser dans des conditions réelles. Notre recherche rencontre ces conditions. Harwell et *al.* (p. 104) soulignent toutefois que l'utilité des résultats de ce type de recherche est fortement dépendante du réalisme de la situation simulée. Les observations, ici différentes valeurs du niveau d'habileté, sont générées au hasard, ce qui caractérise une simulation de type Monte Carlo ou stochastique.

Puisque les résultats obtenus peuvent varier lorsque les caractéristiques des items composant la banque d'items sont modifiées, nous effectuerons une analyse sur des données qui ne tiennent pas compte de la nature de la banque d'items de manière à

neutraliser l'effet de sa composition. La banque d'items sera ainsi composée de façon à ce que le choix des valeurs des paramètres de difficulté, de discrimination ou de pseudo-chance dépende uniquement de la règle de sélection. Urry (1970, p. 82 ; Thissen et Mislevy, 1990, p. 111) propose une règle de sélection qui satisfait cette condition. Selon cette règle, le prochain item administré correspond à un item dont le niveau de difficulté est égal à l'estimateur du niveau d'habileté après l'administration de l'item précédent. Le résultat de cette stratégie n'est d'ailleurs pas incompatible avec des situations réelles de testing adaptatif lorsque sont considérées des modélisations permettant la génération de tous les items d'un univers, telles que le proposent Bejar (1993), Bennett (1999) et Embretson (1999).

De plus, pour éviter l'impact du paramètre de discrimination sur la sélection des items à administrer décrit par Thissen et Mislevy (1990, p. 112-113), on utilisera une modélisation logistique à un paramètre de la réponse à l'item. Cela nous permettra aussi de réaliser l'étude sans que les valeurs des paramètres de discrimination et de pseudo-chance n'aient à être contrôlées. Les résultats seront donc principalement généralisables aux tests où une modélisation logistique à un paramètre de la réponse à l'item est appliquée. Notre attention se dirigera plus spécifiquement sur l'impact de la variation de la valeur des critères dans les règles d'arrêt sur la distribution d'échantillonnage de l'estimateur du niveau d'habileté et de différentes statistiques associées : erreur type, biais, proportion de bonnes réponses, nombre d'items administrés, asymétrie et kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Il nous sera

ainsi possible d'observer la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif et d'émettre des prescriptions quant au nombre d'items à administrer ou à l'erreur type à spécifier.

Nous décrivons d'abord les données qui sont simulées, le déroulement de la simulation, ainsi que la vérification de l'exactitude du calcul de l'estimateur du niveau d'habileté, de l'erreur type, de l'asymétrie et de la kurtose. Nous abordons ensuite les balises d'interprétation des valeurs de l'asymétrie et de la kurtose d'une distribution de probabilité normale ainsi que la méthode d'analyse des résultats. Nous présentons tout au long du chapitre les justifications de choix méthodologiques.

7.1 Caractéristiques des données simulées

Habituellement, les diverses recherches qui recourent à des simulations de tests construits selon la théorie de la réponse à l'item utilisent, en fonction de la problématique étudiée, soit des valeurs aléatoires, soit une ou plusieurs valeurs fixes, soit des valeurs réelles du niveau d'habileté à partir desquelles les simulations sont effectuées (Mcbride, Wetzel et Hetter, 1997, p. 88).

Les recherches où les simulations s'effectuent à partir de valeurs aléatoires du niveau d'habileté (Chen et Thissen, 1997 ; Chen, Hou, Fitzpatrick et Dodd, 1997 ; de Ayala et

Hertzog, 1991 ; de Ayala, Schafer et Sava-Bolesta, 1995 ; Dodd, 1990 ; Dodd, Koch et de Ayala, 1993 ; Jansema, 1974,1977 ; Kinsbury et Weiss, 1983 ; Luecht, 1996 ; Nering, 1997 ; Ramsey, 1991 ; Samejima, 1994 ; Stout, 1987 ; Zwick, Tayer et Wingersky, 1994) permettent de connaître le comportement de diverses statistiques obtenues dans un test adaptatif en fonction des caractéristiques de la distribution de probabilité préalable du niveau d'habileté. Généralement, cette distribution suit une loi normale N(0,1). Certains auteurs tels que Chen, Hou, Fitzpatrick et Dodd (1997), qui s'intéressent à la robustesse des estimateurs, utilisent des distributions de probabilité du niveau d'habileté qui s'éloignent d'un loi normale. L'utilisation de valeurs aléatoires du niveau d'habileté pour effectuer les simulations d'un test adaptatif offre l'avantage de permettre d'étudier la distribution d'échantillonnage de l'estimateur du niveau d'habileté sur toute l'étendue des valeurs possibles du niveau d'habileté et de s'assurer que le niveau d'habileté se distribue réellement selon une loi de probabilité prédéterminée.

D'autres auteurs (Bock et Mislevy, 1982 ; Chang et Yin, 1996 ; de Ayala, 1992a, 1992b ; de Ayala, Dodd et Koch, 1990 ; Hoijtink et Boomsma, 1995, 1996 ; Reckase, 1983 ; Samejima, 1977 ; Wainer et Thissen, 1987 ; Warm, 1989) utilisent une ou des valeurs fixes du niveau d'habileté pour effectuer leurs simulations. Bock et Mislevy (1982), par exemple, effectuent 500 simulations d'un test adaptatif à partir d'une seule valeur du niveau d'habileté fixée à -0,50. Wainer et Thissen (1987), pour leur part, utilisent cinq valeurs du niveau d'habileté variant entre -2 et 2 (-2, -1, 0, 1, 2) et effectuent 100 simulations à chacun de ces niveaux d'habileté. Cette approche permet

de connaître le comportement de l'estimateur du niveau d'habileté aux divers niveaux d'habileté fixés, mais elle ne permet pas de vérifier, par exemple, si, sur toute l'étendue des valeurs du niveau d'habileté, la distribution de probabilité de l'estimateur du niveau d'habileté suit une loi normale N(0,1). Il serait toutefois possible de le vérifier par l'utilisation d'un nombre plus important de valeurs fixes du niveau d'habileté et en pondérant les estimateurs obtenus du niveau d'habileté en fonction de la probabilité théorique d'apparition des valeurs fixes du niveau d'habileté.

Enfin, de Ayala (1989), ainsi que Hetter, Segall et Bloxom (1997) et Vispoel, Wang et Bleiler (1997), préfèrent effectuer les simulations à partir de valeurs du niveau d'habileté obtenues préalablement à des tests. C'est ce que font Vispoel et *al.* lorsqu'ils utilisent les résultats obtenus auprès de 731 élèves de tests visant à estimer le niveau d'habileté de rappel des mélodies. Cette stratégie permet d'étudier la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif à partir d'une distribution de probabilité réelle du niveau d'habileté. La stratégie ne permet pas nécessairement de généraliser les résultats obtenus à l'estimation du niveau d'habileté dans un autre domaine. Elle ne permet surtout pas de contrôler la nature de la distribution de probabilité du niveau d'habileté.

La présente recherche s'intéresse à l'étude de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif dans des conditions où le niveau d'habileté se distribue selon une loi de probabilité normale. Une distribution normale du

niveau d'habileté est justifiée de trois façons. Premièrement, divers auteurs s'intéressant au comportement asymptotique de la distribution d'échantillonnage de l'estimateur du niveau d'habileté montrent ou soulignent que cet estimateur du niveau d'habileté se distribue selon une loi normale N(0,1) (Chang et Stout, 1993 ; Hambleton et Swaminathan, 1987, p. 89 ; Hoijtink et Boomsma, 1995, 1996 ; Klauer, 1990 ; Lord, 1983, p. 234-236 ; Samejima, 1994, p. 232). En second lieu, la plupart des simulations répertoriées dans la littérature, et où sont générées des valeurs aléatoires du niveau d'habileté, utilisent une distribution normale N(0,1) du niveau d'habileté. Enfin, une distribution normale N(0,1) a priori est l'option par défaut à l'intérieur du logiciel PC-BILOG (Mislevy et Bock, 1984, p. 7) lorsque l'estimation du niveau d'habileté est effectuée par la méthode de l'espérance a posteriori.

Pour les fins de ce travail, les valeurs du niveau d'habileté utilisées pour effectuer les simulations d'un test adaptatif sont aléatoires et sont tirées d'une distribution de probabilité qui suit une loi normale N(0,1). Un échantillon constitué de 2000 valeurs différentes du niveau d'habileté est généré de façon aléatoire. C'est une taille d'échantillon convenable considérant que les tailles d'échantillon utilisées dans la littérature consultée varient entre 100 et 10 000, avec une valeur médiane de 500.

7.2 Déroulement de la simulation

Nous décrivons ici le déroulement de la simulation, les règles de départ, de suite et d'arrêt, pour ensuite présenter la méthode d'estimation provisoire du niveau d'habileté, tout comme celle de l'estimation finale du niveau d'habileté.

7.2.1 Simulation des réponses aux items

La simulation d'un test adaptatif est appliquée à chacune des 2000 valeurs aléatoires du niveau d'habileté. Chaque simulation est réalisée selon une méthode de génération d'une réponse aux items fréquemment utilisée à l'intérieur des recherches sur la théorie de la réponse à l'item (de Ayala, 1992a, p. 516 ; de Ayala, Dodd et Koch, 1990, p. 230 ; de Ayala et Herzog, 1991, p. 768 ; de Ayala, Schafer et Sava-Bolesta, 1995, p. 389 ; Dodd, 1990, p. 358 ; Harwell, Stone, Hsu et Kirisci, 1997, p. 116 ; Hoijtink et Boomsma, 1996, p. 317 ; Kinsbury et Weiss, 1983 ; Luecht, 1996, p. 398 ; Nicewander et Thomasson, 1999, p. 244 ; Reckase, 1983 ; Segall, Moreno et Hetter, 1997, p. 129 ; Warm, 1989, p. 432). Selon cette méthode, pour chaque valeur du niveau d'habileté générée au hasard, on obtient la réponse à chacun des items en calculant la probabilité d'obtenir une bonne réponse à l'item, $P(r = 1 | \theta)$, en tenant compte du paramètre de difficulté de l'item ainsi que de la valeur du niveau d'habileté. La façon de déterminer la valeur du paramètre de difficulté de l'item est expliquée à la section traitant de la règle

de départ et de la sélection des items. Cette probabilité est ensuite comparée à un nombre aléatoire x, compris entre 0 et 1, tiré d'une distribution de probabilité uniforme U(0,1). Si la probabilité d'obtenir une bonne réponse à l'item $P(r = 1|\theta)$ est supérieure au nombre aléatoire x, la réponse à l'item prend la valeur 1, soit une bonne réponse. Sinon, la réponse à l'item prend la valeur 0, soit une mauvaise réponse. Ainsi,

$$\text{si } P(r = 1|\theta) \geq x, \text{ alors } r = 1, \text{ sinon } r = 0 \qquad \textbf{(7.1)}$$

7.2.2 Règles de départ et de suite

À tous les niveaux d'habileté simulés, le test débute par l'administration d'un item dont le niveau de difficulté, b_1, est égal à 0, soit la moyenne de la distribution a priori. L'utilisation d'un niveau de difficulté, b_1, constant à tous les niveaux d'habileté simulé permet de s'assurer que l'estimateur du niveau d'habileté obtenu ne varie pas en fonction du niveau de difficulté du premier item administré.

Nous utilisons la méthode de Urry (Thissen et Mislevy, 1990, p. 111 ; Urry, 1970, p. 82) pour sélectionner le prochain item. Selon cette méthode, le prochain item à administrer, b_{j+1}, correspond à un item dont le niveau de difficulté est égal à l'estimateur provisoire du niveau d'habileté, $EAP_j(\theta)$, après l'administration de l'item j.

$$b_{j+1} = EAP_j(\theta) \qquad \textbf{(7.2)}$$

Cette règle de sélection des items, lorsque le modèle à un paramètre est utilisé, permet d'obtenir le prochain item qui fournit l'information maximale ; elle est donc équivalente à la stratégie de maximisation de l'information. De plus, elle permet de faire en sorte que le choix des valeurs du paramètre de difficulté dépende uniquement de la règle de sélection de façon à ce que la composition de la banque d'items ne puisse affecter les valeurs de l'estimateur du niveau d'habileté.

7.2.3 Méthode d'estimation provisoire du niveau d'habileté

L'estimateur provisoire du niveau d'habileté est calculé selon la méthode de l'espérance a posteriori. L'estimateur a posteriori du niveau d'habileté, $EAP_j(\theta)$, est calculé pour chaque valeur j du nombre d'items administrés selon une approximation de l'intégrale (Baker, 1992, p. 211 ; Bock et Mislevy, 1982, p. 433 ; de Ayala, Schafer et Sava-Bolesta, 1995, p. 386) présentée auparavant à l'équation 4.10 :

$$EAP_j(\theta) = \frac{\sum_{k=1}^{q} X_k L_j(X_k) A(X_k)}{\sum_{k=1}^{q} L_j(X_k) A(X_k)} \qquad \textbf{(7.3)}$$

où X_k est un des q points de quadrature équidistants compris entre $\theta = -4$ et $\theta = 4$, $A(X_k)$ est la pondération associée à chacun des points de quadrature selon une loi de probabilité N(0,1) et

$$L_j(\theta) = \prod_{i=1}^{j} P(r_i|\theta,b_i)^{r_i} Q(r_i|\theta,b_i)^{1-r_i} \qquad \textbf{(7.4)}$$

est la vraisemblance (*likelihood*) du patron de réponses, $R = \{r_1 \ldots r_j\}$ après l'administration de j items. La contrainte suivante est de plus imposée :

$$\sum_{k=1}^{q} A(X_k) = 1 \qquad \textbf{(7.5)}$$

L'intégration est ainsi réalisée selon la méthode de l'histogramme de Mislevy (Baker, 1992, p. 187), avec 40 points de quadrature dont la pondération est égale à la probabilité a priori à ces points (Bock et Mislevy, 1982, p. 433 ; de Ayala, Schafer et Sava-Bolesta, 1995, p. 387). En intégration numérique, le nombre de points de quadrature est en lien direct avec le degré de précision obtenue. L'impact du nombre de points de quadrature sur la précision des calculs sera abordé plus loin, à la section 7.3, lors de la vérification de l'exactitude des résultats.

L'erreur type de l'estimateur du niveau d'habileté est calculée selon :

$$S_{EAP_j(\theta)} = \left[\frac{\sum_{k=1}^{q} (X_k - EAP_j(\theta))^2 L_j(X_k) A(X_k)}{\sum_{k=1}^{q} L_j(X_k) A(X_k)} \right]^{1/2} \quad \textbf{(7.6)}$$

Mislevy et Bock (1984, p. 6) indiquent que, pour corriger l'effet du groupement des données en classes de largeur L, ils utilisent la correction de Sheppard (Spiegel, 1961, p. 72) :

$$\sqrt{S^2_{EAP_j(\theta)} - \frac{L^2}{12}} \quad \textbf{(7.7)}$$

Toutefois, Spiegel souligne qu'il n'y a pas accord quant à l'utilisation de la correction de Sheppard. Cette correction pourrait ajouter encore plus d'erreur à l'approximation. De plus, tel que commenté plus loin à la section 7.3, les calculs obtenus par Bock et Mislevy (1982, p. 434), ainsi qu'à partir du logiciel PC-BILOG, ne permettent pas de croire que cette correction est appliquée. Elle ne sera donc pas utilisée.

Les équations 7.8 et 7.9, en concordance avec le calcul des moments centrés proposé par Spiegel (1961, p. 90), sont respectivement utilisées pour réaliser le calcul de l'asymétrie $a3_{EAP(\theta)}$ et de la kurtose $a4_{EAP(\theta)}$ de la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Le calcul de l'asymétrie et de la kurtose est utile pour déterminer

l'allure d'une distribution de probabilité, ici la distribution d'échantillonnage de l'estimateur du niveau d'habileté, et pour illustrer jusqu'à quel point on s'éloigne d'une distribution normale.

$$a3_{EAP_j(\theta)} = \left[\frac{\sum_{k=1}^{q} (X_k - EAP_j(\theta))^3 L_j(X_k) A(X_k)}{\sum_{k=1}^{q} L_j(X_k) A(X_k)} \right] / S^3_{EAP_j(\theta)} \qquad \textbf{(7.8)}$$

$$a4_{EAP_j(\theta)} = -3 + \left[\frac{\sum_{k=1}^{q} (X_k - EAP_j(\theta))^4 L_j(X_k) A(X_k)}{\sum_{k=1}^{q} L_j(X_k) A(X_k)} \right] / S^4_{EAP_j(\theta)} \qquad \textbf{(7.9)}$$

7.2.4 Règles d'arrêt et méthode d'estimation finale du niveau d'habileté

Tous les tests se terminent après l'administration de 60 items. Toutefois, la disponibilité des résultats intermédiaires de 1 à 60 items administrés permet de connaître la valeur de l'estimateur du niveau d'habileté et de son erreur type après l'administration de chacun des 60 items. Sont aussi disponibles, conséquemment, les résultats en ce qui concerne cet estimateur après l'atteinte d'un niveau prédéterminé de l'erreur type de l'estimateur

du niveau d'habileté. L'estimateur final du niveau d'habileté, après l'administration de j items, est égal à l'estimateur provisoire du niveau d'habileté au j^{e} item administré. Le tableau 7.1 présente sommairement le déroulement des tests.

Tableau 7.1
Algorithme décrivant le déroulement des tests adaptatifs utilisés dans la simulation

RÈGLE	ACTION
1. Règle de départ	Administrer un item dont le niveau de difficulté est égal à 0,00
2. Règle de suite	Administrer un item dont le niveau de difficulté est égal à la valeur de l'estimateur provisoire du niveau d'habileté
	L'estimateur provisoire du niveau d'habileté est calculé selon la méthode de l'espérance a posteriori
3. Règle d'arrêt	Terminer les tests après l'administration d'un nombre prédéterminé d'items variant entre 1 et 60 ou lorsqu'une erreur type prédéterminée de l'estimateur du niveau d'habileté variant entre 0,20 et 0,85 est obtenue
	L'estimateur final du niveau d'habileté est calculé selon la méthode de l'espérance a posteriori

7.2.5 Programmation

La simulation est réalisée à partir du langage de programmation du logiciel SAS (1990a) dans sa version 6. Ce langage de programmation est assez flexible et permet d'utiliser des procédures statistiques, qui lui sont déjà incorporées, avec une grande précision dans

les calculs. Le programme complet est présenté à l'annexe III. Étant donné le souci des concepteurs du logiciel SAS d'assurer la stabilité des résultats d'un environnement à un autre, les données brutes sont faciles à reproduire quelle que soit la version de SAS utilisée, en autant qu'elle soit égale ou ultérieure à la version 6, et quelle que soit la plate-forme informatique disponible. De plus, les règles de programmation proposées par Soloway, Adelson et Ehrlich (1988, p. 135) sont considérées (tableau 7.2). La première règle, qui souligne que le nom des variables doit refléter la nature des fonctions, a été l'objet d'une attention particulière. Ainsi, les noms des variables reflètent le plus possible les symboles, notations et indices utilisés dans le texte.

Tableau 7.2
Règles de programmation proposées par Soloway, Adelson et Ehrlich (1988, p. 135)

1	Le nom des variables doit refléter la nature des fonctions
2	Des lignes de code non utilisées ne doivent pas être incluses
3	Si un test de condition est présent, alors la condition doit avoir le potentiel de prendre la valeur *vrai*
4	Une variable initialisée par assignation doit être mise à jour par assignation
5	Il ne faut pas faire double usage de lignes de code
6	Lorsqu'un énoncé est exécuté une seule fois, l'énoncé IF est utilisé, sinon l'énoncé WHILE est appliqué

7.3 Vérification de l'exactitude des calculs

Dans le but de vérifier l'exactitude du calcul de l'estimateur du niveau d'habileté et de son erreur type, une comparaison est effectuée entre des valeurs obtenues par Bock et Mislevy (1982, p. 434) à partir des paramètres d'items qu'ils utilisent dans leur exemple, et celles obtenues en utilisant les procédures de calcul programmées en langage SAS à partir des mêmes paramètres d'items. La comparaison est aussi effectuée avec les valeurs obtenues à partir du logiciel PC-BILOG. Le détail de la programmation de la vérification en langage SAS et le contenu des fichiers nécessaires aux calculs par PC-BILOG sont présentés à l'annexe IV. Les valeurs obtenues à partir des procédures de calcul programmées en langage SAS, celles produites par PC-BILOG ainsi que celles obtenues par Bock et Mislevy, sont présentées au tableau 7.3. Dans tous les cas, l'estimateur du niveau d'habileté et son erreur type sont obtenus en utilisant la constante de Haley et en employant 21 points de quadrature.

Une lecture du tableau 7.3 nous permet de constater que les valeurs obtenues à partir des procédures de calcul programmées en langage SAS et à partir de PC-BILOG sont strictement identiques. Les valeurs obtenues par Bock et Mislevy diffèrent légèrement, quoique très peu, de celles provenant des procédures de calcul programmées en langage SAS et de PC-BILOG, tout au plus de 0,03. Puisque les valeurs correspondent exactement à celles obtenues à l'aide d'un logiciel commercial de référence et sont très près de celles proposées par Bock et Mislevy, le calcul de l'estimateur du niveau

d'habileté et de son erreur type par le biais des procédures de calcul programmées en langage SAS peut être considéré comme exact.

Il faut toutefois souligner que nous avons effectué le calcul de l'erreur type de l'estimateur du niveau d'habileté sans utiliser la correction de Sheppard. Malgré les informations fournies (Bock et Mislevy, 1982, p. 433 ; Mislevy et Bock, 1984, p. 6), il ne semble pas que cette correction soit utilisée dans les calculs. Ainsi, au 20^{e} item, la valeur de l'erreur type modifiée par la correction de Sheppard, telle qu'elle est calculée par l'équation 7.5, est égale à 0,22 plutôt que de 0,25. Considérant ce qui précède, la correction de Sheppard n'est donc pas appliquée.

Tableau 7.3
Calcul de l'estimateur du niveau d'habileté et de l'erreur type de celui-ci obtenus avec 21 points de quadrature à partir de l'exemple de Bock et Mislevy (1982, p. 434)

	SAS		PC-BILOG 1.1		BOCK et MISLEVY (1982)	
ITEM	EAP(θ)	$S_{EAP(\theta)}$	EAP(θ)	$S_{EAP(\theta)}$	EAP(θ)	$S_{EAP(\theta)}$
1	0,38	0,92	0,38	0,92	***0,36**	**0,89**
2	-0,01	0,84	-0,01	0,84	-0,01	**0,83**
3	0,37	0,75	0,37	0,75	0,37	**0,74**
4	0,10	0,70	0,10	0,70	0,10	0,70
5	-0,34	0,57	-0,34	0,57	-0,34	**0,56**
6	-0,58	0,58	-0,58	0,58	-0,58	**0,57**
7	-0,40	0,50	-0,40	0,50	-0,40	0,50
8	-0,67	0,48	-0,67	0,48	**-0,66**	0,48
9	-0,82	0,49	-0,82	0,49	-0,82	**0,48**
10	-0,66	0,42	-0,66	0,42	-0,66	0,42
11	-0,49	0,36	-0,49	0,36	-0,49	0,36
12	-0,56	0,36	-0,56	0,36	-0,56	0,36
13	-0,62	0,36	-0,62	0,36	-0,62	0,36
14	-0,55	0,33	-0,55	0,33	-0,55	0,33
15	-0,49	0,31	-0,49	0,31	-0,49	0,31
16	-0,58	0,30	-0,58	0,30	-0,58	0,30
17	-0,54	0,29	-0,54	0,29	-0,54	0,29
18	-0,49	0,27	-0,49	0,27	-0,49	0,27
19	-0,54	0,26	-0,54	0,26	-0,54	0,26
20	-0,49	0,25	-0,49	0,25	-0,49	0,25

* Les chiffres en caractères gras indiquent que les valeurs obtenues par Bock et Mislevy sont différentes de celles calculées à partir de PC-BILOG 1.1 et des routines programmées dans le langage SAS

Le calcul de l'asymétrie et de la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté par la méthode EAP implique l'approximation numérique de fonctions pour lesquelles la littérature ne fournit aucun exemple. De plus, les logiciels courants, tel que PC-BILOG, ne permettent pas ce calcul. Il semble donc nécessaire de vérifier la stabilité du calcul de l'asymétrie et de la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté par la méthode de l'espérance a posteriori en faisant varier le nombre de points de quadrature. Le tableau 7.4 présente les valeurs de l'asymétrie et de la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté de l'exemple de Bock et Mislevy obtenues lorsque 10, 21, 25, 30, 35 et 80 points de quadrature sont utilisés. Le programme présenté à l'annexe IV permet de réaliser ces calculs.

La lecture du tableau 7.4 permet de constater que les calculs ne sont vraiment stables qu'à partir du moment où 25 points de quadrature sont utilisés. Avec 10 points de quadrature, les valeurs diffèrent considérablement de celles obtenues avec 80 points de quadrature. À partir de 25 points de quadrature, aucune différence supérieure à 0,05 n'est remarquée.

Tableau 7.4

Calcul de l'asymétrie, $a3_{EAP(\theta)}$, et de la kurtose, $a4_{EAP(\theta)}$, de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en fonction du nombre de points de quadrature à partir de l'exemple de Bock et Mislevy (1982, p. 434)

	10 points		21 points		25 points		30 points		35 points		80 points	
ITEM	a3	a4	a3	a4	a3	a4	a3	a4	a3	a4	a3	a4
1	0,00	0,12	0,00	0,12	-0,04	0,12	-0,04	0,12	-0,04	0,12	0,00	0,11
2	-0,10	0,17	-0,10	0,17	-0,10	0,17	-0,10	0,17	-0,10	0,17	-0,10	0,17
3	-0,12	0,46	-0,11	0,44	-0,11	0,44	-0,11	0,44	-0,11	0,44	-0,11	0,44
4	-0,25	0,49	-0,24	0,51	-0,24	0,51	-0,24	0,51	-0,24	0,51	-0,24	0,51
5	***-0,40**	**0,94**	-0,61	1,01	-0,61	1,01	-0,61	1,01	-0,61	1,01	-0,61	1,00
6	-0,64	**1,43**	-0,63	0,82	-0,63	0,82	-0,63	0,82	-0,63	0,82	-0,63	0,82
7	**-0,31**	**1,72**	-0,54	0,92	-0,54	0,92	-0,54	0,92	-0,54	0,92	-0,54	0,92
8	**-1,15**	**2,99**	-0,61	1,02	-0,61	1,02	-0,61	1,02	-0,61	1,02	-0,61	1,01
9	**-0,96**	**0,59**	-0,63	0,97	-0,63	0,96	-0,63	0,96	-0,63	0,96	-0,63	0,96
10	**-1,42**	**2,23**	-0,52	0,96	-0,52	0,95	-0,52	0,95	-0,52	0,95	-0,52	0,95
11	**-1,79**	**11,13**	-0,39	0,98	-0,39	0,97	-0,39	0,97	-0,39	0,97	-0,39	0,97
12	**-2,42**	**8,71**	-0,48	1,15	-0,47	1,15	-0,47	1,15	-0,47	1,15	-0,47	1,15
13	**-2,35**	**5,91**	-0,55	1,31	-0,55	1,31	-0,55	1,31	-0,55	1,31	-0,55	1,31
14	**-2,87**	**11,44**	-0,37	0,80	-0,36	0,79	-0,36	0,79	-0,36	0,79	-0,36	0,79
15	**-2,97**	**22,77**	-0,23	0,49	-0,22	0,47	-0,22	0,47	-0,22	0,47	-0,22	0,47
16	**-3,97**	**16,57**	-0,32	0,57	-0,30	0,60	-0,30	0,60	-0,30	0,60	-0,30	0,60
17	**-4,85**	**29,16**	-0,25	0,41	-0,22	0,40	-0,22	0,41	-0,22	0,41	-0,22	0,41
18	**-5,62**	**61,74**	-0,18	**0,35**	-0,15	0,27	-0,15	0,28	-0,15	0,28	-0,15	0,28
19	**-6,74**	**49,85**	**-0,26**	0,28	-0,20	0,30	-0,20	0,32	-0,20	0,32	-0,20	0,32
20	**-9,54**	**135,03**	**-0,21**	**0,33**	-0,13	0,19	-0,12	0,22	-0,12	0,22	-0,12	0,22

* Les chiffres en caractères gras indiquent que les valeurs obtenues affichent une différence d'au moins 0,05 par rapport aux valeurs calculées avec 80 points de quadrature

La présente recherche vise toutefois la simulation de l'administration d'un nombre d'items supérieur à 20, soit 60 items. Il est possible que, pour assurer la stabilité des calculs lorsque 60 items sont administrés, plus de 25 points de quadrature soient nécessaires. Dans le but de vérifier la stabilité des calculs lorsque plus de 20 items sont administrés, la simulation d'un test adaptatif qui se termine après l'administration de 80 items est réalisée. L'utilisation de 80 items permet de vérifier, au besoin, les résultats obtenus au delà des 60 items qui sont administrés à l'intérieur de cette recherche. Des résultats obtenus antérieurement lors de tests de programmation des routines de calculs sont utilisés. Le niveau d'habileté est de -0,49 et le niveau de difficulté du premier item de -0,01. Les estimations sont réalisées avec 10, 21, 25, 30, 35, 40, 45, 50 et 80 points de quadrature.

Pour les fins de l'analyse de la stabilité des calculs, un résumé des résultats est présenté aux figures 7.1 à 7.4. Ces figures comparent les valeurs obtenues à partir du nombre de points de quadrature, ligne pointillée, où ces valeurs diffèrent de celles obtenues à partir de 80 points de quadrature, ligne pleine. Au delà de ce nombre de points de quadrature, les valeurs ne diffèrent pas de plus de 0,05 par rapport à celles réalisées à partir de 80 points de quadrature. Le détail des résultats est disponible à l'annexe VI.

La figure 7.1 permet de constater qu'avec 21 points de quadrature l'estimateur du niveau d'habileté diffère d'au moins 0,05 à partir du 58^{e} item administré. C'est donc à partir du nombre subséquent de points de quadrature, soit 25, que l'estimateur du niveau

d'habileté concorde jusqu'au 60e item avec celui obtenu à partir de 80 points de quadrature. Ce n'est alors qu'au 72e item qu'il diffère.

Quant à l'erreur type de l'estimateur du niveau d'habileté, une différence de 0,05 entre le résultat obtenu à partir de 10 points de quadrature et celui obtenu à partir de 80 points est remarquée au 9e item administré. C'est ce que décrit la figure 7.2. Un minimum de 21 points de quadrature est nécessaire pour assurer la stabilité des calculs lorsque 60 items sont administrés. Ce n'est qu'à partir du 70e item que les calculs basés sur 21 points de quadrature diffèrent de ceux basés sur 80 points.

En ce qui concerne l'asymétrie de l'estimateur du niveau d'habileté, la figure 7.3 indique que les calculs diffèrent d'au moins 0,05 à partir du 46e item administré lorsque 30 points de quadrature sont utilisés. Un minimum de 35 points de quadrature est donc nécessaire pour remarquer une stabilité dans les calculs jusqu'au 60e item. Ce n'est alors qu'à partir du 71e item administré que les calculs commencent à s'éloigner de ceux obtenus à partir de 80 points de quadrature.

La figure 7.4 compare les calculs de la kurtose réalisés à partir de 35 points de quadrature à ceux obtenus à partir de 80 points de quadrature. Les calculs diffèrent d'au moins 0,05 à partir du 54e item administré. Avec 40 points de quadrature, les calculs diffèrent d'au moins 0,05 seulement à partir du 72e item. Pour les fins de cette recherche, en adoptant une attitude conservatrice, il est donc jugé préférable d'utiliser 40 points de quadrature

pour assurer la stabilité des calculs.

Les résultats obtenus aux tableaux 7.2 et 7.3, ainsi qu'aux figures 7.1 à 7.4, nous permettent de considérer exact le calcul de l'estimateur du niveau d'habileté, de l'erreur type, de l'asymétrie et de la kurtose réalisé à partir des procédures d'estimation programmées en langage SAS lorsque 40 points de quadrature sont utilisés.

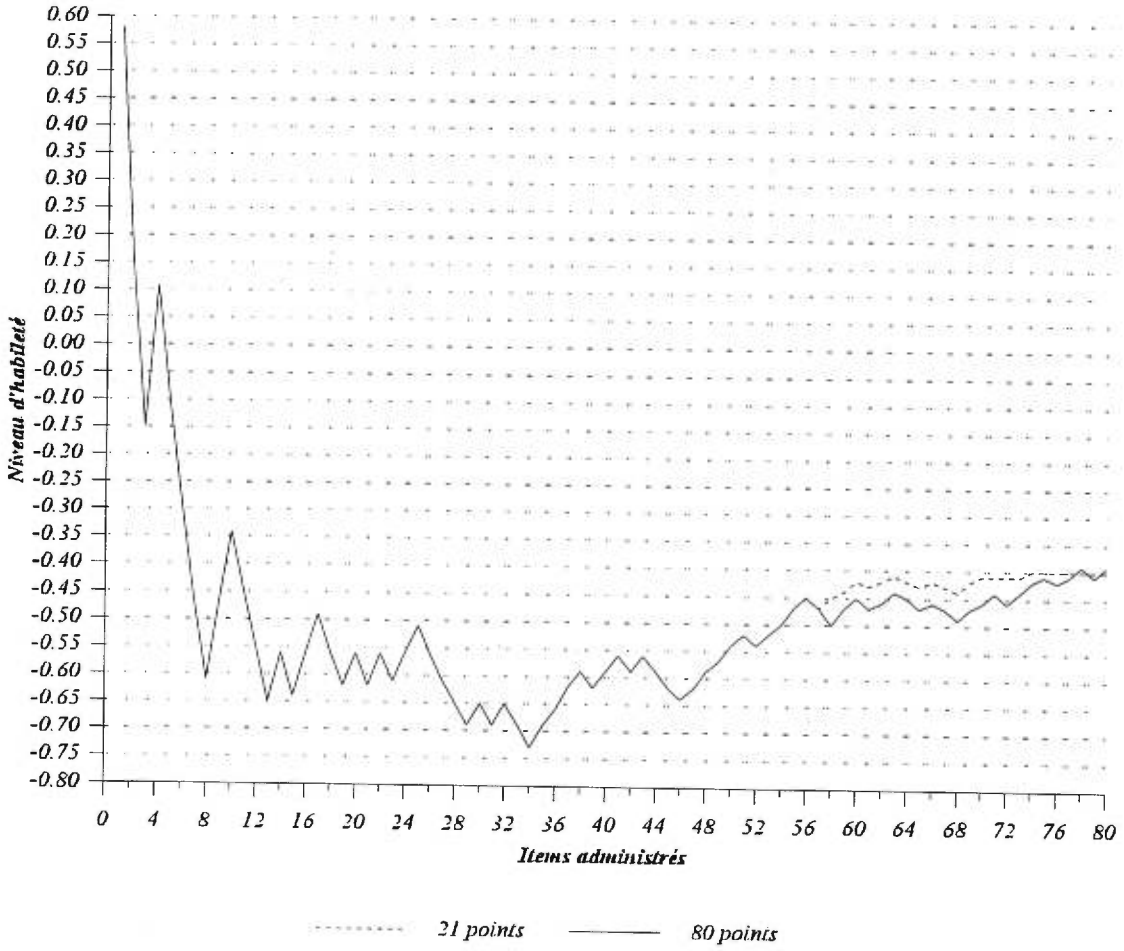

Figure 7.1 Estimateur du niveau d'habileté, EAP(θ), calculé selon la méthode de l'espérance a posteriori en fonction du nombre d'items administrés et du nombre de points de quadrature

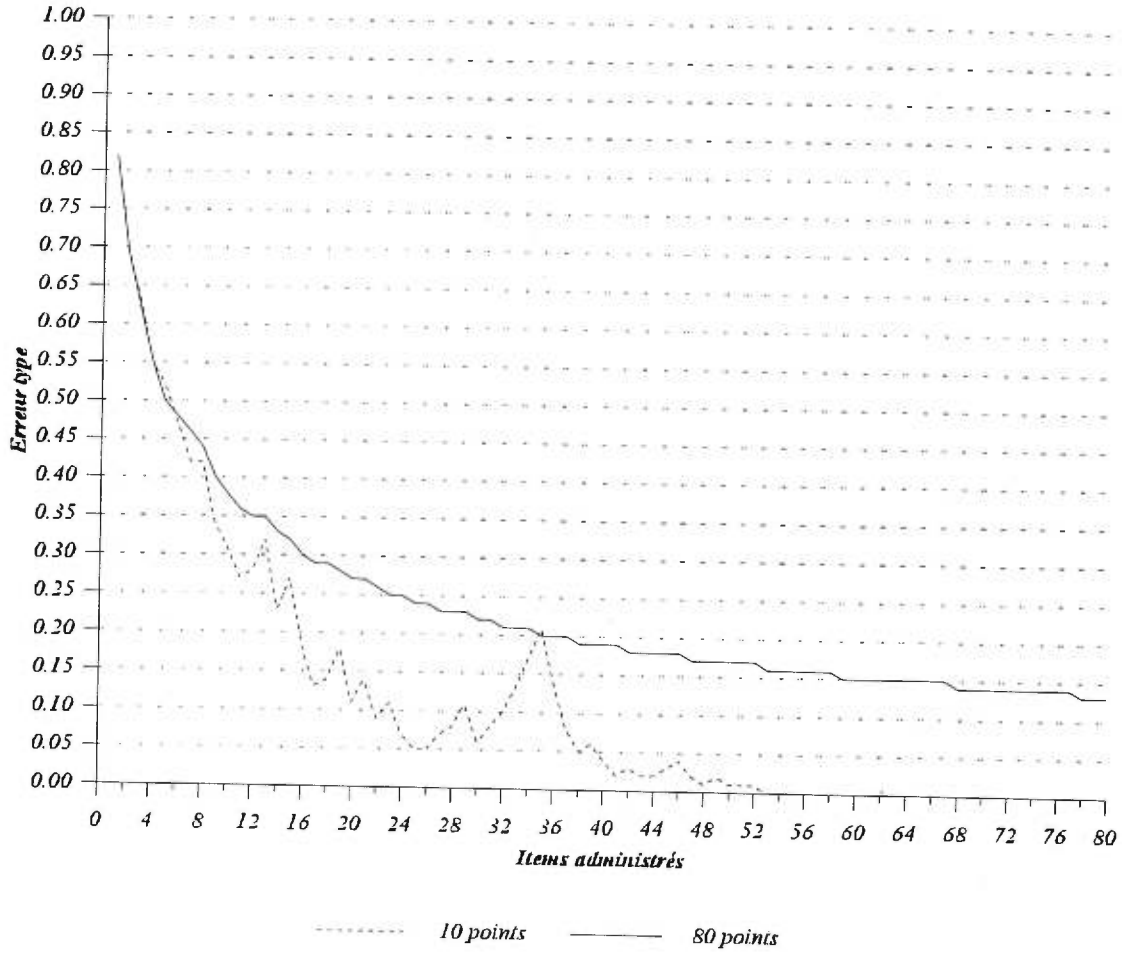

Figure 7.2 Erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, calculée selon la méthode de l'espérance a posteriori en fonction du nombre d'items administrés et du nombre de points de quadrature

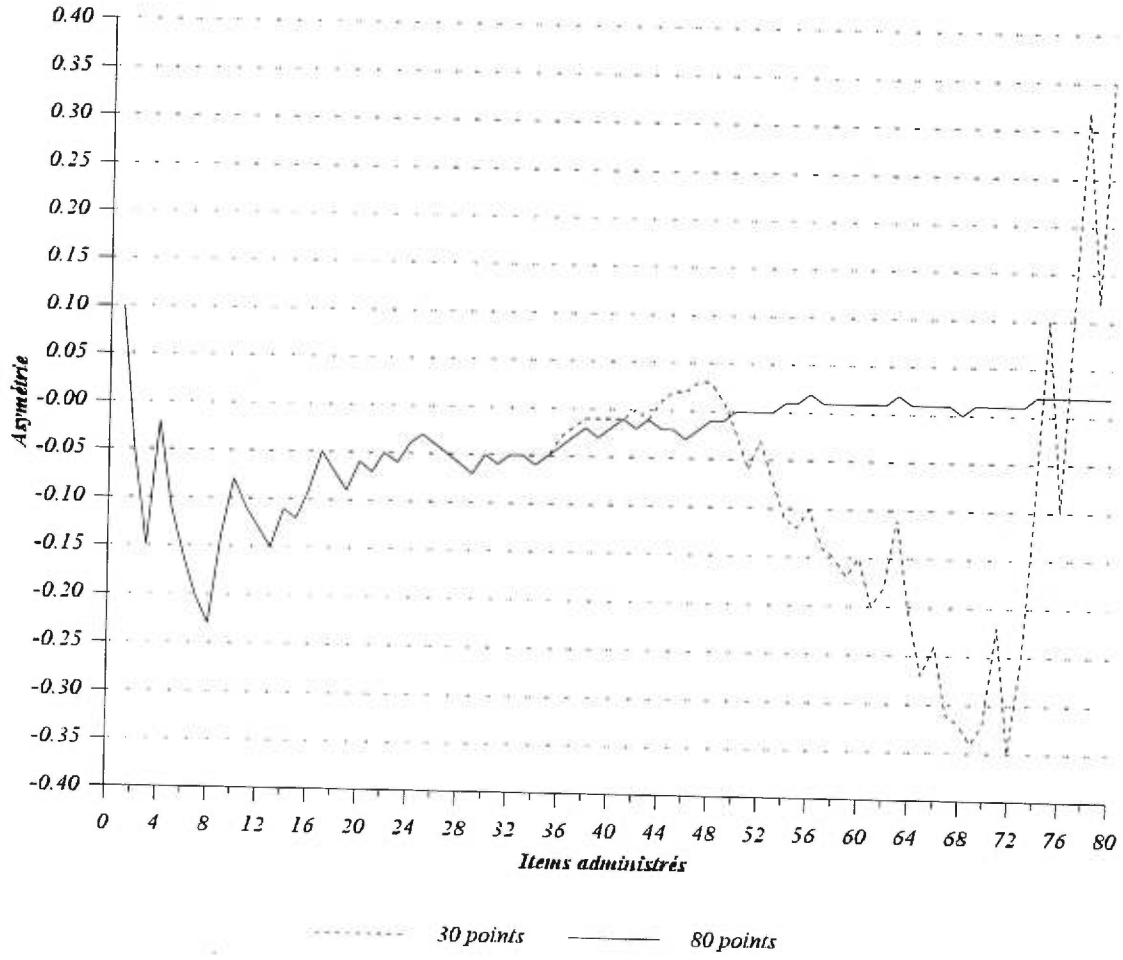

Figure 7.3 Asymétrie de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, calculée selon la méthode de l'espérance a posteriori en fonction du nombre d'items administrés et du nombre de points de quadrature

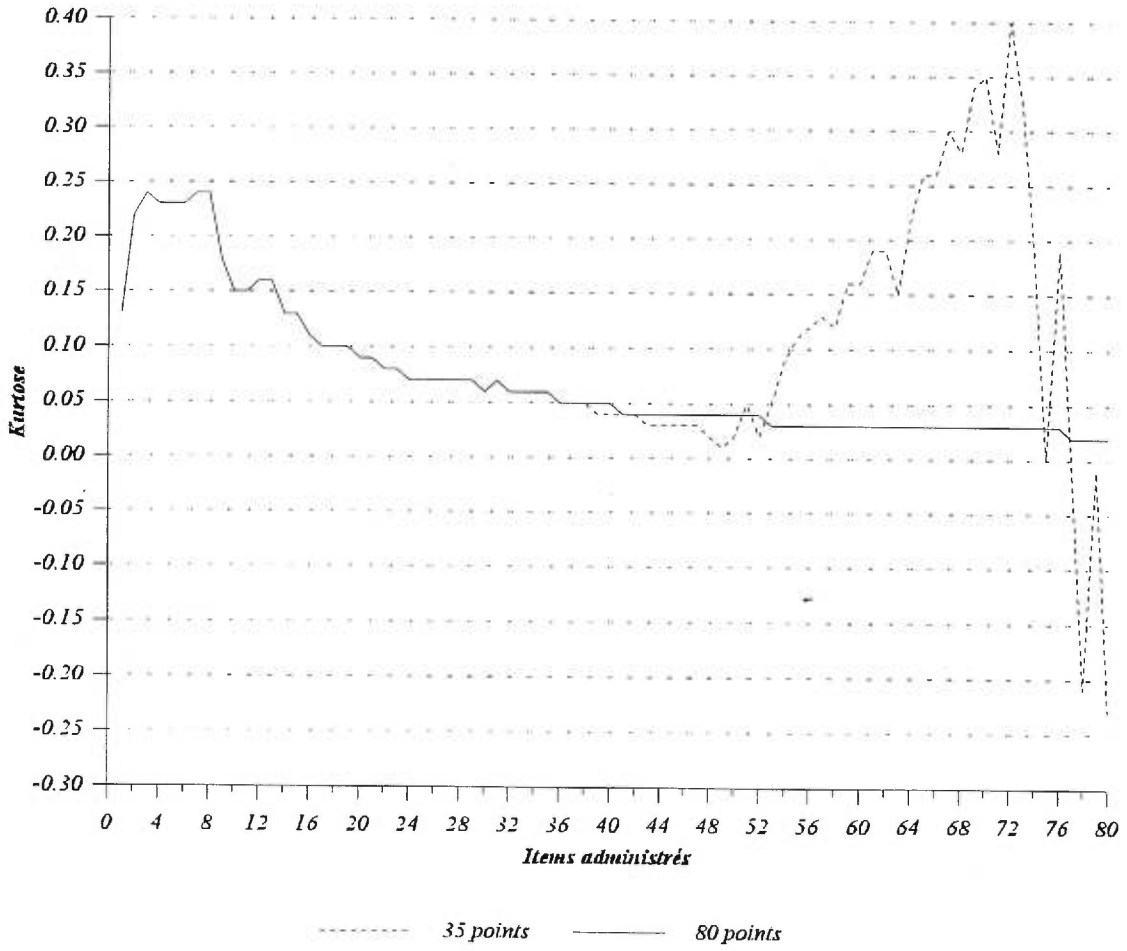

Figure 7.4 Kurtose de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, calculée selon la méthode de l'espérance a posteriori en fonction du nombre d'items administrés et du nombre de points de quadrature

7.4 Interprétation de l'asymétrie et de la kurtose d'une distribution de probabilité

L'asymétrie et la kurtose, comme on l'a souligné précédemment, nous permettent de déterminer l'allure d'une distribution de probabilité et d'illustrer jusqu'à quel point elle s'éloigne d'une distribution normale. Mais à quel moment les valeurs de l'asymétrie et de la kurtose nous permettent-elles de dire qu'une distribution de probabilité s'éloigne vraiment d'une distribution normale ? À quel moment les interprétations basées sur les caractéristiques d'une distribution normale ne tiennent-elles plus ? À notre connaissance, la littérature est plutôt avare et vague à ce sujet. Elle nous est ainsi peu utile pour déterminer à quel moment les valeurs de l'asymétrie et de la kurtose sont importantes. C'est pourquoi, avant de porter un jugement sur les valeurs de l'asymétrie et de la kurtose que nous obtiendrons, nous avons jugé nécessaire de vérifier l'impact de ces valeurs sur les interprétations relatives à une distribution normale.

Dans une distribution normale, la moyenne et la médiane sont de valeurs égales. Lorsque l'asymétrie de la distribution est négative, la valeur de la médiane est supérieure à celle de la moyenne. À l'inverse, lorsque l'asymétrie est positive, la valeur de la médiane est inférieure à celle de la moyenne. La kurtose de la distribution a aussi un impact sur la valeur de la médiane lorsque l'asymétrie est différente de zéro. Quelle est l'importance de la différence entre les valeurs de la médiane et de la moyenne en fonction de la variation de l'asymétrie et de la kurtose ? Une différence importante entre les valeurs de la médiane et de la moyenne indiquerait qu'il faut user de prudence dans l'utilisation de

la moyenne comme estimateur du niveau d'habileté.

De plus, dans une distribution normale, 68,27 % des observations se distribuent dans un intervalle de confiance qui se situe entre plus ou moins un écart type autour de la moyenne ($\mu \pm \sigma$). Dans une distribution leptokurtique, où la kurtose est positive, l'intervalle de confiance à 68,27 % est rétréci et, par le fait même, 68,27 % des observations se distribuent à l'intérieur d'un intervalle inférieur à plus ou moins un écart type autour de la moyenne. À l'opposé, lorsque la distribution est platykurtique, la kurtose est négative et 68,27 % des observations se distribuent à l'intérieur d'un intervalle de confiance supérieur à plus ou moins un écart type autour de la moyenne. Dans les deux cas, on ne peut donc pas considérer l'écart type comme une mesure valide de l'intervalle de confiance à 68,27 % autour de la moyenne. Lorsque la distribution est leptokurtique, il est surestimé, lorsqu'elle est platykurtique, il est sous-estimé. Par exemple, si nous obtenons une valeur de l'erreur type de l'estimateur du niveau d'habileté égale à 0,30 et que la distribution de probabilité de cet estimateur est platykurtique, l'intervalle de confiance à 68,27 % est en fait supérieur à $\pm$0,30. Mais de combien est-il supérieur ? De plus, l'asymétrie de la distribution a un impact sur la grandeur de cet intervalle de confiance. Ce qui nous intéresse ici est de connaître l'importance de la différence entre l'intervalle de confiance à 68,27 % autour de la moyenne d'une distribution qui s'éloigne d'une distribution normale et l'intervalle de confiance associé à une distribution normale, soit son écart type, en fonction des valeurs de l'asymétrie et de la kurtose.

Dans le but de connaître l'importance de la différence entre les valeurs de la médiane et de la moyenne, ainsi que de l'intervalle de confiance à 68,27 % en fonction de la variation de l'asymétrie et de la kurtose dans des distributions qui s'éloignent d'une distribution normale, nous avons simulé des distributions où nous avons fait varier les valeurs de l'asymétrie (a3) et de la kurtose (a4) tout en maintenant constants la moyenne et l'écart type de ces distributions respectivement à 0,00 et à 1,00.

À cette fin, 2000 valeurs du niveau d'habileté sont générées au hasard selon une distribution N(0,1) et une transformation polynomiale par la méthode des puissances (*power method*) de Fleishman (1978 ; Headrick et Sawilowsky, 1999a, 1999b) est appliquée à ces valeurs de façon à simuler des distributions asymétriques, leptokurtiques ou platykurtiques. Puisque cette analyse n'a pour but que de nous aider à interpréter les résultats obtenus à l'intérieur de notre recherche, toute l'étendue des valeurs possibles de l'asymétrie et de la kurtose n'est pas étudiée. Ainsi, uniquement les valeurs les plus extrêmes obtenues de l'asymétrie et de la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté sont utilisées ainsi que les valeurs proposées dans le travail de Fleischman. Les valeurs de l'asymétrie varient entre -0,20 et 1,40, tandis que les valeurs de la kurtose se situent entre -1,20 et 3,20. Éventuellement, une analyse plus approfondie sur une étendue plus grande de l'asymétrie et de la kurtose serait appropriée. Selon Fleischman (p. 526), une distribution qui affiche une kurtose de -1,2 est, à toutes fins utiles, rectangulaire. La transformation polynomiale $T(\theta)$ est de la forme suivante :

$$T(\theta) = a + b\theta + c\theta^2 + d\theta^3 \qquad \textbf{(7.10)}$$

où a, b, c et d sont des coefficients déterminés par les valeurs cibles de l'asymétrie et de la kurtose et où θ est le niveau d'habileté. La valeur de ces coefficients est calculée par une méthode proposée par Fleishman (1978, p. 523-526), soit en solutionnant le système d'équations non linéaires suivant :

$$\left\{\begin{array}{l} a4 = 24\,(bd + c^2(1 + b^2 + 28bd) + d^2(12 + 48bd + 141c^2 + 225d^2)) \\ c = \dfrac{a3}{2(b^2 + 24bd + 105d^2 + 2} \\ 2 = 2b^2 + 12bd + \dfrac{(a3)^2}{(b^2 + 24bd + 105d^2 + 2)^2} + 30d^2 \\ a = -c \end{array}\right. \qquad \textbf{(7.11)}$$

où les coefficients a, b, c et d sont les inconnues qui varient selon les valeurs utilisées de l'asymétrie (a3) et de la kurtose (a4).

Nous ajustons par la suite une fonction de régression qui sert à modéliser la relation entre la différence obtenue entre la médiane et la moyenne et les valeurs de l'asymétrie et de la kurtose de façon à nous permettre de représenter graphiquement la valeur calculée de la différence entre la médiane et la moyenne. L'équation de régression que nous obtenons affiche un coefficient de détermination (R^2) de 0,90 et la fonction correspondante est égale à :

$$\text{MÉDIANE} = -0{,}132860\ a3 + 0{,}002862\ a4 - 0{,}097607\ a3^2 + 0{,}035816\ a3a4 - 0{,}001289\ a4^2 \quad \textbf{(7.12)}$$

Avec un coefficient de détermination de cette importance, il nous est alors possible de calculer de façon suffisamment précise pour nos besoins la différence entre la médiane et la moyenne ($\mu = 0$) à partir des valeurs de l'asymétrie et de la kurtose de la distribution. À titre d'illustration, lorsque l'asymétrie et la kurtose prennent toutes deux une valeur de 0,30, la différence calculée entre la médiane et la moyenne est de -0,04. Il faut souligner que, dans une distribution d'échantillonnage où l'erreur type de l'estimateur de la moyenne est de 0,20, cette différence correspond déjà à 20 % de la valeur de l'erreur type. En fait, plus l'erreur type recherchée est petite, plus cette différence entre la médiane et la moyenne prend de l'importance.

À la figure 7.5, nous pouvons observer la différence calculée entre la valeur de la médiane et la valeur de la moyenne en fonction de différentes valeurs de l'asymétrie et de la kurtose. À l'intérieur des intervalles considérés de l'asymétrie et de la kurtose, la différence calculée entre la médiane et la moyenne varie entre -0,40 et 0,00. Lorsque l'asymétrie est égale, à 0,50 et que la kurtose est nulle, la différence calculée entre la médiane et la moyenne est d'environ -0,10. Cette différence atteint une valeur de -0,25 lorsque l'asymétrie est de 1,00 et que la kurtose est nulle.

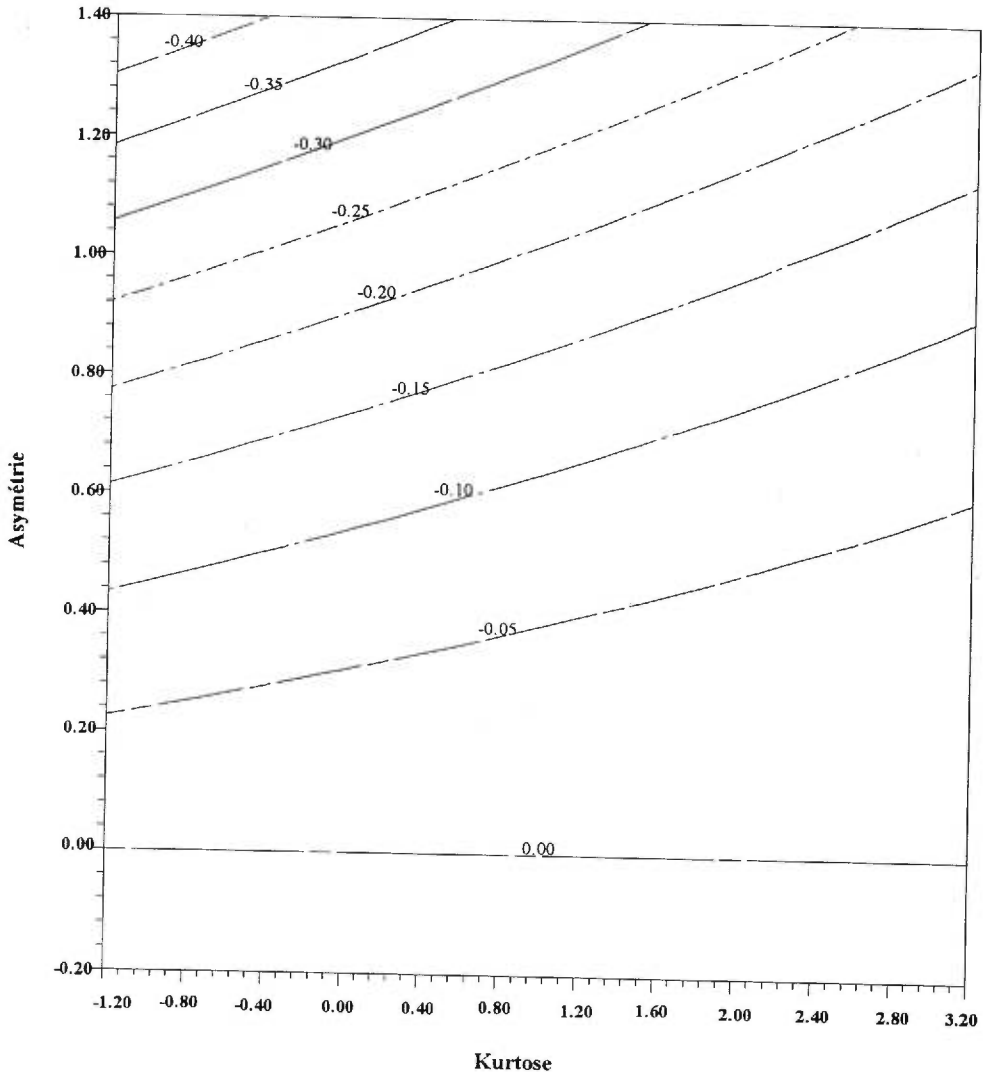

Figure 7.5 Différence calculée entre les valeurs de la médiane et de la moyenne de la distribution de probabilité en fonction de son asymétrie et de sa kurtose ($\mu = 0$, $\sigma = 1$)

Nous ajustons par la suite une fonction de régression qui sert à modéliser la relation entre la différence obtenue entre la valeur théorique de l'écart type, soit $\sigma = 1{,}00$, et la valeur de la variation autour de la moyenne qui correspond à un intervalle de confiance à 68,27 % et les valeurs de l'asymétrie et de la kurtose. L'équation de régression que nous obtenons affiche un coefficient de détermination (R^2) de 0,88 et la fonction correspondante est égale à :

$$\begin{aligned} DIF = & - 0{,}082483 * a3 - 0{,}087388 * a4 + 0{,}370403 * a3^2 \\ & - 0{,}043511 * a3 * a4 + 0{,}010621 * a4^2 \end{aligned} \quad \textbf{(7.13)}$$

Avec un coefficient de détermination de cette importance, il nous est alors possible de calculer de façon suffisamment précise pour nos besoins l'écart entre la différence obtenue entre la valeur théorique de l'écart type et la valeur de la variation autour de la moyenne à partir des valeurs de l'asymétrie et de la kurtose de la distribution. À titre d'illustration, lorsque l'asymétrie et la kurtose prennent toutes deux une valeur de 0,30, la différence calculée entre la valeur théorique de l'écart type et la valeur de la variation autour de la moyenne est de -0,04.

À la figure 7.6, nous pouvons observer la différence calculée entre la valeur théorique de l'écart type et la valeur de la variation autour de la moyenne en fonction de différentes valeurs de l'asymétrie et de la kurtose. À l'intérieur des intervalles considérés de

l'asymétrie et de la kurtose, la différence calculée entre la valeur théorique de l'écart type et la valeur de la variation autour de la moyenne varie entre -0,20 et 0,70, soit entre 20 % au-dessous et 70 % au-dessus de la valeur prévue par une distribution normale. Lorsque la kurtose est égale, à -0,40 et que l'asymétrie est nulle, la différence calculée entre la valeur théorique de l'écart type et la valeur de la variation autour de la moyenne est inférieure à 0,05. Cette différence atteint une valeur de 0,20 lorsque l'asymétrie est d'environ 0,70. Lorsque la valeur de la kurtose est égale à 0,80 et que l'asymétrie est nulle, la différence entre la valeur théorique de l'écart type et la valeur de la variation autour de la moyenne est égale à environ -0,05. Elle prend une valeur d'environ 0,05 lorsque l'asymétrie est de 0,70.

Les analyses que nous avons effectuées et les modèles de régression que nous avons déterminés permettent de connaître l'importance de l'impact de l'asymétrie et de la kurtose sur l'allure d'une distribution de probabilité. L'importance de cet impact, ici sur la médiane et sur l'intervalle de confiance à 68,27 % autour de la moyenne, doit toutefois être toujours interprété en fonction de l'utilisation qui est faite des estimateurs rattachés à une distribution de probabilité. Par exemple, en ce qui nous concerne, lorsque l'estimateur du niveau d'habileté est égal à zéro et que son erreur type est égale à 0,20, un intervalle de confiance à 68,27 % autour de la moyenne de la distribution de probabilité du niveau d'habileté qui correspond à 1,10 plutôt qu'à 1,00 indique que l'écart type est sous-estimé. D'ailleurs, la différence de 0,10 correspond alors à 50 % de la valeur de l'erreur type, ce qui nous semble assez important.

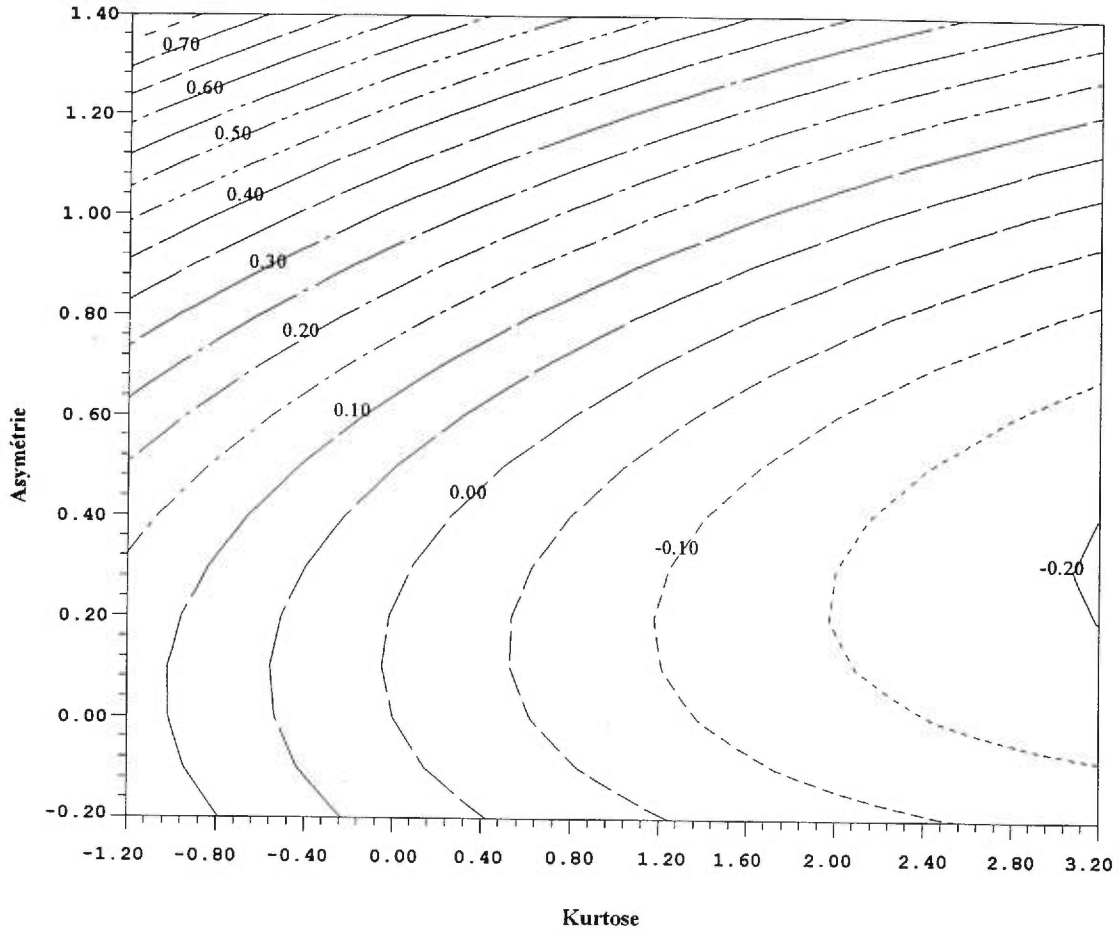

Figure 7.6 Différence entre la valeur théorique de l'écart type, soit de 1,00, et la valeur de la variation autour de la moyenne qui correspond à un intervalle de confiance à 68,27 % autour de la moyenne de la distribution de probabilité en fonction de l'asymétrie et de la kurtose ($\mu = 0$, $\sigma = 1$)

7.5 Méthode d'analyse des résultats

Pour chacune des valeurs retenues pour la règle d'arrêt, les distributions d'échantillonnage associées à chacune des 2000 valeurs du niveau d'habileté générées au hasard sont analysées. Dans le cas de la règle d'arrêt selon le nombre d'items administrés, la règle d'arrêt varie de 1 à 60 items administrés. Conséquemment 120 000 distributions d'échantillonnage de l'estimateur du niveau d'habileté sont générées, soit 2000 distributions d'échantillonnage à chacun des 60 items retenus pour la règle d'arrêt. Lorsque la règle d'arrêt selon l'erreur type de l'estimateur du niveau d'habileté est appliquée, les valeurs retenues pour la règle d'arrêt varient entre 0,20 et 0,85 par sauts de 0,05. Nous générons donc 2000 distributions d'échantillonnage pour chacune des 14 valeurs de l'erreur type retenue pour la règle d'arrêt selon l'erreur type, soit 28 000 distributions d'échantillonnage au total.

Pour chacune des distributions d'échantillonnage produites, l'estimateur du niveau d'habilité, l'erreur type de l'estimateur du niveau d'habileté, l'erreur de mesure du niveau d'habileté, l'asymétrie, la kurtose et la proportion de bonnes réponses sont calculés. Dans le cas de la règle d'arrêt selon l'erreur type, le nombre d'items administrés est aussi calculé. La moyenne, l'écart type, l'asymétrie, la kurtose, le minimum ainsi que le maximum de la distribution de probabilité de chacune des statistiques associées aux 2000 distributions d'échantillonnage de l'estimateur du niveau d'habileté sont calculés pour chacune des valeurs retenues pour la règle d'arrêt. Les résultats sont présentés sous

forme de tableaux et de figures.

Une modélisation selon une régression cubique est aussi réalisée dans le but d'étudier la relation entre le niveau d'habileté et l'estimateur du niveau d'habileté, l'erreur type de l'estimateur du niveau d'habileté, le biais, l'asymétrie, la kurtose, la proportion de bonnes réponses et le nombre d'items administrés associés à la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Un corrélogramme permet de visualiser cette relation.

Au terme de l'analyse des résultats, un jugement est porté sur l'adéquation de la loi normale pour décrire la distribution d'échantillonnage de l'estimateur du niveau d'habileté et sur les valeurs à utiliser comme critères dans les règles d'arrêt selon l'erreur type et selon le nombre d'items administrés.

8. Résultats et analyse des résultats

Dans ce chapitre, nous abordons simultanément les résultats obtenus et leur analyse. Il nous a semblé plus approprié de présenter nos résultats et analyses de cette façon à cause de leur grand nombre ; une section portant uniquement sur l'analyse des résultats à la fin de ce chapitre aurait été trop lourde.

À la section 8.1, la distribution de probabilité du niveau d'habileté est décrite et nous portons un jugement sur sa conformité avec une distribution N(0,1). Cette conformité doit être assurée pour nous permettre de poursuivre nos analyses et de supporter nos interprétations.

La section 8.2 aborde la distribution de probabilité des différentes statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté lorsque la règle d'arrêt utilisée est celle de l'erreur type de l'estimateur du niveau d'habileté. Cette analyse est réalisée en fonction de la variation des valeurs retenues pour la règle d'arrêt selon l'erreur type de l'estimateur du niveau d'habileté. Nous nous préoccupons des aspects qui affectent le déroulement d'un test adaptatif et nous signalons des recommandations quant à l'erreur type à retenir pour la règle d'arrêt.

Nous considérons la distribution de probabilité de diverses statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté en fonction de la règle d'arrêt selon le nombre d'items administrés à la section 8.3. Par souci de cohérence, l'ordre de présentation des diverses analyses est le même à l'intérieur de cette section qu'à l'intérieur de la section 8.2.

Dans le but d'éviter toute confusion, nous tenons à indiquer notre utilisation indifférenciée des expressions en lien avec les statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Par exemple, les expressions suivantes sont pour nous synonymes : "asymétrie de l'estimateur du niveau d'habileté", "asymétrie associée à l'estimateur du niveau d'habileté", "asymétrie de la distribution d'échantillonnage du niveau d'habileté", "asymétrie associée à la distribution d'échantillonnage de l'estimateur du niveau d'habileté". Il en est de même en ce qui concerne les autres statistiques étudiées : erreur type, biais, kurtose, proportion de bonnes réponses et nombre d'items administrés.

Le sommaire des considérations et des recommandations quant à l'erreur type et au nombre d'items à administrer dans un test adaptatif sera réalisé plus tard, soit au chapitre 9, lors de la discussion des résultats.

8.1 Description de la distribution de probabilité du niveau d'habileté

Nous avons généré le niveau d'habileté pour 2000 sujets avec pour objectif d'obtenir une distribution de probabilité du niveau d'habileté de type normale dont la moyenne est nulle et dont l'écart type est égal à 1,00, soit une distribution de probabilité N(0,1). La moyenne que nous avons obtenue est de 0,01 et l'écart type est égal à 0,98 ; ces valeurs se rapprochent respectivement de 0,00 et de 1,00.

La valeur minimale du niveau d'habileté est de -3,13 (z = -3,19) tandis que la valeur maximale est de 3,24 (z = 3,31). L'étendue des valeurs du niveau d'habileté est alors bien couverte puisque, dans une distribution de probabilité N(0,01, 0,98), cette étendue comprend 99,88 % des valeurs probables du niveau d'habileté, 49,93 % au dessous de la moyenne, et 49,95 % au dessus de la moyenne, selon la table de la surface sous la courbe normale N(0,1) proposée par Guilford (1965, p. 576)).

Les coefficients d'asymétrie et de kurtose d'une distribution de probabilité normale doivent être tous deux égaux à 0,00. Les valeurs que nous avons obtenues de ces deux coefficients sont de -0,01, valeurs très rapprochées de 0,00.

Nous avons aussi jugé pertinent de réaliser l'analyse de la normalité de la distribution de probabilité du niveau d'habileté par l'analyse d'un corrélogramme (*normal probability plot*) qui met en relation les scores normalisés (*normal scores*) et les scores z de la

distribution de probabilité du niveau d'habileté, ainsi qu'à partir du test de normalité de Shapiro-Wilk (1965 : voir SAS Institute, 1990b, p. 627). Les scores normalisés sont calculés à partir de la méthode de Blom (1958 : voir SAS Institute, 1990b, p. 495).

À la figure 8.1, le corrélogramme nous permet d'observer que les scores normalisés sont presque identiques aux scores z. Le coefficient de régression linéaire entre ces scores, lorsque l'ordonnée à l'origine est fixée à 0,00, est égal à 0,9994, soit, à toutes fins utiles, égal à 1,00. Quant au test de normalité de Shapiro-Wilks, nous obtenons un coefficient W de 0,99 non significatif au seuil de 0,05 ($W = 0,99$, $p = 0,25$).

À partir de ces constatations, nous pouvons conclure que la distribution de probabilité des valeurs simulées du niveau d'habileté se rapproche de près de la distribution de probabilité N(0,1) tel qu'exigé par cette recherche. Toutefois, nous avons jugé qu'il serait plus réaliste d'effectuer nos analyses à partir de la distribution de probabilité obtenue, soit N(0,01, 0,98).

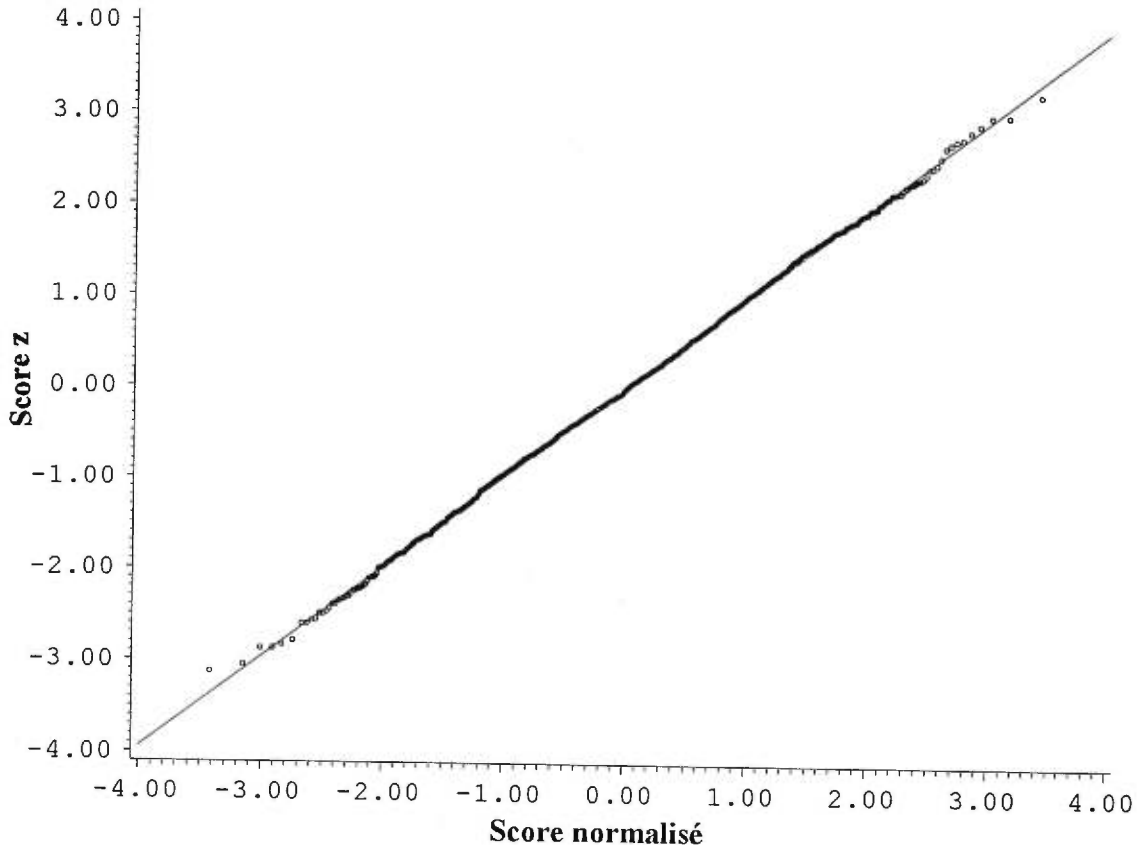

Figure 8.1 Corrélogramme mettant en relation les scores normalisés et les scores z de la distribution de probabilité du niveau d'habileté (R^2 = 1,00)

8.2 Règle d'arrêt selon l'erreur type

Les résultats, ainsi que leur analyse, obtenus en fonction de la valeur retenue dans la règle d'arrêt basée sur l'erreur type de l'estimateur du niveau d'habileté, sont présentés selon la séquence suivante : estimateur du niveau d'habileté, EAP(θ), erreur type, $S_{EAP(\theta)}$, biais, $BIAIS_{EAP(\theta)}$, asymétrie, $a3_{EAP(\theta)}$, kurtose ,$a4_{EAP(\theta)}$, proportion de bonnes réponses et nombre d'items administrés associés à la distribution d'échantillonnage de l'estimateur du niveau d'habileté.

Par souci d'homogénéité et de cohérence, la présentation des résultats et des analyses est réalisée exactement de la même façon pour toutes les statistiques étudiées. Pour cette raison, la section 8.2.1, concernant la distribution de probabilité de l'estimateur du niveau d'habileté, est utilisée comme canevas pour les sections 8.2.2 à 8.2.7 inclusivement.

8.2.1 Caractéristiques de la distribution de probabilité de l'estimateur du niveau d'habileté en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

Nous présentons au tableau 8.1 les caractéristiques de la distribution de probabilité de l'estimateur du niveau d'habileté, EAP(θ), en fonction de chacune des valeurs de l'erreur type retenues pour la règle d'arrêt : de 0,20 à 0,85 par sauts de 0,05. La figure 8.2 permet de visualiser au quadrant (a) les moyennes, minimum et maximum affichés à l'intérieur du tableau 8.1 en fonction de chacune des valeurs de l'erreur type retenues

pour la règle d'arrêt. Aux quadrants (b), (c) et (d), on retrouve l'écart type, l'asymétrie et la kurtose de la distribution de probabilité de l'estimateur du niveau d'habileté.

Selon ce tableau et cette figure, la moyenne de l'estimateur du niveau d'habileté s'écarte peu de la moyenne réelle du niveau d'habileté, soit 0,01. Au plus, elle est de 0,02 lorsque l'erreur type retenue pour la règle d'arrêt est de 0,60, de 0,65 et de 0,70. Pour toutes les autres valeurs retenues de la règle d'arrêt, la moyenne de l'estimateur du niveau d'habileté varie entre -0,01 et 0,01.

En ce qui concerne l'écart type de l'estimateur du niveau d'habileté, nous constatons qu'il semble tendre vers 0,98 avec la diminution de la valeur de l'erreur type retenue pour la règle d'arrêt. Lorsque l'erreur type retenue est égale ou inférieure à 0,30, l'écart type de l'estimateur du niveau d'habileté est égal ou supérieur à 0,94. Lorsque la valeur de l'erreur type retenue est plus importante, l'écart type s'éloigne sensiblement de 0,98. Les résultats obtenus ne nous permettent toutefois pas de connaître à partir de quelle valeur de l'erreur type retenue pour la règle d'arrêt l'écart type serait éventuellement égal à 0,98.

Tableau 8.1
Caractéristiques de la distribution de probabilité de l'estimateur du niveau d'habileté, EAP(θ), en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

ERREUR TYPE	MOYENNE	ÉCART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
0,85	-0,01	0,56	-0,02	-2,00	-0,56	0,56
0,80	0,00	0,71	0,00	-1,11	-0,99	0,99
0,75	0,00	0,71	0,00	-1,11	-0,99	0,99
0,70	0,02	0,76	0,05	-0,39	-1,33	1,33
0,65	0,02	0,80	0,07	-0,35	-1,63	1,63
0,60	0,02	0,82	0,06	-0,28	-1,90	1,90
0,55	0,01	0,85	0,07	-0,07	-2,36	2,36
0,50	0,01	0,87	0,03	-0,02	-2,76	2,76
0,45	0,01	0,89	0,01	0,10	-3,10	2,78
0,40	0,01	0,91	0,00	0,04	-3,38	3,05
0,35	0,01	0,93	0,02	0,01	-3,28	3,04
0,30	0,00	0,94	0,04	0,01	-3,28	3,21
0,25	0,01	0,95	0,01	0,00	-3,10	3,34
0,20	0,00	0,96	0,01	-0,03	-3,16	3,16

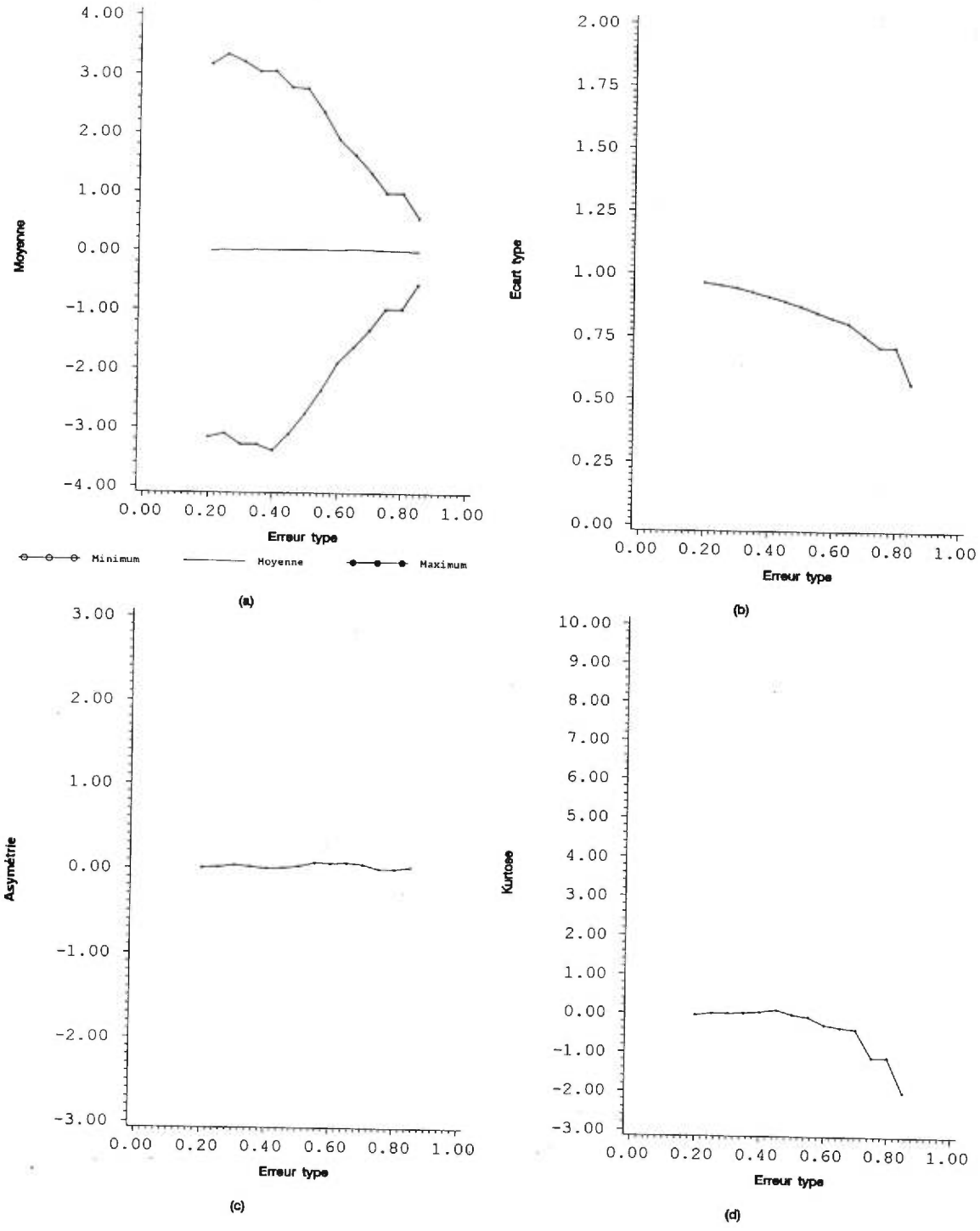

Figure 8.2 Caractéristiques de la distribution de probabilité de l'estimateur du niveau d'habileté, EAP(θ), en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

Nous constatons aussi que l'asymétrie et la kurtose de la distribution de probabilité de l'estimateur du niveau d'habileté affichent généralement des valeurs peu importantes ou près de 0,00 lorsque l'erreur type retenue est inférieure à 0,75.

Enfin, l'analyse des valeurs minimales et maximales obtenues de l'estimateur du niveau d'habileté indique qu'avec la diminution de la valeur de l'erreur type retenue pour la règle d'arrêt ces minimums et maximums se rapprochent de -3,13 et 3,24, valeurs minimales et maximales du niveau d'habileté. Ce n'est toutefois que lorsque l'erreur type retenue pour la règle d'arrêt est égale ou inférieure à 0,40 que les minimums et maximums affichent des valeurs supérieures à 3,00 en valeur absolue, ce qui donne une étendue de l'estimateur du niveau d'habileté qui se rapproche de celle du niveau d'habileté.

Par conséquent, nous concluons que l'estimateur du niveau d'habileté présente une distribution de probabilité similaire à celle du niveau d'habileté, ici une distribution N(0,01, 0,98), à condition que la valeur de l'erreur type de l'estimateur du niveau d'habileté retenue pour la règle d'arrêt soit égale ou inférieure à 0,40. Nous pouvons donc, à cette condition et sans trop d'erreurs, appliquer les interprétations propres à une distribution normale à la distribution de probalilité de l'estimateur du niveau d'habileté.

À la figure 8.3, des corrélogrammes présentent les valeurs obtenues de l'estimateur du niveau d'habileté en fonction du niveau d'habileté pour quatre valeurs de l'erreur type

retenue pour la règle d'arrêt : 0,20, 0,30, 0,35 et 0,40. Ces valeurs, sauf 0,35, sont les mêmes que celles retenues par Bock et Mislevy (1982, p. 439-442) dans le dessein d'analyser le biais de l'estimateur du niveau d'habileté. Nous utilisons la valeur 0,35 dans le seul but d'offrir une valeur supplémentaire à celles proposées par Bock et Mislevy et de nous permettre d'étudier la relation entre le niveau d'habileté et l'erreur type de l'estimateur du niveau d'habileté sur une étendue un peu plus grande de l'erreur type retenue pour la règle d'arrêt.

Une régression cubique est utilisée pour ajuster une courbe de régression à chacun de ces corrélogrammes. Nous présentons les coefficients de régression et de détermination au tableau 8.2.

Remarquons que le coefficient de détermination augmente avec la diminution de la valeur de l'erreur type retenue pour la règle d'arrêt. Il est déjà de 0,84 lorsque la valeur de l'erreur type est de 0,40 et atteint 0,96 lorsque cette valeur est de 0,20. Les modèles de régression expliquent donc assez bien la variabilité observée de l'estimateur du niveau d'habileté et celle-ci, d'après les coefficients de régression que nous observons, est presque entièrement expliquée par une régression linéaire.

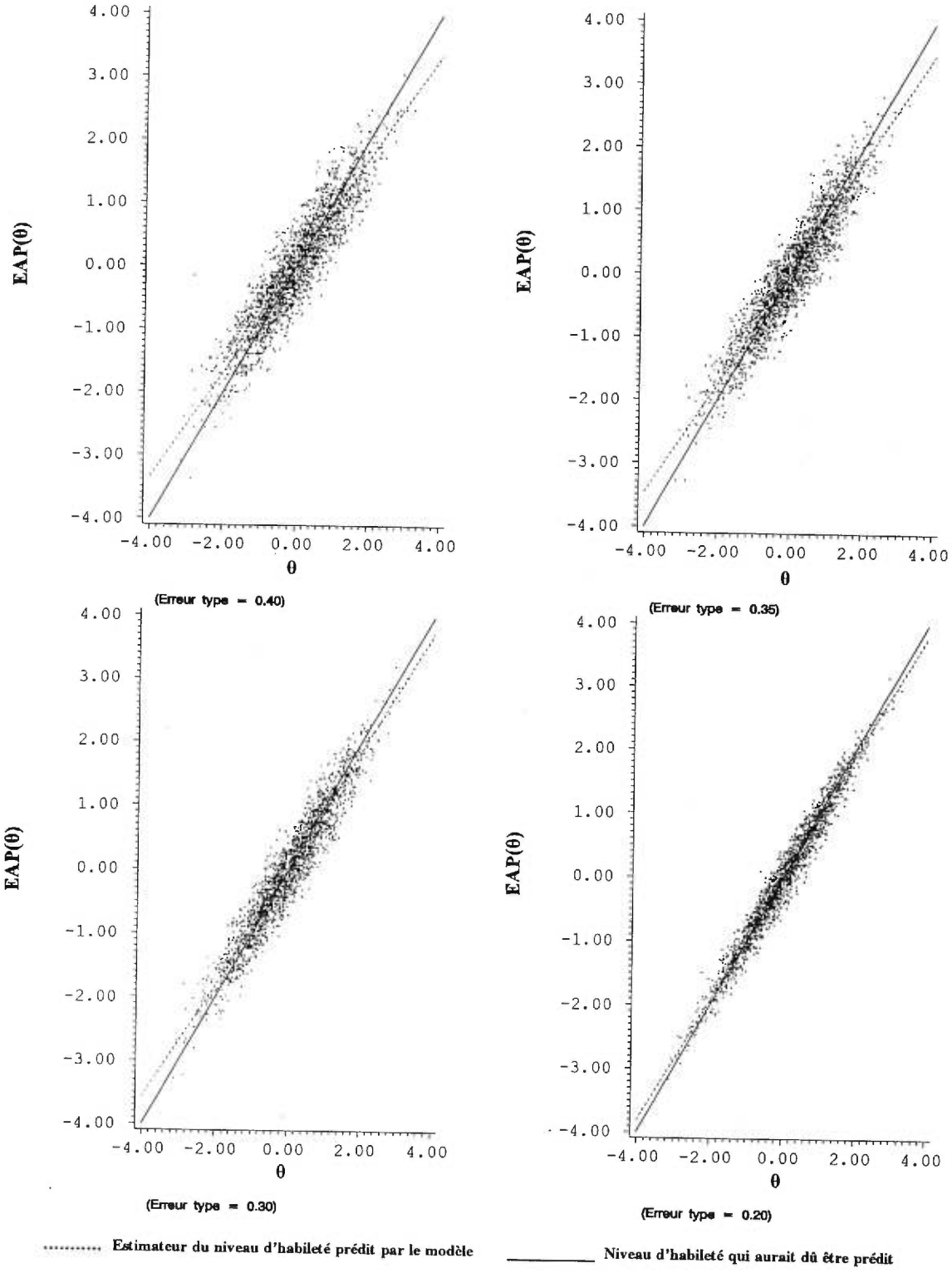

Figure 8.3 Estimateur du niveau d'habileté, EAP(θ), en fonction du niveau d'habileté lorsque la règle d'arrêt selon l'erreur type est égale à 0,20, 0,30, 0,35 et 0,40

Tableau 8.2
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et l'estimateur du niveau d'habileté, EAP(θ), lorsque la règle d'arrêt est basée sur une valeur de l'erreur type de 0,40, 0,35, 0,30 et 0,20

Règle d'arrêt selon l'erreur type	Modèle $(b_0+b_1\theta+b_2\theta^2+b_3\theta^3)$	Coefficient de détermination (R^2)
0,40	0,00206 + 0,84868 θ + 0,00040 θ^2 - 0,00078 θ^3	0,84
0,35	0,00419 + 0,88381 θ + 0,00127 θ^2 - 0,00007 θ^3	0,88
0,30	-0,00962 + 0,91423 θ + 0,00593 θ^2 - 0,00007 θ^3	0,91
0,20	-0,01069 + 0,95981 θ + 0,00096 θ^2 - 0,00041 θ^3	0,96

À la figure 8.3, on peut aussi remarquer que l'estimateur du niveau d'habileté est à peu près égal au niveau d'habileté autour de la moyenne du niveau d'habileté, soit autour de 0,01. Plus le niveau d'habileté s'éloigne de la moyenne, plus nous observons une différence entre l'estimateur du niveau d'habileté et le niveau d'habileté. Lorsque le niveau d'habileté est inférieur à la moyenne, l'estimateur du niveau d'habileté a tendance à être surestimé. À l'inverse, lorsque le niveau d'habileté est supérieur à la moyenne, l'estimateur du niveau d'habileté est généralement sous-estimé. Nous constatons cette situation à toutes les valeurs de l'erreur type retenue pour la règle d'arrêt. Son importance devient toutefois moins grande avec la diminution de la valeur de l'erreur type retenue.

Notons aussi que la variabilité de l'estimateur du niveau d'habileté est moins importante lorsque l'erreur type retenue pour la règle d'arrêt diminue. Cela reflète ce que nous observerons à la section suivante concernant la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté, soit la diminution de l'erreur type de l'estimateur du niveau d'habileté en fonction de la variation de la règle d'arrêt selon l'erreur type.

À titre d'exemple, selon les fonctions de régression, lorsque le niveau d'habileté est égal à -3,00 et que l'erreur type retenue correspond à 0,40 et à 0,20, l'estimateur du niveau d'habileté serait respectivement égal à -2,52 et à -2,87. La différence entre le niveau d'habileté et l'estimateur du niveau d'habileté est alors assez importante ; elle est respectivement de 0,48 et de 0,13. Soulignons que, dans le premier cas, cette différence est supérieure à l'erreur type retenue pour la règle d'arrêt et que, dans le second cas, elle s'en rapproche beaucoup. La précision de l'estimateur du niveau d'habileté est donc diminuée de manière importante aux valeurs extrêmes du niveau d'habileté. Lorsque le niveau d'habileté est égal à 1,00, l'estimateur du niveau d'habileté est égal à 0,85 et à 0,95. À ce moment, la différence entre le niveau d'habileté et l'estimateur du niveau d'habileté est de 0,15 et de 0,05. C'est donc uniquement lorsque la valeur de l'erreur type retenue pour la règle d'arrêt correspond à 0,20 et que la valeur du niveau d'habileté n'est pas trop extrême que nous jugeons raisonnable la précision de l'estimateur du niveau d'habileté. Un concept en lien avec ces dernières observations est d'ailleurs traité plus en profondeur à l'intérieur d'une prochaine section (8.2.3) ; il s'agit du biais de l'estimateur du niveau d'habileté.

8.2.2 Caractéristiques de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

La figure 8.4 présente au quadrant (a) les moyennes, minimum et maximum affichés à l'intérieur du tableau 8.3 en fonction de chacune des valeurs de l'erreur type retenues pour la règle d'arrêt. Elle présente respectivement aux quadrants (b), (c) et (d), l'écart type, l'asymétrie et la kurtose de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté.

Il est important tout au long de cette section de ne pas confondre l'erreur type retenue pour la règle d'arrêt et l'erreur type obtenue de l'estimateur du niveau d'habileté, cette dernière n'étant pas forcément égale à la première. En fait, la valeur de l'erreur type retenue doit être considérée comme la limite supérieure à atteindre de l'erreur type de l'estimateur du niveau d'habileté.

Selon le tableau 8.3 et la figure 8.4, plus la valeur de l'erreur type retenue pour la règle d'arrêt est petite, plus l'erreur type de l'estimateur d'habileté se rapproche de la valeur retenue pour la règle d'arrêt. Ainsi, lorsque la valeur de l'erreur type retenue pour la règle d'arrêt se situe entre 0,85 et 0,45, une différence d'au plus 0,09 est remarquée entre cette valeur et celle de la moyenne de l'erreur type obtenue. Entre une valeur de l'erreur type retenue pour la règle d'arrêt de 0,40 et 0,30, la différence n'est que de 0,01. Ce n'est qu'à partir d'une valeur de 0,25 qu'on ne remarque aucune différence entre l'erreur type de l'estimateur du niveau d'habileté et la valeur de l'erreur type retenue pour la règle

d'arrêt. Toutefois, si l'on exclut la situation où l'erreur type retenue est égale à 0,85, ce n'est que lorsque la valeur de l'erreur type retenue pour la règle d'arrêt est égale à 0,20 que les maximums et minimums de l'erreur type de l'estimateur du niveau d'habileté sont tous deux égaux à la valeur de l'erreur type retenue pour la règle d'arrêt.

Notons aussi que les maximums et minimums sont toujours situés à l'intérieur d'un intervalle peu étendu, soit au plus 0,07, et que l'écart type est, à toutes fins utiles, toujours nul. Enfin, la valeur maximale de l'erreur type est toujours égale à l'erreur type retenue pour la règle d'arrêt lorsque cette dernière est inférieure à 0,60. Lorsque l'erreur type retenue pour la règle d'arrêt est égale ou supérieure à 0,60, le maximum est toujours inférieur à l'erreur type retenue pour la règle d'arrêt. Cette situation était prévisible puisque l'administration du test adaptatif est stoppée au moment où l'erreur type de l'estimateur du niveau d'habileté atteint au plus la valeur de l'erreur type retenue pour la règle d'arrêt. Le test adaptatif peut donc se terminer par l'obtention d'une valeur de l'erreur type égale à l'erreur type retenue pour la règle d'arrêt ou encore lorsque la valeur de l'erreur type est inférieure à l'erreur type retenue pour la règle d'arrêt.

Tableau 8.3
Caractéristiques de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

ERREUR TYPE	MOYENNE	ÉCART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
0,85	0,82	0,00	nd.*	nd.	0,82	0,82
0,80	0,71	0,02	-0,02	-2,00	0,69	0,73
0,75	0,71	0,02	-0,02	-2,00	0,69	0,73
0,70	0,67	0,03	-0,77	-1,03	0,62	0,69
0,65	0,61	0,02	-1,93	2,18	0,57	0,62
0,60	0,56	0,01	0,23	-0,70	0,54	0,59
0,55	0,53	0,02	-0,12	-1,48	0,49	0,55
0,50	0,47	0,01	0,30	-0,29	0,45	0,50
0,45	0,43	0,01	-0,64	0,22	0,41	0,45
0,40	0,39	0,01	-0,53	-0,01	0,37	0,40
0,35	0,34	0,00	-0,02	-0,66	0,33	0,35
0,30	0,29	0,00	-0,35	-0,32	0,29	0,30
0,25	0,25	0,00	-0,25	-0,66	0,24	0,25
0,20	0,20	0,00	-0,45	-0,66	0,20	0,20

* Valeur non disponible car indéterminée (division par zéro)

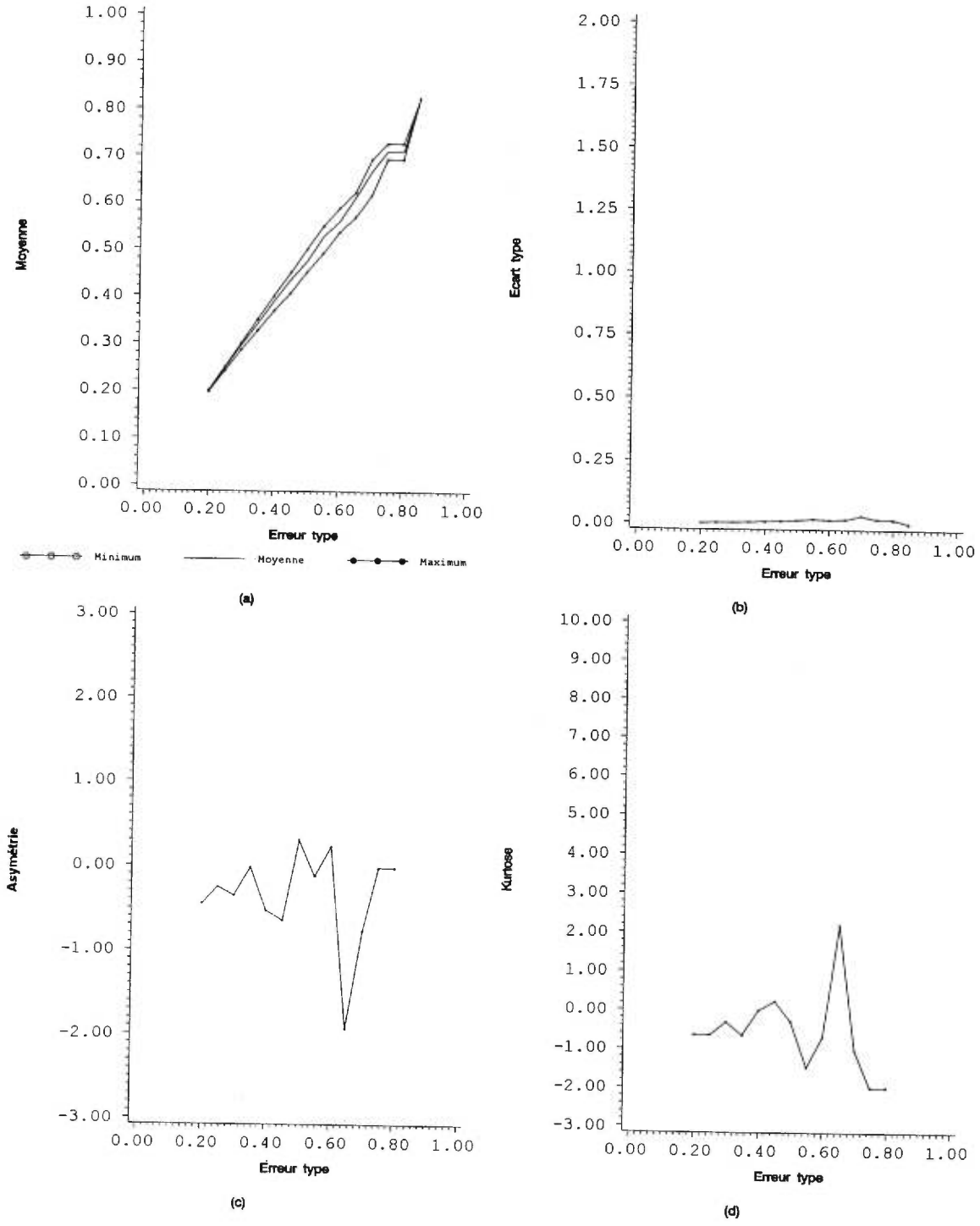

Figure 8.4 Caractéristiques de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

L'asymétrie et la kurtose de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté sont presque toujours négatives, ce qui indique que la distribution est généralement platykurtique, sauf lorsque l'erreur type est égale à 0,45 ou à 0,65, et que la médiane est généralement supérieure à la moyenne, sauf lorsque l'erreur type est égale à 0,50 ou à 0,60. Le cheminement en dent de scie des valeurs d'asymétrie et de kurtose est toutefois surprenant.

Nous remarquons que ce cheminement affiche une plus grande variabilité lorsque l'erreur type retenue pour la règle d'arrêt est élevée. Le nombre restreint de valeurs possibles de l'estimateur du niveau d'habileté lorsque l'erreur type retenue est élevée peut expliquer ce phénomène. Ainsi, quand l'erreur type retenue est égale à 0,70 seulement trois valeurs de l'erreur type sont observées, soient 0,62, 0,67 et 0,69 dans respectivement 25 %, 26 % et 49 % des cas. Cela est bien sûr normal, car le nombre de patrons de réponses possibles est déterminé par le nombre d'items administrés : quand deux items sont administrés, seulement quatre patrons de réponses différents sont possibles (2^2) et nous obtenons tout au plus quatre valeurs différentes de l'estimateur du niveau d'habileté. De plus, l'erreur type peut être la même pour des valeurs de l'estimateur du niveau d'habileté qui ne diffèrent que par le fait qu'elles sont négatives ou positives ; cela diminue d'autant le nombre de valeurs possibles de l'erreur type de l'estimateur du niveau d'habileté. Quand l'erreur type retenue est égale à 0,65, nous obtenons cinq valeurs de l'erreur type : 0,5686, 0,6111, 0,6163, 0,6217 et 0,6218 dans respectivement 13 %, 26 %, 25 %, 13 % et 23 % des cas. Ce petit nombre de valeurs différentes de l'erreur

type peut faire en sorte que l'asymétrie et la kurtose affichent une importante variabilité lorsque l'erreur type retenue de l'estimateur du niveau d'habileté est élevée. Quand l'erreur type retenue est inférieure ou égale à 0,45, le nombre de valeurs différentes de l'erreur type est toujours supérieur à 100, atteignant un maximum de 303 lorsque l'erreur type retenue est de 0,35.

Lorsque les minimums et maximums de l'erreur type de l'estimateur du niveau d'habileté sont égaux et qu'en conséquence l'écart type de l'erreur type de l'estimateur du niveau d'habileté est nul, il est impossible de calculer l'asymétrie ni la kurtose ; c'est que leur valeur est indéterminée puisque le diviseur est alors nul. Il en est ainsi lorsque l'erreur type retenue pour la règle d'arrêt est égale à 0,80. Lorsque la valeur de l'erreur type retenue pour la règle d'arrêt est égale ou inférieure à 0,80, l'écart type obtenu peut être égal à 0,00 quant sa valeur est arrondie à deux décimales ; dans les faits, il est tout de même supérieur à zéro et c'est pourquoi les valeurs de l'asymétrie et de la kurtose ne sont pas indéterminées.

À la figure 8.5, des corrélogrammes présentent les valeurs de l'erreur type de l'estimateur du niveau d'habileté obtenues en fonction du niveau d'habileté pour quatre valeurs de l'erreur type retenue pour la règle d'arrêt : 0,20, 0,30, 0,35 et 0,40. De plus, une régression cubique est utilisée pour ajuster une ligne de régression à chacun de ces corrélogrammes. Les coefficients de régression et de détermination sont présentés au tableau 8.4.

Dans tous les cas, le coefficient de détermination est soit égal à 0,00 ou de très faible importance. Cette situation, prévisible selon nous, peut s'expliquer par le fait que l'erreur type de l'estimateur du niveau d'habileté est à toutes fins utiles constante sur toute l'étendue du niveau d'habileté. Il n'y a donc presque aucune variabilité de l'erreur type et, par conséquent, aucune covariabilité.

À la figure 8.5, on peut aussi remarquer que l'erreur type de l'estimateur du niveau d'habileté varie très peu et qu'elle est plutôt constante sur toute l'étendue étudiée du niveau d'habileté ; et ce quelle que soit l'erreur type retenue pour la règle d'arrêt. On note aussi que moins la valeur de l'erreur type retenue pour la règle d'arrêt est grande, moins importante est la variabilité de l'erreur type de l'estimateur du niveau d'habileté.

Enfin, comme nous avons pu le constater au tableau 8.3, il y a une différence d'au plus 0,11 (0,80 - 0,69) entre l'erreur type retenue pour la règle d'arrêt et le minimum de l'erreur type de l'estimateur du niveau d'habileté. Lorsque l'erreur type retenue pour la règle d'arrêt est égale ou inférieure à 0,30, donc à l'intérieur d'un intervalle utilisé dans la pratique de l'administration des tests adaptatifs, cette différence est d'au plus 0,01. Ces considérations permettent de constater que la précision de l'estimateur du niveau d'habileté, telle que mesurée par son erreur type, est quasi constante quel que soit le niveau d'habileté.

Cette caractéristique de l'erreur type de l'estimateur du niveau d'habileté en testing adaptatif fait en sorte que, lorsque la règle d'arrêt selon l'erreur type est utilisée, on obtient, à toutes fins utiles, l'homogénéité de l'erreur type postulée dans la théorie classique des tests. Cette dernière constatation est intéressante, car il est ainsi possible d'assurer une égale précision de l'estimateur du niveau d'habileté quel que soit le niveau d'habileté de l'élève.

Sur la base de ces observations concernant la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté, il nous semble pertinent de recommander de ne pas utiliser une valeur supérieure à 0,30 comme règle d'arrêt dans un test adaptatif pour que l'erreur type soit à toutes fins utiles constante quel que soit le niveau d'habileté. Notons toutefois qu'au tableau 8.4 on remarquait une constance absolue de l'erreur type du niveau d'habileté seulement lorsque l'erreur type retenue était égale ou inférieure à 0,20.

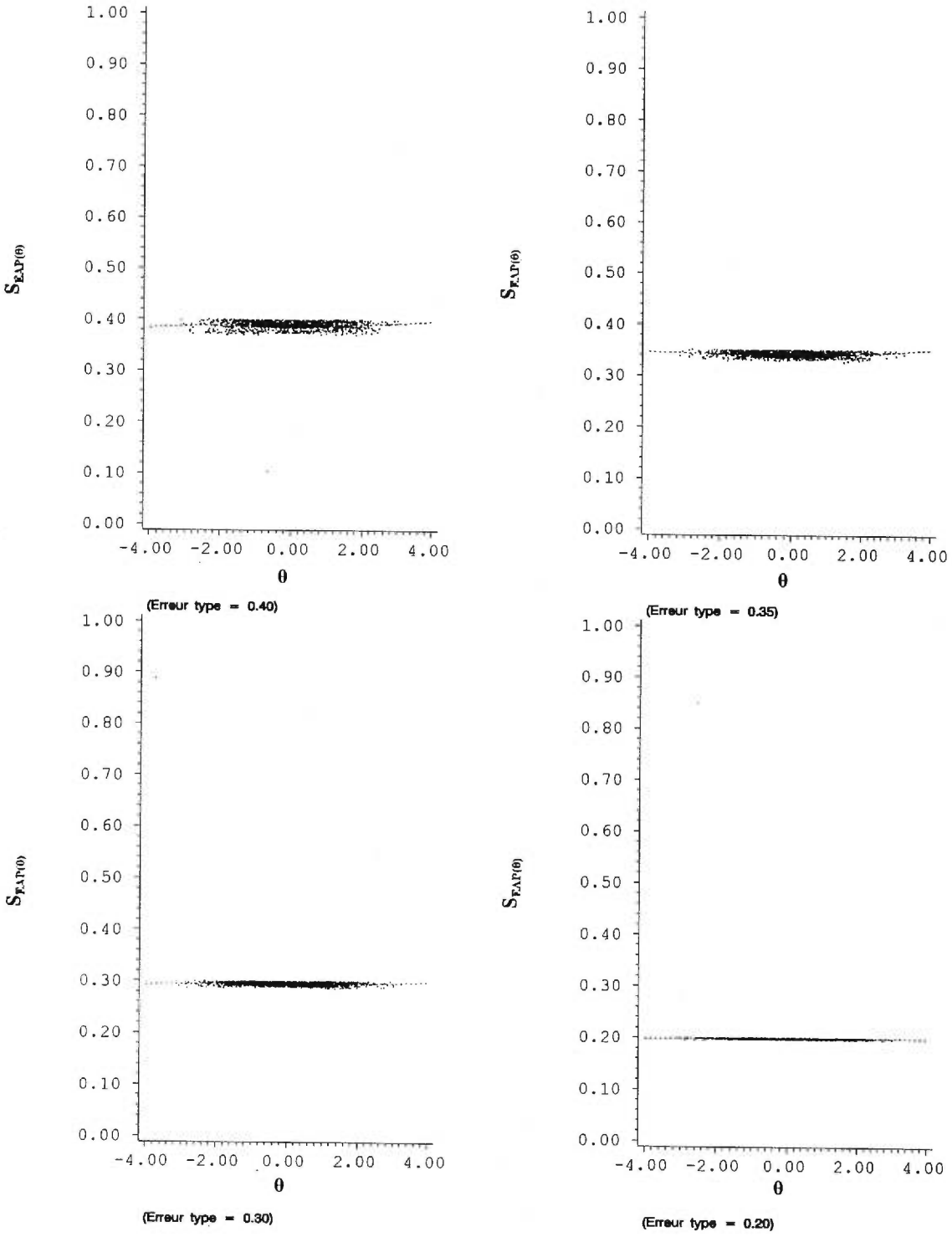

Figure 8.5 Erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, en fonction du niveau d'habileté lorsque la règle d'arrêt selon l'erreur type est égale à 0,20, 0,30, 0,35 et 0,40

Tableau 8.4
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et l'erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, lorsque la règle d'arrêt est basée sur une valeur de l'erreur type de 0,40, 0,35, 0,30 et 0,20

Règle d'arrêt selon l'erreur type	Modèle $(b_0+b_1\theta+b_2\theta^2+b_3\theta^3)$	Coefficient de détermination (R^2)
0,40	$0,38881 - 0,00007\,\theta + 0,00010\,\theta^2 + 0,00010\,\theta^3$	0,00
0,35	$0,34099 - 0,00010\,\theta + 0,00045\,\theta^2 + 0,00005\,\theta^3$	0,02
0,30	$0,29483 - 0,00010\,\theta + 0,00007\,\theta^2 + 0,00005\,\theta^3$	0,00
0,20	$0,19861 - 0,00000\,\theta + 0,00003\,\theta^2 + 0,00000\,\theta^3$	0,00

8.2.3 Caractéristiques de la distribution de probabilité de l'erreur de mesure du niveau d'habileté et du biais de l'estimateur du niveau d'habileté en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

Nous présentons au tableau 8.5 les caractéristiques de la distribution de probabilité de l'erreur de mesure du niveau d'habileté, (EAP(θ) - θ), selon chacune des valeurs de l'erreur type retenue pour la règle d'arrêt. Ces mêmes informations sont reproduites sous forme graphique à la figure 8.6. Nous traiterons un peu plus loin, à l'intérieur de cette même section, du biais de l'estimateur du niveau d'habileté, $BIAIS_{EAP(\theta)}$. À titre de rappel, le biais correspond à la différence moyenne obtenue entre l'estimateur du niveau d'habileté et le niveau d'habileté (équation 4.7, p. 62).

Nous pouvons observer au tableau 8.5 que la moyenne de l'erreur de mesure du niveau d'habileté est à peu près égale à 0,00, et ce quelle que soit la valeur de l'erreur type retenue pour la règle d'arrêt, tout au plus atteint-elle 0,01 en valeur absolue. L'écart type de cette distribution de probabilité est pour sa part presque égal, mais jamais supérieur, à l'erreur type retenue pour la règle d'arrêt. Nous notons que la différence entre ces deux dernières valeurs est d'au plus 0,10 lorsque l'erreur type retenue pour la règle d'arrêt est élevée. Lorsque l'erreur type retenue pour la règle d'arrêt est plutôt petite, soit égale ou inférieure à 0,40, la différence entre l'écart type de l'erreur de mesure et l'erreur type retenue est d'au plus 0,01. La figure 8.6 montre qu'il existe une relation presque linéaire entre l'écart type de l'erreur de mesure du niveau d'habileté et l'erreur type retenue pour la règle d'arrêt.

Tableau 8.5
Caractéristiques de la distribution de probabilité de l'erreur de mesure du niveau d'habileté, (EAP(θ) - θ), en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

ERREUR TYPE	MOYENNE	ÉCART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
0,85	-0,01	0,81	0,04	0,19	-2,68	3,43
0,80	-0,01	0,70	-0,03	0,17	-2,51	3,01
0,75	-0,01	0,70	-0,03	0,17	-2,51	3,01
0,70	0,01	0,65	-0,04	0,20	-2,17	3,01
0,65	0,01	0,60	-0,02	0,10	-2,17	2,70
0,60	0,01	0,55	0,02	0,13	-1,91	2,43
0,55	0,01	0,51	0,03	0,11	-1,69	2,19
0,50	0,01	0,47	-0,01	-0,07	-1,52	1,59
0,45	0,00	0,43	-0,05	-0,19	-1,33	1,22
0,40	0,00	0,39	0,05	-0,24	-1,16	1,19
0,35	0,00	0,34	-0,01	-0,09	-1,12	1,09
0,30	0,00	0,29	-0,01	-0,11	-0,88	1,07
0,25	0,00	0,25	0,04	-0,12	-0,71	0,97
0,20	-0,01	0,19	-0,09	0,20	-0,65	0,73

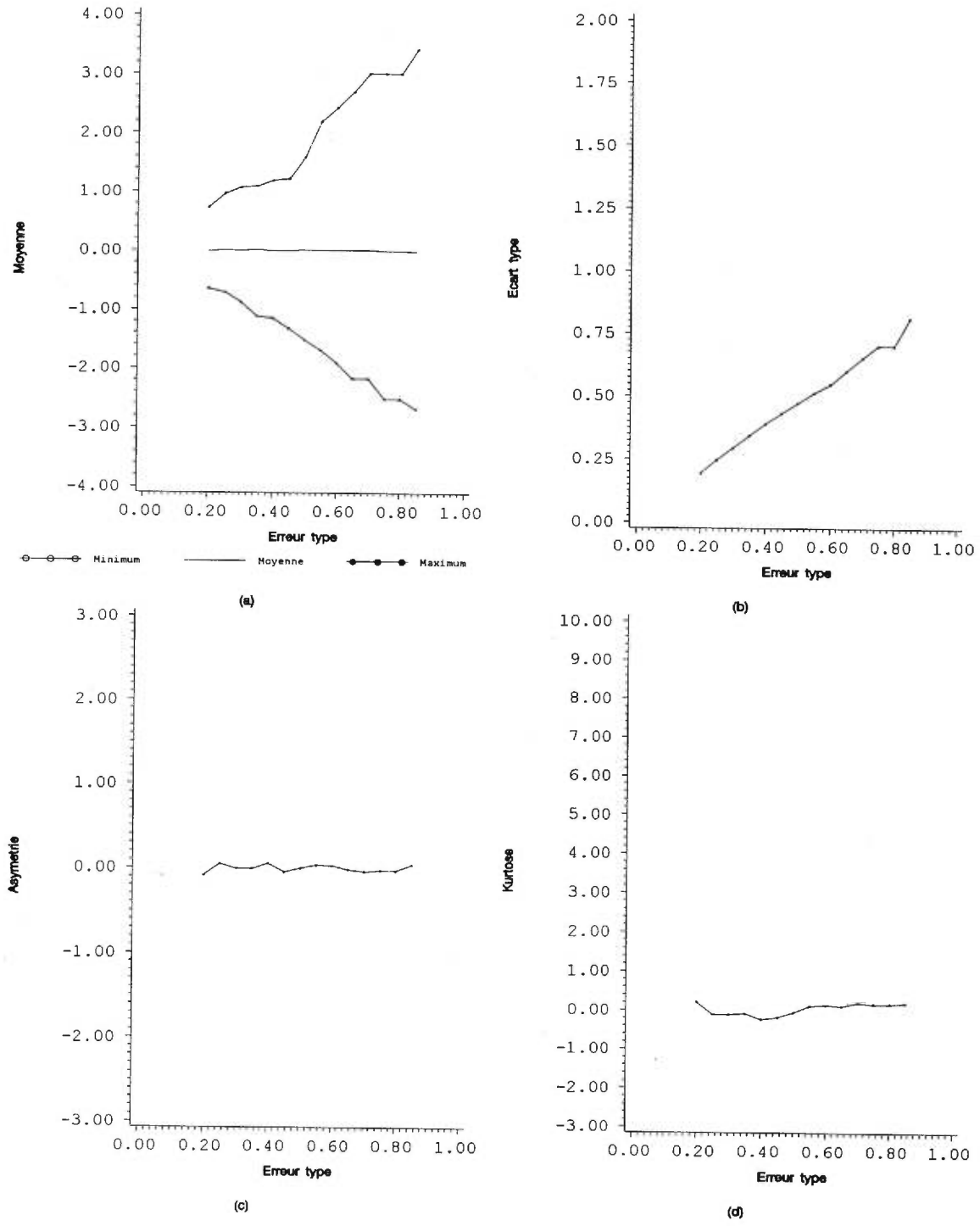

Figure 8.6 Caractéristiques de la distribution de probabilité de l'erreur de mesure du niveau d'habileté, (EAP(θ) - θ), en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

L'asymétrie de la distribution de probabilité de l'erreur de mesure du niveau d'habileté affiche des valeurs assez peu importantes, soit entre -0,09 et 0,05. La distribution de probabilité peut donc être considérée comme symétrique. Nous remarquons cependant que la kurtose présente une plus grande étendue de valeurs, même si celles-ci ne sont généralement pas très importantes, soit entre -0,24 et 0,20. Ces valeurs fluctuent entre des valeurs positives et négatives ; il nous est donc impossible de dire que la distribution de probabilité de l'erreur de mesure est constamment leptokurtique ou platykurtique. La distribution de probabilité de l'erreur de mesure du niveau d'habileté, considérant les valeurs obtenues de la kurtose, semble donc s'éloigner légèrement d'une distribution normale.

Enfin, les valeurs minimales et maximales de l'erreur de mesure du niveau d'habileté sont assez importantes à tous les niveaux de l'erreur type retenue pour la règle d'arrêt. Lorsque l'erreur type retenue pour la règle d'arrêt est égale à 0,20, l'erreur de mesure du niveau d'habileté peut atteindre -0,65 et 0,73. Ces valeurs sont respectivement 3,25 et 3,65 fois supérieures à la valeur de l'erreur type retenue pour la règle d'arrêt ; ce sont des valeurs probables des minimums et maximums d'une distribution normale N(0,00, 0,20). À toutes les valeurs de l'erreur type retenue pour la règle d'arrêt, les minimums et maximums de l'erreur de mesure du niveau d'habileté affichent des valeurs entre -3,35 et 4,15 fois supérieures à cette dernière.

Nous retenons de cette analyse que, malgré quelques fluctuations de sa kurtose, tout autant négatives que positives, la distribution de probabilité de l'erreur de mesure du niveau d'habileté se comporte à peu près comme une distribution normale $N(0,00, S_{EAP(\theta)})$. De plus, selon cette analyse, puisque l'erreur de mesure du niveau d'habileté calculée sur toute l'étendue des valeurs du niveau d'habileté est à toutes fins utiles égale à 0,00, la moyenne de l'estimateur du niveau d'habileté est considérée comme un estimateur non biaisé de la moyenne du niveau d'habileté.

Nous retenons aussi que l'erreur de mesure du niveau d'habileté, malheureusement inconnue dans les situations réelles, peut s'avérer très importante et faire en sorte que l'estimateur du niveau d'habileté s'éloigne considérablement du niveau d'habileté. En ce sens, l'erreur type de l'estimateur du niveau d'habileté quantifie l'intervalle de confiance à 68,27 % autour de l'estimateur du niveau d'habileté ; un intervalle de confiance qui s'approcherait de 99,00 % serait alors beaucoup plus important, comme le montrent les minimums et maximums obtenus de l'erreur de mesure du niveau d'habileté.

On trouve à la figure 8.7 une représentation graphique des corrélogrammes correspondant aux valeurs obtenues pour l'erreur de mesure du niveau d'habileté en fonction du niveau d'habileté pour quatre valeurs de l'erreur type retenue pour la règle d'arrêt. Nous utilisons, de plus, une régression cubique dans le but d'ajuster une ligne de régression à chacun des corrélogrammes.

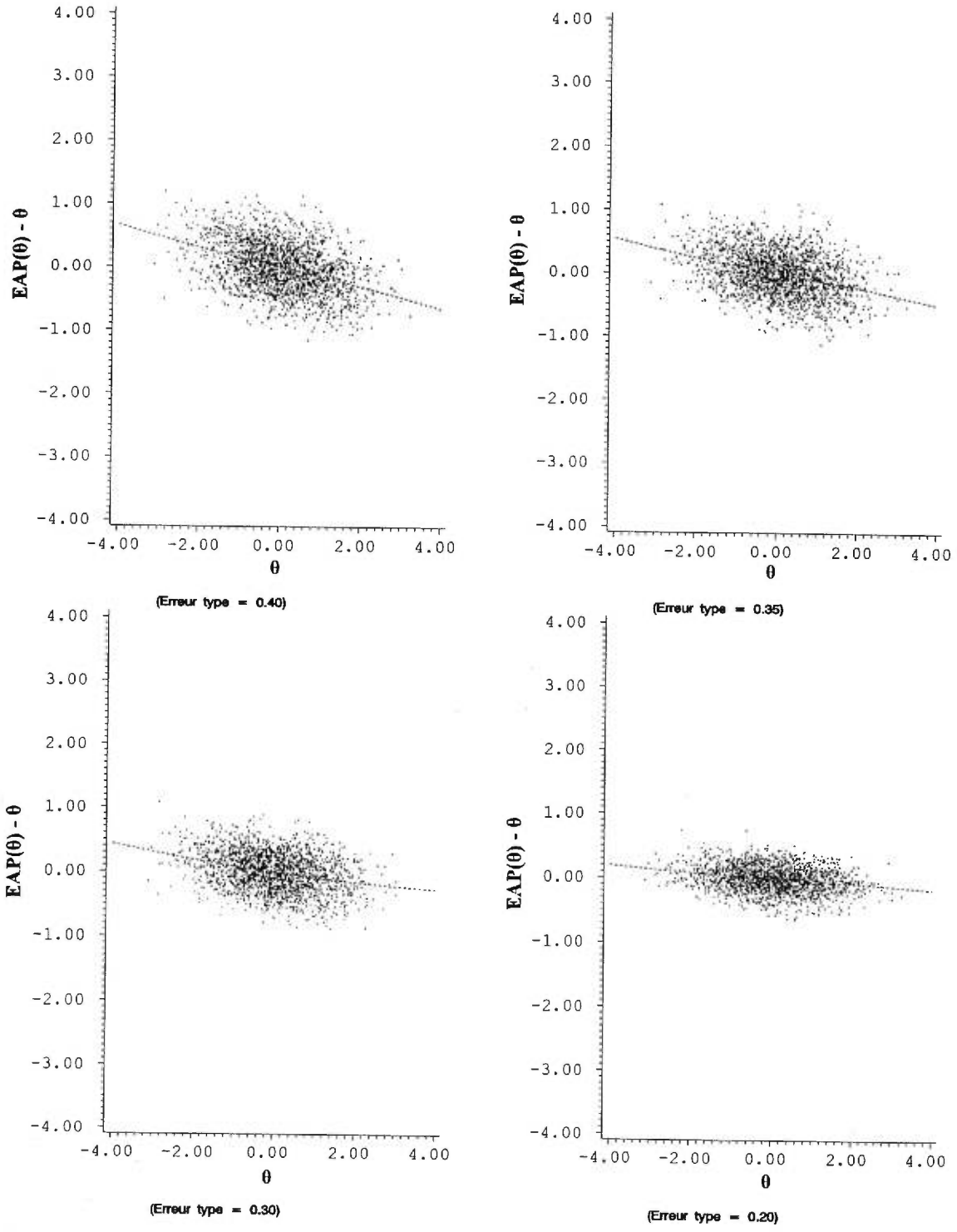

Figure 8.7 Erreur de mesure du niveau d'habileté, EAP(θ) - θ, en fonction du niveau d'habileté lorsque la règle d'arrêt selon l'erreur type est égale à 0,20, 0,30, 0,35 et 0,40

Les coefficients de régression et les coefficients de détermination sont présentés au tableau 8.6. Le coefficient de détermination est généralement peu important ; tout au plus atteint-il 0,15. Nous pouvons donc affirmer que l'erreur de mesure du niveau d'habileté n'est pas prédite de manière vraiment exacte par la valeur du niveau d'habileté et qu'ainsi la variabilité de l'erreur de à chaque valeur du niveau d'habileté est assez importante. Dans le cas présent, l'erreur de mesure prédite correspond aussi à la valeur moyenne de l'erreur de mesure à des valeurs spécifiques du niveau d'habileté, soit le biais de l'estimateur du niveau d'habileté. La valeur du biais de l'estimateur du niveau d'habileté prédite par le modèle de régression permettra de juger si l'estimateur du niveau d'habileté est biaisé ou non à différentes valeurs du niveau d'habileté.

Tableau 8.6
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et l'erreur de mesure du niveau d'habileté, EAP(θ) - θ, lorsque la règle d'arrêt est basée sur une valeur de l'erreur type de 0,40, 0,35, 0,30 et 0,20

Règle d'arrêt selon l'erreur type	Modèle ($b_0+b_1\theta+b_2\theta^2+b_3\theta^3$)	Coefficient de détermination (R^2)
0,40	0,00206 - 0,15132 θ + 0,00040 θ^2 - 0,00078 θ^3	0,15
0,35	0,00419 - 0,11620 θ + 0,00127 θ^2 - 0,00088 θ^3	0,12
0,30	-0,00962 - 0,08577 θ + 0,00593 θ^2 - 0,00007 θ^3	0,08
0,20	-0,01069 - 0,04019 θ + 0,00096 θ^2 - 0,00041 θ^3	0,04

Nous notons que plus l'erreur type retenue pour la règle d'arrêt est élevée, plus le coefficient de régression linéaire est élevé. Ainsi, lorsque l'erreur type retenue pour la règle d'arrêt est égale à 0,40, le coefficient de régression linéaire est égal à 0,15. Lorsque cette dernière est égale à 0,20, le coefficient de régression affiche une valeur de 0,04 seulement. Nous constatons donc que le biais de l'estimateur du niveau d'habileté devient de plus en plus constant sur toute l'étendue du niveau d'habileté avec la diminution de l'erreur type retenue pour la règle d'arrêt. Ce résultat, que l'on peut aussi constater à la figure 8.7, était prévisible.

Lorsque le niveau d'habileté est très faible, le biais de l'estimateur du niveau d'habileté a tendance à être positif, indication de surestimation du niveau d'habileté. À l'opposé, lorsque le niveau d'habileté est élevé, la tendance est à la sous-estimation du niveau d'habileté. Cela correspond exactement à la situation que nous avons rencontrée à la section précédente (8.2.1) où nous analysions la distribution de probabilité de l'estimateur du niveau d'habileté. Il s'ensuit que les valeurs ajustées du biais de l'estimateur du niveau d'habileté par les fonctions de régression sont exactement les mêmes que celles de la différence entre le niveau d'habileté et l'estimateur du niveau d'habileté obtenues à la section précédente.

Le biais de l'estimateur du niveau d'habileté n'est pas constant sur toute l'étendue du niveau d'habileté. Il peut même être assez important aux valeurs extrêmes du niveau d'habileté. Par exemple, il est de 0,13 lorsque le niveau d'habileté est de -3,00 et que

l'erreur type retenue pour la règle d'arrêt est de 0,20. Le biais est alors égal à un peu plus de 50 % de la valeur de l'erreur type de l'estimateur du niveau d'habileté. À ce moment, il serait sûrement plus prudent d'utiliser une formule permettant l'ajustement de l'erreur type de l'estimateur du niveau d'habileté en fonction du biais, similaire à la formule proposée par Samejima (1994) lorsque l'estimateur du niveau d'habileté est obtenu par la méthode de vraisemblance maximale, et similaire à celle suggérée par Warm (1989, p. 430) lorsqu'il est obtenu par la méthode de maximisation a posteriori. Bock et Mislevy (1982, p. 439-442) proposent d'ailleurs une stratégie de correction du biais de l'estimateur du niveau d'habileté lorsque celui-ci est obtenu à partir de la méthode de l'espérance a posteriori. Mais il ne faut pas confondre la correction utilisée par Bock et Mislevy avec celle proposée Sheppard et abordée lors de la description de notre méthodologie. Bock et Mislevy effectuent cette correction en divisant l'estimateur du niveau d'habileté par une approximation du coefficient de fidélité :

$$1 - S^2_{EAP(\theta)} \qquad \textbf{(8.1)}$$

L'estimateur corrigé du niveau d'habileté devient alors égal à :

$$\frac{EAP(\theta)}{1 - S^2_{EAP(\theta)}} \qquad \textbf{(8.2)}$$

Selon leurs observations, cette correction réduit considérablement le biais de l'estimateur du niveau d'habileté à toutes les valeurs du niveau d'habileté. Nous avons vérifié l'impact de cet ajustement : lorsque l'erreur type retenue est égale à 0,40 et que le niveau d'habileté est de -2,64, le biais de l'estimateur du niveau d'habileté, au départ de 0,43, est alors ramené à 0,04. À la même valeur du niveau d'habileté, lorsque l'erreur type retenue est de 0,20, le biais de l'estimateur du niveau d'habileté passe de 0,08 à -0,02. La diminution du biais à partir de la correction de Bock et Mislevy est alors très importante.

Notons toutefois, comme le souligne d'ailleurs ces auteurs, que la réduction du biais s'accompagne toujours d'une augmentation de la variabilité des valeurs obtenues pour l'estimateur du niveau d'habileté. En effet, nous observons que, lorsque l'erreur type retenue est égale à 0,40 et que le niveau d'habileté est de -2,64, l'erreur de mesure, qui variait entre -0,51 et 1,19, affiche maintenant des valeurs qui se situent entre -1,08 et 0,92 ; l'étendue de ces valeurs passe donc de 1,70 à 2,00. Quand l'erreur type retenue est égale à 0,20, l'erreur de mesure, variant au départ entre -0,14 et 0,37, se situe, après correction, entre -0,27 et 0,27, une étendue qui passe de 0,51 à 0,54. Cette augmentation de la variabilité des valeurs obtenues de l'estimateur du niveau d'habileté se traduit par une augmentation de l'erreur type de l'estimateur du niveau d'habileté ; de 0,40, elle passe à 0,48 et de 0,20, elle devient égale à 0,21. L'effet de la correction de l'estimateur du niveau d'habileté par la méthode de Bock et Mislevy affecte assez peu l'erreur type lorsque celle-ci est assez petite, 0,20 par exemple.

Puisque la correction de Bock et Mislevy fait en sorte que la variabilité de l'estimateur du niveau d'habileté est plus grande, l'écart type de cet estimateur sur toute l'étendue du niveau d'habileté devrait aussi augmenter. En effet, nous observons que, lorsque l'erreur type est de 0,40, l'écart type de la distribution de probabilité de l'estimateur du niveau d'habileté qui était de 0,91 (tableau 8.1) prend maintenant une valeur de 1.08. Lorsque l'erreur type retenue est égale à 0,20, l'écart type passe de 0,96 à 1,00. Il faut rappeler que l'écart type de la distribution simulée du niveau d'habileté est au départ de 0,98. La correction de Bock et Mislevy augmente donc assez peu l'écart type de la distribution de probabilité de l'estimateur du niveau d'habileté lorsque l'erreur type retenue pour la règle d'arrêt est peu importante.

Ces constatations nous amènent à conclure que l'estimateur du niveau d'habileté ne peut pas être considéré comme un estimateur non biaisé sur toute l'étendue du niveau d'habileté. Selon le modèle de régression obtenu, dans la situation où l'erreur type retenue pour la règle d'arrêt est égale à 0,20, nous jugeons que l'estimateur du niveau d'habileté est un estimateur plutôt non biaisé lorsque le niveau d'habileté varie entre -1,50 et 1,50. À ce moment, le biais de l'estimateur du niveau d'habileté varie entre -0,07 et 0,05. Lorsque l'erreur type retenue pour la règle d'arrêt est égale à 0,30, le niveau d'habileté devrait se situer entre −1,00 et 1,00 ; le biais varie entre -0,05 et 0,03. En dehors de ces valeurs du niveau d'habileté, l'estimateur du niveau d'habileté est un estimateur biaisé. Il nous semble ainsi pertinent de suggérer d'utiliser une valeur de l'erreur type d'au plus 0,20 en tant que règle d'arrêt pour qu'il soit possible d'obtenir un

estimateur non biaisé du niveau d'habileté à l'intérieur de l'intervalle de -1,50 à 1,50. Pour obtenir un estimateur non biaisé sur un intervalle plus grand, l'erreur type retenue pour la règle d'arrêt devrait être encore plus petite. Il serait, cependant, prudent d'ajuster l'estimateur du niveau d'habileté en appliquant la formule de correction de l'estimateur du niveau d'habileté proposée par Bock et Mislevy. Le biais de l'estimateur du niveau d'habileté est alors diminué sur toute l'étendue du niveau d'habileté. Pour conclure, soulignons qu'il est possible d'obtenir un estimateur du niveau d'habileté à toutes fins utiles non biaisé lorsque l'erreur type retenue pour la règle d'arrêt est inférieure ou égale à 0,40, que la correction de Bock et Mislevy est utilisée et que le niveau d'habileté est compris dans l'intervalle [-3,00, 3,00].

8.2.4 Caractéristiques de la distribution de probabilité de l'asymétrie de l'estimateur du niveau d'habileté en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

Au tableau 8.7 et à la figure 8.8, nous observons le comportement de l'asymétrie, $a3_{EAP(\theta)}$, de la distribution d'échantillonnage de l'estimateur du niveau d'habileté selon chacune des valeurs de l'erreur type retenue pour la règle d'arrêt.

La moyenne de l'asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté peut être considérée, à toutes fins utiles, comme étant constamment égale à 0,00. L'écart type de l'asymétrie, pour sa part, diminue graduellement avec la réduction de l'erreur type retenue pour la règle d'arrêt. De plus, l'écart type de l'asymétrie semble tendre vers 0,00 avec la diminution de la valeur du critère de la règle d'arrêt.

La distribution de probabilité de l'asymétrie de l'estimateur du niveau d'habileté affiche des valeurs d'asymétrie qui sont toujours peu importantes, soit entre -0,07 et 0,06. Toutefois, la distribution de probabilité de l'asymétrie est platykurtique à toutes les valeurs de l'erreur type retenue pour la règle d'arrêt. La kurtose de la distribution de probabilité de l'asymétrie augmente lentement avec la réduction de l'erreur type retenue pour la règle d'arrêt. Cet aplatissement de la distribution de probabilité de l'asymétrie fait en sorte que nous ne pouvons pas considérer que cette distribution de probabilité se comporte comme une distribution normale.

Tableau 8.7
Caractéristiques de la distribution de probabilité de l'asymétrie de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

ERREUR TYPE	MOYENNE	ÉCART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
0,85	0,00	0,11	-0,02	-2,00	-0,11	0,11
0,80	0,00	0,13	-0,02	-1,21	-0,18	0,18
0,75	0,00	0,13	-0,02	-1,21	-0,18	0,18
0,70	0,01	0,12	0,03	-0,20	-0,22	0,22
0,65	0,00	0,13	0,03	-0,85	-0,24	0,24
0,60	0,00	0,13	-0,02	-0,97	-0,24	0,24
0,55	0,00	0,12	-0,07	-0,44	-0,25	0,25
0,50	0,00	0,12	-0,02	-0,73	-0,29	0,29
0,45	0,00	0,11	-0,05	-0,64	-0,26	0,28
0,40	0,00	0,10	-0,06	-0,67	-0,27	0,29
0,35	0,00	0,08	0,03	-0,60	-0,22	0,23
0,30	0,00	0,07	0,06	-0,59	-0,19	0,17
0,25	0,00	0,05	-0,04	-0,54	-0,15	0,15
0,20	0,00	0,04	0,01	-0,42	-0,12	0,10

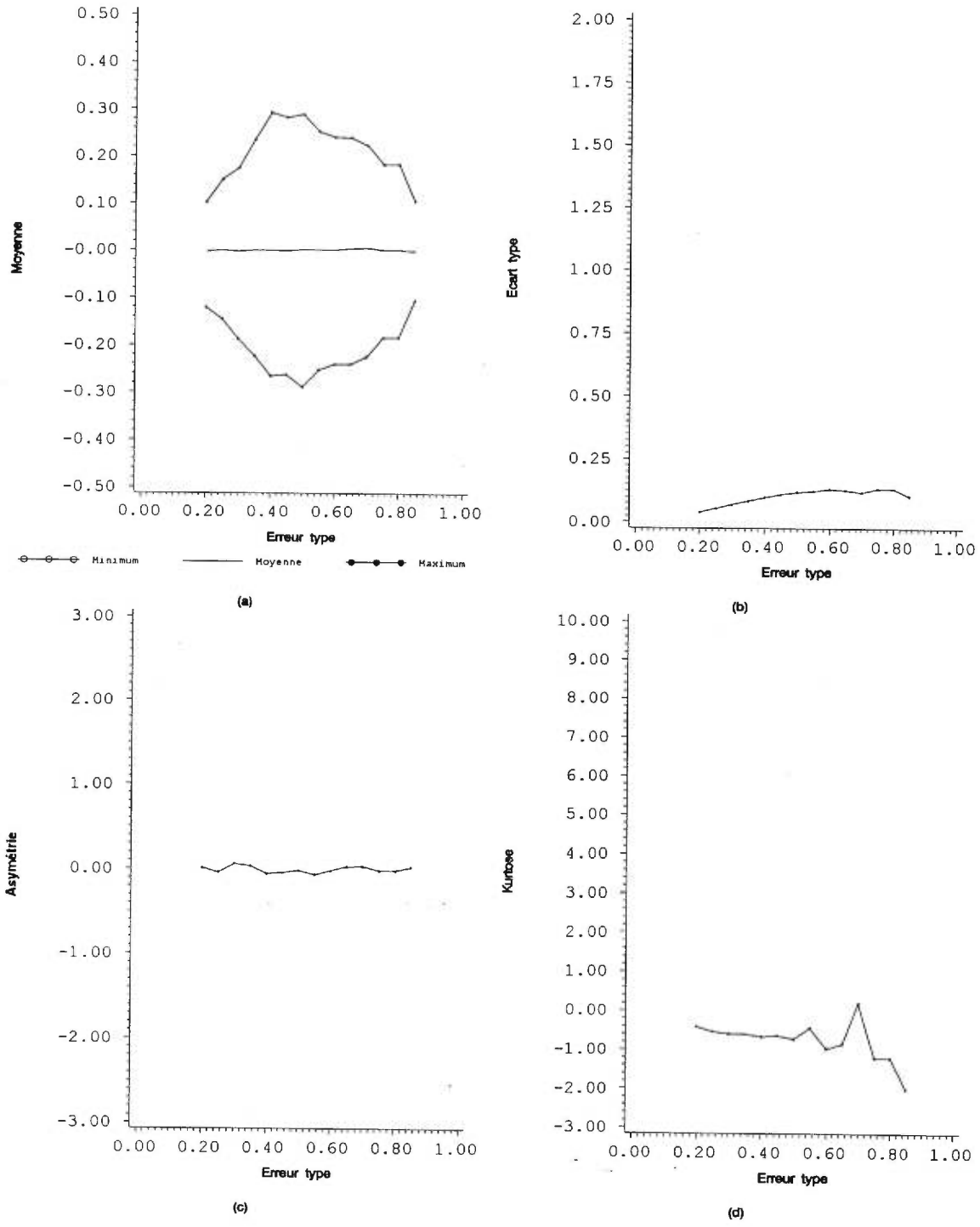

Figure 8.8 Caractéristiques de la distribution de probabilité de l'asymétrie de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

Pour ce qui est des minimums et maximums de l'asymétrie associée à l'estimateur du niveau d'habileté, nous constatons qu'ils diminuent lorsque l'erreur type retenue est inférieure à 0,40. Les valeurs minimales et maximales deviennent alors peu importantes et ont peu d'effets sur la différence entre la médiane et la moyenne de la distribution d'échantillonnage de l'estimateur du niveau d'habileté.

À titre d'illustration, selon la figure 7.5 et l'équation 7.12 (section 7.4), quand l'asymétrie affiche une valeur de 0,29 et que la kurtose est nulle, la différence entre la médiane et la moyenne correspond à 5 % de l'écart type. Lorsque l'erreur type de l'estimateur du niveau d'habileté est de 0,40 la valeur maximale de l'asymétrie est égale à 0.29 ; ce 5 % correspond alors à une différence entre médiane et moyenne de seulement -0,02. Lorsque l'erreur type est égale à 0,20 la valeur la plus extrême de l'asymétrie est de -0,12 et la différence entre la médiane et la moyenne de l'estimateur du niveau d'habileté n'est plus que de 1 % de l'erreur type, soit aussi peu que 0,002. Nous pouvons donc conclure qu'aux valeurs utilisées dans la pratique de l'erreur type retenue pour la règle d'arrêt, la différence entre l'estimateur du niveau d'habileté, calculé selon la méthode de l'espérance a posteriori, et la médiane de sa distribution d'échantillonnage peut être considérée comme étant sans grande importance puisqu'elle est alors d'au plus $\pm 0,02$.

Nous présentons au tableau 8.8 les coefficients de régression et de détermination relatifs à la relation entre l'asymétrie de la distribution d'échantillonnage de l'estimateur du

niveau d'habileté et le niveau d'habileté selon différentes valeurs de l'erreur type retenue pour la règle d'arrêt. La représentation graphique de cette relation peut être observée à la figure 8.9.

Le coefficient de détermination varie entre 0,23 et 0,34. Il diminue légèrement avec la réduction de l'erreur type retenue pour la règle d'arrêt. La relation est non linéaire, principalement aux valeurs extrêmes du niveau d'habileté, soit lorsque le niveau d'habileté est inférieur à -2,00 et lorsqu'il est supérieur à 2,00. Entre ces valeurs du niveau d'habileté, plus le niveau d'habileté est élevé, plus l'asymétrie de l'estimateur du niveau d'habileté est élevée. Lorsque l'erreur type retenue est de 0,40 et que le niveau d'habileté est de -2,00 et de 2,00, l'asymétrie calculée selon notre modèle n'est que de -0,11 et 0,11 respectivement, valeurs affectant très peu la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Lorsque l'erreur type retenue pour la règle d'arrêt n'est que de 0,20, l'asymétrie calculée de la distribution d'échantillonnage est encore beaucoup moins importante, soit de -0,04 et 0,03 seulement.

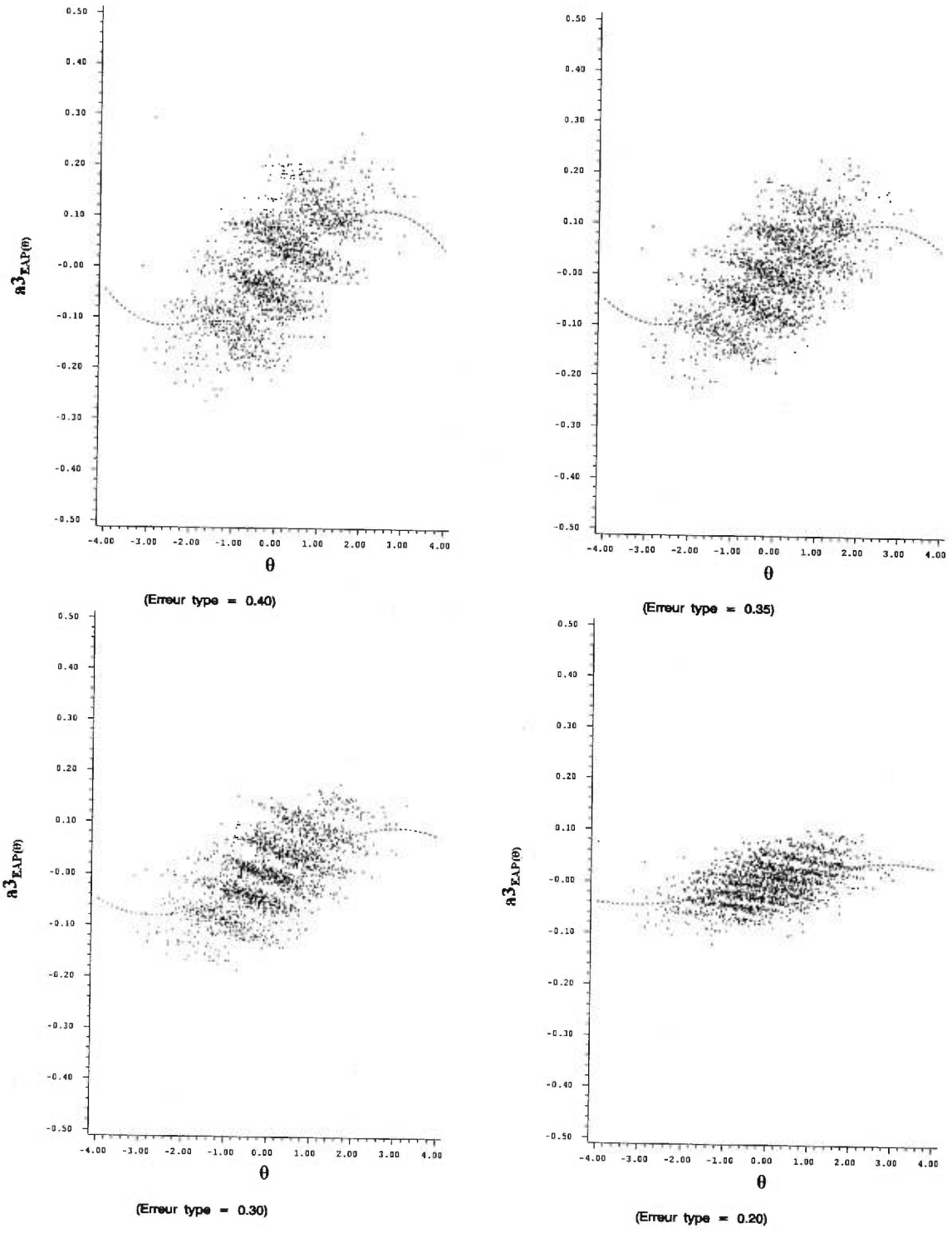

Figure 8.9 Asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, en fonction du niveau d'habileté lorsque la règle d'arrêt selon l'erreur type est égale à 0,20, 0,30, 0,35 et 0,40

Tableau 8.8
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et l'asymétrie de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, lorsque la règle d'arrêt est basée sur une valeur de l'erreur type de 0,40, 0,35, 0,30 et 0,20

Règle d'arrêt selon l'erreur type	Modèle $(b_0+b_1\theta+b_2\theta^2+b_3\theta^3)$	Coefficient de détermination (R^2)
0,40	$-0,00366 + 0,06807\,\theta - 0,00026\,\theta^2 - 0,00361\,\theta^3$	0,34
0,35	$0,00029 + 0,05743\,\theta - 0,00005\,\theta^2 - 0,00281\,\theta^3$	0,34
0,30	$-0,00316 + 0,04492\,\theta + 0,00101\,\theta^2 - 0,00183\,\theta^3$	0,33
0,20	$-0,00163 + 0,02037\,\theta - 0,00022\,\theta^2 - 0,00073\,\theta^3$	0,23

À partir de ces observations et des résultats présentés au tableau 8.7, nous estimons que la distribution d'échantillonnage de l'estimateur du niveau d'habileté est très peu affectée par la variation de son asymétrie. En ce sens, puisque la valeur de l'asymétrie ne devient jamais assez importante pour affecter substantiellement la distribution d'échantillonnage de l'estimateur du niveau d'habileté, l'estimateur du niveau d'habileté peut être considéré comme un estimateur du niveau d'habileté tiré d'une distribution normale $N(EAP(\theta), S_{EAP(\theta)})$.

8.2.5 Caractéristiques de la distribution de la kurtose de l'estimateur du niveau d'habileté en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

Nous abordons maintenant la kurtose, $a4_{EAP(\theta)}$, de la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Les caractéristiques de sa distribution de probabilité peuvent être observées au tableau 8.9 et à la figure 8.10.

Nous notons que la moyenne de la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté est constamment supérieure à 0,00 sur toute l'étendue des valeurs étudiées de l'erreur type retenue pour la règle d'arrêt, soit une kurtose qui se situe entre 0,05 et 0,22. La distribution d'échantillonnage de l'estimateur du niveau d'habileté tend donc à être leptokurtique. De plus, la kurtose de cette distribution d'échantillonnage diminue constamment avec la réduction de l'erreur type retenue pour la règle d'arrêt ; on peut penser qu'elle pourrait tendre éventuellement vers 0,00 avec une valeur encore plus petite de l'erreur type retenue.

L'écart type de la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté est presque nul, indépendamment de l'erreur type retenue pour la règle d'arrêt. L'asymétrie et la kurtose de la distribution de probabilité de la kurtose de l'estimateur du niveau d'habileté affichent toutefois des valeurs très importantes. La plupart du temps l'asymétrie est fortement négative et atteint jusqu'à -9,46 ; la médiane est alors supérieure à la moyenne.

Tableau 8.9
Caractéristiques de la distribution de probabilité de la kurtose de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

ERREUR TYPE	MOYENNE	ÉCART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
0,85	0,13	0,00	nd.*	nd.	0,13	0,13
0,80	0,19	0,03	0,02	-2,00	0,16	0,22
0,75	0,19	0,03	0,02	-2,00	0,16	0,22
0,70	0,20	0,02	-1,02	-0,85	0,17	0,22
0,65	0,22	0,03	-0,99	-0,10	0,15	0,24
0,60	0,22	0,03	-1,54	2,68	0,12	0,25
0,55	0,21	0,04	-3,48	19,47	-0,01	0,26
0,50	0,20	0,03	-5,95	67,84	-0,20	0,26
0,45	0,18	0,03	-6,97	119,64	-0,37	0,24
0,40	0,16	0,03	-9,46	172,04	-0,39	0,23
0,35	0,13	0,02	-9,44	194,42	-0,28	0,19
0,30	0,11	0,01	-7,35	133,58	-0,15	0,15
0,25	0,08	0,01	-4,15	95,81	-0,06	0,11
0,20	0,05	0,00	0,41	1,63	0,04	0,07

* Valeur non disponible car indéterminée (division par zéro)

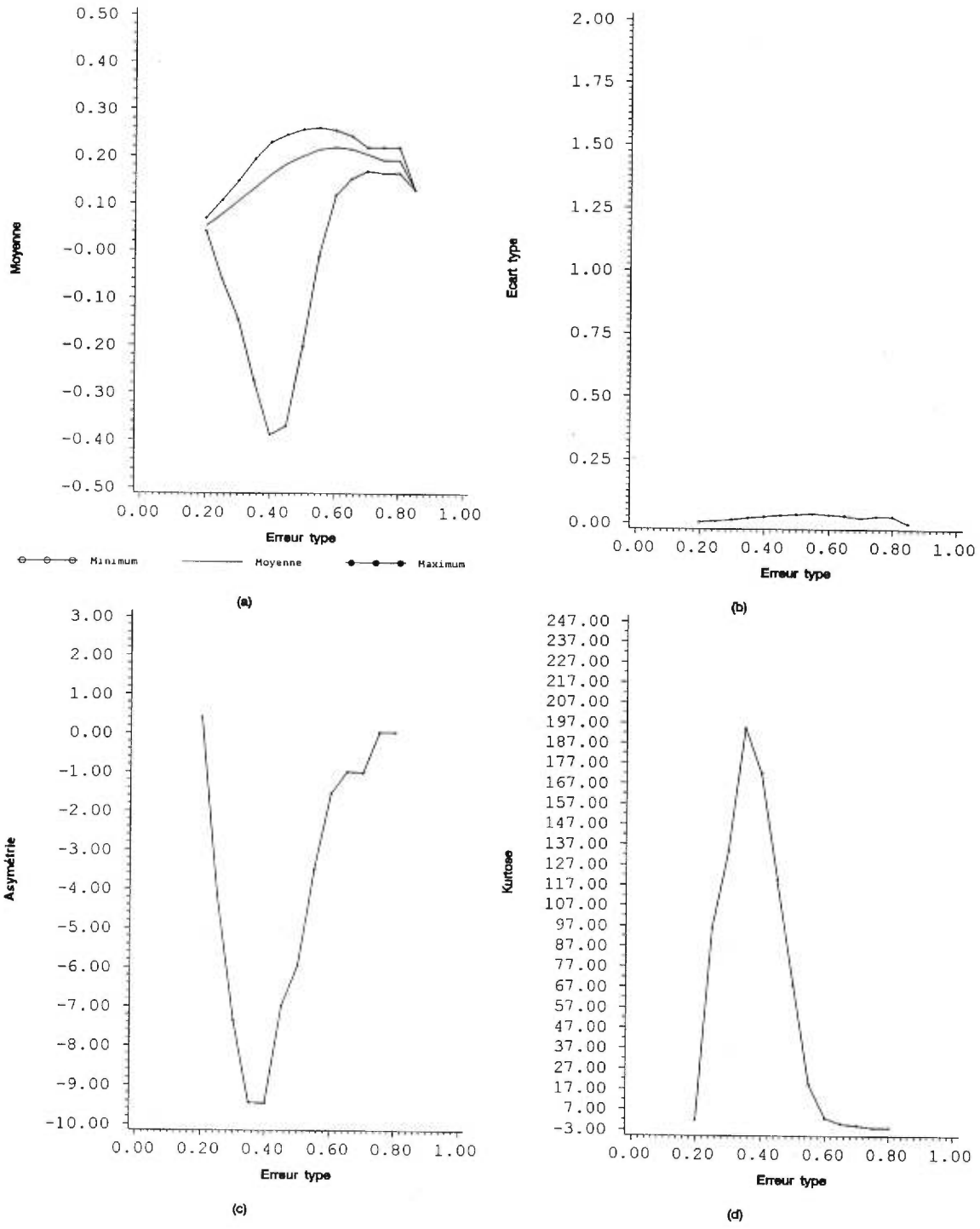

Figure 8.10 Caractéristiques de la distribution de probabilité de la kurtose de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

La kurtose affiche presque toujours des valeurs positives très importantes, jusqu'à 194,42, ce qui indique que la distribution de probabilité de la kurtose est très leptokurtique et que l'intervalle de confiance autour de la moyenne est alors sérieusement surestimé par l'écart type, lui-même déjà peu important. Quand l'erreur type retenue pour la règle d'arrêt est de 0,20, l'asymétrie et la kurtose affichent cependant des valeurs beaucoup plus petites, soit 0,41 et 1,63 respectivement. Est-ce que les valeurs de l'asymétrie et de la kurtose pourraient alors diminuer de façon importante avec une réduction plus substantielle de l'erreur type retenue ? Des simulations avec des valeurs plus petites de l'erreur type retenue seraient nécessaires pour répondre à cette interrogation.

Les minimums et maximums de la kurtose de l'estimateur du niveau d'habileté varient entre -0,39 et 0,26. La kurtose de cette distribution d'échantillonnage affecte donc légèrement l'interprétation de l'estimateur du niveau d'habileté. À titre d'illustration, selon la figure 7.6 et l'équation 7.13 (section 7.4), quand la kurtose d'une distribution de probabilité est égale à -0,39 et que son asymétrie est de 0,29, l'intervalle de confiance à 68,27 % augmente de 5 %. Nous retrouvons justement cette situation lorsque l'erreur type retenue est égale à 0,40 (tableaux 8.7 et 8.9) ; cette augmentation de l'intervalle de confiance à 68,27 % correspond alors à une augmentation de 5 % de l'erreur type retenue de 0,40, soit une augmentation peu importante de seulement 0,02. Lorsque l'erreur type retenue pour la règle d'arrêt est de 0,20, la kurtose est au maximum de 0,07 tandis que la valeur la plus extrême de l'asymétrie est de -0,12 (tableau 8.7). Ces valeurs

correspondent à une augmentation de 1 % de l'intervalle de confiance, soit seulement 0,002. Lorsque l'erreur type retenue est peu importante, la kurtose de la distribution d'échantillonnage affecte donc très peu les interprétaions associées à l'estimateur du niveau d'habileté.

Le tableau 8.10 et la figure 8.11 permettent d'étudier la relation entre le niveau d'habileté et la kurtose de l'estimateur du niveau d'habileté. Le tableau 8.10 donne les coefficients de régression et de détermination de la modélisation de cette relation selon quatre valeurs de l'erreur type retenue pour la règle d'arrêt.

Tableau 8.10
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et la kurtose de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, lorsque la règle d'arrêt est basée sur une valeur de l'erreur type de 0,40, 0,35, 0,30 et 0,20

Règle d'arrêt selon l'erreur type	Modèle $(b_0+b_1\theta+b_2\theta^2+b_3\theta^3)$	Coefficient de détermination (R^2)
0,40	0,16646 - 0,00163 θ - 0,00660 θ^2 + 0,00042 θ^3	0,13
0,35	0,13662 - 0,00185 θ - 0,00377 θ^2 + 0,00086 θ^3	0,08
0,30	0,10733 - 0,00060 θ - 0,00225 θ^2 + 0,00035 θ^3	0,07
0,20	0,05241 + 0,00011 θ - 0,00036 θ^2 - 0,00005 θ^3	0,03

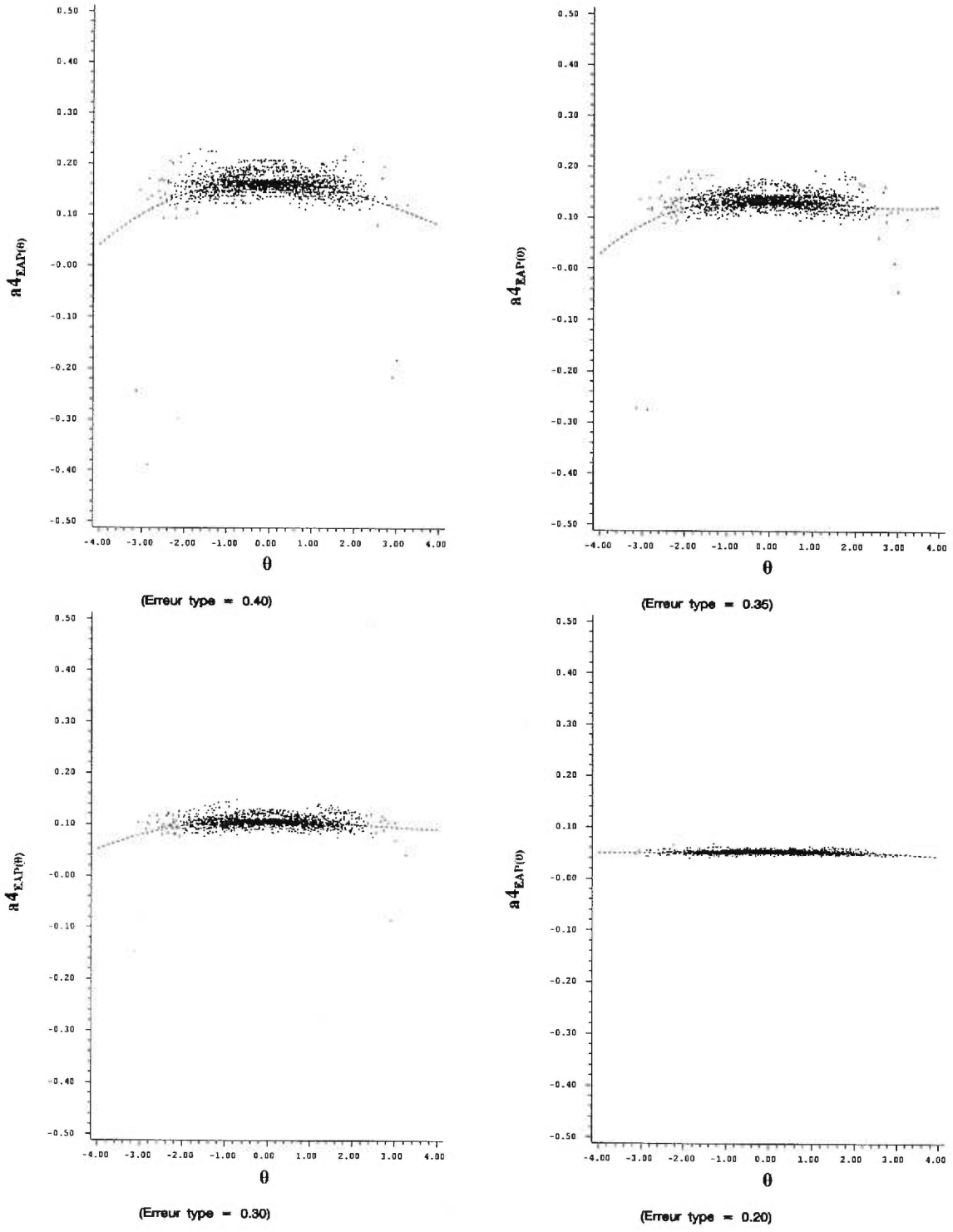

Figure 8.11 Kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, en fonction du niveau d'habileté lorsque la règle d'arrêt selon l'erreur type est égale à 0,20, 0,30, 0,35 et 0,40

Nous notons que la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté est à toutes fins utiles constante sur toute l'étendue du niveau d'habileté. Cette constatation est de plus en plus vraie avec la réduction de l'erreur type retenue pour la règle d'arrêt. Les coefficients de détermination sont d'ailleurs toujours peu importants, soit entre 0,03 et 0,13. Il nous semble toutefois nécessaire de souligner que, selon les modélisations de la relation entre la kurtose et le niveau d'habileté, la moyenne de la kurtose est légèrement plus élevée lorsque le niveau d'habileté est autour de 0,00 tandis qu'elle afiche les valeurs moyennes les plus basses aux valeurs extrêmes du niveau d'habileté. Dans tous les cas, la valeur calculée de la kurtose en fonction du niveau d'habileté est positive, ce qui dénote une distribution d'échantillonnage de l'estimateur du niveau d'habileté à tendance leptokurtique.

Nous en concluons que la kurtose affecte peu la distribution d'échantillonnage de l'estimateur du niveau d'habileté lorsque les valeurs de l'erreur type retenue pour la règle d'arrêt utilisées dans la pratique sont considérées.

8.2.6 Caractéristiques de la distribution de probabilité de la proportion de bonnes réponses associée à l'estimateur du niveau d'habileté en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

Le tableau 8.11 et la figure 8.12 révèlent les caractéristiques de la distribution de probabilité de la proportion de bonnes réponses associée à l'estimateur du niveau d'habileté en fonction de différentes valeurs de l'erreur type retenue pour la règle d'arrêt. Nous y notons que la moyenne de la proportion de bonnes réponses est à toutes fins utiles constamment égale à 0,50, et ce indépendamment de la valeur de l'erreur type retenue pour la règle d'arrêt. Pour sa part, l'écart type de la proportion de bonnes réponses diminue avec la réduction de l'erreur type retenue pour la règle d'arrêt. Cet écart type est approximativement égal à l'erreur type d'une proportion sur toute l'étendue des valeurs de l'erreur type retenue pour la règle d'arrêt, soit :

$$\sqrt{\frac{pq}{n}} \qquad \textbf{8.3}$$

où p est la proportion de bonnes réponses, q est égal à (1 - p) et n est le nombre maximal d'items administrés. À titre indicatif, le nombre maximal d'items administrés à chacune des valeurs de l'erreur type retenue pour la règle d'arrêt est présenté à la section suivante au tableau 8.13. Enfin, à la figure 8.12, nous notons aussi que, lorsque l'erreur type retenue pour la règle d'arrêt est inférieure à 0,70, la relation entre la proportion de bonnes réponses et l'erreur type retenue pour la règle d'arrêt est à toutes fins utiles linéaire.

Tableau 8.11
Caractéristiques de la distribution de probabilité de la proportion de bonnes réponses en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

ERREUR TYPE	MOYENNE	ÉCART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
0,85	0,49	0,50	0,02	-2,00	0,00	1,00
0,80	0,50	0,36	-0,01	-1,02	0,00	1,00
0,75	0,50	0,36	-0,01	-1,02	0,00	1,00
0,70	0,51	0,27	0,06	0,19	0,00	1,00
0,65	0,51	0,25	0,06	-0,54	0,00	1,00
0,60	0,50	0,22	0,03	-0,63	0,00	1,00
0,55	0,50	0,20	0,01	-0,47	0,00	1,00
0,50	0,50	0,18	-0,02	-0,49	0,00	1,00
0,45	0,50	0,16	-0,05	-0,49	0,00	0,91
0,40	0,50	0,14	-0,06	-0,54	0,00	0,92
0,35	0,50	0,13	0,00	-0,50	0,12	0,82
0,30	0,50	0,11	0,05	-0,50	0,19	0,82
0,25	0,50	0,09	-0,03	-0,42	0,21	0,76
0,20	0,50	0,08	0,01	-0,33	0,27	0,72

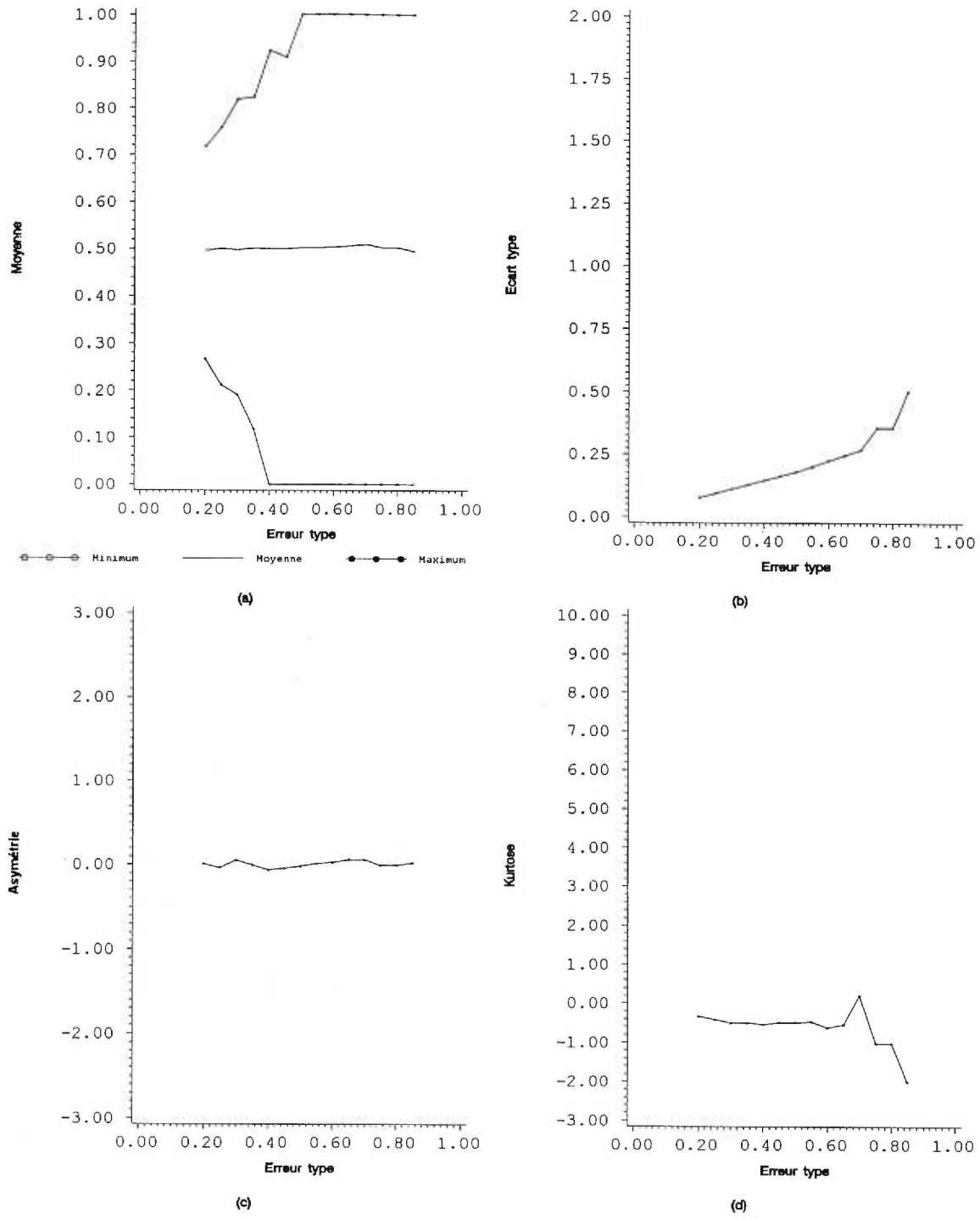

Figure 8.12 Caractéristiques de la distribution de probabilité de la proportion de bonnes réponses en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

Nous pouvons aussi constater que l'asymétrie de la distribution de probabilité de la proportion de bonnes réponses est à peu près nulle, tandis que la kurtose de cette distribution affiche des valeurs presque toujours négatives et d'importance moyenne lorsque l'erreur type retenue est inférieure à 0,75. La distribution de probabilité de la proportion de bonnes réponses est donc plutôt platykurtique et sa kurtose a tendance à diminuer avec la réduction de l'erreur type retenue pour la règle d'arrêt. Il faut tout de même noter que la kurtose est de 0,19 lorsque l'erreur type retenue est de 0,70. À ce moment, la proportion de bonnes réponses est égale à 0,00, 0,33, 0,50, 0,67 et 1,00 dans respectivement 12 %, 13 %, 49 %, 11 % et 14 % des cas ; il y a donc prépondérance des observations autour de 0,50, ce qui explique cette kurtose plus élevée.

Pour ce qui est des valeurs extrêmes observées de la proportion de bonnes réponses, elles tendent à se rapprocher de la moyenne, soit 0,50, avec la diminution de la valeur de l'erreur type retenue pour la règle d'arrêt. Cette dernière observation conjuguée à l'importance de l'écart type de la proportion de bonnes réponses indique que, dans un test adaptatif, la proportion de bonnes réponses peut fréquemment s'écarter de 0,50. Pour obtenir une proportion de bonnes réponses constamment égale à 0,50, la valeur de l'erreur type retenue pour la règle d'arrêt devrait être très petite, à condition d'ailleurs que cette situation soit asymptomatiquement possible. Pour le vérifier, des analyses devraient être réalisées à partir de valeurs beaucoup plus petites de l'erreur type retenue pour la règle d'arrêt, des valeurs toutefois non appliquées dans la pratique de l'administration des tests adaptatifs.

Nous étudions maintenant la relation entre le niveau d'habileté et la proportion de bonnes réponses selon différentes valeurs de l'erreur type retenue pour la règle d'arrêt. La figure 8.13 et le tableau 8.12 démontrent que la proportion moyenne de bonnes réponses varie en fonction du niveau d'habileté. L'importance du coefficient de détermination diminue toutefois avec la réduction de l'erreur type retenue pour la règle d'arrêt. Il est de 0,65 lorsque la règle d'arrêt retenue est de 0,40 et de 0,47 lorsqu'elle est de 0,20. Cette relation est à toutes fins utiles linéaire pour toutes les valeurs de l'erreur type retenue et la proportion de bonnes réponses augmente avec l'accroissement du niveau d'habileté ; il s'agit là d'un résultat prévisible. La pente de la courbe de régression diminue toutefois avec la réduction de l'erreur type retenue pour la règle d'arrêt, caractéristique reliée au fait, remarqué plus tôt, que la proportion de bonnes réponses tend à se rapprocher de 0,50 avec la diminution de l'erreur type retenue.

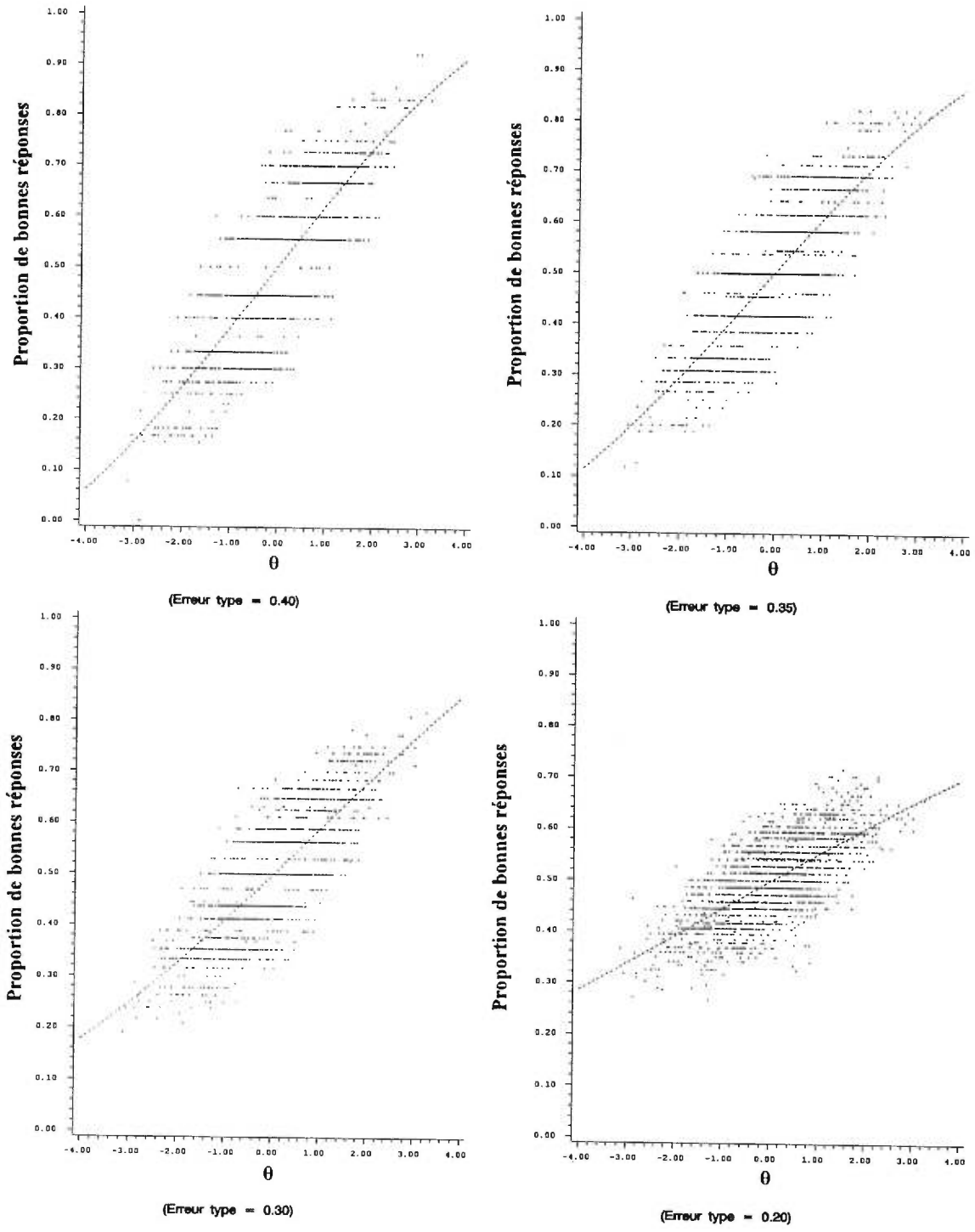

Figure 8.13 Proportion de bonnes réponses en fonction du niveau d'habileté lorsque la règle d'arrêt selon l'erreur type est égale à 0,20, 0,30, 0,35 et 0,40

Tableau 8.12
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et la proportion de bonnes réponses lorsque la règle d'arrêt est basée sur une valeur de l'erreur type de 0,40, 0,35, 0,30 et 0,20

Règle d'arrêt selon l'erreur type	Modèle $(b_0+b_1\theta+b_2\theta^2+b_3\theta^3)$	Coefficient de détermination (R^2)
0,40	$0,50053 + 0,12040\,\theta - 0,00084\,\theta^2 - 0,00087\,\theta^3$	0,65
0,35	$0,50114 + 0,10547\,\theta - 0,00073\,\theta^2 - 0,00074\,\theta^3$	0,64
0,30	$0,49637 + 0,08924\,\theta + 0,00103\,\theta^2 - 0,00034\,\theta^3$	0,62
0,20	$0,49711 + 0,05335\,\theta - 0,00038\,\theta^2 - 0,00012\,\theta^3$	0,47

À titre d'illustration, lorsque la valeur de l'erreur type retenue pour la règle d'arrêt est égale à 0,40 et que le niveau d'habileté est de -3,00, -2,00, 2,00 et 3,00, la proportion de bonnes réponses calculée à partir du modèle de régression est respectivement de 0,16, 0,26,0,73 et 0,83. Lorsque la valeur de l'erreur type retenue pour la règle d'arrêt est de 0,20, la proportion de bonnes réponses calculée est alors de 0,34, 0,39, 0,60 et 0,65. Les valeurs extrêmes passent donc de 0,16 et 0,83 à 0,34 et 0,65 et tendent donc à se rapprocher de 0,50.

Cette dernière constatation nous semble importante puisque, en testing adaptatif, on peut avoir l'impression que la proportion de bonnes réponses devrait être toujours égale à 0,50. Certains auteurs, de par leur manière de présenter ce qu'est un test adaptatif, nous le laissent croire. Par exemple, Weiss affirme que :

> *«Le résultat d'une bonne procédure adaptative de testing consiste en une banque d'items sélectionnés pour chaque examiné qui ont une probabilité de 0,50 d'obtenir une bonne réponse de cet individu étant entendu qu'il n'y a pas de pseudo-chance* (1985, p. 776).»[1]

De leur côté Kingsbury et Weiss soulignent que :

> *«En utilisant la théorie de la réponse à l'item, il est possible de reproduire une forme de test adaptatif analogue à un test de maîtrise basé sur le pourcentage de bonnes réponses de la théorie classique des test, alors même que l'utilisation de la stratégie de maximisation de l'information tend à ce que chaque personne obtienne 50 % de bonnes réponses étant donné une banque d'items suffisamment importante (car les items administrés auront fort probablement un niveau de difficulté à peu près égal au niveau d'habileté de la personne). (Kinsgbury et Weiss, 1983, p. 262).»*[2]

---

1 *«The result of a good adaptive testing procedure is a set of items selected for each examinee that have 0,50 probability of a correct response (assuming no guessing) for that individual.»*

2 *«Using IRT, it is possible to generate an analog to the percentage-correct mastery level of classical theory for use in adaptive testing, even though the use of MISS (maximum information search and selection) will tend to result in each person answering about 50 % of the items correctly given a large enough item pool (because items administered will most probably have difficulty levels very close to the individual's level of θ).»*

Nos résultats indiquent que la probabilité de bonnes réponses peut s'éloigner considérablement de 0,50, même lorsque l'erreur type retenue pour la règle d'arrêt est égale à 0,20, donc assez peu importante.

8.2.7 Caractéristiques de la distribution de probabilité du nombre d'items administrés associé à l'estimateur du niveau d'habileté en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

Nous abordons maintenant l'étude de la distribution de probabilité du nombre d'items administrés. Le tableau 8.13 et la figure 8.14 livrent plusieurs informations au sujet de cette distribution. En premier lieu, caractéristique prévisible, le nombre d'items administrés augmente avec la réduction de la valeur de l'erreur type retenue pour la règle d'arrêt. Cette progression, non linéaire, est de plus en plus rapide, principalement lorsque l'erreur type est égale ou inférieure à 0,30. Ainsi, nous observons que, lorsque l'erreur type retenue pour la règle d'arrêt est égale à 0,30, 0,25 et 0,20, le nombre moyen d'items administrés correspond respectivement à approximativement 17, 24 et 37 items.

Pour sa part, l'écart type du nombre d'items administrés augmente constamment avec la diminution de l'erreur type retenue pour la règle d'arrêt. À titre d'illustration, lorsque l'erreur type retenue pour la règle d'arrêt est égale à 0,20, l'écart type du nombre d'items administrés est de 1,58. Dans 95 % des cas, le nombre d'items administrés serait donc théoriquement compris à l'intérieur de l'intervalle de confiance qui s'étend de 33,53 à 39,73 (soit $36,63 \pm 1,96 * 1,58$). Dans 99 % des cas, toujours théoriquement, l'intervalle de confiance du nombre d'items administrés serait compris entre 32,55 et 40,71 (soit $36,63 \pm 2,58 * 1,58$). La relation entre l'écart type du nombre d'items administrés et l'erreur type retenue pour la règle d'arrêt est de plus presque linéaire (figure 8.14).

Tableau 8.13
Caractéristiques de la distribution de probabilité du nombre d'items administrés en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

ERREUR TYPE	MOYENNE	ÉCART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
0,85	1,00	0,00	nd.*	nd.	1,00	1,00
0,80	2,00	0,00	nd.	nd.	2,00	2,00
0,75	2,00	0,00	nd.	nd.	2,00	2,00
0,70	2,51	0,50	-0,02	-2,00	2,00	3,00
0,65	3,26	0,44	1,10	-0,78	3,00	4,00
0,60	4,13	0,33	2,24	3,00	4,00	5,00
0,55	4,84	0,62	0,95	3,11	4,00	7,00
0,50	6,18	0,60	1,30	3,86	5,00	9,00
0,45	7,50	0,79	1,92	4,37	7,00	13,00
0,40	9,53	0,86	1,95	4,19	9,00	14,00
0,35	12,47	0,92	1,88	4,51	11,00	17,00
0,30	16,76	1,08	1,93	5,11	16,00	24,00
0,25	23,85	1,28	2,00	6,17	22,00	33,00
0,20	36,63	1,58	1,93	5,89	35,00	48,00

* Valeur non disponible car indéterminée (division par zéro)

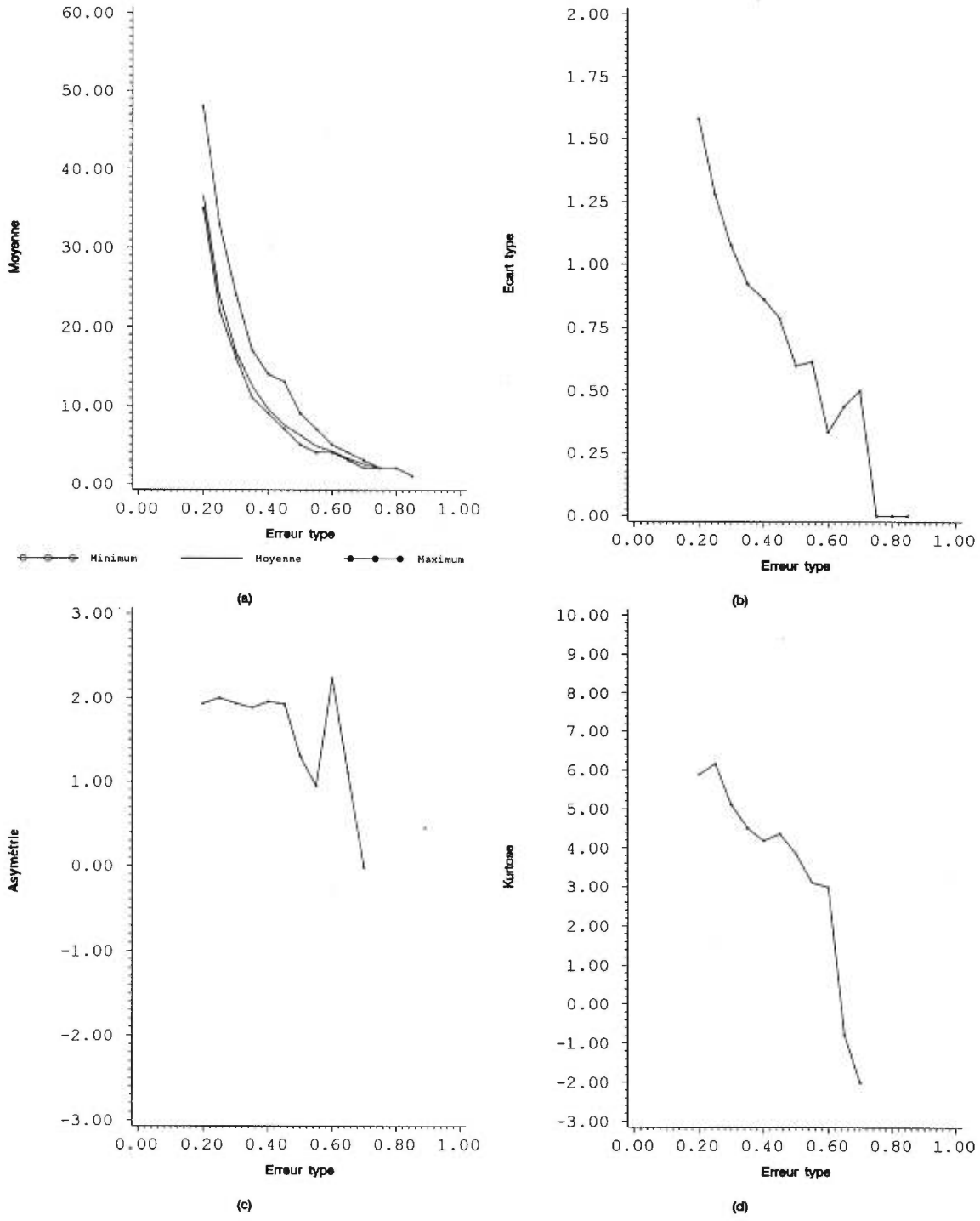

Figure 8.14 Caractéristiques de la distribution de probabilité du nombre d'items administrés en fonction de la valeur de l'erreur type retenue pour la règle d'arrêt

Nous observons toutefois que l'asymétrie de la distribution de probabilité du nombre d'items administrés, sauf lorsque l'erreur type retenue est égale à 0,70, prend toujours une valeur positive, la plupart du temps très importante. Cette valeur tend à se stabiliser autour de 2,00 à partir d'une erreur type retenue pour la règle d'arrêt de 0,45. Cela nous informe que la médiane de la distribution est constamment inférieure à la moyenne de cette distribution. La moyenne du nombre d'items administrés surestime ainsi le nombre médian d'items, soit le nombre d'items qui sont administrés dans 50 % des cas.

On note aussi que la kurtose de la distribution de probabilité affiche constamment une valeur positive très importante, sauf lorsque l'erreur type retenue est égale ou supérieure à 0,65, ce qui indique que la distribution de probabilité est très fortement leptokurtique. L'écart type du nombre d'items administrés est donc une mesure qui surestime l'intervalle de confiance autour de la moyenne. De plus, la kurtose de la distribution de probabilité du nombre d'items administrés augmente avec la diminution de l'erreur type retenue pour la règle d'arrêt ; la relation entre la kurtose et l'erreur type retenue est presque linéaire.

En ce qui concerne les minimums et maximums du nombre d'items administrés, tout comme l'intervalle entre eux, ils augmentent constamment avec la réduction de l'erreur type retenue pour la règle d'arrêt. À partir de nos observations au sujet de l'asymétrie et de la kurtose, il est évident que la distribution de probabilité du nombre d'items administrés ne se comporte pas selon les caractéristiques d'une loi normale.

Il nous semble important de souligner que les caractéristiques observées de la distribution de probabilité du nombre d'items administrés correspondent à celles présentées par Bock et Mislevy (1982, p. 438). En effet, ces deux auteurs remarquent que, lorsqu'une modélisation de la réponse à l'item à partir du modèle logistique à deux paramètres est appliquée et que l'erreur type retenue pour la règle d'arrêt est de 0,40, 0,30 et 0,20, le nombre moyen d'items administrés est respectivement de 9,17 et 37. Nous obtenons ici les mêmes résultats à partir d'une modélisation logistique à un paramètre, soit 9,53, 16,77 et 36,63. Bock et Mislevy, par leurs résultats concernant la fréquence observée du nombre d'items administrés, laissent aussi supposer que l'asymétrie de la distribution de probabilité du nombre d'items administrés tend à être fortement positive lorsque la modélisation à trois paramètres est utilisée, ce que nous observons à partir de nos résultats basés sur une modélisation logistique à un paramètre.

Le tableau 8.14 présente les coefficients de régression et de détermination qui caractérisent la relation entre le nombre d'items administrés et le niveau d'habileté en fonction de différentes valeurs de l'erreur type retenue pour la règle d'arrêt. À la figure 8.15, nous pouvons observer la représentation graphique des corrélogrammes correspondants.

Le coefficient de détermination varie entre 0,29 et 0,37 et la relation observée est clairement non linéaire. Plus on s'éloigne du niveau d'habileté moyen, soit un niveau d'habileté de 0,01, plus le nombre d'items nécessaires pour obtenir la valeur attendue de

l'erreur type retenue pour la règle d'arrêt est élevé. Ce résultat était prévisible.

La figure 8.15 permet aussi de constater que la variation du nombre d'items administrés est moins importante autour de la moyenne du niveau d'habileté. À titre d'exemple, l'écart type du nombre d'items administrés, lorsque l'erreur type retenue est de 0,20 et que le niveau d'habileté est égal à 0,00 $\pm$0,50, -1,00 $\pm$0,50, et -2,00 $\pm$0,50 est respectivement de 1,16, 1,42 et 1,83. Quand l'erreur type retenue est égale à 0,40 l'écart type est de 0,37, 0,84 et 1,13. Il est donc évident que plus on s'éloigne de la moyenne du niveau d'habileté, plus le nombre d'items administrés varie. Ces constatations sont valides à toutes les valeurs analysées de l'erreur type retenue pour la règle d'arrêt.

Tableau 8.14
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et le nombre d'items administrés lorsque la règle d'arrêt est basée sur une valeur de l'erreur type de 0,40, 0,35, 0,30 et 0,20

Règle d'arrêt selon l'erreur type	Modèle $(b_0+b_1\theta+b_2\theta^2+b_3\theta^3)$	Coefficient de détermination (R^2)
0,40	9,18214 - 0,03173 θ + 0,35772 θ^2 - 0,00121 θ^3	0,32
0,35	12,09788 - 0,01264 θ + 0,39186 θ^2 - 0,00024 θ^3	0,34
0,30	16,30019 + 0,00008 θ + 0,48180 θ^2 - 0,00270 θ^3	0,37
0,20	36,03057 + 0,05924 θ + 0,62047 θ^2 - 0,03751 θ^3	0,29

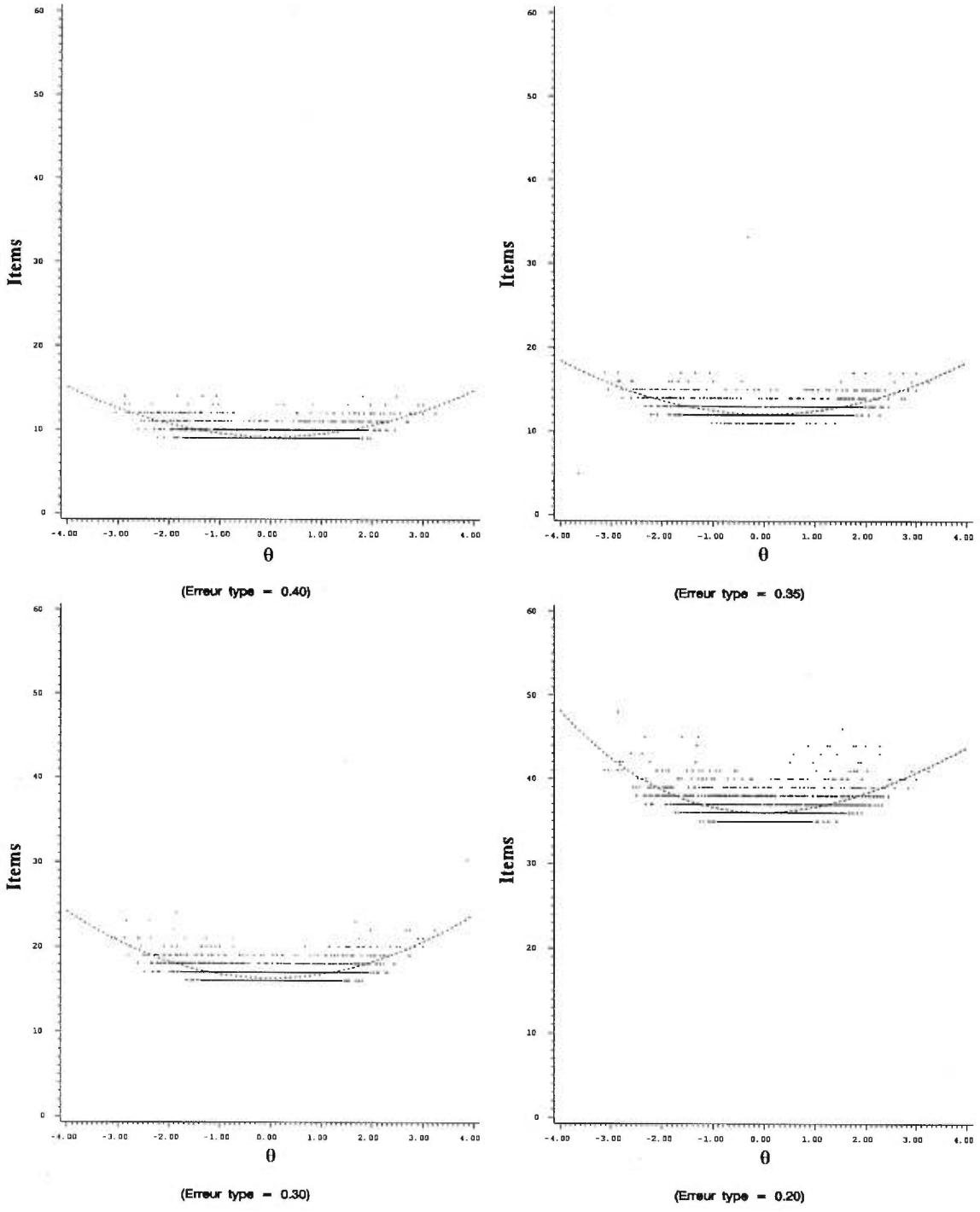

Figure 8.15 Nombre d'items administrés en fonction du niveau d'habileté lorsque la règle d'arrêt selon l'erreur type est égale à 0,20, 0,30, 0,35 et 0,40

8.3 Règle d'arrêt selon le nombre d'items administrés

Pour présenter les résultats en fonction des diverses valeurs du nombre d'items administrés retenues pour la règle d'arrêt, soit entre 1 et 60, nous avons analysé l'estimateur du niveau d'habileté, EAP(θ), l'erreur type, $S_{EAP(\theta)}$, le biais, $BIAIS_{EAP(\theta)}$, l'asymétrie, $a3_{EAP(\theta)}$, la kurtose, $a4_{EAP(\theta)}$, ainsi que la proportion de bonnes réponses associés à la distribution d'échantillonnage de l'estimateur du niveau d'habileté.

Notons que plusieurs de nos observations et remarques quant à la règle d'arrêt selon le nombre d'items administrés reflètent exactement ce qui a été souligné dans notre analyse précédente de la règle d'arrêt selon l'erreur type. C'est pourquoi certaines de nos constatations reçoivent des explications un peu moins élaborées.

8.3.1 Caractéristiques de la distribution de probabilité de l'estimateur du niveau d'habileté en fonction du nombre d'items administrés retenu pour la règle d'arrêt

En fonction de chacune des valeurs retenues de la règle d'arrêt selon le nombre d'items administrés, les caractéristiques de la distribution de probabilité de l'estimateur du niveau d'habileté, EAP(θ), sont présentées au tableau 8.15 et illustrées à la figure 8.16. À la figure 8.16, au quadrant (a), on retrouve les moyennes, minimums et maximums affichés au tableau 8.15 en fonction de chacune des valeurs retenues pour la règle d'arrêt selon le nombre d'items administrés. Aux quadrants (b), (c) et (d) on retrouve l'écart type,

l'asymétrie et la kurtose de la distribution de probabilité de l'estimateur du niveau d'habileté.

Nous observons au tableau 8.15 et à la figure 8.16 que la moyenne de l'estimateur du niveau d'habileté affiche toujours des valeurs rapprochées de la moyenne de la distribution du niveau d'habileté, soit 0,01 : elle est d'au plus 0,02 et jamais inférieure à -0,01. L'écart type de l'estimateur du niveau d'habileté, pour sa part, augmente constamment avec l'augmentation du nombre d'items administrés. Il s'éloigne sensiblement de la valeur de l'écart type de la distribution de probabilité du niveau d'habileté quant le nombre d'items administrés est peu important. Lorsque 60 items ont été administrés, il atteint 0,97, ce qui nous permet de croire qu'il tend vers la valeur de l'écart type de la distribution de probabilité de l'estimateur du niveau d'habileté, soit 0,98, quand le nombre d'items administrés devient plus important.

L'asymétrie de la distribution de probabilité de l'estimateur du niveau d'habileté est peu importante quelle que soit le nombre d'items administrés. Elle est toujours positive et atteint tout au plus 0,07 lorsque le nombre d'items administrés est égal à 5 et 6.

Tableau 8.15
Caractéristiques de la distribution de probabilité de l'estimateur du niveau d'habileté, EAP(θ), en fonction du nombre d'items administrés retenu pour la règle d'arrêt

ITEM	MOYENNE	ÉCART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
1	-0,01	0,56	0,02	-2,00	-0,56	0,56
2	0,00	0,71	0,00	-1,11	-0,99	0,99
3	0,02	0,79	0,05	-0,78	-1,33	1,33
4	0,02	0,82	0,06	-0,48	-1,63	1,63
5	0,01	0,85	0,07	-0,36	-1,90	1,90
6	0,02	0,87	0,07	-0,19	-2,14	2,14
7	0,01	0,88	0,03	-0,12	-2,36	2,36
8	0,01	0,89	0,02	-0,09	-2,57	2,57
9	0,01	0,91	0,02	-0,04	-2,76	2,76
10	0,01	0,92	0,01	-0,03	-2,94	2,94
11	0,01	0,92	0,00	-0,01	-3,10	2,78
12	0,01	0,92	0,00	-0,06	-3,25	2,92
13	0,00	0,93	0,02	-0,04	-3,38	3,05
14	0,00	0,93	0,04	-0,04	-3,27	3,13
15	0,00	0,94	0,05	0,00	-3,38	3,25
20	0,01	0,95	0,02	0,02	-3,35	3,29
25	0,00	0,96	0,03	0,00	-3,16	3,35
30	0,00	0,96	0,02	-0,04	-3,10	3,29
40	0,00	0,96	0,00	-0,04	-3,13	3,20
50	0,00	0,97	0,00	0,01	-3,31	3,13
60	-0,01	0,97	0,01	0,01	-3,40	3,16

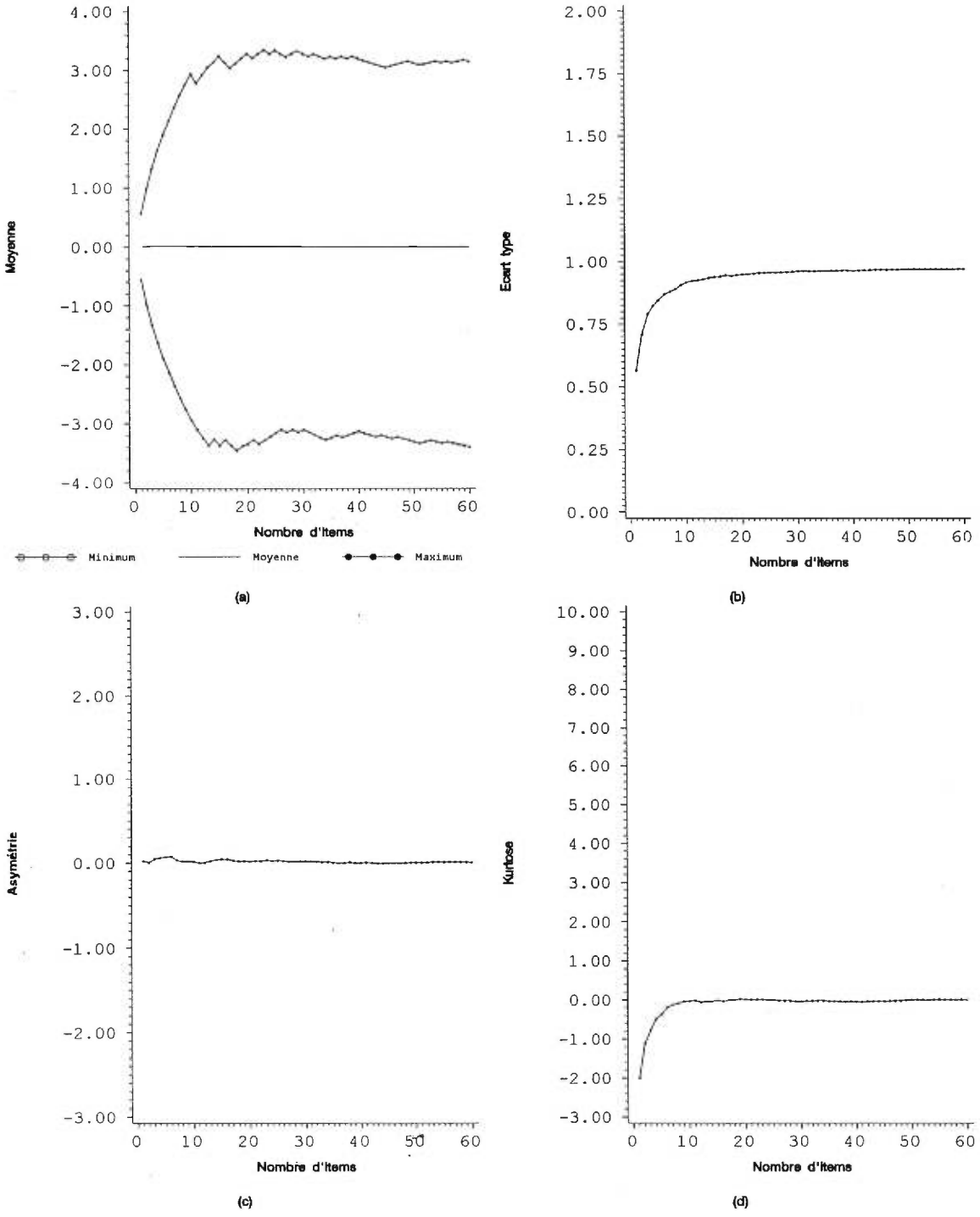

Figure 8.16 Caractéristiques de la distribution de probabilité de l'estimateur du niveau d'habileté, EAP(θ), en fonction de la règle d'arrêt basée sur le nombre d'items administrés

Remarquons que la kurtose de la distribution de probabilité de l'estimateur du niveau d'habileté présente des valeurs peu importantes lorsque le nombre d'items administrés est supérieur à 3. En valeur absolue, la kurtose est de moins en moins grande avec l'augmentation du nombre d'items administrés. Déjà, à partir du 8^{e} item, elle est n'est que de -0,09.

Nous observons aussi que plus le nombre d'items administrés est grand, plus les minimums et maximums de l'estimateur du niveau d'habileté sont près des minimums et maximums du niveau d'habileté, -3,13 et 3,24. C'est à partir de l'administration du 13^{e} item, cependant, que les minimums et maximums présentent des valeurs constamment supérieures à 3,00 en valeur absolue.

À partir de ces observations, nous pouvons conclure que la distribution de probabilité de l'estimateur du niveau d'habileté se rapproche de la distribution du niveau d'habileté quand le nombre d'items administrés est égal ou supérieur à 13. Comme nous le remarquerons au tableau 8.17, à la section suivante traitant de l'erreur type de l'estimateur du niveau d'habilité, la moyenne de l'erreur type de l'estimateur du niveau d'habileté est à ce moment inférieure à 0,34.

Une régression cubique est utilisée pour effectuer une modélisation de la relation entre l'estimateur du niveau d'habileté et le niveau d'habileté. Cette modélisation est effectuée selon quatre valeurs du nombre d'items administrés retenu pour la règle d'arrêt : 10, 20,

40 et 60. Ces valeurs permettent d'analyser cette relation lorsque le nombre d'items administrés est très petit (10), moyen (20 et 40) et assez grand (60). La valeur de 10 est retenue à cause de son utilisation dans les travaux de Hoijtink et Boomsma (1995, 1996) et pour nous permettre ainsi de comparer nos résultats aux leurs. Nous présentons les coefficients de régression et de détermination obtenus au tableau 8.16 et les corrélogrammes correspondants à la figure 8.17.

Tableau 8.16
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et l'estimateur du niveau d'habileté, EAP(θ), lorsque la règle d'arrêt selon le nombre d'items administrés est égale à 10, 20, 40 et 60

Règle d'arrêt selon le nombre d'items administrés	Modèle $(b_0+b_1\theta+b_2\theta^2+b_3\theta^3)$	Coefficient de détermination (R^2)
10	$0,00029 + 0,87453\ \theta + 0,00049\ \theta^2 - 0,00523\ \theta^3$	0,85
20	$-0,00221 + 0,92795\ \theta + 0,00370\ \theta^2 - 0,00014\ \theta^3$	0,93
40	$-0,00833 + 0,96412\ \theta + 0,00011\ \theta^2 - 0,00010\ \theta^3$	0,97
60	$-0,01465 + 0,97210\ \theta + 0,00188\ \theta^2 - 0,00116\ \theta^3$	0,98

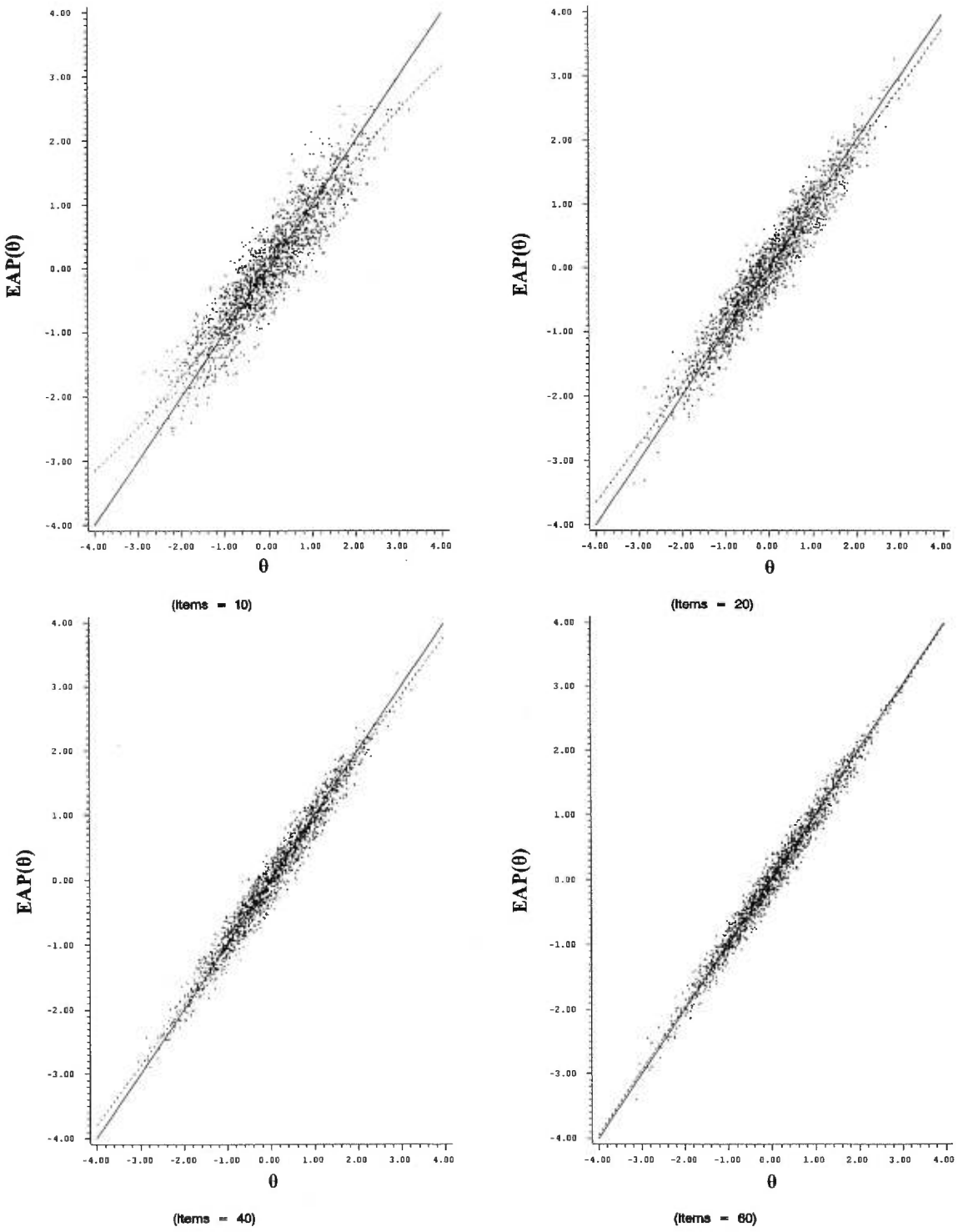

Figure 8.17 Estimateur du niveau d'habileté, EAP(θ), en fonction du niveau d'habileté lorsque la règle d'arrêt selon le nombre d'items administrés est égale à 10, 20, 40 et 60

Les coefficients de détermination sont toujours très élevés. Même après l'administration de 10 items, le coefficient de détermination est de 0,85, et la valeur de ce coefficient devient plus importante avec l'augmentation du nombre d'items administrés, ce qui indique que le niveau d'habileté est de mieux en mieux estimé avec l'augmentation du nombre d'items. Le coefficient de détermination atteint 0,98 avec l'administration de 60 items. La relation est à toutes fins utiles linéaire, les coefficients de régression de deuxième et troisième ordres étant peu importants.

L'estimateur du niveau d'habileté est à peu près égal au niveau d'habileté autour de la moyenne du niveau d'habileté. Toutefois, plus le niveau d'habileté s'éloigne de la moyenne, plus la différence s'accroît entre l'estimateur du niveau d'habileté et le niveau d'habileté. Ainsi, lorsque le niveau d'habileté est inférieur à la moyenne, il est surestimé. À l'inverse, lorsqu'il est supérieur à la moyenne, il est sous-estimé. L'augmentation du nombre d'items réduit cette différence observée aux valeurs extrêmes du niveau d'habileté. Nous abordons plus spécifiquement cette problématique dans une prochaine section (section 8.3.3) qui traite du biais de l'estimateur du niveau d'habileté.

8.3.2 Caractéristiques de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté en fonction du nombre d'items administrés retenu pour la règle d'arrêt

Au tableau 8.17, nous présentons les caractéristiques de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, en fonction de chacune des valeurs retenues pour la règle d'arrêt selon le nombre d'items administrés. La figure 8.18 permet de visualiser au quadrant (a) les moyennes, minimums et maximums affichés au tableau 8.17 en fonction de chacune des valeurs retenues pour la règle d'arrêt selon le nombre d'items administrés. Aux quadrants (b), (c) et (d), on retrouve l'écart type, l'asymétrie et la kurtose de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté.

Résultat prévisible, la valeur de l'erreur type de l'estimateur du niveau d'habileté diminue avec l'augmentation du nombre d'items administrés. Ainsi, à partir de l'administration du 40^{e} item, la moyenne de l'erreur type est inférieure à 0,20. Pour que l'erreur type puisse diminuer de 0,05, le nombre d'items administrés doit atteindre 60. Comme on le voit à la figure 8.18, plus l'erreur type est petite, plus le nombre d'items qui doivent être administrés devient important pour obtenir une diminution, elle-même peu importante, de l'erreur type de l'estimateur du niveau d'habileté.

L'écart type de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté est à toutes fins utiles égal à 0,00, au plus il atteint 0,02. L'erreur type varie donc très peu. L'asymétrie, pour sa part, est constamment positive après l'administration

du 3^{e} item. Elle se situe généralement autour de 2,00 et tend à augmenter légèrement lorsque le nombre d'items administrés est plus important. La médiane se trouve alors, conséquemment, inférieure à la moyenne.

Quant à la kurtose, elle est également constamment positive, mais seulement après l'administration du 4^{e} item. De plus, elle augmente de façon assez importante avec l'augmentation du nombre d'items administrés, pour atteindre une valeur supérieure à 8,52 à partir de l'administration du 30^{e} item. La distribution de probabilité de l'erreur type est donc à la fois très asymétrique et fortement leptokurtique. Notons que ces valeurs observées de l'asymétrie et de la kurtose étaient bien moins importantes lorsque la règle d'arrêt selon l'erreur type était appliquée (tableau 8.3).

Signalons aussi que les minimums et maximums de l'erreur type ne sont jamais égaux, même après l'administration de 60 items. Toutefois, à partir du 25^{e} item, la différence entre le maximum et le minimum n'est plus que de 0,05, différence que nous considérons ici peu importante. À partir du 50^{e} item, cette différence n'est que de 0,02. Cette observation est corollaire au fait que l'écart type de l'erreur type de l'estimateur du niveau d'habileté est très petit et que la kurtose de la distribution de probabilité est très importante.

Tableau 8.17
Caractéristiques de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, en fonction du nombre d'items administrés retenu pour la règle d'arrêt

ITEM	MOYENNE	ECART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
1	0,82	0,00	nd.*	nd.	0,82	0,82
2	0,71	0,02	-0,02	-2,00	0,69	0,73
3	0,63	0,02	1,01	-0,85	0,61	0,67
4	0,57	0,02	1,43	1,08	0,55	0,62
5	0,52	0,02	1,68	2,58	0,50	0,59
6	0,48	0,02	1,85	3,60	0,46	0,56
7	0,45	0,02	1,91	4,21	0,43	0,53
8	0,42	0,02	1,95	4,37	0,40	0,51
9	0,40	0,02	2,05	4,89	0,38	0,49
10	0,38	0,02	2,00	4,80	0,36	0,47
11	0,36	0,02	2,03	5,16	0,35	0,46
12	0,35	0,01	2,12	6,14	0,33	0,45
13	0,34	0,01	2,06	5,86	0,32	0,42
14	0,32	0,01	1,89	4,33	0,31	0,38
15	0,31	0,01	1,96	4,86	0,30	0,38
20	0,27	0,01	1,98	5,95	0,26	0,34
25	0,24	0,01	2,06	6,13	0,23	0,28
30	0,22	0,01	2,26	8,12	0,21	0,26
40	0,19	0,00	2,25	8,32	0,19	0,23
50	0,17	0,00	2,39	8,59	0,17	0,19
60	0,15	0,00	2,30	8,10	0,15	0,17

* Valeur non disponible car indéterminée (division par zéro)

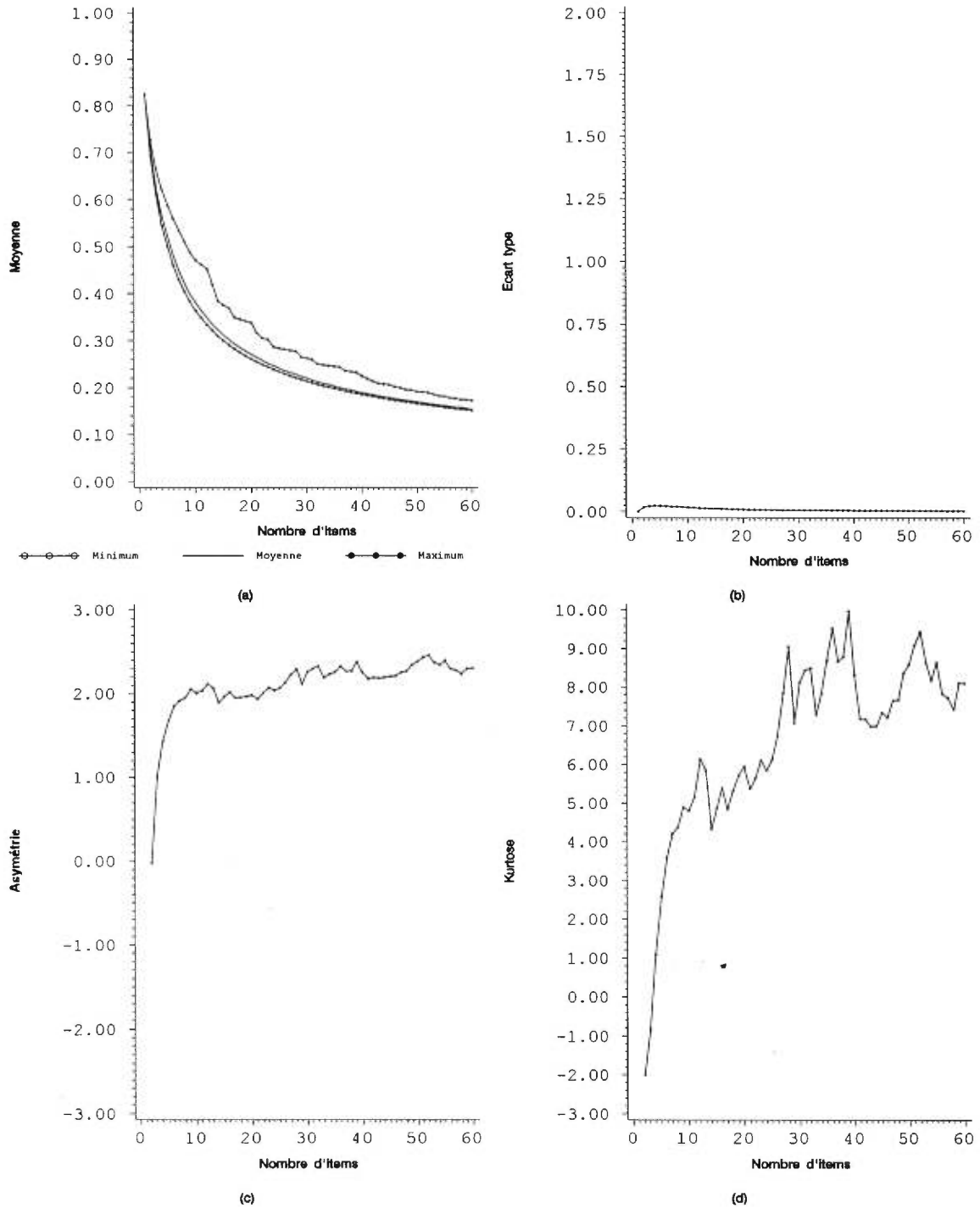

Figure 8.18 Caractéristiques de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, en fonction de la règle d'arrêt basée sur le nombre d'items administrés

La figure 8.19 présente les valeurs obtenues de l'erreur type de l'estimateur du niveau d'habileté en fonction du niveau d'habileté pour quatre valeurs du nombre d'items administrés retenues pour la règle d'arrêt : 10, 20, 40 et 60. Comme nous le soulignons à la section précédente qui traite de l'estimateur du niveau d'habileté, la valeur de 10 est retenue à cause de son utilisation dans les travaux de Hoijtink et Boomsma (1995, 1996). Ceux-ci laissent supposer que l'administration de 10 items pourrait s'avérer suffisante pour obtenir un estimateur du niveau d'habileté non biaisé et dont l'erreur type n'est pas trop importante[1]. En ce qui concerne l'erreur type, nous avons vu au tableau 8.17 et à la figure 8.18 que celle-ci affiche, au contraire, une valeur plutôt importante, soit 0,38, lorsque seulement 10 items sont administrés. Quant au biais de l'estimateur du niveau d'habileté, sujet abordé à l'intérieur de la section suivante, l'hypothèse est à vérifier.

De plus, une régression cubique est utilisée pour modéliser la relation entre l'erreur type et le niveau d'habileté. Les résultats de cette modélisation sont présentés au tableau 8.18. Le coefficient de détermination de la régression cubique diminue avec le nombre d'items administrés, ce qui indique que l'erreur type de l'estimateur du niveau d'habileté est de moins en moins bien prédite par le niveau d'habileté avec l'augmentation du nombre d'items administrés. Le coefficient de détermination est d'importance moyenne lorsque le nombre d'items administrés n'est que de 10, soit de 0,40. Quand le nombre d'items

---

1 *«A general recommendation would be to estimate person parameters only if the number of items is at least ten. For smaller item sets both the biais and the error variance will be very large and we cannot even rely on asymptotic results to estimate these quantities (Hoijtink et Boomsma, 1995, p. 68).»*

administrés atteint 60, le coefficient de détermination devient peu important, il affiche alors une valeur de 0,26.

Tableau 8.18
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et l'erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, lorsque la règle d'arrêt selon le nombre d'items administrés est égale à 10, 20, 40 et 60

Règle d'arrêt selon le nombre d'items administrés	Modèle $(b_0+b_1\theta+b_2\theta^2+b_3\theta^3)$	Coefficient de détermination (R^2)
10	$0,37397 - 0,00031\,\theta + 0,00764\,\theta^2 - 0,00008\,\theta^3$	0,40
20	$0,26682 + 0,00001\,\theta + 0,00382\,\theta^2 + 0,00000\,\theta^3$	0,38
40	$0,18842 + 0,00023\,\theta + 0,00158\,\theta^2 - 0,00012\,\theta^3$	0,31
60	$0,15362 + 0,00010\,\theta + 0,00093\,\theta^2 - 0,00005\,\theta^3$	0,26

Dans tous les cas, si l'on en juge par les coefficients de régression, l'erreur type de l'estimateur du niveau d'habileté augmente lorsque le niveau d'habileté s'éloigne du niveau d'habileté moyen (0,01). C'est donc aux valeurs extrêmes du niveau d'habileté que l'erreur type devient plus importante. Résultat attendu : plus le nombre d'items administrés augmente, moins l'erreur type de l'estimateur du niveau d'habileté est affectée par l'éloignement du niveau d'habileté de sa moyenne (figure 8.19).

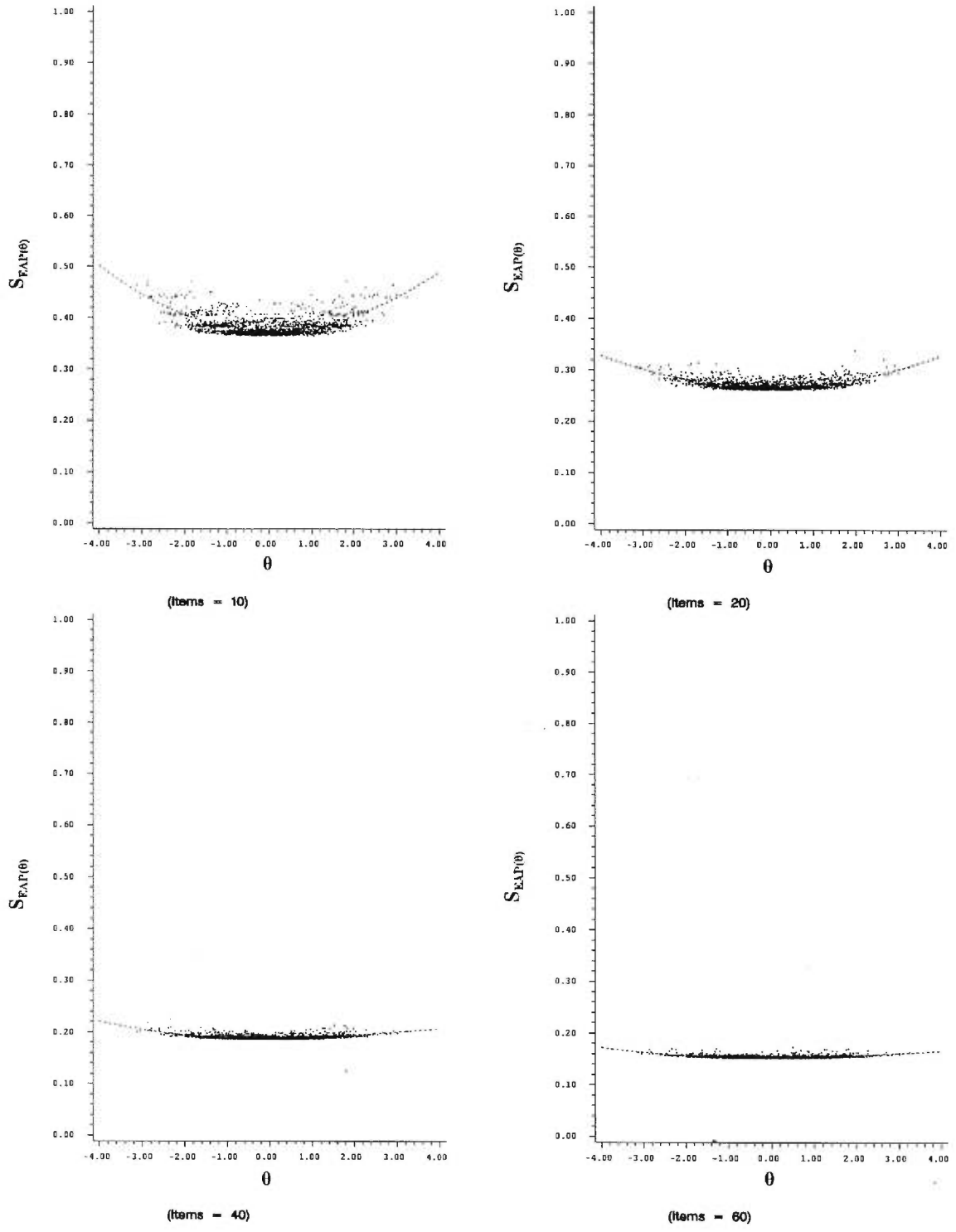

Figure 8.19 Erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, en fonction du niveau d'habileté lorsque la règle d'arrêt selon le nombre d'items administrés est égale à 10, 20, 40 et 60

8.3.3 Caractéristiques de la distribution de probabilité de l'erreur de mesure du niveau d'habileté et du biais de l'estimateur du niveau d'habileté en fonction du nombre d'items administrés retenu pour la règle d'arrêt

Nous présentons au tableau 8.19 et à la figure 8.20 les résultats obtenus en lien avec les caractéristiques de la distribution de probabilité de l'erreur de mesure du niveau d'habileté, (EAP(θ) - θ), en fonction de chacune des valeurs retenues pour la règle d'arrêt. Nous traiterons un peu plus loin, à l'intérieur de cette même section, du biais de l'estimateur du niveau d'habileté, $BIAIS_{EAP(\theta)}$.

L'erreur de mesure du niveau d'habileté est d'au plus 0,01 en valeur absolue, quelle que soit la valeur du nombre d'items administrés retenue pour la règle d'arrêt. À toutes fins utiles, elle est à peu près égale à 0,00. Tout comme pour la règle d'arrêt selon l'erreur type (section 8.2.3), l'écart type de la distribution de probabilité de l'erreur de mesure du niveau d'habileté est presque égal, jamais supérieur, à l'erreur type de l'estimateur du niveau d'habileté (tableau 8.17). La différence entre ces deux valeurs est d'au plus 0,01. Également, il existe une relation curvilinéaire presque parfaite entre le nombre d'items administrés et l'écart type de l'erreur de mesure du niveau d'habileté (figure 8.20).

Tableau 8.19
Caractéristiques de la distribution de probabilité de l'erreur de mesure du niveau d'habileté, (EAP(θ) - θ), en fonction du nombre d'items administrés retenu pour la règle d'arrêt

ITEM	MOYENNE	ÉCART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
1	-0,01	0,81	0,04	0,19	-2,68	3,43
2	-0,01	0,70	-0,03	0,17	-2,51	3,01
3	0,01	0,62	-0,03	0,10	-2,17	2,70
4	0,01	0,56	0,01	0,12	-1,91	2,43
5	0,01	0,51	0,00	0,11	-1,69	2,19
6	0,01	0,48	0,00	0,01	-1,98	1,52
7	0,00	0,45	-0,03	-0,16	-1,35	1,78
8	0,00	0,43	-0,01	-0,16	-1,20	1,59
9	0,00	0,40	0,06	-0,18	-1,16	1,41
10	0,00	0,38	0,03	-0,16	-1,13	1,25
11	0,00	0,36	0,02	-0,18	-1,06	1,12
12	0,00	0,35	0,04	-0,07	-1,12	1,10
13	-0,01	0,34	0,01	-0,08	-1,13	1,09
14	0,00	0,32	0,01	-0,08	-1,05	1,19
15	-0,01	0,31	0,04	-0,12	-0,97	1,08
20	0,00	0,27	0,01	-0,22	-0,86	1,00
25	0,00	0,24	0,01	-0,17	-0,70	0,97
30	-0,01	0,22	-0,02	0,06	-0,74	0,84
40	-0,01	0,18	-0,08	0,02	-0,59	0,63
50	-0,01	0,16	0,01	-0,23	-0,49	0,47
60	-0,01	0,15	0,02	-0,04	-0,45	0,47

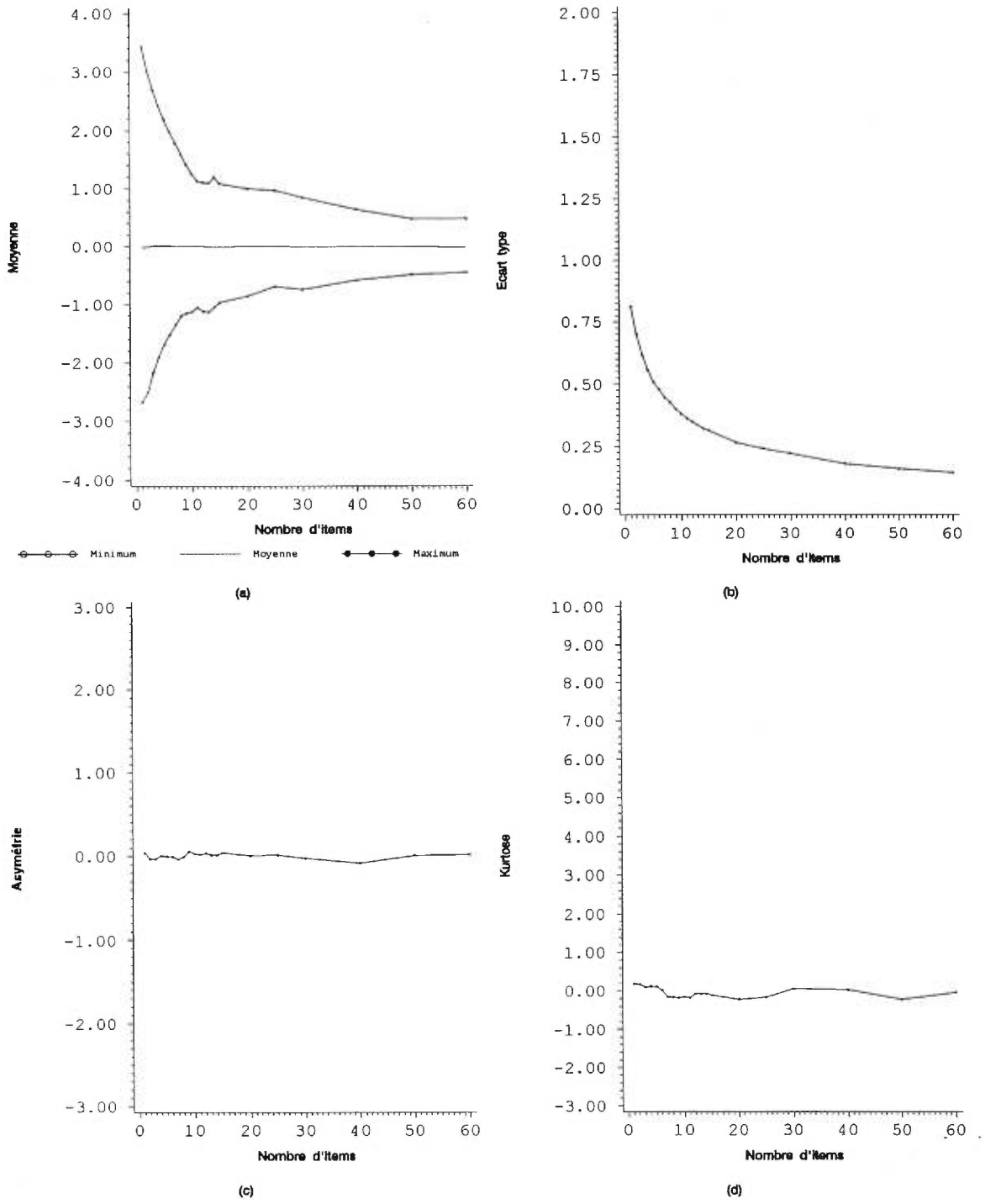

Figure 8.20 Caractéristiques de la distribution de probabilité de l'erreur de mesure du niveau d'habileté, (EAP(θ) - θ), en fonction de la règle d'arrêt basée sur le nombre d'items administrés

Les valeurs obtenues pour l'asymétrie de la distribution de probabilité de l'erreur de mesure du niveau d'habileté sont de peu d'importance. Elles ne sont pas inférieures à -0,08 ou supérieures à 0,06. La distribution de probabilité peut donc être considérée comme symétrique. Les valeurs de la kurtose de la distribution de probabilité de l'erreur de mesure du niveau d'habileté sont également peu importantes, elles se situent entre -0,23 et 0,19. Puisque la kurtose peut être légèrement positive ou légèrement négative, nous ne pouvons pas conclure que la distribution de probabilité est toujours leptokurtique ou platykurtique.

Le fait que l'erreur de mesure observée est à toutes fins utiles égale à 0,00 ne doit pas nous faire oublier que les minimums et maximums de l'erreur de mesure du niveau d'habileté sont importants quel que soit le nombre d'items administrés retenu pour la règle d'arrêt. Après l'administration de 60 items, les minimums et maximums atteignent -0,45 et 0,47. À ce moment, ces minimums et maximums de l'erreur de mesure correspondent tout de même à des valeurs entre -3,00 et 3,13 fois supérieures à l'erreur type observée au tableau 8.17.

En conséquence, nous concluons que la distribution de probabilité de l'erreur de mesure du niveau d'habileté se comporte à toutes fins utiles comme une distribution normale $N(0,01, S_{EAP(\theta)})$, tout comme dans le cas de la règle d'arrêt selon l'erreur type (section 8.2.3). Quant on tient compte de toute l'étendue du niveau d'habileté, la moyenne de l'estimateur du niveau d'habileté peut donc être considérée comme un estimateur non

biaisé de la moyenne du niveau d'habileté.

Au tableau 8.20, on trouve les coefficients de régression et de détermination de la modélisation par une régression cubique de la relation entre l'erreur de mesure du niveau d'habileté en fonction du niveau d'habileté pour quatre valeurs du nombre d'items administrés retenu pour la règle d'arrêt. Les corrélogrammes correspondants sont présentés à la figure 8.21. Comme nous l'avions remarqué dans le cas de la règle d'arrêt selon l'erreur type, au tableau 8.6, le coefficient de détermination est peu important et affiche tout au plus une valeur de 0,13 lorsque le nombre d'items administrés est égal à 10. Ce coefficient diminue avec l'augmentation du nombre d'items administrés pour n'être que de 0,03 après l'administration du 60^{e} item. La prédiction de l'erreur de mesure associée à l'estimateur du niveau d'habileté par le niveau d'habileté est donc peu précise quel que soit le nombre d'items administrés.

Puisque dans le cas présent, l'erreur de mesure prédite correspond aussi à la valeur moyenne de l'erreur de mesure à des valeurs spécifiques du niveau d'habileté, soit le biais de l'estimateur du niveau d'habileté, $BIAIS_{EAP(\theta)}$, les modèles de régression obtenus permettent de vérifier la valeur du biais de l'estimateur du niveau d'habileté en fonction du niveau d'habileté et du nombre d'items administrés.

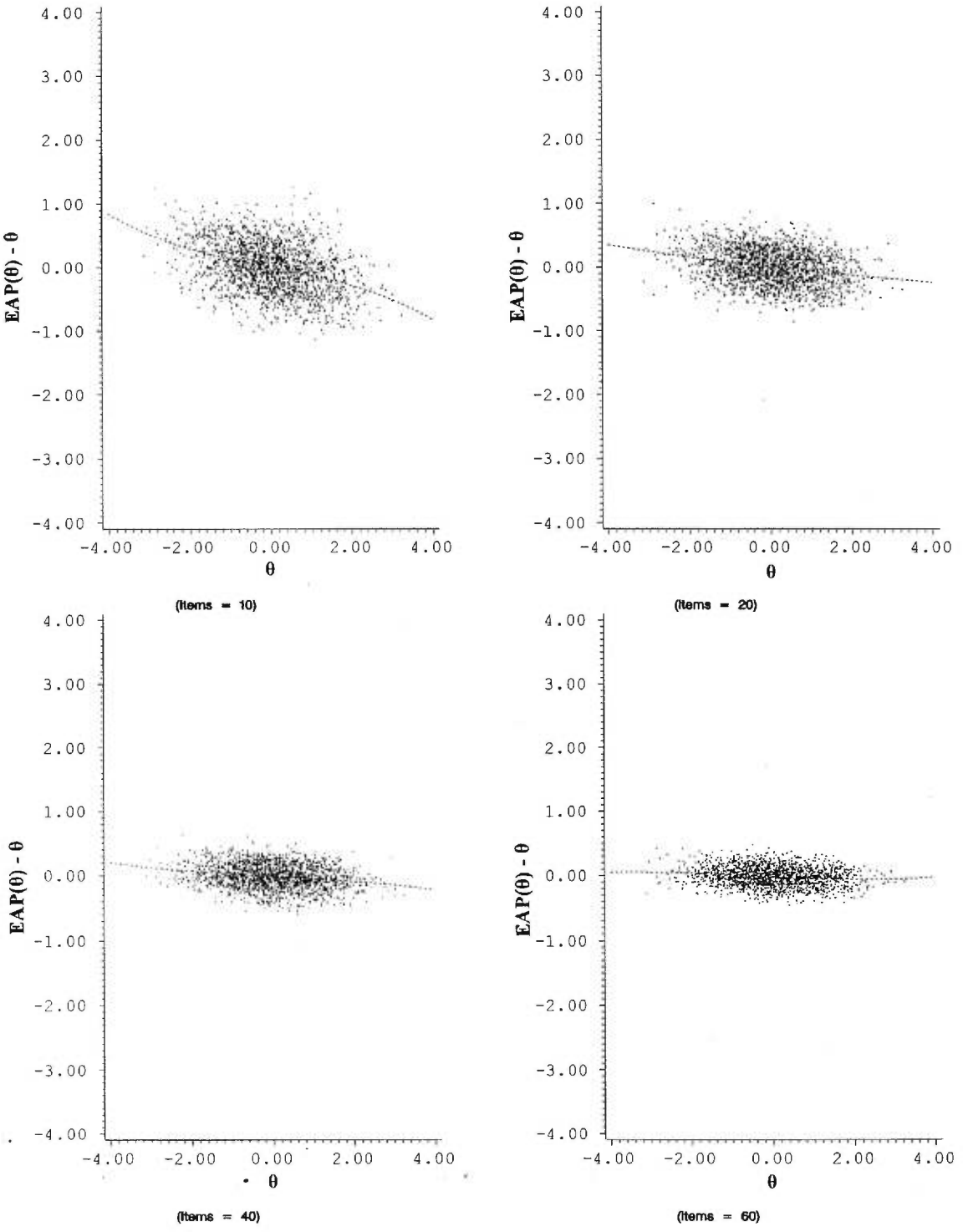

Figure 8.21 Erreur de mesure associée à l'estimateur du niveau d'habileté, (EAP(θ) - θ), en fonction du niveau d'habileté lorsque la règle d'arrêt selon le nombre d'items administrés est égale à 10, 20, 40 et 60

Tableau 8.20
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et l'erreur de mesure associée à l'estimateur du niveau d'habileté, EAP(θ) - θ, lorsque la règle d'arrêt selon le nombre d'items administrés est égale à 10, 20, 40 et 60

Règle d'arrêt selon le nombre d'items administrés	Modèle ($b_0+b_1\theta+b_2\theta^2+b_3\theta^3$)	Coefficient de détermination (R^2)
10	0,00028 - 0,12547 θ + 0,00050 θ^2 - 0,00522 θ^3	0,13
20	-0,00221 - 0,07206 θ + 0,00371 θ^2 - 0,00014 θ^3	0,07
40	-0,00833 - 0,03588 θ + 0,00012 θ^2 - 0,00010 θ^3	0,04
60	-0,01465 - 0,02789 θ + 0,00188 θ^2 + 0,00115 θ^3	0,03

La valeur du biais de l'estimateur du niveau d'habileté prédite par les modèles de régression permet de juger si l'estimateur du niveau d'habileté est biaisé ou non à différentes valeurs du niveau d'habileté. Ainsi, le biais est positif quand le niveau d'habileté est faible, une indication de surestimation du niveau d'habileté. Quand le niveau d'habileté est élevé, le biais est négatif et le niveau d'habileté a tendance à être sous-estimé. Cela nous indique que le biais de l'estimateur du niveau d'habileté n'est pas constant sur toute l'étendue du niveau d'habileté.

À titre indicatif, selon nos modélisations, quand le niveau d'habileté est égal à -3,00 et que le nombre d'items administrés est égal à 10, 20, 40 et 60, le biais est respectivement

de 0,52, 0,25, 0,13 et 0,05. Avec des erreurs types correspondantes de 0,38, 0,27, 0,19 et 0,15 (tableau 8.17), le biais de l'estimateur du niveau d'habileté est alors égal à 136,84 %, 92,59 %, 52,63 % et 33,33 % de la valeur de l'erreur type. Nous avons souligné à la section 8.3.2, où sont étudiées les caractéristiques de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté en fonction de la règle d'arrêt selon le nombre d'items administrés, que Hoijtink et Boomsma (1995, 1996) laissent croire que le biais de l'estimateur du niveau d'habileté est peu important lorsque seulement 10 items sont administrés. Les résultats que nous obtenons sont loin de confirmer les propos de ces auteurs ; au contraire, le biais peut à ce moment devenir extrêmement important.

L'estimateur du niveau d'habileté ne peut donc pas être considéré comme un estimateur non biaisé sur toute l'étendue du niveau d'habileté et il serait plus prudent d'appliquer la correction suggérée par Bock et Mislevy (1982, p. 439-442). Le biais de l'estimateur du niveau d'habileté deviendrait alors relativement nul sur toute l'étendue du niveau d'habileté en autant que le nombre d'items administrés soit supérieur ou égal à 10 et que le niveau d'habileté soit compris à l'intérieur de l'intervalle [-3,00, 3,00].

8.3.4 Caractéristiques de la distribution de probabilité de l'asymétrie de l'estimateur du niveau d'habileté en fonction du nombre d'items administrés retenu pour la règle d'arrêt

Au tableau 8.21 et à la figure 8.22, on trouve plusieurs informations relatives à la distribution de probabilité de l'asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Leur analyse démontre, en premier lieu, que la moyenne de l'asymétrie de l'estimateur du niveau d'habileté est toujours nulle quel que soit le nombre d'items administrés. Ensuite, l'écart type de l'asymétrie diminue constamment avec l'augmentation du nombre d'items administrés et semble tendre vers 0,00. Nous avions observé des caractéristiques similaires de l'asymétrie en ce qui a trait à la règle d'arrêt selon l'erreur type (section 8.2.4).

L'asymétrie de cette distribution de probabilité est d'au plus 0,09 en valeur absolue et peut être aussi bien négative que positive. Elle est, de plus, à peu près constante quel que soit le nombre d'items administrés. Nous concluons donc à la symétrie de cette distribution de probabilité. Pour sa part, la valeur de la kurtose est toujours négative, révélant ainsi que la disribution de probabilité de l'asymétrie est plutôt platykurtique. La kurtose devient toutefois peu importante, quoique très progressivement, avec l'augmentation du nombre d'items administrés.

Tableau 8.21
Caractéristiques de la distribution de probabilité de l'asymétrie de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, en fonction du nombre d'items administrés retenu pour la règle d'arrêt

ITEM	MOYENNE	ÉCART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
1	0,00	0,11	0,02	-2,00	-0,11	0,11
2	0,00	0,13	-0,02	-1,21	-0,18	0,18
3	0,00	0,14	0,00	-1,16	-0,22	0,22
4	0,00	0,14	-0,01	-0,99	-0,24	0,24
5	0,00	0,14	-0,05	-0,89	-0,25	0,25
6	0,00	0,13	0,00	-0,85	-0,27	0,27
7	0,00	0,13	-0,04	-0,73	-0,28	0,28
8	0,00	0,12	-0,05	-0,51	-0,29	0,29
9	0,00	0,11	-0,05	-0,43	-0,29	0,29
10	0,00	0,11	-0,06	-0,51	-0,29	0,29
11	0,00	0,10	-0,05	-0,37	-0,29	0,28
12	0,00	0,10	-0,02	-0,34	-0,29	0,28
13	0,00	0,09	0,06	-0,33	-0,28	0,29
14	0,00	0,09	0,07	-0,38	-0,25	0,23
15	0,00	0,08	0,05	-0,38	-0,22	0,23
20	0,00	0,07	0,03	-0,28	-0,21	0,23
25	0,00	0,06	0,00	-0,27	-0,19	0,17
30	0,00	0,05	-0,02	-0,08	-0,16	0,16
40	0,00	0,04	-0,06	-0,12	-0,12	0,11
50	0,00	0,03	-0,09	0,02	-0,12	0,10
60	0,00	0,03	0,01	-0,23	-0,10	0,08

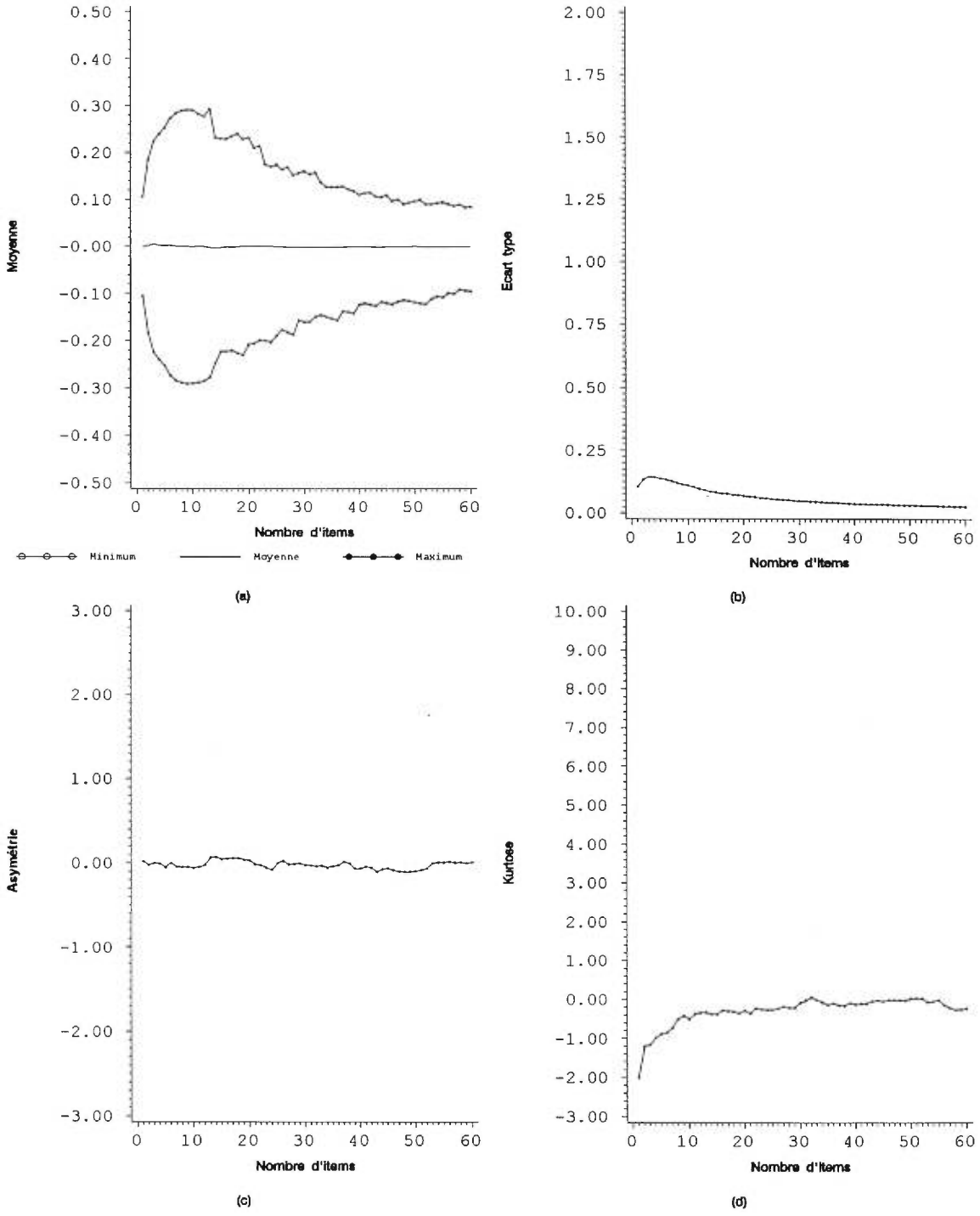

Figure 8.22 Caractéristiques de la distribution de probabilité de l'asymétrie de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, en fonction de la règle d'arrêt basée sur le nombre d'items administrés

Quant aux minimums et maximums de l'asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, nous observons qu'ils se rapprochent graduellement de la moyenne avec l'augmentation du nombre d'items administrés. Au delà de l'administration du 20^{e} item, l'écart avec la moyenne de l'asymétrie devient d'ailleurs assez peu important. Les valeurs maximales et minimales de l'asymétrie ont donc peu d'effet sur la différence entre la médiane et la moyenne de la distribution d'échantillonnage de l'estimateur du niveau d'habileté.

Au tableau 8.22, nous présentons le détail de la modélisation de la relation entre l'asymétrie de l'estimateur du niveau d'habileté et le niveau d'habileté en fonction de quatre valeurs du nombre d'items administrés retenu pour la règle d'arrêt. Les coefficients de régression et de détermination de ces modélisations qui y sont présentés sont également illustrés à la figure 8.23.

Selon les informations présentées au tableau 8.22, le coefficient de détermination varie entre 0,38 et 0,17 et diminue avec l'augmentation du nombre d'items administrés. Les modélisations sont plutôt non linéaires lorsque le niveau d'habileté affiche des valeurs extrêmes, soit des valeurs supérieures à 2,00 en valeur absolue. Entre ces valeurs du niveau d'habileté, la relation est pratiquement linéaire : plus le niveau d'habileté est élevé, plus l'asymétrie est importante. Lorsque le nombre d'items administrés est égal à 60, la relation entre le niveau d'habileté et l'asymétrie du niveau d'habileté est toutefois presque linéaire sur toute l'étendue étudiée du niveau d'habileté.

L'asymétrie obtenue à partir de ces modélisations, quand 10 et 60 items sont administrés et que le niveau d'habileté est égal à -2,00, correspond respectivement à -0,18 et -0,04. Ces valeurs calculées de l'asymétrie sont peu importantes ; nous en concluons que la distribution d'échantillonnage de l'estimateur du niveau d'habileté est peu affectée par son asymétrie, quelle que soit la valeur du niveau d'habileté quand le niveau d'habileté est compris entre -3,00 et 3,00.

Tableau 8.22
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et l'asymétrie de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, lorsque la règle d'arrêt selon le nombre d'items administrés est égale à 10, 20, 40 et 60

Règle d'arrêt selon le nombre d'items administrés	Modèle $(b_0+b_1\theta+b_2\theta^2+b_3\theta^3)$	Coefficient de détermination (R^2)
10	$- 0,00068 + 0,07434\ \theta - 0,00136\ \theta^2 - 0,00189\ \theta^3$	0,38
20	$- 0,00029 + 0,04075\ \theta + 0,00031\ \theta^2 - 0,00094\ \theta^3$	0,30
40	$- 0,00079 + 0,01946\ \theta - 0,00059\ \theta^2 - 0,00050\ \theta^3$	0,22
60	$- 0,00152 + 0,01184\ \theta - 0,00006\ \theta^2 - 0,00004\ \theta^3$	0,17

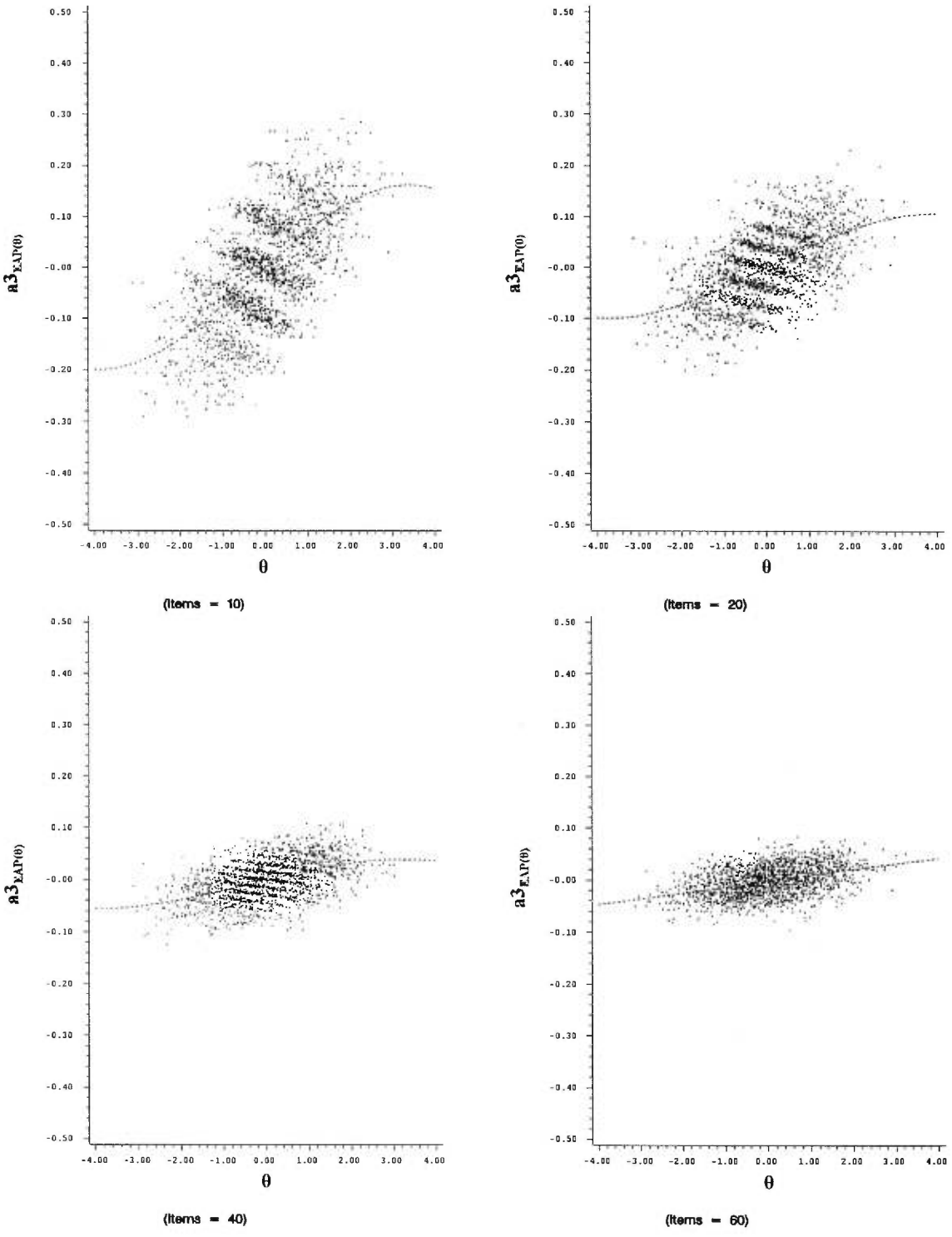

Figure 8.23 Asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, en fonction du niveau d'habileté lorsque la règle d'arrêt selon le nombre d'items administrés est égale à 10, 20, 40 et 60

8.3.5 Caractéristiques de la distribution de probabilité de la kurtose de l'estimateur du niveau d'habileté en fonction du nombre d'items administrés retenu pour la règle d'arrêt

Au tableau 8.23 et à la figure 8.24 sont présentés les résultats relatifs aux caractéristiques de la distribution de probabilité de la kurtose de l'estimateur du niveau d'habileté.

La moyenne de la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté est supérieure à 0,00 quel que soit le nombre d'items administrés retenu pour la règle d'arrêt. La kurtose est d'au plus 0,22 lorsque le nombre d'items administrés est égal à quatre. La valeur minimale observée est de 0,03, cela au moment où 60 items sont administrés. La valeur de la kurtose diminue d'ailleurs avec l'augmentation du nombre d'items administrés et il est possible qu'elle puisse éventuellement devenir nulle avec l'augmentation du nombre d'items. Les valeurs les plus importantes de la kurtose, soit celles supérieures ou égales à 0,20, sont obtenues lorsque le nombre d'items administrés se situe entre trois et six. Comme nous l'avions souligné lors de l'étude de la règle d'arrêt selon l'erreur type à la section 8.2.5, la distribution de probabilité de la kurtose a tendance à être leptokurtique. Pour sa part, l'écart type de la distribution de probabilité de la kurtose est à toutes fins utiles nul, quel que soit le nombre d'items administrés ; il est d'au plus 0,03 et pourrait éventuellement devenir nul avec l'augmentation du nombre d'items administrés.

Tableau 8.23
Caractéristiques de la distribution de probabilité de la kurtose de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, en fonction du nombre d'items administrés retenu pour la règle d'arrêt

ITEM	MOYENNE	ÉCART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
1	0,13	0,00	nd.*	nd.	0,13	0,13
2	0,19	0,03	0,02	-2,00	0,16	0,22
3	0,21	0,03	-0,61	-1,24	0,17	0,24
4	0,22	0,03	-0,94	0,05	0,15	0,25
5	0,21	0,03	-1,31	2,44	0,12	0,26
6	0,20	0,03	-1,92	6,69	0,06	0,26
7	0,19	0,03	-2,68	15,81	-0,01	0,26
8	0,18	0,03	-3,51	31,02	-0,10	0,26
9	0,17	0,03	-4,82	58,88	-0,20	0,25
10	0,16	0,03	-5,19	84,97	-0,30	0,24
11	0,15	0,03	-5,86	111,50	-0,37	0,24
12	0,14	0,03	-6,16	127,92	-0,41	0,24
13	0,13	0,03	-7,69	143,94	-0,39	0,24
14	0,12	0,02	-7,14	145,51	-0,34	0,21
15	0,12	0,02	-9,39	189,01	-0,36	0,19
20	0,09	0,02	-8,98	185,23	-0,26	0,18
25	0,07	0,01	-5,18	154,00	-0,15	0,14
30	0,06	0,01	1,67	13,03	0,00	0,11
40	0,05	0,00	2,26	7,46	0,04	0,08
50	0,04	0,00	0,89	2,61	0,03	0,06
60	0,03	0,01	-0,05	-1,31	0,01	0,06

* Valeur non disponible car indéterminée (division par zéro)

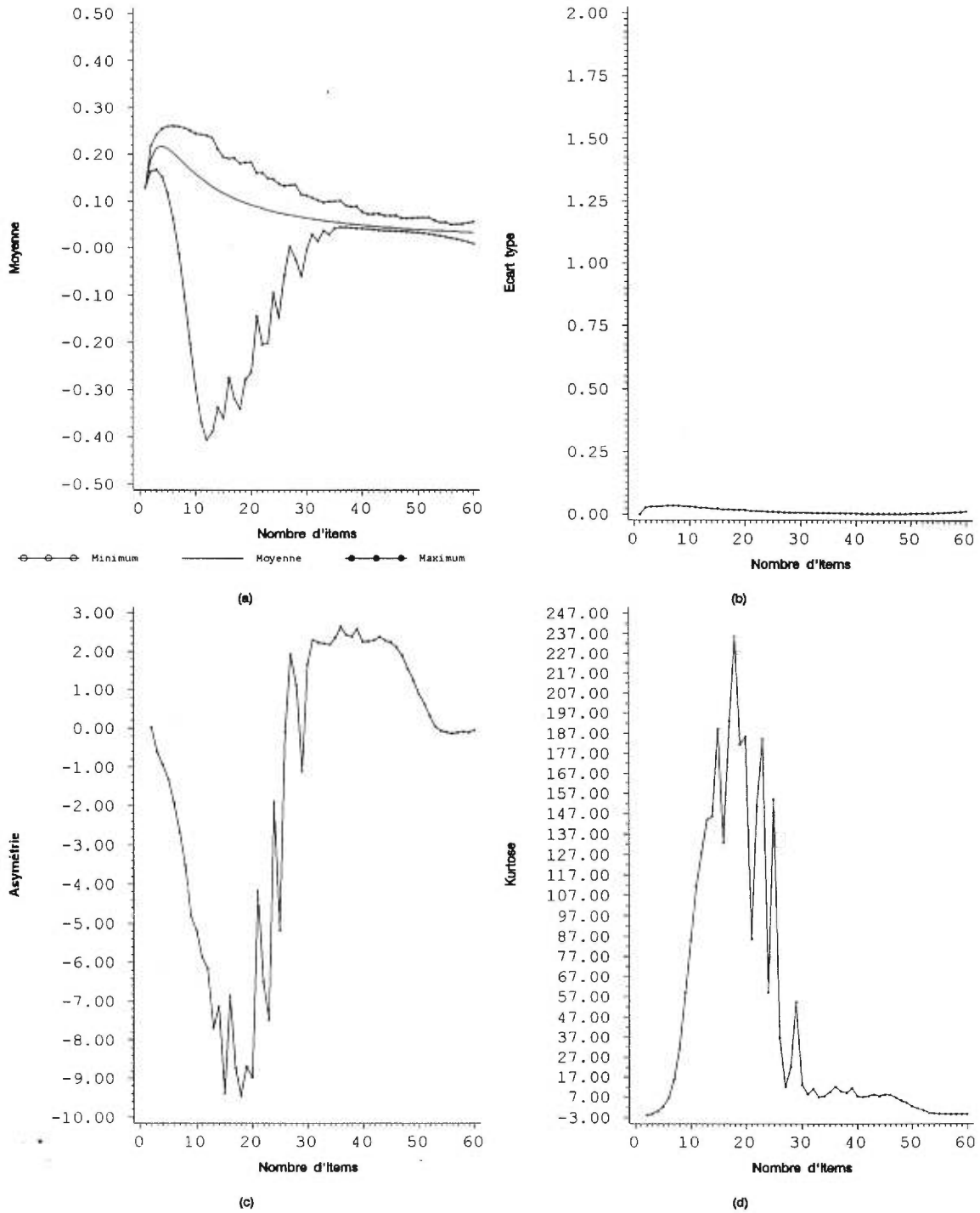

Figure 8.24 Caractéristiques de la distribution de probabilité de la kurtose de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, en fonction de la règle d'arrêt basée sur le nombre d'items administrés

L'asymétrie de la distribution de probabilité de la kurtose présente généralement des valeurs très importantes, la plupart du temps négatives. Ce n'est qu'à partir du 50^{e} item que l'asymétrie semble diminuer de façon significative ; elle est à ce moment égale à 0,89. Une analyse plus détaillée des valeurs de l'asymétrie après l'administration du 50^{e} item indique que l'asymétrie n'est plus que de -0,05 après l'administration de 53 items et qu'elle ne dépasse pas 0,13 en valeur absolue par la suite. En ce qui concerne la kurtose de cette distribution de probabilité, elle présente des valeurs positives importantes après l'administration du 5^{e} item jusqu'à ce que 52 items aient été administrés. La distribution de probabilité de la kurtose est alors fortement leptokurtique. Par la suite, la kurtose devient négative et rapidement platykurtique. Ce comportement est difficile à expliquer.

Nous observons aussi que les minimums et maximums de la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté se situent entre -0,41 et 0,26. Nous en concluons que les interprétations relatives à la distribution d'échantillonnage de l'estimateur du niveau d'habileté sont légèrement affectées par les valeurs de la kurtose. Notons aussi qu'au delà de l'administration du 30^{e} item, le minimum de la kurtose est constamment positif ; c'est une indication que l'intervalle de confiance autour de l'estimateur du niveau d'habileté est toujours surestimé. Toutefois, à partir de l'administration du 25^{e} item, les minimums et maximums de la kurtose ne sont que de -0,15 et 0,14 et, à ce moment, celle-ci n'affecte que très peu les interprétations relatives à l'estimateur du niveau d'habileté.

À partir du tableau 8.24 et de la figure 8.25, il nous est possible d'analyser la relation entre le niveau d'habileté et la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Selon le tableau 8.24, les coefficients de détermination sont de peu d'importance : ils ne sont jamais supérieurs à 0,05. La kurtose est d'ailleurs presque constante sur toute l'étendue du niveau d'habileté. C'est ce que nous avions déjà remarqué lors de l'étude de la règle d'arrêt selon l'erreur type (section 8.2.5). Nous y remarquions aussi que la moyenne de la kurtose est légèrement plus élevée quand le niveau d'habileté se situe autour de la moyenne du niveau d'habileté. Tel que nous pouvons l'observer à la figure 8.25, ce n'est d'ailleurs qu'aux valeurs extrêmes du niveau d'habileté que se trouvent les valeurs minimales de la kurtose.

Tableau 8.24
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et la kurtose de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, lorsque la règle d'arrêt selon le nombre d'items administrés est égale à 10, 20, 40 et 60

Règle d'arrêt selon le nombre d'items administrés	Modèle $(b_0+b_1\theta+b_2\theta^2+b_3\theta^3)$	Coefficient de détermination (R^2)
10	$0{,}16198 - 0{,}00179\,\theta - 0{,}00366\,\theta^2 + 0{,}00034\,\theta^3$	0,03
20	$0{,}09224 - 0{,}00095\,\theta - 0{,}00038\,\theta^2 + 0{,}00054\,\theta^3$	0,01
40	$0{,}04760 + 0{,}00037\,\theta + 0{,}00052\,\theta^2 - 0{,}00017\,\theta^3$	0,05
60	$0{,}03238 - 0{,}00112\,\theta + 0{,}00014\,\theta^2 + 0{,}00021\,\theta^3$	0,00

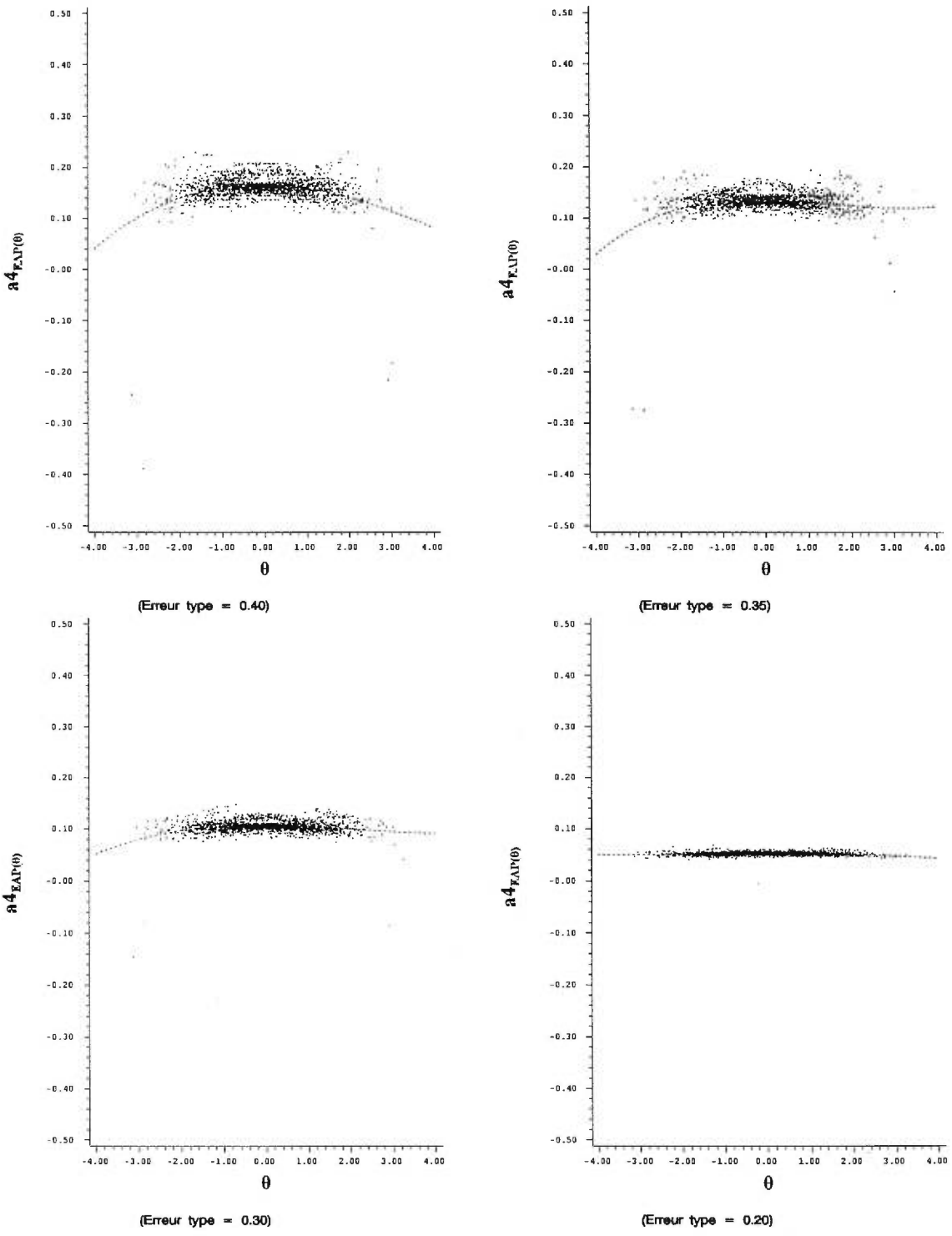

Figure 8.25 Kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, en fonction du niveau d'habileté lorsque la règle d'arrêt selon le nombre d'items administrés est égale à 10, 20, 40 et 60

Par rapport à notre objectif de recherche, une remarque importante est à faire. L'étude de la distribution de probabilité de la kurtose ainsi que de l'asymétrie associées à l'estimateur du niveau d'habileté, nous permet de conclure que les interprétations relatives à la distribution d'échantillonnage de l'estimateur du niveau d'habileté ne sont que très peu affectées par les valeurs de la kurtose et de l'asymétrie quel que soit le nombre d'items administrés. Nous pouvons ainsi conclure que la distribution d'échantillonnage de l'estimateur du niveau d'habileté présente les caractéristiques d'une distribution de probabilité normale $N(EAP(\theta), S_{EAP(\theta)})$.

8.3.6 Caractéristiques de la distribution de probabilité de la proportion de bonnes réponses associée à l'estimateur du niveau d'habileté en fonction du nombre d'items administrés retenu pour la règle d'arrêt

Le tableau 8.25 et la figure 8.26, présentent les caractéristiques de la distribution de probabilité de la proportion de bonnes réponses selon les valeurs retenues de la règle d'arrêt basée sur le nombre d'items administrés. Selon ces résultats, la moyenne de la proportion de bonnes réponses est presque toujours égale à 0,50. Quand ce n'est pas le cas, elle ne s'éloigne de 0,50 que de $\pm 0{,}01$, et ce seulement lorsque le nombre d'items administrés est inférieur à cinq. L'écart type de la proportion de bonnes réponses, pour sa part, est pratiquement toujours égal à l'erreur type d'une proportion. Nous l'avions déjà remarqué à la section 8.2.6 à propos de la règle d'arrêt basée sur l'erreur type de l'estimateur du niveau d'habileté. À ce moment, puisque le nombre d'items administrés à chaque valeur de l'erreur type retenue était variable, l'écart type de la proportion de bonnes réponses était approximativement égal à l'erreur type d'une proportion. Les seules exceptions à cette égalité entre l'écart type de la proportion de bonnes réponses et l'écart type d'une proportion sont celles qui correspondent à l'administration de 2, 8 et 9 items où l'erreur type s'éloigne de seulement $\pm 0{,}01$ par rapport à l'erreur type d'une proportion ; il s'agit donc d'un écart peu important. Signalons aussi, ce qui était prévisible, que l'écart type de la proportion de bonnes réponses diminue avec l'augmentation du nombre d'items administrés : la relation est curvilinéaire.

Tableau 8.25
Caractéristiques de la distribution de probabilité de la proportion de bonnes réponses en fonction du nombre d'items administrés retenu pour la règle d'arrêt

ITEM	MOYENNE	ÉCART TYPE	ASYMÉTRIE	KURTOSE	MINIMUM	MAXIMUM
1	0,49	0,50	0,02	-2,00	0,00	1,00
2	0,50	0,36	-0,01	-1,02	0,00	1,00
3	0,51	0,29	0,04	-0,70	0,00	1,00
4	0,51	0,25	0,05	-0,41	0,00	1,00
5	0,50	0,22	0,03	-0,34	0,00	1,00
6	0,51	0,20	0,05	-0,24	0,00	1,00
7	0,50	0,19	-0,03	-0,23	0,00	1,00
8	0,50	0,17	-0,04	-0,18	0,00	1,00
9	0,50	0,16	-0,04	-0,12	0,00	1,00
10	0,50	0,16	-0,05	-0,23	0,00	1,00
11	0,50	0,15	-0,06	-0,19	0,00	0,91
12	0,50	0,14	-0,05	-0,24	0,00	0,92
13	0,50	0,14	0,01	-0,21	0,00	0,92
14	0,50	0,13	0,04	-0,24	0,07	0,93
15	0,50	0,13	0,05	-0,19	0,07	0,93
20	0,50	0,11	0,01	-0,24	0,15	0,85
25	0,50	0,10	0,01	-0,23	0,20	0,80
30	0,50	0,09	-0,02	-0,11	0,20	0,77
40	0,50	0,08	-0,04	-0,11	0,22	0,72
50	0,50	0,07	-0,04	-0,04	0,26	0,72
60	0,50	0,06	0,02	-0,16	0,30	0,70

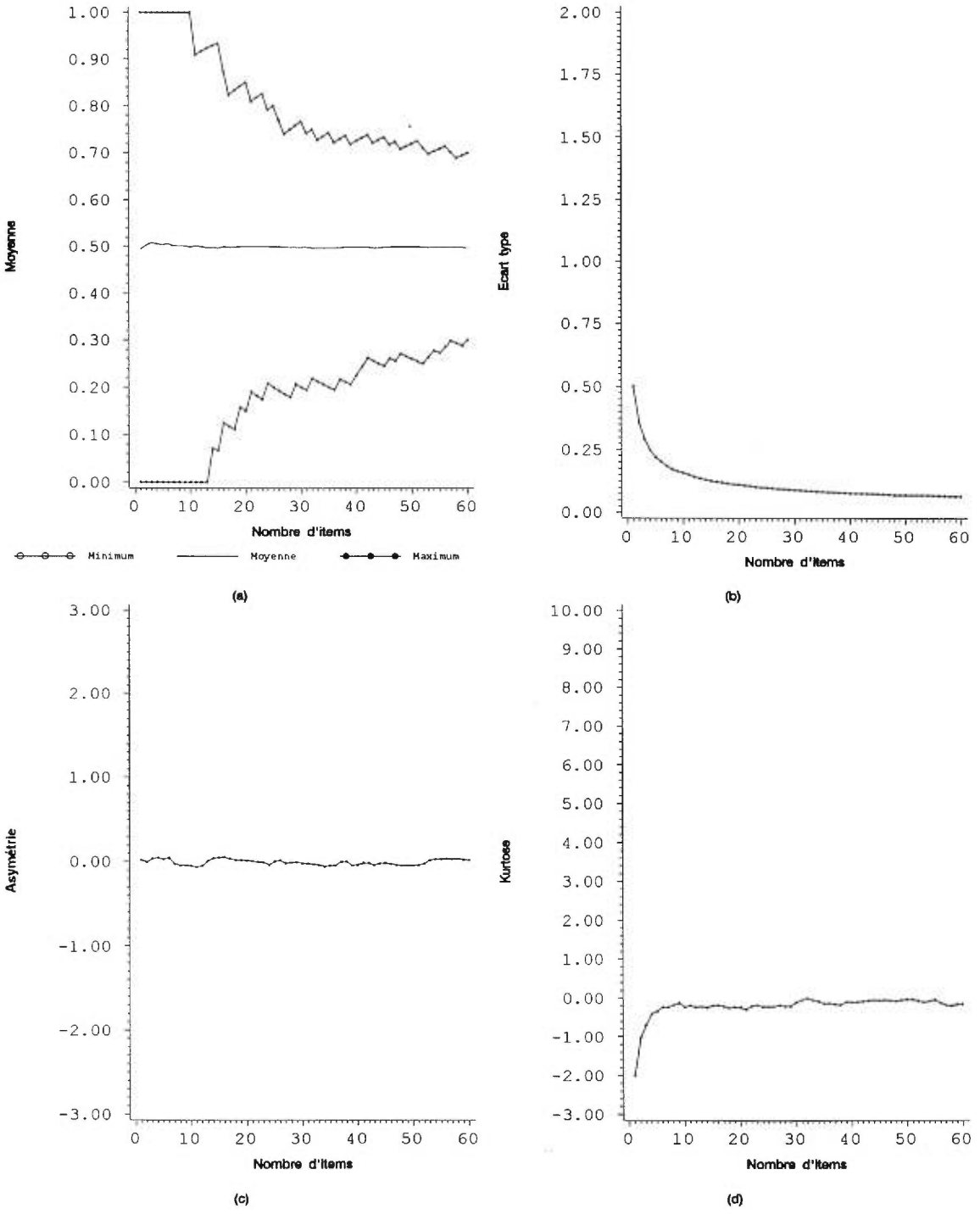

Figure 8.26 Caractéristiques de la distribution de probabilité de la proportion de bonnes réponses en fonction de la règle d'arrêt basée sur le nombre d'items administrés

L'asymétrie de la distribution de probabilité de la proportion de bonnes réponses est peu importante, sinon presque nulle, sur toute l'étendue étudiée du nombre d'items administrés : elle varie entre -0,06 et 0,05. Nous pouvons donc considérer que cette distribution de probabilité est symétrique. La kurtose de la distribution de probabilité de la proportion de bonnes réponses est constamment négative, ce qui dénote une légère tendance à être platykurtique, et peu importante lorsque le nombre d'items administrés est supérieur à cinq.

Nous remarquons aussi que les minimums et maximums de la proportion de bonnes réponses se rapprochent de la moyenne avec l'augmentation du nombre d'items administrés. De façon similaire à ce que nous avions indiqué à la section 8.2.6 concernant la règle d'arrêt selon l'erreur type, nous signalons que, pour obtenir une proportion de bonnes réponses constamment égale à 0,50, dans la mesure où cette situation est possible, le nombre d'items administrés devrait être très grand.

Le tableau 8.26 et la figure 8.27 nous permettent de constater que la proportion de bonnes réponses varie en fonction du niveau d'habileté. La relation observée, plutôt linéaire, est toutefois moins importante avec l'augmentation du nombre d'items administrés, comme l'indique la diminution du coefficient de détermination avec l'augmentation du nombre d'items. Ce coefficient est ainsi de 0,65 quant 10 items sont administrés et de 0,45 lorsque le nombre d'items retenu est égal à 60. La pente de la courbe de régression est aussi de moins en moins importante avec l'augmentation du nombre d'items administrés,

puisque la proportion de bonnes réponses tend à se rapprocher de 0,50 sur toute l'étendue du niveau d'habileté avec l'augmentation du nombre d'items administrés.

La moyenne de la proportion de bonnes réponses en fonction du nombre d'items administrés et du niveau d'habileté peut donc être calculée de façon assez précise à partir des modélisations que nous proposons au tableau 8.26. Par exemple, lorsque le niveau d'habileté est égal à -4,00, -3,00, -2,00, 2,00, 3,00 et 4,00 et que 60 items sont administrés, la proportion de bonnes réponses calculée à partir du modèle de régression est respectivement de 0,30, 0,36, 0,42, 0,58, 0,63 et 0,69. Toutefois, contrairement aux résultats obtenus à partir du modèle logistique à un paramètre à un test conventionnel fixe et invariable (Molenaar, 1995, p. 10), la proportion de bonnes réponses n'est plus une statistique suffisante (*sufficient statistic*) pour permettre l'estimation du niveau d'habileté. Si la proportion de bonnes réponses était une statistique suffisante, l'estimateur du niveau d'habileté serait totalement déterminé par cette proportion (Fischer, 1995, p. 15), ce qui n'est évidemment pas le cas dans un test adaptatif. Enfin, les valeurs extrêmes de 0,30 et 0,69 sont à peu près égales à celles que nous avons observées au tableau 8.25, soit 0,30 et 0,70.

Tableau 8.26
Coefficients de détermination et de régression de la modélisation de la relation entre le niveau d'habileté et la proportion de bonnes réponses lorsque la règle d'arrêt selon le nombre d'items administrés est égale à 10, 20, 40 et 60

Règle d'arrêt selon le nombre d'items administrés	Modèle $(b_0+b_1\theta+b_2\theta^2+b_3\theta^3)$	Coefficient de détermination (R^2)
10	$0,49976 + 0,12441\ \theta - 0,00135\ \theta^{2+} + 0,00150\ \theta^3$	0,65
20	$0,49989 + 0,08006\ \theta - 0,00004\ \theta^2 + 0,00159\ \theta^3$	0,57
40	$0,49842 + 0,05104\ \theta - 0,00095\ \theta^2 + 0,00032\ \theta^3$	0,46
60	$0,49653 + 0,03700\ \theta - 0,00005\ \theta^2 + 0,00080\ \theta^3$	0,36

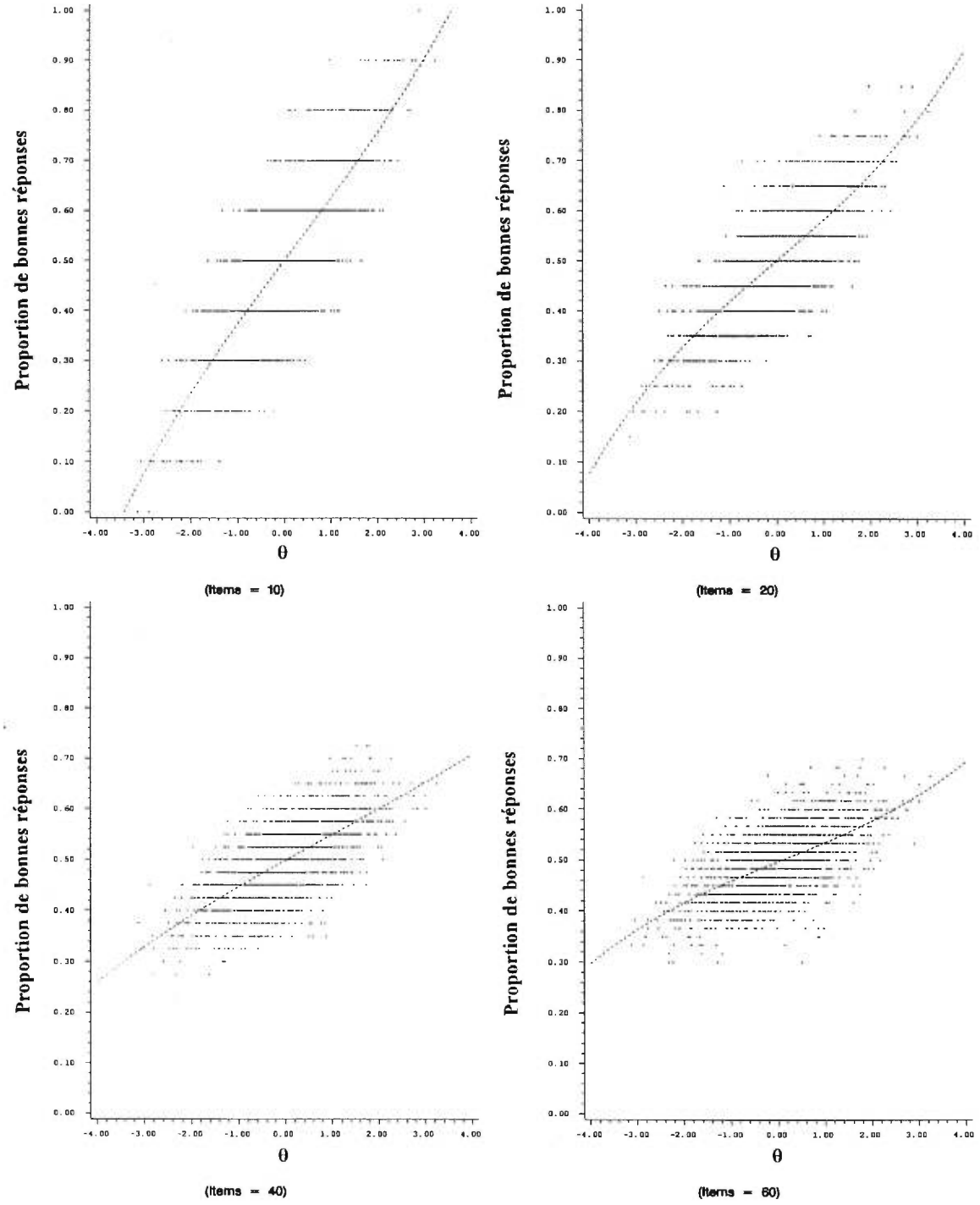

Figure 8.27 Proportion de bonnes réponses en fonction du niveau d'habileté lorsque la règle d'arrêt selon le nombre d'items administrés est égale à 10, 20, 40 et 60

9. Discussion des résultats

Tout au long du chapitre précédent, nous avons abordé plusieurs aspects de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif. Il est maintenant opportun de discuter des résultats que nous avons obtenus et de les mettre en relation avec les résultats produits à l'intérieur des recherches sur la distribution d'échantillonnage de l'estimateur du niveau d'habileté que nous avons analysés au chapitre 6. Par la même occasion, nous faisons la synthèse des constatations que nous avons faites et nous apportons certaines recommandations appropriées à un test adaptatif dont la règle d'arrêt est basée, soit sur l'erreur type de l'estimateur du niveau d'habileté, soit sur le nombre d'items administrés.

Nous débutons la discussion des résultats en abordant, en premier lieu, la règle d'arrêt selon l'erreur type. La discussion des résultats en lien avec la règle d'arrêt selon le nombre d'items administrés vient en second lieu.

9.1 Règle d'arrêt selon l'erreur type

Les résultats obtenus par rapport à la règle d'arrêt selon l'erreur type de l'estimateur du niveau d'habileté sont discutés dans le même ordre qu'ils apparaissent à l'intérieur du chapitre précédent où ils ont été présentés et analysés : estimateur du niveau d'habileté,

erreur type, biais, asymétrie, kurtose, proportion de bonnes réponses et nombre d'items administrés associés à la distribution d'échantillonnage de l'estimateur du niveau d'habileté.

Au tableau 9.1, nous avons repris et concentré nos observations quant aux minimums et maximums des diverses statistiques étudiées en nous appuyant sur les valeurs les plus susceptibles d'être utilisées dans la pratique, soit des valeurs de l'erreur type retenue pour la règle d'arrêt d'au plus 0,40. Nous les commenterons plus loin. De façon générale, comme nous l'expliquerons d'ailleurs plus en détail dans ce qui suit, plus l'erreur type retenue pour la règle d'arrêt est petite, plus la distribution d'échantillonnage de l'estimateur du niveau d'habileté se comporte selon une loi normale.

Tableau 9.1
Minimums (↓) et maximums (↑) des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté lorsque la règle d'arrêt selon l'erreur type est utilisée

Critère	$EAP(\theta)$		$S_{EAP(\theta)}$		$EAP(\theta)-\theta$		Bonnes réponses		Items		$a3_{EAP(\theta)}$		$a4_{EAP(\theta)}$	
	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑
0,40	-3,38	3,05	0,37	0,40	-1,16	1,19	0,00	0,92	9	14	-0,27	0,29	-0,39	0,23
0,35	-3,28	3,04	0,33	0,35	-1,12	1,09	0,12	0,82	11	17	-0,22	0,23	-0,28	0,19
0,30	-3,28	3,21	0,29	0,30	-0,88	1,07	0,19	0,82	16	24	-0,19	0,17	-0,15	0,15
0,25	-3,10	3,34	0,24	0,25	-0,71	0,97	0,21	0,76	22	33	-0,15	0,15	-0,06	0,11
0,20	-3,16	3,16	0,20	0,20	-0,65	0,73	0,27	0,72	35	48	-0,12	0,10	0,04	0,07

9.1.1 Estimateur du niveau d'habileté, EAP(θ)

Nous avons observé que la moyenne de l'estimateur du niveau d'habileté est égale à la moyenne du niveau d'habileté quelle que soit la valeur de l'erreur type retenue pour la règle d'arrêt. Toutefois, lorsque l'erreur type retenue est élevée, l'écart type de l'estimateur du niveau d'habileté s'éloigne considérablement de l'écart type du niveau d'habileté, ici de 0,98. À titre d'illustration, comme nous l'avons fait remarquer au tableau 8.1, lorsque l'erreur type retenue est de 0,40, l'écart type de la distribution de probabilité de l'estimateur du niveau d'habileté est de 0,91. Lorsque l'erreur type retenue est égale à 0,20, cet écart type est de 0,96. Avec la réduction de l'erreur type retenue pour la règle d'arrêt, l'écart type de l'estimateur du niveau d'habileté semble donc tendre vers l'écart type du niveau d'habileté.

Pour leur part, l'asymétrie et la kurtose de la distribution de probabilité de l'estimateur du niveau d'habileté affichent des valeurs peu importantes lorsque l'erreur type retenue est inférieure à 0,75. La distribution de probabilité de l'estimateur du niveau d'habilité est donc très peu affectée par ces valeurs.

Tel que noté au tableau 9.1, lorsque l'erreur type de l'estimateur du niveau d'habileté est inférieure ou égale à 0,40, les minimums et maximums de l'estimateur du niveau d'habileté se rapprochent des minimums et maximums du niveau d'habileté. Toutefois, lorsque le niveau d'habileté s'éloigne considérablement de la moyenne, l'estimateur du

niveau d'habileté tend à être fortement surestimé ou sous-estimé. Ce n'est que lorsque la valeur de l'erreur type retenue est de 0,20 et que la valeur du niveau d'habileté n'est pas située à l'extérieur de l'intervalle [-1,50, 1,50] que la précision de l'estimateur du niveau d'habileté est jugée raisonnable. Cette caractéristique est abordée de façon plus spécifique lors de l'analyse de l'erreur de mesure du niveau d'habileté et du biais de l'estimateur du niveau d'habileté (section 9.1.3).

9.1.2 Erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$

Selon les résultats obtenus en testing adaptatif par Dodd, Koch et de Ayala (1993) à partir d'une modélisation de la réponse à l'item selon le modèle de crédit partiel de Master, lorsque l'erreur type retenue est égale à 0,30, l'erreur type moyenne de l'estimateur du niveau d'habileté varie entre 0,29 et 0,40. Notons qu'ils obtiennent une valeur maximale de l'erreur type qui est supérieure à l'erreur type retenue pour la règle d'arrêt du fait que l'administration du test cesse lorsque au maximum 20 items sont administrés. À l'intérieur de notre recherche, comme nous le rappelons au tableau 9.1, ces valeurs sont comprises entre 0,29 et 0,30, valeurs comparables à celles obtenues par Dodd et *al.* si l'on tient compte du fait que la limite d'items administrés est de 60 ; il nous est alors possible d'obtenir une valeur maximale de l'erreur type toujours égale à 0,30.

Nous avons constaté que plus l'erreur type retenue pour la règle d'arrêt est petite, plus

l'erreur type de l'estimateur du niveau d'habileté se rapproche de l'erreur type retenue. L'écart type de la distribution de probabilité de l'erreur type de l'estimateur du niveau d'habileté est à toutes fins utiles toujours nul et sa distribution de probabilité est asymétrique ainsi que platykurtique.

Nous observons aussi que l'erreur type de l'estimateur du niveau d'habileté a tendance à devenir constante sur toute l'étendue du niveau d'habileté avec la réduction de l'erreur type retenue pour la règle d'arrêt. Ce n'est toutefois que lorsque l'erreur type retenue est égale à 0,20 que les minimums et maximums de l'erreur type sont égaux à deux décimales près. Cette dernière remarque nous conduit à conclure que lorsque l'erreur type retenue pour la règle d'arrêt est petite, la précision de l'estimateur du niveau d'habileté est constante quel que soit le niveau d'habileté, tout comme le postule la théorie classique des tests. Bock et Mislevy (1982, p. 437), d'ailleurs, en arrivent à la même conclusion.

9.1.3 Erreur de mesure du niveau d'habileté et biais de l'estimateur du niveau d'habileté, $BIAIS_{EAP(\theta)}$

Quand nous tenons compte de toute l'étendue du niveau d'habileté, quelle que soit la valeur de l'erreur type retenue pour la règle d'arrêt, nous pouvons considérer que la moyenne de l'erreur de mesure du niveau d'habileté est à peu près égale à 0,00. Nous

remarquons aussi que l'écart type de l'erreur de mesure affiche toujours une valeur rapprochée de l'erreur type retenue. La distribution de probabilité de l'erreur de mesure du niveau d'habileté s'éloigne toutefois légèrement d'une distribution normale si nous considérons les valeurs observées de la kurtose de cette distribution. Celle-ci peut être aussi bien platykurtique que leptokurtique.

Nous constatons aussi que le biais de l'estimateur du niveau d'habileté n'est pas nul sur toute l'étendue du niveau d'habileté. L'estimateur du niveau d'habileté ne peut donc pas être considéré comme un estimateur non biaisé du niveau d'habileté lorsque celui-ci affiche des valeurs extrêmes. En fait, lorsque l'erreur type retenue pour la règle d'arrêt est égale à 0,20, l'estimateur du niveau d'habileté est plutôt non biaisé seulement si le niveau d'habileté est compris dans l'intervalle qui s'étend de -1,50 à 1,50. Quand l'erreur type retenue est de 0,30, cet intervalle est diminué considérablement, soit de -1,00 à 1,00.

Comme nous l'avons souligné, il serait alors plus prudent de corriger la valeur de l'estimateur du niveau d'habileté par une fonction qui en diminue le biais. C'est d'ailleurs ce que suggèrent Bock et Mislevy (1982, p 441). Ils nous offrent une stratégie pour effectuer cette correction. Il s'agit tout simplement de diviser l'estimateur du niveau d'habileté par $(1 - S^2_{EAP(\theta)})$; cela conduit donc à une correction selon l'erreur type de l'estimateur du niveau d'habileté. Nos analyses montrent que cette correction est très efficace et que le biais devient alors négligeable lorsque l'erreur type retenue est

inférieure ou égale à 0,40 et que le niveau d'habileté est compris dans l'intervalle [-3,00, 3,00].

Les valeurs du bais de l'estimateur du niveau d'habileté que nous obtenons à partir des formules de régression cubique du tableau 8.6 sont comparables à celle que Bock et Mislevy (1982, p. 440) présentent à l'intérieur de leur recherche. À titre d'illustration, lorsque l'erreur type retenue pour la règle d'arrêt est égale à 0,30 et que le niveau d'habileté est égal à -3,00, -2,00 et -1,00, le biais que nous obtenons à partir de la formule de régression appropriée est respectivement de 0,30, 0,19 et 0,08. Les valeurs que Bock et Mislevy présentent sont égales à 0,35, 0,23 et 0,07.

La variabilité des valeurs obtenues de l'estimateur du niveau d'habileté est toutefois haussée par la correction proposée par Bock et Mislevy. Ainsi, l'erreur type de l'estimateur du niveau d'habileté, l'écart type de l'estimateur du niveau d'habileté ainsi que l'étendue des valeurs estimées deviennent plus importants après l'utilisation de cette correction. Quand l'erreur type retenue est égale à 0,20, l'augmentation de la variabilité des valeurs obtenues de l'estimateur du niveau d'habileté est cependant négligeable. Il est donc fortement suggéré, dans un contexte de testing adaptatif, d'appliquer la correction suggérée par Bock et Mislevy à l'estimateur du niveau d'habileté obtenu par la méthode de l'espérance a posteriori.

9.1.4 Asymétrie de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$

À notre connaissance, les recherches effectuées sur la distribution d'échantillonnage de l'estimateur du niveau d'habileté ne nous fournissent pas d'indications précises au sujet de l'asymétrie de cette distribution. Les résultats que nous obtenons comblent cette lacune. Ils sont donc importants puisqu'ils nous permettent de porter un jugement sur l'adéquation des interprétations relatives à une loi normale à la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Bien sûr, la même remarque s'appliquera à la kurtose associée à la distribution d'échantillonnage de l'estimateur du niveau d'habileté que nous traiterons un peu plus loin (section 9.1.5).

L'asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté est peu importante quelle que soit l'erreur type retenue pour la règle d'arrêt, mais elle est tout de même présente. Comme nous le rappelle le tableau 9.1, elle varie entre -0,27 et 0,29, selon l'erreur type retenue. Son importance diminue avec la réduction de l'erreur type retenue pour la règle d'arrêt. Quand l'erreur type retenue est égale à 0,20, l'asymétrie se situe entre -0,12 et 0,10. Les formules de régression obtenues au tableau 8.8 nous indiquent qu'elle est cependant moins marquée lorsque le niveau d'habileté se rapproche de la moyenne, ici de 0,01. Nous concluons que la distribution d'échantillonnage de l'estimateur du niveau d'habileté est peu affectée par ces valeurs de l'asymétrie. À toutes fins utiles, la distribution d'échantillonnage de l'estimateur du niveau d'habileté se comporte donc comme une distribution normale $N(EAP(\theta), \sigma_{EAP(\theta)})$

et toutes les interprétations qui y sont rattachées s'appliquent. Plus particulièrement, la moyenne est égale à la médiane et l'intervalle de confiance autour de la moyenne est bien représenté par l'écart type de cette distribution.

9.1.5 Kurtose de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$

Comme pour l'asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, la kurtose de cette distribution est toujours peu importante. Elle varie entre -0,39 et 0,26, selon l'erreur type retenue pour la règle d'arrêt et est à peu près constante sur toute l'étendue du niveau d'habileté. Sa moyenne est constamment positive, ce qui nous indique qu'elle est généralement leptokurtique et qu'il y a une tendance plutôt conservatrice à surestimer l'intervalle de confiance autour de l'estimateur du niveau d'habileté à partir de l'erreur type de l'estimateur du niveau d'habileté. C'est principalement lorsque l'erreur type retenue est inférieure à 0,40 que la kurtose affiche des valeurs vraiment négligeables. Quand l'erreur type retenue est égale à 0,20, la kurtose se situe entre 0,04 et 0,07. Ces valeurs sont à ce moment minimes et n'ont à toutes fins utiles aucun impact sur les interprétations associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté.

Ainsi, la kurtose affecte peu la distribution d'échantillonnage de l'estimateur du niveau d'habileté même si l'on tient compte, de surcroît, de l'asymétrie de cette distribution.

Cette propriété de la distribution d'échantillonnage de l'estimateur du niveau d'habileté est importante puisqu'elle assure une adéquation avec une distribution de probabilité dont les caractéristiques sont bien connues, soit la distribution normale.

9.1.6 Proportion de bonnes réponses

La moyenne de la proportion de bonnes réponses est toujours égale à 0,50, quelle que soit l'erreur type retenue pour la règle d'arrêt. Cela n'est toutefois pas vrai sur toute l'étendue du niveau d'habileté, car la proportion de bonnes réponses augmente avec l'accroissement du niveau d'habileté. Avec la diminution de l'erreur type retenue pour la règle d'arrêt, cette relation entre la proportion de bonnes réponses et le niveau d'habileté devient moins importante. La proportion de bonnes réponses tend alors de plus en plus à tourner autour de 0,50 sur toute l'étendue du niveau d'habileté. En effet, lorsque l'erreur type retenue diminue, le nombre d'items administrés augmente avec pour conséquence directe de provoquer une baisse de l'écart type de la proportion de bonnes réponses, l'écart type de la proportion de bonnes réponses étant en relation inversement proportionnelle avec la racine carrée du nombre d'items administrés. Au tableau 9.1, nous indiquons que lorsque l'erreur type retenue est égale à 0,40 la proportion de bonnes réponses varie entre 0,00 et 0,92. Elle se situe entre 0,27 et 0,72 quand la règle d'arrêt retenue est de 0,20.

Nous croyons, cependant, que la proportion de bonnes réponses pourrait éventuellement devenir constamment égale à 0,50, mais seulement à la condition que l'erreur type retenue pour la règle d'arrêt soit très petite, soit à des valeurs trop petites pour être utilisées dans la pratique.

Cette dernière constatation nous semble assez importante puisque, en testing adaptatif, on peut avoir l'impression que la proportion de bonnes réponses devrait être toujours égale à 0,50. Certains auteurs (Kinsbury et Weiss, 1983, p. 262 ; Weiss, 1985, p. 776), de par leur manière de présenter ce qu'est un test adaptatif, nous le laissent croire. Nos résultats indiquent toutefois que la proportion de bonnes réponses peut s'éloigner considérablement de 0,50, même lorsque l'erreur type retenue pour la règle d'arrêt est égale à 0,20, donc assez peu importante.

9.1.7 Nombre d'items administrés

Comme nous le rappelle le tableau 9.1, lorsque l'erreur type retenue est égale à 0,40, le nombre nécessaire d'items à administrer varie entre 9 et 14 ; il est donc, à ce moment, très petit. Quand l'erreur type retenue pour la règle d'arrêt est de 0,30 et de 0,20, ce nombre varie respectivement entre 16 et 24 et entre 35 et 48. Le nombre d'items administrés ne peut donc pas considéré comme négligeable quand l'erreur type retenue est plutôt petite, ici 0,20. Avec une erreur type retenue pour la règle d'arrêt inférieure

à 0,20, le nombre d'items à administrer pourrait devenir éventuellement assez important. De plus, le nombre d'items administrés augmente lorsque le niveau d'habileté affiche des valeurs extrêmes.

Il est intéressant de comparer les résultats que nous avons obtenus avec ceux présentés dans des recherches sur la distribution d'échantillonnage de l'estimateur du niveau d'habileté décrites auparavant (chapitre 6).

Dodd, Koch et de Ayala (1993) obtiennent un nombre moyen d'items administrés qui varie entre 13,61 et 18,23 lorsque l'erreur type retenue pour la règle d'arrêt est égale à 0,30 (tableau 6.2). Selon nos résultats, les minimum et maximum sont toutefois de 16 et de 24 respectivement. Dans la recherche de Dodd et *al.*, le nombre maximal d'items administrés est fixé à 20, ce qui explique en partie qu'ils n'ont pas pu obtenir un nombre maximal de 24, comme dans notre recherche. Il faut aussi souligner qu'ils ont utilisé pour la modélisation de la réponse à l'item le modèle de crédit partiel de Master (1982). Cette modélisation peut nécessiter l'administration de moins d'items pour obtenir la même valeur de l'erreur type de l'estimateur du niveau d'habileté. De plus, le plus petit nombre moyen d'items administrés est obtenu lorsque la banque d'items est composée d'items dont la fonction d'information est, selon leur terminologie, légèrement surélevée. Ces deux derniers points peuvent expliquer en partie pourquoi ces auteurs obtiennent une valeur moindre du nombre minimal d'items administrés. Ne perdons pas de vue, non plus qu'il s'agit, dans leur cas, de la valeur moyenne du nombre d'items administrés ; il

ne s'agit donc pas vraiment du nombre minimum d'items administrés.

Nos résultats sont aussi très comparables à ceux obtenus par Bock et Mislevy (1982, p. 438). Lorsque la règle d'arrêt selon l'erreur type est égale à 0,40, 0,30 et 0,20 et qu'une modélisation logistique à deux paramètres est utilisée, ils observent qu'en moyenne 9, 17 et 37 items doivent être administrés. Nous obtenons, pour notre part, des valeurs de 9,53, 16,76 et de 36,63. Comme eux, de plus, nous observons aussi que la distribution de probabilité du nombre d'items administrés est asymétrique.

9.1.8 Recommandations quant à l'application de la règle d'arrêt selon l'erreur type

Lorsque l'erreur type retenue pour la règle d'arrêt est égale ou inférieure à 0,40 il n'est pas du tout nécessaire de tenir compte de l'asymétrie et de la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté. C'est aussi à ce moment que la distribution de probabilité de l'estimateur du niveau d'habileté est à toutes fins utiles identique à celle du niveau d'habileté.

Nous jugeons qu'il est nécessaire de toujours appliquer la correction de Bock et Mislevy dans le but de réduire le biais de l'estimateur du niveau d'habileté sur toute l'étendue du niveau d'habileté. Il faut indiquer, cependant, que cette correction ramène le biais de l'estimateur du niveau d'habileté pratiquement à zéro seulement sur une étendue du niveau

d'habileté limitée à l'intervalle [-3,00, 3,00] quand l'erreur type retenue pour la règle d'arrêt est inférieure ou égale à 0,40.

Lorsque que nous désirons obtenir une erreur type strictement constante sur tout le continuum du niveau d'habileté, comme le postule la théorie classique des tests, il est toutefois nécessaire que la valeur retenue pour la règle d'arrêt soit égale ou inférieure à 0,20.

Considérant ces recommandations, nous préconisons l'utilisation d'une valeur de l'erreur type d'au plus 0,40 accompagnée de la correction de Bock et Mislevy lorsque la règle d'arrêt selon l'erreur type est appliquée. À ce moment, du moins dans les conditions relatives à cette recherche, nous sommes assurés de l'équivalence des distributions de probabilité du niveau d'habileté et de l'estimateur du niveau d'habileté ainsi que d'une valeur à peu près nulle du biais de l'estimateur du niveau d'habileté quel que soit le niveau d'habileté. Si, de plus, il devient nécessaire de satisfaire à la condition que l'erreur type de l'estimateur du niveau d'habileté soit constante sur toute l'étendue du niveau d'habileté, l'erreur type retenue pour la règle d'arrêt doit être d'au plus 0,20.

9.2 Règle d'arrêt selon le nombre d'items administrés

Il convient de faire ici une synthèse de nos commentaires ainsi que des recommandations quant à l'application de la règle d'arrêt selon le nombre d'items administrés. Comme

point de départ, nous présentons un bref résumé des résultats présentés à la section 8.3, au tableau 9.2. Les diverses valeurs du nombre d'items administrés retenu pour la règle d'arrêt permettent de se rapprocher des situations les plus courantes observées dans la littérature sur les tests adaptatifs et de la pratique de leur administration.

Comme nous l'avons fait à la section précédente, nous terminons la discussion des résultats qui ont trait à la règle d'arrêt selon le nombre d'items administrés par l'exposition de prescriptions générales qui découlent de nos constatations.

Tableau 9.2
Minimums (↓) et maximums (↑) des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté lorsque la règle d'arrêt selon le nombre d'items administrés est utilisée

Items	$EAP(\theta)$		$S_{EAP(\theta)}$		$EAP(\theta)-\theta$		$a3_{EAP(\theta)}$		$a4_{EAP(\theta)}$		Bonnes réponses	
	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑
10	-2,94	2,94	0,36	0,47	-1,13	1,25	-0,29	0,29	-0,30	0,24	0,00	1,00
15	-3,38	3,25	0,30	0,38	-0,97	1,08	-0,22	0,23	-0,36	0,39	0,07	0,93
20	-3,35	3,29	0,26	0,34	-0,86	1,00	-0,21	0,23	-0,26	0,18	0,15	0,85
25	-3,16	3,35	0,23	0,28	-0,70	0.97	-0,19	0,17	-0,15	0,14	0,20	0,80
30	-3,10	3,29	0,21	0,26	-0,74	0,84	-0,16	0,16	0,00	0,11	0,20	0,77
40	-3,13	3,20	0,19	0,23	-0,59	0,63	-0,12	0,11	0,04	0,08	0,22	0,72
60	-3,40	3,16	0,15	0,17	-0,45	0,47	-0,10	0,08	0,01	0,06	0,30	0,70

9.2.1 Estimateur du niveau d'habileté, EAP(θ)

Quel que soit le nombre d'items administrés retenu pour la règle d'arrêt, la moyenne de l'estimateur du niveau d'habileté est égale à la moyenne du niveau d'habileté. Tout comme nous l'avions remarqué lors de l'étude de la règle d'arrêt selon l'erreur type, l'écart type de l'estimateur du niveau d'habileté s'éloigne substantiellement de l'écart type du niveau d'habileté ($\sigma = 0,98$) quant le nombre d'items administrés est peu important. Quand le nombre d'items administrés est égal à 13, l'écart type de l'estimateur du niveau d'habileté est de 0,93, soit une différence de 0,05. Pour obtenir une différence de 0,02, il est nécessaire d'administrer au moins 25 items.

Nous avons aussi noté que l'asymétrie et la kurtose de la distribution de probabilité de l'estimateur du niveau d'habileté affichent toutes deux des valeurs peu importantes quand le nombre d'items administrés est supérieur à trois. À partir du 8^{e} item administré, déjà la kurtose est d'au plus -0,09.

Ce n'est toutefois que lorsque 13 items ont été administrés que les minimums et maximums de l'estimateur du niveau d'habileté se rapprochent des minimums et maximums du niveau d'habileté. Nous jugeons utile de souligner que, lorsque l'erreur type retenue était de 0,40, la règle d'arrêt selon l'erreur type exigeait l'administration d'au moins 9 et d'au plus 14 items pour satisfaire à cette condition (tableau 9.1). L'utilisation de la règle d'arrêt selon l'erreur type favorise donc une économie du nombre

d'items administrés.

Selon nous, l'administration de 10 items seulement, comme le suggèrent Hoijtink et Boomsma (1995, p. 68), est nettement insuffisante puisque les minimums et maximums obtenus de l'estimateur du niveau d'habileté sont alors inférieurs à 3,00 en valeur absolue.

9.2.2 Erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$

Nous avons observé que l'erreur type de l'estimateur du niveau d'habileté diminue avec l'augmentation du nombre d'items administrés et que l'écart type de cette erreur type est généralement presque nul quel que soit le nombre d'items administrés. L'erreur type de l'estimateur du niveau d'habileté devient de plus en plus constante sur toute l'étendue du niveau d'habileté avec l'augmentation du nombre d'items administrés. Déjà, à partir de l'administration du 25^{e} item, les minimums et maximums de l'erreur type ne diffèrent que de 0,05. Ces minimums et maximums ne deviennent toutefois jamais égaux, contrairement à ce que nous avons observé en ce qui a trait à la règle d'arrêt selon l'erreur type, même après l'administration de 60 items.

Il nous semble pertinent de comparer les valeurs que nous avons obtenues avec celles que l'on retrouve dans la littérature discutée au chapitre 6. Ainsi, Hoijtink et Boomsma

(1996, p. 317-322) obtiennent des valeurs de l'erreur type de l'estimateur du niveau d'habileté de 0,40 et de 0,39 lorsqu'ils appliquent une méthode bayésienne d'estimation du niveau d'habileté (ML), que le niveau d'habileté est égal à 3,00 et que le nombre d'items administrés est respectivement de 15 et 25. Pour notre part, pour le même nombre d'items administrés, nous obtenons des valeurs maximales de l'erreur type égales à 0,38 et 0,28. Les valeurs que nous obtenons sont donc inférieures à celles que présentent Hoijtink et Boomsma. Ces derniers ont toutefois dû appliquer une stratégie pour pallier le fait que la méthode d'estimation du niveau d'habileté qu'ils ont utilisée ne permettait pas d'estimer le niveau d'habileté lorsque toutes les réponses sont bonnes ou qu'elles sont toutes mauvaises. À ce moment, ils ont tout simplement assigné une valeur de -5,00 à l'estimateur du niveau d'habileté lorsque toutes les réponses étaient mauvaises et une valeur de 5,00 dans le cas contraire. Cette stratégie augmente de manière exagérée l'étendue de l'estimateur du niveau d'habileté et a probablement provoqué la surestimation de l'erreur type.

Pour leur part, Vispoel, Wang et Bleuler (1997, p. 49, 53 et 56) présentent des valeurs de la fidélité qui, par transformation, soit la fonction inverse associée à l'équation 8.1, nous permettent de calculer l'erreur type de l'estimateur du niveau d'habileté. Les valeurs de l'erreur type de l'estimateur du niveau d'habileté ainsi calculées diffèrent généralement très peu de celles que nous avons obtenues. Ainsi, quand le nombre d'items administrés est supérieur à 10, la différence entre nos résultats et les leurs n'est que de $\pm 0,07$. Lorsque le nombre d'items administrés est égal ou inférieur à 10, seul

un test adaptatif sur trois affiche des valeurs qui peuvent être aussi inférieures que de 0,16 avec les valeurs que nous obtenons dans le cadre de notre recherche. Les deux autres tests adaptatifs analysés par ces auteurs présentent des valeurs de l'erreur type qui se rapprochent de celles que nous observons ici.

Enfin, comme nous l'avons vu à la section 8.3.2, l'erreur type de l'estimateur du niveau d'habileté n'est pas constante sur toute l'étendue du niveau d'habileté quel que soit le nombre d'items administrés. Il n'est donc pas possible, contrairement à ce que permet la règle d'arrêt selon l'erreur type, d'obtenir une même précision de l'estimateur du niveau d'habileté à tous les niveaux d'habileté tel que le postule la théorie classique des tests. C'est du moins ce que nous avons vérifié avec l'administration d'au plus 60 items. Au delà de ce nombre, la variation de l'erreur type deviendrait éventuellement très petite et, de ce fait, l'erreur type serait alors pratiquement constante sur toute l'étendue du niveau d'habileté.

9.2.3 Erreur de mesure de l'estimateur du niveau d'habileté et biais de l'estimateur du niveau d'habileté, $BIAIS_{EAP(\theta)}$

Comme le montre le tableau 8.19, quand nous considérons toute l'étendue du niveau d'habileté, l'erreur de mesure du niveau d'habileté ne dépasse pas 0,01 en valeur absolue indépendamment du nombre d'items administrés. L'écart type de l'erreur de mesure du

niveau d'habileté diminue avec l'augmentation du nombre d'items administrés et est à toutes fins utiles égal à l'erreur type de l'estimateur du niveau d'habileté ; la différence est d'au plus 0,01 en valeur absolue. La distribution de probabilité de l'erreur de mesure s'éloigne légèrement d'une distribution normale à certaines valeurs du nombre d'items administrés. La kurtose peut alors affecter, quoique de façon peu importante, la distribution de probabilité de l'erreur de mesure du niveau d'habileté ; cette kurtose se situe entre -0,23 et 0,19.

Le biais de l'estimateur du niveau d'habileté n'est toutefois pas nul sur toute l'étendue du niveau d'habileté quel que soit le nombre d'items administrés. Ainsi, lorsque le niveau d'habileté présente des valeurs extrêmes, le biais peut devenir assez important ; l'estimateur du niveau d'habileté a alors tendance à surestimer le niveau d'habileté quand ce dernier est très faible et à le sous-estimer quand il est très élevé. D'ailleurs, les résultats que nous obtenons invalident les propos de Hoijtink et Boomsma (1995, p. 68) qui laissent croire que le biais de l'estimateur du niveau d'habileté est de peu d'importance lorsque seulement 10 items sont administrés. Au contraire, le biais peut être assez important quel que soit le nombre d'items administrés.

L'utilisation de la correction de l'estimateur du niveau d'habileté par la stratégie de Bock et Mislevy (1982) est de mise pour diminuer le biais. Le biais de l'estimateur du niveau d'habileté devient alors pratiquement nul sur toute l'étendue du niveau d'habileté si au moins 10 items sont administrés et que le niveau d'habileté est compris à l'intérieur de

l'intervalle [-3,00, 3,00]. Toutefois, comme nous l'avons souligné en ce qui a trait à la règle d'arrêt selon l'erreur type, la variabilité de l'estimateur du niveau d'habileté augmente lorsque la correction de Bock et Mislevy est appliquée. Cette augmentation devient cependant négligeable quand l'erreur type de l'estimateur du niveau d'habileté est égale ou inférieure à 0,20. Cette dernière remarque implique que l'administration d'au moins 40 items est nécessaire pour assurer que l'erreur type de l'estimateur du niveau d'habileté ne soit pas, à toutes fins utiles, affectée.

9.2.4 Asymétrie de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$

Tout comme dans le cas de la règle d'arrêt selon l'erreur type, la distribution de probabilité de l'asymétrie, ainsi que de la kurtose, associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté n'ont pas reçu d'attention dans la littérature consultée. L'étude de ces statistiques en est d'autant plus importante qu'il s'agit d'une lacune.

L'asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, comme l'indique le tableau 9.2, affiche des minimums et maximums qui varient entre -0,29 et 0,29. L'asymétrie est donc peu importante, mais est tout de même présente. Ces minimums et maximums diminuent en importance avec l'augmentation du nombre d'items administrés retenu pour la règle d'arrêt. Lorsque 60 items sont administrés,

l'asymétrie varie entre -0,10 et 0,08. L'asymétrie a aussi tendance à augmenter quand le niveau d'habileté s'éloigne de la moyenne. Elle tend à être négative quand le niveau d'habileté est inférieur à la moyenne et à être positive quand le niveau d'habileté est supérieur à la moyenne. Même lorsque peu d'items sont administrés, dix items par exemple, cette variation de l'asymétrie en fonction du niveau d'habileté est peu importante et n'affecte que très peu l'interprétation de la distribution d'échantillonnage de l'estimateur du niveau d'habileté.

9.2.5 Kurtose de l'estimateur du niveau d'habileté, $a4_{EAP(0)}$

La kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté varie entre -0,41 et 0,26. Plus le nombre d'items administrés est grand, moins les minimums et maximums de la kurtose sont importants. Quand le nombre d'items administrés retenu pour la règle d'arrêt est égal à 60, la kurtose varie en fait très peu, soit entre 0,01 et 0,06. Nous en concluons que l'interprétation de la distribution d'échantillonnage de l'estimateur du niveau d'habileté est peu affectée par sa kurtose quel que soit le nombre d'items administrés.

C'est aux valeurs se situant autour de la moyenne du niveau d'habileté que nous retrouvons les valeurs les plus importantes de la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Toutefois, la kurtose est plutôt

constante sur toute l'étendue du niveau d'habileté. Nous en concluons donc que les valeurs observées de la kurtose, même combinées avec celles de l'asymétrie, affectent très peu les interprétations rattachées à une distribution de probabilité normale. L'adéquation de la distribution d'échantillonnage de l'estimateur du niveau d'habileté à une distribution de probabilité normale est donc, à toutes fins utiles, assurée.

9.2.6 Proportion de bonnes réponses

Quel que soit le nombre d'items administrés retenu pour la règle d'arrêt, la moyenne de la proportion de bonnes réponses est égale à 0,50 $\pm$0,01. Cependant la proportion de bonnes réponses est de plus en plus grande avec l'augmentation du niveau d'habileté. Elle n'est donc pas constamment égale à 0,50 sur toute l'étendue du niveau d'habileté et affiche un minimum de 0,00 ainsi qu'un maximum de 1,00.

Quand le nombre d'items administrés devient plus important, la proportion de bonnes réponses présente des minimums et maximums qui se rapprochent de plus en plus de 0,50. La proportion de bonnes réponses devient alors plus constante sur toute l'étendue du niveau d'habileté. Il semble bien toutefois que, pour que nous puissions obtenir une proportion de bonnes réponses constante de 0,50 sur toute l'étendue du niveau d'habileté, le nombre d'items administrés devrait être assez important. Cette éventualité serait d'ailleurs à vérifier par l'augmentation du nombre d'items administrés dans l'application

de cette règle d'arrêt.

L'écart type de la proportion de bonnes réponses, pour sa part, est identique à celui d'une proportion et peut ainsi être calculé de la même façon. La distribution de probabilité de la proportion de bonnes réponses est symétrique quel que soit le nombre d'items administrés. De plus, à partir du 6^{e} item administré, la valeur de la kurtose de cette distribution, constamment platykurtique, n'est que de -0,24. La kurtose semble se rapprocher de 0,00 avec l'augmentation du nombre d'items administrés.

9.2.7 Recommandations quand à l'application de la règle d'arrêt selon le nombre d'items administrés

Certaines recommandations peuvent maintenant être faites quant à l'application de la règle d'arrêt selon le nombre d'items administrés. Il est à noter que celles-ci sont en étroite relation avec celles qui ont été faites à propos de la règle d'arrêt selon l'erreur type.

Nous recommandons de fixer la règle d'arrêt selon le nombre d'items administrés à un minimum de 13 items pour permettre d'obtenir une distribution de probabilité de l'estimateur du niveau d'habileté équivalente à celle du niveau d'habileté.

Nous recommandons également l'application de la correction proposée par Bock et Mislevy (1982) pour ramener le biais de l'estimateur du niveau d'habileté à zéro pour

autant que le niveau d'habileté soit compris dans l'intervalle [-3,00, 3,00] et que le nombre d'items administrés soit au moins égal à 10. Nous suggérons toutefois l'application de cette correction seulement lorsque le nombre d'items administrés est égal ou supérieur à 40. Une telle pratique est recommandée dans le but d'éviter une augmentation trop importante de l'écart type associé à la distribution de probabilité de l'estimateur du niveau d'habileté ainsi que de l'erreur type associée à la distribution d'échantillonnage de l'estimateur du niveau d'habileté.

En conclusion, la règle d'arrêt selon le nombre d'items administrés devrait être appliquée seulement si au moins 13 items sont administrés. Il est toutefois préférable d'administrer un minimum de 40 items pour obtenir un biais de l'estimateur du niveau d'habileté de peu d'importance sur une étendue suffisamment large du niveau d'habileté, soit de -3,00 à 3,00, sans trop affecter les distributions de probabilité et d'échantillonnage de l'estimateur du niveau d'habileté.

10. Conclusion

Nous concluons cette recherche par la présentation, en premier lieu, d'un rappel de l'objectif de recherche, des grands thèmes et des propositions de développement qui ont été abordés. Ensuite, nous confirmons l'atteinte de l'objectif de recherche et nous exposons les recommandations quant à l'application des règles d'arrêt étudiées. Enfin, de nouvelles pistes de recherche sont proposées.

10.1 Rappel

L'objectif de cette recherche était d'étudier l'impact de la variation de la règle d'arrêt selon le nombre d'items administrés et de la règle d'arrêt selon la détermination a priori de l'erreur type de l'estimateur du niveau d'habileté sur la distribution de probabilité des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif. Nous désirions ainsi apporter des prescriptions quant aux critères d'application de ces règles d'arrêt. Le premier chapitre avait pour but de nous introduire à cet objectif de recherche.

Le chapitre deux présentait la description et les étapes d'un test adaptatif tout en abordant les développements qui ont présidé à son évolution en éducation. Différentes formes de tests adaptatifs papier crayon y ont été décrites : test de Binet, test à deux étapes, test à

niveaux flexibles, test pyramidal et test stratifié. La transition vers l'administration de tests adaptatifs par ordinateur utilisant des modélisations issues de la théorie de la réponse à l'item a été ensuite abordée.

La théorie de la réponse à l'item et ses différentes modélisations ont par la suite été l'objet spécifique du chapitre trois : modèles logistiques à un, à deux, à trois et à quatre paramètres. Les méthodes d'estimation du niveau d'habileté étaient présentées au chapitre suivant : méthodes de vraisemblance maximale, bayésienne de maximisation a posteriori et de l'espérance a posteriori. Les avantages de chacune de ces méthodes y étaient considérées.

Le déroulement d'un test adaptatif basé sur la théorie de la réponse à l'item, et plus spécifiquement les règles de départ, de sélection des items et d'arrêt étaient ensuite abordés au chapitre cinq. Une place plus importante était accordée à la description des règles d'arrêt puisqu'elles devaient être plus spécifiquement l'objet de la recherche.

La présentation détaillée des études qui ont apporté une contribution à la description des caractéristiques de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction des règles d'arrêt retenues a été l'objet du chapitre six. Des travaux qui permettent de comparer les caractéristiques de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction des règles d'arrêt ainsi que des recherches sur les caractéristiques de la distribution

d'échantillonnage de l'estimateur du niveau d'habileté lorsque le nombre d'items administrés est petit, comme en testing adaptatif, y étaient abordés. Ce chapitre se terminait sur une présentation de précisions au sujet de l'objectif de recherche, soit l'étude de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en fonction de la variation des critères dans les règles d'arrêt basées sur le nombre d'items administrés et sur l'erreur type de l'estimateur du niveau d'habileté en testing adaptatif.

Enfin, la méthodologie proposée pour atteindre l'objectif du projet de recherche ainsi que la méthode d'analyse des résultats étaient présentées au chapitre sept. Selon Thissen et Mislevy, le choix d'une règle d'arrêt a un impact sur la valeur obtenue de l'estimateur du niveau d'habileté ; ils soulignent qu'il serait utile d'étudier, par une simulation informatisée, les effets des règles d'arrêt sur cet estimateur en testing adaptatif. La présente recherche avait pour objectif de vérifier cette hypothèse. L'utilisation d'une simulation a été retenue puisqu'aucune solution analytique exacte n'existe et que cette stratégie permet la manipulation, qui n'est pas possible à réaliser dans des conditions réelles, de différents facteurs simultanément. L'analyse des résultats permet de mettre en relation la variation des valeurs des critères des règles d'arrêt avec différentes statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Plus précisément, nous étudions l'impact du nombre d'items fixés à l'avance, dans la première règle d'arrêt, et de l'erreur type de l'estimateur du niveau d'habileté, dans la seconde règle, sur la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Les résultats obtenus nous ont permis d'identifier les conditions d'utilisations

optimales de ces règles : nombre minimal d'items à administrer et erreur type maximale de l'estimateur du niveau d'habileté à atteindre.

10.2 Atteinte de l'objectif de recherche et recommandations

Les résultats que nous avons obtenus, présentés au chapitre huit et discutés au chapitre neuf, nous permettent de décrire, comme prévu, la distribution de probabilité de quelques statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Ils nous permettent également d'émettre certaines recommandations. Les résultats de cette recherche s'avéreront utiles dans l'application des tests adaptatifs par ordinateur.

Ainsi, quelle que soit la règle d'arrêt utilisée, nous recommandons d'appliquer la correction de Bock et Mislevy dans le but de diminuer de manière importante le biais de l'estimateur du niveau d'habileté à zéro sur toute l'étendue du niveau d'habileté.

De façon générale, il est préférable de ne pas utiliser cette règle d'arrêt avec une erreur type retenue supérieure à 0,40. Pour que l'augmentation de la variabilité de l'estimateur du niveau d'habileté provoquée par la correction de Bock et Mislevy soit négligeable, il est toutefois recommandé que l'erreur type retenue pour la règle d'arrêt soit égale ou inférieure à 0,20. Il est aussi recommandé de ne pas appliquer cette règle d'arrêt avec

une erreur type retenue supérieure à 0,20 si l'on désire que l'erreur type de l'estimateur du niveau d'habileté soit strictement constante sur toute l'étendue du niveau d'habileté.

La règle d'arrêt selon le nombre d'items administrés devrait être appliquée seulement si au moins 13 items sont administrés. Il est toutefois préférable d'administrer au moins 40 items pour obtenir un biais négligeable de l'estimateur du niveau d'habileté sur une étendue suffisamment large du niveau d'habileté, soit de -3,00 à 3,00, sans trop affecter les distributions de probabilité et d'échantillonnage de l'estimateur du niveau d'habileté.

Le nombre d'items administré, que la règle d'arrêt soit basée sur l'erreur type de l'estimateur du niveau d'habileté ou sur le nombre d'items administrés, peut paraître important lorsqu'on cherche à ce que l'erreur type de l'estimateur du niveau d'habileté soit constante sur toute l'étendue du niveau d'habileté. Bock et Mislevy (1982), soulignent toutefois que cette propriété de l'estimateur du niveau d'habileté n'est vraiment utile que lorsqu'on désire appliquer des analyses statistiques standards à l'estimateur du niveau d'habileté ou que ce dernier est utilisé à des fins de classification et de sélection. La constance de l'erreur type de l'estimateur du niveau d'habileté n'est généralement pas requise dans les applications usuelles des tests en éducation et, de ce fait, le nombre d'items administrés peut alors n'être que de 13 et une valeur de 0,40 de l'erreur type retenue pour la règle d'arrêt peut être utilisée. Il nous faut relever qu'il s'agit là d'une économie appréciable quant au nombre d'items administrés par rapport à un test papier crayon fixe et invariable.

Nous tenons aussi à souligner que si une modélisation logistique à deux paramètres avait été utilisée plutôt que la modélisation logistique à un paramètre, le nombre d'items à administrer pour obtenir une même erreur type aurait pu être sensiblement supérieur à celui que nous avons obtenu. En fait, si la valeur moyenne du paramètre de discrimination, a, du modèle logistique à deux paramètres était inférieure à 1,00, la quantité moyenne d'information apportée par les items aurait été moindre ; en conséquence, le nombre d'items administrés aurait été plus important. La même remarque s'applique en ce qui a trait au modèle logistique à trois paramètres où l'augmentation de la valeur moyenne du paramètre de pseudo chance, c, occasionne aussi une diminution la quantité d'information fournie par les items.

Enfin, notons que l'utilisation d'une modélisation de la réponse à l'item qui tient compte de réponses polytomiques permettrait éventuellement de diminuer de façon importante le nombre d'items administrés ; ce qui serait pratique lorsque nous devons garder constante l'erreur type sur tout le continuum du niveau d'habileté. Pour le moment, toutefois, les caractéristiques de la distribution d'échantillonnage de l'estimateur du niveau d'habileté sont peu documentées lorsque de telles modélisations sont appliquées en testing adaptatif. Une étude comparable à celle que nous avons réalisée ici devrait, à cet effet, être effectuée.

10.3 Validité de l'estimateur du niveau d'habileté en testing adaptatif

Dans notre étude, nous ne nous sommes pas penchés sur les problèmes éventuels de validité associés à l'administration d'un test adaptatif. Qu'il s'agisse de validité apparente (*face validity*), de validité de contenu, de validité prédictive ou de validité de construit, la pertinence des inférences faites à partir des résultats obtenus à un test adaptatif peut présenter des limites.

En ce sens, Wainer, Dorans, Green, Mislevy, Steinberg et Thissen (1990, p. 260-261) soulignent que le nombre d'items administrés dans un test adaptatif peut paraître insuffisant aux yeux de certaines personnes au point de susciter des poursuites légales de leur part. Il s'agit d'un problème de validité apparente qui exige l'établissement de normes reconnues de la part des organismes qui administrent les tests adaptatifs.

La structure de la banque d'items doit aussi faire en sorte que le contenu des items soit bien équilibré de manière à ce que le domaine des situations en lien avec l'habileté concernée soit bien couvert par la banque d'items. Comme le soulignent Thissen et Mislevy (1990, p. 113), il faut éviter de se retrouver dans une position où des contenus particuliers sont sur-représentés au détriments d'autres contenus, principalement quant ces contenus sont reliés étroitement aux paramètres des items. Certaines stratégies de sélection des items permettent d'obtenir un meilleur équilibre du contenu des items (Hetter et Sympson, 1997, p. 141-144 ; Kingsbury et Houser, 1999, p. 104-105).

Évidemment, plus la taille de la banque d'items est importante, plus la représentativité de contenu est renforcée.

La validité de construit peut être affectée lorsque les paramètres des items qui composent la banque d'items ont été obtenus suite à l'administration d'une version papier crayon du test (Wainer et *al.*, 1990, p. 258-260). On doit alors être prudent quant à l'équivalence des versions papier crayon et des versions adaptatives des tests. Cela est particulièrement vrai à l'intérieur de populations où les individus ont été peu en contact avec les ordinateurs.

Enfin, l'utilisation d'une modélisation de la réponse à l'item inappropriée peut aussi affecter la validité de construit. Par exemple, il est tout à fait inadéquat d'utiliser le modèle logistique à un paramètre lorsqu'en fait le construit est multidensionnel (Wainer, et *al.*, 1990, p. 240-242).

10.4 Nouvelles pistes de recherche

Pour terminer, soulignons la nécessité d'étudier d'autres caractéristiques importantes associées à l'application de la méthode de l'espérance a posteriori en tant que stratégie d'estimation du niveau d'habileté en testing adaptatif. En ce sens, il serait pertinent d'étudier plus à fond la variation du nombre de points de quadrature ainsi que de

l'intervalle d'intégration dans cette méthode d'estimation. Nos essais préliminaires laissent croire qu'il serait très intéressant de modifier en cours d'administration du test adaptatif les limites d'intégration en fonction de l'estimateur provisoire du niveau d'habileté, principalement quand cet estimateur provisoire affiche des valeurs extrêmes : l'erreur type de l'estimateur du niveau d'habileté et le biais de celui-ci pourraient éventuellement être sensiblement diminués.

En ce qui a trait au nombre de points de quadrature, les observations que nous obtenons au chapitre sept (section 7.3), nous laissent supposer qu'il n'est pas nécessaire d'utiliser le même nombre de points de quadrature quel que soit le nombre d'items administrés. En ce sens, une économie de temps de calcul pourrait être réalisée en utilisant seulement 10 points de quadrature au début du test, pour ensuite augmenter ce nombre de points en fonction du nombre d'items administrés.

Des études préliminaires permettent aussi d'envisager de faire diminuer plus rapidement l'erreur type de l'estimateur du niveau d'habileté, ainsi que le biais qui lui est associé, en variant la moyenne de la distribution a priori en cours d'administration du test en fonction de la valeur de l'estimateur provisoire du niveau d'habileté ; c'est d'ailleurs une stratégie qui est signalée par Vispoel (1999, p. 168). Considérant le lien qui existe entre le biais et l'asymétrie associée à l'estimateur du niveau d'habileté, cette dernière pourrait être alors diminuée. Ainsi, la distribution d'échantillonnage de l'estimateur du niveau d'habileté pourrait éventuellement présenter les caractéristiques d'une distribution

normale à partir de l'administration d'un nombre encore mois important d'items. Conséquemment, cette stratégie devrait aussi permettre de diminuer l'impact du premier item administré et de l'estimateur a priori sur l'estimateur final du niveau d'habileté.

Selon ces mêmes études préliminaires, en appliquant la correction de Bock et Mislevy en cours d'administration du test adaptatif, il serait aussi possible de diminuer à la fois l'erreur type et le biais de l'estimateur du niveau d'habileté sur toute l'étendue du niveau d'habileté.

Enfin, l'objet de notre recherche pourrait être élargi aux modélisations logistiques à deux et trois paramètres de la réponse à l'item. Dans ce cas, pour nous permettre d'appliquer une méthodologie de recherche similaire à celle que nous avons utilisée ici, la valeur des paramètres de discrimination et de pseudo-chance pourrait être déterminée aléatoirement pour chaque item administré selon des distributions de probabilité appropriées à ces paramètres : par exemple Baker (1992, p. 203-207) suggère l'utilisation d'une distribution lognormale pour le paramètre de discrimination ($0 \leq a \geq \infty$) et d'une distribution beta pour le paramètre de pseudo-chance ($0 \leq c \geq 1$).

Références

Ackerman, T.A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied measurement in education*, *7*-4, 255-278.

Almond, R.G., Mislevy, R.J. (1999). Graphical models and computerized adaptive testing. *Applied psychological measurement*, *23*-3, 223-237.

Andersen, E.B. (1995). Polytomous Rasch models and their estimation. *In* G.H. Fischer et I.W. Molenaar (Éds) : *Rasch models - Foundations, recent developments, and applications*. New York : Springer-Verlag.

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., 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.

Baker, F.B. (1992). *Item response theory : parameter estimation techniques*. New York : Marcel Dekker.

Baker, F.B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied psychological measurement*, *22*-2, 153-169.

Bejar, I.I. (1993). A generative approach to psychological and educational measurement. *In* N. Fredericksen, R.J. Mislevy et I.I. Bejar (Éds) : *Test theory for a new generation of tests*. Hillsdale : Lawrence Erlbaum Associates.

Bennett, R.E. (1999). Using new technology to improve assessment. *Educational Measurement : Issues and Practice*, *18*-3, 5-12.

Bertrand, R. (1986). *Pratique de l'analyse statistique des données*. Sillery : Presses de l'Université du Québec.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *In* F.M. Lord et M.R. Novick (Éds) : *Statistical theories of mental test scores*. Reading : Addison-Wesley.

Blais, J.-G. (1987). Effets de la violation du postulat d'unidimensionnalité dans la théorie des réponses aux items. Thèse de doctorat non publiée. Montréal :

Université de Montréal.

Blais, J.-G., Ajar, D. (1992). Théorie des réponses aux items et modélisation. *Mesure et évaluation en éducation. 14*-4, 5-18.

Bock, D.R. (1997a). A brief history of item response theory. *Educational measurement : issues and practice, 16*-4, 21-33.

Bock, R.D. (1997b). The nominal categories model. *In* W.J. van der Linden et R.K. Hambleton (Eds.) : *Handbook of modern item response theory*. New York : Springer-Verlag.

Bock, R.D., Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a micro computer environment. *Applied psychological measurement, 6*-4, 431-444.

Brennan, R.L. (1992). *Elements of generalizability theory*. Iowa City, Iowa : American College Testing.

Bunderson, C.V., Inouye, D.K., Olsen, J.B. (1989). The four generations of computerized educational measurement. *In* R.L. Linn (Éd.) : *Educational measurement*. New York : Macmillan (3^{e} édition).

Camilli, G. (1994). Origin of the scaling constant d = 1,7 in item response theory. *Journal of educational and behavioral statistics*, *19*-3, 293-295.

Chang, H.H., Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*-1, 37-52.

Chang, H.H., Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied psychological measurement*, *20*-3, 213-229.

Chang, H.H., Ying, Z. (1999). α-stratified multistage computerized adaptive testing. *Applied psychological measurement*, *23*-3, 211-222.

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

Chen, S.K., Hou, L., Fitzpatrick, S.J., 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., Thissen, D. (1997). Local dependence indexes for item pairs using item

response theory. *Journal of educational and behavioral statistics*, *22*-3, 265-289.

Cowles, M.K., Carlin, B.P. (1996). Markov chain Monte Carlo convergence diagnostics : A comparative review. *Journal of the american statistical association*, *91*-434, 883-903.

Davaud, C., Cardinet, J. (1992). La problématique des épreuves communes. *In.* D. Laveault (Éd.) : *Les pratiques d'évaluation en éducation - Textes rédigés en vue de la XV^e session d'études de l'Association pour le développement de la mesure et de l'évaluation en éducation* . Ottawa : Association pour le développement de la mesure et de l'évaluation en éducation.

Davey, T., Godwin, J., Mittelholtz, D. (1997). Developing and scoring an innovative computerized writing assessment. *Journal of educational measurement*, *34*-1, 21-41.

de Ayala, R.J. (1989). A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. *Educational and psychological measurement*, *49*-4, 789-805.

de Ayala, R.J. (1992a). The influence of dimensionality on CAT ability estimation. *Educational and psychological measurement*, *52*-3, 513-528.

de Ayala, R.J. (1992b). The nominal response model in computerized adaptive testing. *Applied psychological measurement*, *16*-4, 327-343.

de Ayala, R.J., Dodd, B.G., Koch, W.R. (1990). A simulation and comparison of flexilevel and bayesian computerized adaptive testing. *Journal of educational measurement*, *27*-3, 227-239.

de Ayala, R.J., Hertzog, M.A. (1991). The assessment of dimensionality for use in item response theory. *Multivariate behavioral research*, *26*-4, 765-792.

de Ayala, R.J., Schafer, W.D., Sava-Bolesta, M. (1995). An investigation of the standard errors of expected a posteriori ability estimates. *British journal of mathematical and statistical psychology*, *48*-2, 385-405.

Dechef, H., Laveault, D. (1999). Le testing adaptatif par ordinateur. *Psychologie et psychométrie*, *20*-2/3, 151-179.

de Gruijter, D.N.M. (1980). A two-stage testing procedure. *In* L.J.T. van der Kamp, W.F. Langerak et D.N.M. de Gruijter (Éds) : *Psychometrics for educational debates*. New York : John Wiley and Sons.

Dodd, B.G. (1990). The effect of item selection procedure and stepsize on computerized

adaptive attitude measurement using the rating scale model. *Applied psychological measurement*, *14*-4, 355-366.

Dodd, B.G., de Ayala, R.J., Koch, W.R. (1995). Computerized adaptive testing with polytomous items. *Applied psychological measurement*, *19*-1, 5-22.

Dodd, B.G., Koch, W.R., 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*, *5*3-1, 61-77.

Eggen, T.J.H.M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied psychological measurement*, *23*-3, 249-261.

Embretson, S.E. (1999). Generating items during testing : Psychometric issues and models. *Psychometrika*, *64*-4, 407-433.

Fischer, G.H. (1995). Derivations of the Rasch model. *In* G.H. Fischer et I.W. Molenaar (Éds) : *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.

Fismen, M. (1997). Exact simulation using Markov chains. Mémoire de maîtrise présenté au département des sciences mathématiques. Trondheim, Norvège : Université norvégienne des sciences et de la technologie.

Fleishman, A.I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*-4, 521-532.

Freund, J.E. et Walpole, R.E. (1980). *Mathematical statistics*. Englewood Cliffs : Prentice-Hall (3e édition).

Gitomer, D.H., Rock, D. (1993). Addressing process variables in test analysis. *In* N. Frederiksen, R.J. Mislevy et I.I. Bejar (Éds) : *Test theory for a new generation of tests*. Hillsdale : Lawrence Erlbaum Associates.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British journal of mathematical and statistical psychology*, *33*, 234-246.

Goldstein, H. (1994a). Mathematical and ideological assumptions in the modelling of test item responses. *In* D. Laveault, B.D. Zumbo, M.E. Gessaroli et M.W. Boss (Éds) : *Modern theories of measurement - Problems and issues*. Ottawa : Université d'Ottawa.

Goldstein, H. (1994b). Présupposés mathématiques et idéologiques des modèles de réponses aux items. *Mesure et évaluation en éducation*, *17*-2, 107-114.

Goldstein, H., Wood, R. (1989). Five decades of item response modelling. *British journal of mathematical and statistical psychology*, *42*, 139-167.

Green, B.F. (1990). System design and operations. *In* H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (Éds) : *Computerized adaptive testing - A primer*. Hillsdale : Lawrence Erlbaum Associates.

Hambleton, R.K., Swaminathan, H. (1987). *Item response theory : principles and applications*. Boston : Kluwer.

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

Hambleton, R.K., Zaal, J.N., Pieters, J.M.P. (1991). Computerized adaptive testing : theory, applications, and standards. *In* R.K. Hambleton et J.N. Zaal (Éds) : *Advances in educational and psychological testing - Theory and applications*. Boston : Kluwer.

Harwell, M.R. (1997). Analyzing the results of Monte Carlo studies in item response theory. *Educational and psychological measurement*, *57*-2, 266-279.

Harwell, M.R., Stone, C.A., Hsu, T.C., Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied psychological measurement*, *20*-2, 101-125.

Headrick, T.C., Sawilowsky, S.S. (1999a). Errata for "Simulating correlated multivariate non-normal distributions : Extending the Fleishman power method". *Psychometrika*, *64*-2, 251.

Headrick, T.C., Sawilowsky, S.S. (1999b). Simulating correlated multivariate non-normal distributions : Extending the Fleishman power method. *Psychometrika*, *64*-1, 25-35.

Hetter, R.D., Segall, D.O., Bloxom, B.M. (1997). Evaluating item calibration medium in computerized adaptive testing. *In* W.A. Sands, B.K. Waters et J.R. McBride (Éds) : *Computer adaptive testing - From inquiry to operation*. Washington : American Psychological Association.

Hetter, R.D., Sympson, J.B. (1997). Item exposure control in CAT-ASVAB. *In* W.A. Sands, B.K. Waters et J.R. McBride (Éds) : *Computer adaptive testing - From inquiry to operation*. Washington : American Psychological Association.

Hoijtink, H., Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. *In* G.H. Fischer et I.W. Molenaar (Éds) : *Rasch models - Foundations, recent developments, and applications*. New York : Springer-Verlag.

Hoijtink, H., Boomsma, A. (1996). Statistical inference based on latent ability estimates. *Psychometrika*, *61*-2, 313-330.

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. (1977). Bayesian tailored testing and the influence of item bank characteristics. *Applied psychological measurement*, *1*-1, 111-120.

Junker, B.W., Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied psychological Measurement*, *24*-1, 65-81.

Kass, R.E., Carlin, B.P., Gelman, A., Neal, R.M. (1997). Markov chain Monte Carlo in practice : a roundtable discussion. Pittsburg, PA : Carnegie Mellon University.

Kingsbury, G.G., Houser, R.L. (1999). Developing CATs for children. In F. Drasgow

et J.B. Olson-Buchanan (Éds) : Innovations in computerized assessment. Mahwah : Lawrence Erlbaum Associates.

Kingsbury, G.G., Weiss, D.J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. *In* D.J. Weiss (Éd.) : *New horizons in testing - Latent trait test theory and computerized adaptive testing*. New York : Academic Press.

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. *In* R. Hivon (Éd.) : *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. *In.* C. Depover et B. Noël (Éds) : *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.

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

Lawley, D.N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburg*, *62*-A, 74-82.

Lord, F.M. (1952). A theory of test scores. *Psychometric monographs*, no 7.

Lord, F.M. (1971). A theoretical study of two-stage testing. *Psychometrika*, *36*-3, 227-242.

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

Lord, F.M. (1980b). Some how and which for practical tailored testing. *In* L.J.T. van der Kamp, W.F. Langerak et D.N.M. de Gruijter (Éds) : *Psychometrics for educational debates.* New York : John Wiley and Sons.

Lord, F.M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48*-2, 233-245.

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

Luecht, R.M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied psychological measurement, 20*-4, 389-404.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*-2, 149-174.

McBride, J.R. (1997). Technical perspective. *In* W.A. Sands, B.K. Waters et J.R. McBride (Éds) : *Computer adaptive testing - From inquiry to operation.* Washington : American Psychological Association.

McBride, J.R., Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. *In* D.J. Weiss (Éd.) : *New horizons in testing - Latent trait test theory and computerized adaptive testing*. New York : Academic Press.

McBride, J.R., Wetzel, C.D., Hetter, R.D. (1997). Preliminary psychometric research for CAT-ASVAB : Selecting an adaptive testing strategy. *In* W.A. Sands, B.K. Waters et J.R. McBride (Éds) : *Computer adaptive testing - From inquiry to operation*. Washington : American Psychological Association.

McDonald, R.P. (1967). Non-linear factor analysis. *Psychometric monographs*, no 15.

McDonald, R.P. (1982). Linear versus nonlinear models in item response theory. *Applied psychological measurement*, *6*-4, 379-396.

McDonald, R.P. (1997). Normal-ogive multidimensional model. *In* W.J. van der Linden et R.K. Hambleton (Eds.) : *Handbook of modern item response theory*. New York : Springer-Verlag.

McDonald, R.P. (2000). A basis for multidimensional item response theory. *Applied psychological measurement*, *24*-2, 99-114.

Meijer, R.R., Nering, M.L. (1999). Computerized adaptive testing : overview and introduction. *Applied psychological measurement*, *23*-3, 187-194.

Microsoft (2000, 27 août). Adaptive testing. Disponibilité : *http://www.windowsgalore.com/cert/adaptive_testing/index.htm*.

Mislevy, R.J., Bock, R.D. (1982). Biweight estimates of latent ability. *Educational and psychological measurement*, *42*-2, 725-737.

Mislevy, R.J., Bock, R.D. (1984). *Bilog : item analysis and test scoring with binary logistic models*. Mooresville : Scientific Software.

Mislevy, R.J., Chang, H.H. (1998). Does adaptive testing violate local independence ? Rapport technique 476 du Center for the Study of Evaluation. Los Angeles : Université de Californie.

Mislevy, R.J., Chang, H.H. (2000). Does adaptive testing violate local independence ? *Psychometrika*, *65*-2, 149-156.

Mislevy, R.J., Stocking, M.L. (1987). *A consumer's guide to LOGIST and BILOG.* Research Report RR-87-4. Princeton : Educational Testing Service.

Mokken, R.J. (1997). Nonparametric models for dichotomous responses. *In* W.J. van der Linden et R.K. Hambleton (Eds.) : *Handbook of modern item response theory*. New York : Springer-Verlag.

Mokken, R.J., Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied psychological measurement*, *6*-4, 417-430.

Molenaar, I.W. (1995). Some background for item response theory and the Rasch model. *In* G.H. Fischer et I.W. Molenaar (Éds) : *Rasch models - Foundations, recent developments, and applications*. New York : Springer-Verlag.

Molenaar, I.W. (1997). Nonparametric models for polytomous responses. *In* W.J. van der Linden et R.K. Hambleton (Eds.) : *Handbook of modern item response theory*. New York : Springer-Verlag.

Moore, D.S. et McCabe, G.P. (1993). *Introduction to the practice of statistics*. New York : W.H. Freeman.

Morissette, D. (1984). *La mesure et l'évaluation en enseignement*. Ste-Foy : Presses de l'Université Laval.

Nadeau, M.A. (1988). *L'évaluation de programme : théorie et pratique*. Québec :

Presses de l'Université Laval.

Nering, M.L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied psychological measurement*, *21*-2, 115-127.

Nicewander, W.A., Thomasson, G.L. (1999). Some reliability estimates for computerized adaptive tests. *Applied psychological measurement*, *23*-3, 239-247.

Novick, M.R. (1966). The axioms and principal results of classical test theory. *Journal of mathematical psychology*, *3*, 1-18.

Owen, R.J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the american statistical association : theory and methods section*, *70*-350, 351-356.

Patz, R.J., Junker, B.W. (1997a). Applications and extensions of MCMC in IRT ; multiple item types, missing data, and rated responses. Pittsburg, PA : Carnegie Mellon University.

Patz, R.J., Junker, B.W. (1997b). A straightforward approach to Markov chain Monte Carlo methods for item response models. Pittsburg, PA : Carnegie Mellon

University.

Patz, R.J., Junker, B.W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of educational and behavioral statistics*, *24*-2, 146-178.

Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*-4, 611-630.

Ramsay, J.O. (1993a). TESTGRAF : a program for the graphical analysis of multiple choice test and questionnaire data. Montréal : Département de psychologie, Université McGill.

Ramsay, J.O. (1993b). TESTGRAF : some graphics tools for the analysis of examination data. 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.

Ramsay, J.O. (1997). A functional approach to modelling test data. *In* W.J. van der Linden et R.K. Hambleton (Eds.) : *Handbook of modern item response theory*. New York : Springer-Verlag.

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

Reckase, M.D. (1983). A procedure for decision making using tailored testing. *In* D.J. Weiss (Éd.) : *New horizons in testing - Latent trait test theory and computerized adaptive testing*. New York : Academic Press.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied psychological measurement*, *9*-4, 401-412.

Reckase, M.D. (1989). Adaptive testing : the evolution of a good idea. *Educational measurement : issues and practice*, *12*-1, 11-15.

Reckase, M.D. (1997). The past and future of multidimensional item response theory. *Applied psychological measurement*, *21*-1, 25-36.

Roberts, J.S., Donoghue, J.R., Laughlin, J.E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied psychological measurement*, *24*-1, 3-32.

Roos, L.L., Wise, S.L., Plake, B.S. (1997). The role of item feedback in self-adapted testing. *Educational and psychological measurement*, *57*-1, 85-98.

Samejima, F. (1972). A general model for free-response data. *Psychometrika monograph supplement, 18.*

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

Samejima, F. (1973b). Homogeneous case of the continuous response model. *Psychometrika, 38*-2, 203-219.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39*-1, 111-121.

Samejima, F. (1977). A method of estimating item characteristic functions using the maximum likelihood estimate of ability. *Psychometrika, 42*-2, 163-188.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied psychological measurement, 18*-3, 229-244.

Samejima, F. (1997a). Departure from normal assumptions : a promise for future psychometrics with substantive mathematical modeling. *Psychometrika, 62*-4, 471-493.

Samejima. F. (1997b). Graded response model. *In* W.J. van der Linden et R.K. Hambleton (Eds.) : *Handbook of modern item response theory*. New York : Springer-Verlag.

Sands, W.A., Waters, B.R. (1997). Introduction to Asvab and CAT. *In* W.A. Sands, B.K. Waters et J.R. McBride (Éds) : *Computerized adaptive testing - From inquiry to operation*. Washington : American Psychological Association.

SAS Institute (1990a). *SAS language : reference version 6 first edition*. Cary : SAS Institute.

SAS Institute (1990b). *SAS procedure guide : version 6 third edition*. Cary : SAS Institute.

Schnipke, D.L., Green, B.F. (1995). A comparison of item selection routines in linear and adaptive tests. *Journal of educational measurement*, *32*-3, 227-242.

Segall, D.O., Moreno, K.E., Hetter, R.D. (1997). Item pool development and evaluation. *In* W.A. Sands, B.K. Waters et J.R. McBride (Éds) : *Computer adaptive testing - From inquiry to operation*. Washington : American Psychological Association.

Séguin, S.P., Auger, R. (1986). Une introduction à la théorie des réponses aux items (TRI). *Mesure et évaluation en éducation*, *9*-1, 7-44.

Soloway, E, Adelson, B., Ehrlich, K. (1988). Knowledge and processes in the comprehension of computer programs. *In* M.T.H. Chi, R. Glaser et M.J. Farr (Éds) : *The nature of expertise*. Hillsdale : Lawrence Erlbaum Associates.

Spiegel, M.R. (1961). *Theory and problems of statistics*. New York : McGraw-Hill.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, *103*, 677-680. *In* Lieberman, B. (Éd.) : *Contemporary problems in statistics - A book of readings for the behavioral sciences*. New York : Oxford University Press, 1971.

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

Stout, W. (1990). A new item response theory modelling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*-2, 293-325.

Tanner, M.A. (1996). *Tools for statistical inference : methods for the exploration of posterior distributions and likelihood functions*. New York : Springer.

Thissen, D. (1988). *MULTILOG : multiple, categorical item analysis and test scoring using item response theory*. Mooresville, IN : Scientific Software.

Thissen, D. (1990). Reliability and measurement precision. *In* H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (Éds) : *Computerized adaptive testing - A primer*. Hillsdale : Lawrence Erlbaum Associates.

Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. *In* N. Frederiksen, R.J. Mislevy et I.I.Bejar (Éds) : *Test theory for a new generation of tests*. Hillsdale : Lawrence Erlbaum Associates.

Thissen, D., Mislevy, R.J. (1990). Testing algorithms. *In* H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (Éds) : *Computerized adaptive testing - A primer*. Hillsdale : Lawrence Erlbaum Associates.

Thissen, D., Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*-4, 567-577.

Urry, V.W. (1970). A Monte Carlo investigation of logistic mental models. Thèse de doctorat non publiée. West Lafayette : Purdue University.

Urry, V.W. (1977). Tailored testing : a successful application of latent trait theory. *Journal of educational measurement*, *14*, 181-196.

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., Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York : Springer-Verlag.

van Krimpen-Stoop, E.M.L.A., Meijer, R.R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied psychological measurement*, *23*-4, 327-345.

Veerkamp, W.J.J., Berger, M.P.F. (1997). Some new item selection criteria for adaptive testing. *Journal of educational and behavioral statistics*, *22*-2, 203-226.

Veerkamp, W.J.J., Berger, M.P.F. (1999). Optimal item discrimination and maximum information for logistic IRT models. *Applied psychological measurement*, *23*-1, 31-40.

Vispoel, W.P. (1999). Creating computerized adaptive tests of music aptitude : problems, solutions, and future directions. *In* F. Drasgow et J.B. Olson-Buchanan

(Éds) : *Innovations in computerized assessment*. Mahwah : Lawrence Erlbaum Associates.

Vispoel, W.P., Rocklin, T.R., Wang, T., Bleiler, T. (1999). Can examinees use a review option to obtain positively biased ability estimates on a computerized adaptive test ? *Journal of educational measurement*, *36*-2, 141-157.

Vispoel, W.P., Wang, T., Bleiler, T. (1997). Computerized adaptive and fixed-item testing of music listening skill : a comparison of efficiency, precision, and concurrent validity. *Journal of educational measurement*, *34*-1, 43-63.

Wainer, H. (1983). Are we correcting for guessing in the wrong direction ? *In* D.J. Weiss (Éd.) : *New horizons in testing - Latent trait test theory and computerized adaptive testing*. New York : Academic Press.

Wainer, H. (1990). Introduction and history. *In* H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (Éds) : *Computerized adaptive testing - A primer*. Hillsdale : Lawrence Erlbaum Associates.

Wainer, H., Dorans, N.J., Green, B.F., Mislevy, R.J., Steinberg, L., Thissen, D. (1990). Future challenges. *In* H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (Éds) : *Computerized adaptive*

testing - A primer. Hillsdale : Lawrence Erlbaum Associates.

Wainer, H., Kiely, G.L. (1987). Item clusters and computerized adaptive testing : a case for testlets. *Journal of educational measurement*, *24*-3, 185-201.

Wainer, H., Mislevy, R.J. (1990). Item response theory, item calibration and proficiency estimation. *In* H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (Éds) : *Computerized adaptive testing - A primer*. Hillsdale : Lawrence Erlbaum Associates.

Wainer, H., Thissen, D. (1987). Estimating ability with the wrong model. *Journal of educational statistics*, *12*-4, 339-368.

Wainer, H., Wright, B. (1980). Robust estimation of ability in the Rash model. *Psychometrika*, *45*, 370-390.

Wang, T., Hanson, B.A., Lau, C.M.A. (1999). Reducing bias in CAT trait estimation : a comparison of approaches. *Applied psychological measurement*, *23*-3, 263-278.

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

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied psychological measurement*, *6*-4, 473-492.

Weiss, D.J. (1983). *New horizons in testing - Latent trait test theory and computerized adaptive testing*. New York : Academic Press.

Weiss, D.J. (1985). Adaptive testing by computer. *Journal of consulting and clinical psychology*, *53*-6, 774-789.

Weiss, D.J., Yoes, M.E. (1990). Item response theory. *In* R.K. Hambleton et J.N. Zaal (Éds) : *Advances in educational and psychological testing - Theory and applications*. Boston : Kluwer.

Wilson, M. (1992). The ordered partition model : an extension of the partial credit model. *Applied psychological measurement*, *16*-4, 309-325.

Wright, B.D. (1977). A history of social science measurement. *Educational measurement : issues and practice*, *16*-4, 33-45, 52.

Wright, B.D., Stone, M.H. (1979). *Best test design*. Chicago : Mesa.

Yakowitz, S., Szidarovszky, F. (1986). *An introduction to numerical computations*.

New York : Macmillan Publishing Company.

Yamamoto, K., Gitomer, D.H. (1993). Application of a HYBRID model to a test of cognitive skill representation. *In* N. Frederiksen, R.J. Mislevy et I.I. Bejar (Éds) : *Test theory for a new generation of tests*. Hillsdale : Lawrence Erlbaum Associates.

Zwick, R. (1997). The effect of adaptive administration on the variability of the Mantel-Haenszel measure of differential item functioning. *Educational and psychological measurement*, *57*-3, 412-421.

Zwick, R., Thayer, D.T., Wingersky, M. (1994). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied psychological measurement*, *18*-2, 121-140.

ANNEXES

Annexe I

Comparaison des fonctions de probabilité logistique et normale cumulées (modèle logistique à un paramètre utilisant la constante D)

Tableau I.1
Comparaison des fonctions de probabilité logistique et normale cumulées (modèle logistique à un paramètre utilisant la constante D)

Z	LOI LOGISTIQUE	LOI NORMALE N(0,1)	|ÉCART|
-3,00	0,01	0,00	0,01
-2,50	0,01	0,01	0,00
-2,00	0,03	0,02	0,01
-1,50	0,07	0,07	0,00
-1,00	0,15	0,16	0,01
-0,50	0,30	0,31	0,01
0,00	0,50	0,50	0,00
0,50	0,70	0,69	0,01
1,00	0,85	0,84	0,01
1,50	0,93	0,93	0,00
2,00	0,97	0,98	0,01
2,50	0,99	0,99	0,00
3,00	0,99	1,00	0,01

Annexe II

Équations utilisées dans la méthode bayésienne d'Owen

Considérant que

$$\text{erf}(D) = \frac{2}{\sqrt{\pi}} \int_0^D e^{-t^2} dt \qquad \textbf{(II.1)}$$

et que

$$D = \frac{b_i - \text{OWEN}_i(\theta)}{\sqrt{2(a_i^{-2} + S^2_{\text{OWEN}_i(\theta)})}} \qquad \textbf{(II.2)}$$

l'estimateur du niveau d'habileté, $\text{OWEN}_{i+1}(\theta)$, après l'administration d'un nouvel item, i, et en fonction de la réponse donnée à l'item, est égal à (Jansema, 1977, p. 111-113 ; Owen, 1975) :

$$\text{OWEN}_{i+1}(\theta|r_i=1) = \text{OWEN}_i(\theta) + \frac{\sqrt{2}\ S^2_{\text{OWEN}_i(\theta)}}{\sqrt{\pi(a_i^{-2} + S^2_{\text{OWEN}_i(\theta)})}} \frac{e^{-D^2}}{(1 - \text{erf}(D))} \qquad \textbf{(II.3)}$$

ou à

$$\text{OWEN}_{i+1}(\theta|r_i=0) = \text{OWEN}_i(\theta) - \frac{\sqrt{2}\ S^2_{\text{OWEN}_i(\theta)}}{\sqrt{\pi(a_i^{-2} + S^2_{\text{OWEN}_i(\theta)})}} \frac{e^{-D^2}}{(1 + \text{erf}(D))} \qquad \textbf{(II.4)}$$

L'erreur type de l'estimateur du niveau d'habileté, $S_{OWEN_{i+1}(\theta)}$, après l'administration d'un nouvel item, en fonction de la réponse donnée à l'item, est estimée à partir d'une des fonctions suivantes (Jansema, 1977, p. 111-113 ; Owen, 1975) :

$$S^2_{OWEN_{i+1}(\theta|r_i=1)} = S^2_{OWEN_i(\theta)}\left(1-\frac{(2/\sqrt{\pi})-2De^{D^2}(1-erf(D))}{\sqrt{\pi}(1+a_i^{-2}S^{-2}_{OWEN_i(\theta)})[e^{D^2}(1-erf(D)]^2}\right) \quad \textbf{(II.5)}$$

ou

$$S^2_{OWEN_{i+1}(\theta|r_i=0)} = S^2_{OWEN_i(\theta)}\left(1-\frac{(2/\sqrt{\pi})+2De^{D^2}(1+erf(D))}{\sqrt{\pi}(1+a_i^{-2}S^{-2}_{OWEN_i(\theta)})[e^{D^2}(1+erf(D)]^2}\right) \quad \textbf{(II.6)}$$

Annexe III

Programmation de la simulation en langage SAS 6 (SAS Institute, 1990a)

```
/*************************************************************************/;
/*************************************************************************/;
/**                                                                     **/;
/**       Simulation d'un test adaptatif selon deux règles d'arrêt      **/;
/**                                                                     **/;
/**                            Gilles Raîche                            **/;
/**                                                                     **/;
/**                       Université de Montréal                        **/;
/**       Faculté d'éducation et d'administration de l'éducation        **/;
/**                                                                     **/;
/**                             Juin 2000                               **/;
/**                                                                     **/;
/*************************************************************************/;
/*************************************************************************/;


 Options replace nodate pageno=1 nomtrace;
 Libname doc "d:\data\doctorat\these";run;
 Title1 "Simulation d'un test adaptatif selon deux règles d'arrêt";

/* Initialisation des constantes ****************************************/;
 %Let nitem=3    ;  /* Nombre d'items                                    */;
 %Let qpoint=20  ;  /* Nombre de points de quadrature                    */;
 %Let D=1.702    ;  /* Constante de Haley                                */;
 %Let thetamin=-4;  /* Limite inférieure d'intégration dans EAP          */;
 %Let thetamax= 4;  /* Limite supérieure d'intégration dans EAP          */;
 %Let bayes=0    ;  /* Niveau d'habileté a priori                        */;
 %Let sujet=2000 ;  /* Nombre de valeurs aléatoires du niveau d'habileté */;
/* *********************************************************************/;


 %Macro Prior (Theta,mu);
 %* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 %* Fonction de probabilité a priori N(mu,1) telle qu`utilisée par Bock   *;
 %* et Mislevy (1982) ainsi que dans PC-BILOG (Baker, 1992, p. 187, Mislevy *;
 %* et Bock, 1984, p. 6)                                                 *;
 %* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 %* Variables locales  : theta, mu                                       *;
 %* Variables globales : step                                            *;
 %* Appel de fonction  : aucun                                           *;
 %* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
  Probnorm(&Theta-&mu + step/2) - Probnorm(&Theta-&mu - step/2);;
 %Mend;


 %Macro logist (Theta,b,D,r);
 %* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 %* Fonction logistique à un paramètre                                   *;
 %* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 %* Variables locales  : theta, b, r                                     *;
 %* Variables globales : D                                               *;
 %* Appel de fonction  : aucun                                           *;
 %* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
  ((((1+exp(-&D*(&Theta-&b)))**-1))**&r) *
  (1-(((1+exp(-&D*(&Theta-&b)))**-1)))**(1-&r);
 %Mend;
```

```
%Macro Reponse (i,habilete);
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Génération d`une réponse à un item en fonction du niveau d`habileté     *;
%* préalablement tiré au hasard et du niveau de difficulté du prochain     *;
%* item selon la méthode utilisée par Harwell, Stone, Hsu et Kirisci       *;
%* (1996, p. 116)                                                          *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Variables locales  : i, habilete                                        *;
%* Variables globales : D, b                                               *;
%* Appel de fonction  : aucun                                              *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 plog= exp(&D*(&habilete-b{&i})) / (1+exp(&D*(&habilete-b{&i})));
 puni=ranuni(1);
 r{&i}=plog > puni;
%Mend;


%Macro step;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Calcul de la largeur de classe en fonction du nombre de points de       *;
%* quadrature ainsi que de la surface de la somme de la fonction a priori  *;
%* sur l`intervalle Thetamin à Thetamax                                    *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Variables locales  : surface, j                                         *;
%* Variables globales : step, thetamax, thetamin, qpoint                   *;
%* Appel de fonction  : prior                                              *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 surface=0;
 Do j=1 to &nitem;
    %* Largeur de classe en fonction du nombre de points de quadrature .....;
    step=(&thetamax-&thetamin)/(&qpoint-1);
    end;
 %* Calcul de la surface sous la fonction a priori .......................;
 Do theta=&thetamin to &thetamax by step;
    surface=surface + %Prior(theta,0);
    end;
 %* Put "surface = " surface; *;
%Mend;


%Macro MU2(habilete);
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Estimation du niveau d`habileté par la méthode EAP                      *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Variables locales  : b1,j,num,p,lnp,i,logist,prior                      *;
%* Variables globales : nitem,habilete,thetamin,thetamax,step,theta,b,D,r  *;
%*                      bayes,te,denom                                     *;
%* Appel de fonction  : reponse,logist,prior                               *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 %* Paramètre de difficulté du premier item ..............................;
 b1=0;
 %* Initialisation a 0 des numérateurs et dénominateurs ..................;
 Do j=1 to &nitem;
    num(j)=0;denom(j)=0;
    end;
 %* Estimation des j niveaux d`habileté ..................................;
 Do j=1 to &nitem;
    %Reponse(j,&habilete);
    %* Calcul de l`espérance a posteriori du niveau d`habileté ............;
    Do theta=&thetamin to &thetamax by step;
       p=1;lnp=log(p);
       Do i=1 to j;
          logist=%logist(theta,b(i),&D,r(i));
          lnp=lnp+log(logist);
          end;
```

```
        prior=%prior(theta,&bayes);
        num(j)=num(j)+      (theta)*exp(lnp+log(prior));
        denom(j)=denom(j)+  exp(lnp+log(prior));
        end;
     %* jième estimation du niveau d`habileté ..............................;
     te(j)=num(j)/denom(j);
     if j<&nitem then do;
        b(j+1)=te(j);
        end;
     end;
%Mend;


%Macro VAR2;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Estimation de l`erreur type, de l`asymétrie et de la                      *;
%* kurtose  par la méthode EAP                                               *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Variables locales  : b1,j,num2,num3,num4,p,lnp,i,logist,prior,rc          *;
%* Variables globales : nitem,habilete,thetamin,thetamax,step,theta,b,D,r    *;
%*                      bayes,prc,se,as,ku                                   *;
%* Appel de fonction  : reponse,logist,prior                                 *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 %* Initialisation a 0 des numérateurs et dénominateurs .....................;
 %* et calcul de la proportion de bonnes réponses (prc) .....................;
 rc=0;
 Do j=1 to &nitem;
    num2_(j)=0;num3_(j)=0;num4_(j)=0;
    %* Proportion de bonnes réponses ......................................;
    rc=rc+r(j);prc(j)=rc/j;
    end;
 %* Estimation des j valeurs d`erreur type, asymétrie et kurtose ...........;
 Do j=1 to &nitem;
    %* Calcul de l`espérance a posteriori .................................;
    Do theta=&thetamin to &thetamax by step;
       p=1;lnp=log(p);
       Do i=1 to j;
          logist=%logist(theta,b(i),&D,r(i));
          lnp=lnp+log(logist);
          end;
       prior=%prior(theta,&bayes);
       num2_(j)=num2_(j)+((theta-te(j))**2)  *exp(lnp+log(prior));
       num3_(j)=num3_(j)+((theta-te(j))**3)  *exp(lnp+log(prior));
       num4_(j)=num4_(j)+((theta-te(j))**4)  *exp(lnp+log(prior));
       end;
    %* jième estimation de l`erreur type du niveau d`habileté ..............;
    se(j)=(num2_(j)/denom(j))**.5;
    %* jième estimation de l`asymétrie ....................................;
    as(j)=(num3_(j)/denom(j))/se(j)**3;
    %* jième estimation de la kurtose .....................................;
    ku(j)=((num4_(j)/denom(j))/se(j)**4)-3;
    denom(j)=denom(j)/surface;
    end;
%Mend;


Data work.thesd_&nitem;
     array te[&nitem];      /* Estimation de l'habileté ..................*/;
     array r[&nitem];       /* Vecteur de réponses .......................*/;
     array b[&nitem];       /* Paramètres de difficulté ..................*/;
     array num[&nitem];     /* Numérateur dans EAP (MU) ..................*/;
     array num2_[&nitem];   /* Numérateur dans EAP (SE) ..................*/;
     array num3_[&nitem];   /* Numérateur dans EAP (MC3) .................*/;
     array num4_[&nitem];   /* Numérateur dans EAP (MC4) .................*/;
     array denom[&nitem];   /* Dénominateur dans EAP .....................*/;
```

```
      array se[&nitem];      /* Erreur type ..............................*/;
      array as[&nitem];      /* Asymétrie ................................*/;
      array ku[&nitem];      /* Kurtose ..................................*/;
      array prc[&nitem];     /* Proportion de bonnes réponses ............*/;
      length default=3       /* Format de storage des variables ..........*/;
      b1=0;                  /* Initialisation du niveau de difficulté de b1*/;
      %STEP;
      Do sujet = 1 to &sujet;
         habilete=rannor(1);/* Génération d'un niveau d'habileté N(0,1) ...*/;
          %MU2(habilete);
          %VAR2;
          /* Valeurs des différentes statistiques selon le nombre .........*/;
          /* d'items administrés ..........................................*/;
          Do item=1 to &nitem;
             ro=r{item};bo=b{item};teo=te{item};prco=prc{item};
             seo=se{item};aso=as{item};kuo=ku{item};biais = teo-habilete;
             output;
             end;
          End;
      Keep sujet item ro bo habilete teo prco seo aso kuo biais;
      run;
Proc print;run;
```

Annexe IV

Routines destinées à vérifier l'exactitude du calcul de l'estimateur du niveau d'habileté et de l'erreur type de l'estimateur du niveau d'habileté

```
/*************************************************************************/;
/*************************************************************************/;
/**                                                                     **/;
/**        Test de l'exactitude des calculs dans la méthode EAP         **/;
/**                                                                     **/;
/**                          Gilles Raîche                              **/;
/**                                                                     **/;
/**                      Université de Montréal                         **/;
/**      Faculté d'éducation et d'administration de l'éducation         **/;
/**                                                                     **/;
/**                          Juin 2000                                  **/;
/**                                                                     **/;
/*************************************************************************/;
/*************************************************************************/;


 Options replace nodate pageno=1 nomtrace;
 Libname doc "d:\data\doctorat\these";run;
 Title1 "Simulation d'un test adaptatif selon deux règles d'arrêt";

/* Initialisation des constantes ****************************************/;
 %Let nitem=20    ;  /* Nombre d'items                                   */;
 %Let qpoint=21   ;  /* Nombre de points de quadrature                   */;
 %Let D=1.702     ;  /* Constante de Haley                               */;
 %Let thetamin=-3;   /* Limite inférieure d'intégration dans EAP         */;
 %Let thetamax= 3;   /* Limite supérieure d'intégration dans EAP         */;
 %Let bayes=0     ;  /* Niveau d'habileté a priori                       */;
 %Let sujet=1     ;  /* Nombre de valeurs du niveau d'habileté           */;
/* *********************************************************************/;


 %Macro Bock;
 %* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 %* Initialisation des parametres d items tel qu`utilisés par Bock       *;
 %* et Mislevy (1982, p.  434)                                           *;
 %* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
  %* Parametres de discrimination ..................................;
   a1=0.71; a2=0.65; a3=1.11; a4=0.69; a5=2.10;
   a6=1.00; a7=0.93; a8=1.68; a9=0.92;a10=1.23;
  a11=1.96;a12=0.88;a13=0.82;a14=1.03;a15=1.23;
  a16=1.68;a17=0.95;a18=1.34;a19=1.63;a20=1.68;
  %* Parametres de difficulte ....................................;
   b1=-0.05; b2=00.38; b3=00.04; b4=00.39; b5=00.14;
   b6=-0.42; b7=-0.54; b8=-0.42; b9=-0.64;b10=-0.84;
  b11=-0.69;b12=-0.19;b13=-0.16;b14=-0.90;b15=-0.90;
  b16=-0.26;b17=-1.01;b18=-0.91;b19=-0.13;b20=-0.86;
  %* Parametres de pseudo chance ..................................;
  Do j=1 to &nitem;
     c(j)=0.10;
     end;
  %* Reponses aux items ...........................................;
   r1=1; r2=0; r3=1; r4=0; r5=0;
   r6=0; r7=1; r8=0; r9=0;r10=1;
  r11=1;r12=0;r13=0;r14=1;r15=1;
  r16=0;r17=1;r18=1;r19=0;r20=1;
  %Mend;
```

```
%Macro Prior (Theta,mu);
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Fonction de probabilité a priori N(mu,1) telle qu'utilisée par Bock     *;
%* et Mislevy (1982) ainsi que dans PC-BILOG (Baker, 1992, p. 187, Mislevy *;
%* et Bock, 1984, p. 6)                                                    *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Variables locales  : theta, mu                                          *;
%* Variables globales : step                                               *;
%* Appel de fonction  : aucun                                              *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 Probnorm(&Theta-&mu + step/2) - Probnorm(&Theta-&mu - step/2);
%Mend;


%Macro logist (Theta,a,b,c,D,r);
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Fonction logistique à un paramètre                                      *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Variables locales  : theta,a,b,c,r                                      *;
%* Variables globales : D                                                  *;
%* Appel de fonction  : aucun                                              *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 (&c+(1-&c)*(((1+exp(-&D*&a*(&Theta-&b)))**-1))**&r) *
 (1-(&c+(1-&c)*((1+exp(-&D*&a*(&Theta-&b)))**-1)))**(1-&r);
%Mend;


%Macro step;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Calcul de la largeur de classe en fonction du nombre de points de       *;
%* quadrature ainsi que de la surface de la somme de la fonction a priori  *;
%* sur l'intervalle Thetamin à Thetamax                                    *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Variables locales  : surface, j                                         *;
%* Variables globales : step, thetamax, thetamin, qpoint                   *;
%* Appel de fonction  : prior                                              *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 surface=0;
 Do j=1 to &nitem;
    %* Largeur de classe en fonction du nombre de points de quadrature .....;
    step=(&thetamax-&thetamin)/(&qpoint-1);
    end;
 %* Calcul de la surface sous la fonction a priori .......................;
 Do theta=&thetamin to &thetamax by step;
    surface=surface + %Prior(theta,0);
    end;
 %* Put "surface = " surface;
%Mend;


%Macro MU2(habilete);
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Estimation du niveau d'habileté par la méthode EAP                      *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Variables locales  : b1,j,num,p,lnp,i,logist,prior                      *;
%* Variables globales : nitem,habilete,thetamin,thetamax,step,theta,b,D,r  *;
%*                      bayes,te,denom                                     *;
%* Appel de fonction  : reponse,logist,prior                               *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 %* Initialisation a 0 des numérateurs et dénominateurs ....................;
 Do j=1 to &nitem;
    num(j)=0;denom(j)=0;
    end;
```

```
 %* Estimation des j niveaux d`habileté .....................................;
 Do j=1 to &nitem;
    %* %Reponse(j,&habilete);
    %* Calcul de l`espérance a posteriori du niveau d`habileté .............;
    Do theta=&thetamin to &thetamax by step;
       p=1;lnp=log(p);
       Do i=1 to j;
          logist=%logist(theta,a(i),b(i),c(i),&D,r(i));
          lnp=lnp+log(logist);
          end;
       prior=%prior(theta,&bayes);
       num(j)=num(j)+(theta)*exp(lnp+log(prior));
       denom(j)=denom(j)+  exp(lnp+log(prior));
       end;
    %* jième estimation du niveau d`habileté ...............................;
    te(j)=num(j)/denom(j);
    end;
%Mend;


%Macro VAR2;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Estimation de l`erreur type, de l`asymétrie et de la                      *;
%* kurtose  par la méthode EAP                                               *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
%* Variables locales  : b1,j,num2,num3,num4,p,lnp,i,logist,prior,rc          *;
%* Variables globales : nitem,habilete,thetamin,thetamax,step,theta,b,D,r    *;
%*                      bayes,prc,se,as,ku                                   *;
%* Appel de fonction  : reponse,logist,prior                                 *;
%* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
 %* Initialisation a 0 des numérateurs et dénominateurs .....................;
 %* et calcul de la proportion de bonnes réponses (prc) .....................;
 rc=0;
 Do j=1 to &nitem;
    num2_(j)=0;num3_(j)=0;num4_(j)=0;
    %* Proportion de bonnes réponses ........................................;
    rc=rc+r(j);prc(j)=rc/j;
    end;
 %* Estimation des j valeurs d`erreur type, asymétrie et kurtose ...........;
 Do j=1 to &nitem;
    %* Calcul de l`espérance a posteriori ..................................;
    Do theta=&thetamin to &thetamax by step;
       p=1;lnp=log(p);
       Do i=1 to j;
          logist=%logist(theta,a(i),b(i),c(i),&D,r(i));
          lnp=lnp+log(logist);
          end;
       prior=%prior(theta,&bayes);
       num2_(j)=num2_(j)+((theta-te(j))**2)  *exp(lnp+log(prior));
       num3_(j)=num3_(j)+((theta-te(j))**3)  *exp(lnp+log(prior));
       num4_(j)=num4_(j)+((theta-te(j))**4)  *exp(lnp+log(prior));
       end;
    %* jième estimation de l`erreur type du niveau d`habileté ..............;
    se(j)=(num2_(j)/denom(j))**.5;
    %* jième estimation de l`asymétrie .....................................;
    as(j)=(num3_(j)/denom(j))/se(j)**3;
    %* jième estimation de la kurtose ......................................;
    ku(j)=((num4_(j)/denom(j))/se(j)**4)-3;
    denom(j)=denom(j)/surface;
    end;
%Mend;
```

```
Data work.thesd_&nitem;
     array te[&nitem];      /* Estimateur de l'habileté ..................*/;
     array r[&nitem];       /* Vecteur de réponses .......................*/;
     array a[&nitem];       /* Paramètres de discrimination ..............*/;
     array b[&nitem];       /* Paramètres de difficulté ..................*/;
     array c[&nitem];       /* Paramètres de pseudo chance ...............*/;
     array num[&nitem];     /* Numérateur dans EAP (MU) ..................*/;
     array num2_[&nitem];   /* Numérateur dans EAP (SE) ..................*/;
     array num3_[&nitem];   /* Numérateur dans EAP (MC3) .................*/;
     array num4_[&nitem];   /* Numérateur dans EAP (MC4) .................*/;
     array denom[&nitem];   /* Dénominateur dans EAP .....................*/;
     array se[&nitem];      /* Erreur type ...............................*/;
     array as[&nitem];      /* Asymétrie .................................*/;
     array ku[&nitem];      /* Kurtose ...................................*/;
     array prc[&nitem];     /* Proportion de bonnes réponses .............*/;
     length default=3       /* Format de storage des variables ...........*/;
     %Bock;
     %STEP;
     Do sujet = 1 to &sujet;
        habilete=-0.50;/* Niveau d'habileté chez Bock et Mislevy ......*/;
        %MU2(habilete);
        %VAR2;
        /* Valeurs des différentes statistiques selon le nombre .......*/;
        /* d'items administrés ........................................*/;
        Do item=1 to &nitem;
           ro=r{item};bo=b{item};teo=te{item};prco=prc{item};
           seo=se{item};aso=as{item};kuo=ku{item};biais = teo-habilete;
           output;
           end;
        End;
     Keep sujet item ro bo habilete teo prco seo aso kuo biais;
     run;


Proc print;run;
```

PC-BILOG 1.1

Fichier : BOCK.IF

```
  BOCK et MISLEVY (1982, p. 434) : TAO SIMULE AU TABLEAU 1
                       20 ITEMS.

   1  20
  20
0001B  .03550    .00000    .71000    .00000   -.05000    .00000   1.40845    .00000    .10000    .00000  111
0002B -.24700    .00000    .65000    .00000    .38000    .00000   1.53846    .00000    .10000    .00000  211
0003B -.04400    .00000   1.11000    .00000    .04000    .00000    .90090    .00000    .10000    .00000  311
0004B -.26910    .00000    .69000    .00000    .39000    .00000   1.44928    .00000    .10000    .00000  411
0005B -.29400    .00000   2.10000    .00000    .14000    .00000    .47619    .00000    .10000    .00000  511
0006B  .42000    .00000   1.00000    .00000   -.42000    .00000   1.00000    .00000    .10000    .00000  611
0007B  .50220    .00000    .93000    .00000   -.54000    .00000   1.07527    .00000    .10000    .00000  711
0008B  .70560    .00000   1.68000    .00000   -.42000    .00000    .59524    .00000    .10000    .00000  811
0009B  .58880    .00000    .92000    .00000   -.64000    .00000   1.08696    .00000    .10000    .00000  911
0010B 1.03320    .00000   1.23000    .00000   -.84000    .00000    .81301    .00000    .10000    .00000 1011
0011B 1.35240    .00000   1.96000    .00000   -.69000    .00000    .51020    .00000    .10000    .00000 1111
0012B  .16720    .00000    .88000    .00000   -.19000    .00000   1.13636    .00000    .10000    .00000 1211
0013B  .13120    .00000    .82000    .00000   -.16000    .00000   1.21951    .00000    .10000    .00000 1311
0014B  .92700    .00000   1.03000    .00000   -.90000    .00000    .97089    .00000    .10000    .00000 1411
0015B 1.10700    .00000   1.23000    .00000   -.90000    .00000    .81301    .00000    .10000    .00000 1511
0016B  .43680    .00000   1.68000    .00000   -.26000    .00000    .59524    .00000    .10000    .00000 1611
0017B  .95950    .00000    .95000    .00000  -1.01000    .00000   1.05263    .00000    .10000    .00000 1711
0018B 1.21940    .00000   1.34000    .00000   -.91000    .00000    .74627    .00000    .10000    .00000 1811
0019B  .21190    .00000   1.63000    .00000   -.13000    .00000    .61350    .00000    .10000    .00000 1911
0020B 1.44800    .00000   1.68000    .00000   -.86000    .00000    .59524    .00000    .10000    .00000 2011
```

Fichier : BOCK.BLG

```
  BOCK ET MISLEVY (1989, p. 434) : TAO SIMULE AU TABLEAU 1
                        20 ITEMS.


>GLOBAL  DFDISP=1, NIDW=5, SFDISP=1, CFDISP=0, IFDISP=1,
IFNAME='BOCK.IF', SFNAME='BOCK.SF', NPARM=3;
>LENGTH NITEMS=20;
>INPUT NTOT=20;
(5A1,5X,20A1)
>TEST TNAME=BOCK;
49-25     10100010011001101101

>SCORE, METHOD=2,PRINT,NQPT=21,IDIST=0;
```

Annexe V

Trois exemples de résultats de la simulation

Tableau V.1
Exemple de résultats de la simulation pour le sujet 1 ($\theta = 0,81$)

Item	r	b	EAP(θ)	p	$S_{EAP(\theta)}$	$a3_{EAP(\theta)}$	$a4_{EAP(\theta)}$	$ERREUR_{EAP(\theta)}$
1	1	0,00	0,56	1,00	0,82	0,11	0,13	-0,24
2	1	0,56	0,99	1,00	0,73	0,18	0,16	0,18
3	0	0,99	0,64	0,67	0,62	0,04	0,21	-0,16
4	1	0,64	0,91	0,75	0,56	0,13	0,23	0,10
5	0	0,91	0,68	0,60	0,50	0,02	0,20	-0,13
6	0	0,68	0,50	0,50	0,47	-0,06	0,19	-0,31
7	0	0,50	0,33	0,43	0,44	-0,12	0,19	-0,47
8	1	0,33	0,48	0,50	0,41	-0,04	0,17	-0,33
9	1	0,48	0,61	0,56	0,39	0,02	0,15	-0,20
10	1	0,61	0,73	0,60	0,37	0,06	0,15	-0,08
11	1	0,73	0,83	0,64	0,36	0,09	0,15	0,02
12	0	0,83	0,73	0,58	0,34	0,04	0,13	-0,08
13	0	0,73	0,64	0,54	0,32	0,01	0,12	-0,17
14	0	0,64	0,56	0,50	0,31	-0,02	0,11	-0,25
15	0	0,56	0,48	0,47	0,30	-0,05	0,11	-0,33
16	1	0,48	0,55	0,50	0,29	-0,02	0,10	-0,25
17	0	0,55	0,49	0,47	0,29	-0,04	0,10	-0,32
18	1	0,49	0,55	0,50	0,28	-0,02	0,09	-0,26
19	1	0,55	0,61	0,53	0,27	0,01	0,09	-0,20
20	0	0,61	0,55	0,50	0,26	-0,01	0,09	-0,25
21	1	0,55	0,61	0,52	0,26	0,00	0,08	-0,20
22	1	0,61	0,66	0,55	0,25	0,02	0,08	-0,14
23	0	0,66	0,61	0,52	0,25	0,00	0,08	-0,20
24	1	0,61	0,66	0,54	0,24	0,02	0,07	-0,15
25	1	0,66	0,71	0,56	0,24	0,03	0,07	-0,10
26	1	0,71	0,75	0,58	0,23	0,04	0,07	-0,05
27	1	0,75	0,80	0,59	0,23	0,05	0,07	-0,01
28	0	0,80	0,76	0,57	0,22	0,04	0,07	-0,05
29	1	0,76	0,80	0,59	0,22	0,05	0,07	-0,01
30	0	0,80	0,76	0,57	0,22	0,04	0,06	-0,05
31	1	0,76	0,80	0,58	0,21	0,04	0,06	-0,01
32	0	0,80	0,76	0,56	0,21	0,03	0,06	-0,05
33	0	0,76	0,72	0,55	0,21	0,02	0,06	-0,09
34	1	0,72	0,76	0,56	0,20	0,03	0,06	-0,05
35	0	0,76	0,72	0,54	0,20	0,02	0,05	-0,09
36	0	0,72	0,69	0,53	0,20	0,01	0,05	-0,12
37	0	0,69	0,66	0,51	0,19	0,00	0,05	-0,15
38	1	0,66	0,69	0,53	0,19	0,01	0,05	-0,12
39	0	0,69	0,66	0,51	0,19	0,00	0,05	-0,15
40	0	0,66	0,63	0,50	0,19	-0,01	0,05	-0,18
41	0	0,63	0,60	0,49	0,18	-0,01	0,04	-0,21
42	1	0,60	0,63	0,50	0,18	-0,01	0,04	-0,18
43	1	0,63	0,66	0,51	0,18	0,00	0,04	-0,15
44	1	0,66	0,68	0,52	0,18	0,01	0,04	-0,13
45	1	0,68	0,71	0,53	0,18	0,01	0,04	-0,10

Tableau V.1 (suite)
Exemple de résultats de la simulation pour le sujet 1 ($\theta = 0,81$)

Item	r	b	EAP(θ)	p	$S_{EAP(\theta)}$	$a3_{EAP(\theta)}$	$a4_{EAP(\theta)}$	ERREUR $_{EAP(\theta)}$
46	1	0,71	0,73	0,54	0,17	0,02	0,04	-0,07
47	1	0,73	0,76	0,55	0,17	0,02	0,04	-0,05
48	0	0,76	0,73	0,54	0,17	0,02	0,04	-0,07
49	0	0,73	0,71	0,53	0,17	0,01	0,04	-0,10
50	0	0,71	0,69	0,52	0,17	0,01	0,04	-0,12
51	1	0,69	0,71	0,53	0,17	0,01	0,04	-0,10
52	0	0,71	0,69	0,52	0,16	0,00	0,04	-0,12
53	0	0,69	0,67	0,51	0,16	0,00	0,04	-0,14
54	1	0,67	0,69	0,52	0,16	0,00	0,04	-0,12
55	1	0,69	0,71	0,53	0,16	0,01	0,04	-0,10
56	0	0,71	0,69	0,52	0,16	0,00	0,04	-0,12
57	1	0,69	0,71	0,53	0,16	0,01	0,05	-0,10
58	1	0,71	0,73	0,53	0,15	0,01	0,05	-0,08
59	0	0,73	0,71	0,53	0,15	0,01	0,05	-0,10
60	0	0,71	0,69	0,52	0,15	0,00	0,05	-0,12

Tableau V.2
Exemple de résultats de la simulation pour le sujet 2 (θ = -0,02)

Item	r	b	EAP(θ)	p	$S_{EAP(\theta)}$	$a3_{EAP(\theta)}$	$a4_{EAP(\theta)}$	ERREUR $_{EAP(\theta)}$
1	0	0,00	-0,56	0,00	0,82	-0,11	0,13	-0,54
2	1	-0,56	-0,15	0,50	0,69	0,04	0,22	-0,12
3	1	-0,15	0,17	0,67	0,62	0,15	0,24	0,19
4	0	0,17	-0,09	0,50	0,55	0,01	0,23	-0,07
5	0	-0,09	-0,31	0,40	0,50	-0,08	0,22	-0,28
6	0	-0,31	-0,49	0,33	0,47	-0,14	0,22	-0,47
7	0	-0,49	-0,66	0,29	0,45	-0,19	0,23	-0,63
8	1	-0,66	-0,50	0,38	0,41	-0,10	0,18	-0,48
9	1	-0,50	-0,37	0,44	0,39	-0,03	0,16	-0,35
10	1	-0,37	-0,26	0,50	0,37	0,02	0,14	-0,23
11	1	-0,26	-0,15	0,55	0,35	0,06	0,14	-0,13
12	0	-0,15	-0,25	0,50	0,34	0,01	0,13	-0,23
13	0	-0,25	-0,34	0,46	0,32	-0,02	0,12	-0,31
14	0	-0,34	-0,42	0,43	0,31	-0,05	0,12	-0,40
15	0	-0,42	-0,50	0,40	0,31	-0,07	0,12	-0,48
16	1	-0,50	-0,43	0,44	0,29	-0,04	0,11	-0,40
17	0	-0,43	-0,50	0,41	0,29	-0,06	0,11	-0,47
18	1	-0,50	-0,43	0,44	0,28	-0,03	0,10	-0,41
19	1	-0,43	-0,37	0,47	0,27	-0,01	0,09	-0,34
20	1	-0,37	-0,31	0,50	0,26	0,01	0,09	-0,28
21	1	-0,31	-0,25	0,52	0,26	0,03	0,08	-0,23
22	1	-0,25	-0,20	0,55	0,25	0,04	0,08	-0,17
23	0	-0,20	-0,25	0,52	0,25	0,02	0,08	-0,23
24	1	-0,25	-0,20	0,54	0,24	0,04	0,08	-0,18
25	1	-0,20	-0,15	0,56	0,24	0,05	0,07	-0,13
26	1	-0,15	-0,11	0,58	0,23	0,06	0,07	-0,08
27	1	-0,11	-0,06	0,59	0,23	0,07	0,07	-0,04
28	0	-0,06	-0,11	0,57	0,23	0,05	0,07	-0,08
29	1	-0,11	-0,06	0,59	0,22	0,06	0,07	-0,04
30	0	-0,06	-0,11	0,57	0,22	0,05	0,06	-0,08
31	1	-0,11	-0,07	0,58	0,21	0,06	0,06	-0,04
32	0	-0,07	-0,10	0,56	0,21	0,04	0,06	-0,08
33	0	-0,10	-0,14	0,55	0,21	0,03	0,06	-0,12
34	1	-0,14	-0,11	0,56	0,20	0,04	0,06	-0,08
35	0	-0,11	-0,14	0,54	0,20	0,03	0,05	-0,12
36	0	-0,14	-0,17	0,53	0,20	0,02	0,05	-0,15
37	0	-0,17	-0,21	0,51	0,19	0,01	0,05	-0,18
38	1	-0,21	-0,17	0,53	0,19	0,02	0,05	-0,15
39	0	-0,17	-0,20	0,51	0,19	0,01	0,05	-0,18
40	0	-0,20	-0,23	0,50	0,19	0,00	0,05	-0,21
41	0	-0,23	-0,26	0,49	0,18	-0,01	0,04	-0,24
42	1	-0,26	-0,24	0,50	0,18	0,00	0,04	-0,21
43	1	-0,24	-0,21	0,51	0,18	0,01	0,04	-0,18
44	1	-0,21	-0,18	0,52	0,18	0,01	0,04	-0,16
45	1	-0,18	-0,15	0,53	0,18	0,02	0,04	-0,13

Tableau V.2 (suite)
Exemple de résultats de la simulation pour le sujet 2 (θ = -0,02)

Item	r	b	EAP(θ)	p	$S_{EAP(\theta)}$	$a3_{EAP(\theta)}$	$a4_{EAP(\theta)}$	ERREUR $_{EAP(\theta)}$
46	1	-0,15	-0,13	0,54	0,17	0,02	0,04	-0,10
47	1	-0,13	-0,10	0,55	0,17	0,03	0,04	-0,08
48	0	-0,10	-0,13	0,54	0,17	0,02	0,04	-0,10
49	0	-0,13	-0,15	0,53	0,17	0,02	0,04	-0,13
50	0	-0,15	-0,18	0,52	0,17	0,01	0,03	-0,15
51	1	-0,18	-0,15	0,53	0,17	0,02	0,04	-0,13
52	0	-0,15	-0,18	0,52	0,16	0,01	0,03	-0,15
53	0	-0,18	-0,20	0,51	0,16	0,01	0,03	-0,17
54	1	-0,20	-0,18	0,52	0,16	0,01	0,03	-0,15
55	1	-0,18	-0,15	0,53	0,16	0,01	0,03	-0,13
56	0	-0,15	-0,18	0,52	0,16	0,01	0,03	-0,15
57	1	-0,18	-0,16	0,53	0,16	0,01	0,03	-0,13
58	1	-0,16	-0,13	0,53	0,16	0,01	0,04	-0,11
59	0	-0,13	-0,15	0,53	0,15	0,01	0,03	-0,13
60	0	-0,15	-0,17	0,52	0,15	0,01	0,02	-0,15

Tableau V.3
Exemple de résultats de la simulation pour le sujet 3 (θ = -0,60)

Item	r	b	EAP(θ)	p	$S_{EAP(\theta)}$	$a3_{EAP(\theta)}$	$a4_{EAP(\theta)}$	ERREUR $_{EAP(\theta)}$
1	0	0,00	-0.56	0.00	0.82	-0.11	0.13	0.04
2	0	-0.56	-0.99	0.00	0.73	-0.18	0.16	-0.38
3	1	-0.99	-0.64	0.33	0.62	-0.04	0.21	-0.04
4	1	-0.64	-0.38	0.50	0.55	0.07	0.21	0.22
5	0	-0.38	-0.60	0.40	0.50	-0.03	0.20	0.00
6	0	-0.60	-0.79	0.33	0.47	-0.10	0.21	-0.18
7	0	-0.79	-0.95	0.29	0.45	-0.15	0.21	-0.35
8	1	-0.95	-0.80	0.38	0.41	-0.07	0.17	-0.20
9	1	-0.80	-0.67	0.44	0.39	-0.01	0.15	-0.07
10	1	-0.67	-0.55	0.50	0.37	0.04	0.15	0.05
11	1	-0.55	-0.45	0.55	0.36	0.07	0.14	0.15
12	0	-0.45	-0.55	0.50	0.34	0.03	0.13	0.05
13	0	-0.55	-0.64	0.46	0.32	-0.01	0.12	-0.04
14	0	-0.64	-0.72	0.43	0.31	-0.04	0.12	-0.12
15	0	-0.72	-0.80	0.40	0.31	-0.06	0.12	-0.20
16	1	-0.80	-0.73	0.44	0.29	-0.03	0.10	-0.12
17	0	-0.73	-0.79	0.41	0.29	-0.05	0.10	-0.19
18	1	-0.79	-0.73	0.44	0.28	-0.03	0.09	-0.13
19	1	-0.73	-0.67	0.47	0.27	0.00	0.09	-0.06
20	0	-0.67	-0.73	0.45	0.26	-0.02	0.09	-0.12
21	1	-0.73	-0.67	0.48	0.26	0.00	0.08	-0.07
22	1	-0.67	-0.62	0.50	0.25	0.01	0.08	-0.01
23	0	-0.62	-0.67	0.48	0.25	0.00	0.08	-0.06
24	1	-0.67	-0.62	0.50	0.24	0.01	0.07	-0.02
25	1	-0.62	-0.57	0.52	0.24	0.03	0.07	0.03
26	0	-0.57	-0.62	0.50	0.23	0.01	0.07	-0.01
27	1	-0.62	-0.57	0.52	0.23	0.02	0.07	0.03
28	0	-0.57	-0.61	0.50	0.22	0.01	0.06	-0.01
29	1	-0.61	-0.57	0.52	0.22	0.02	0.06	0.03
30	0	-0.57	-0.61	0.50	0.21	0.01	0.06	-0.01
31	1	-0.61	-0.58	0.52	0.21	0.02	0.06	0.03
32	0	-0.58	-0.61	0.50	0.21	0.01	0.06	-0.01
33	0	-0.61	-0.65	0.48	0.20	0.00	0.06	-0.05
34	1	-0.65	-0.61	0.50	0.20	0.01	0.05	-0.01
35	0	-0.61	-0.65	0.49	0.20	0.00	0.05	-0.04
36	0	-0.65	-0.68	0.47	0.20	-0.01	0.05	-0.08
37	0	-0.68	-0.71	0.46	0.19	-0.02	0.05	-0.11
38	1	-0.71	-0.68	0.47	0.19	-0.01	0.05	-0.08
39	0	-0.68	-0.71	0.46	0.19	-0.02	0.05	-0.11
40	0	-0.71	-0.74	0.45	0.19	-0.02	0.05	-0.14
41	0	-0.74	-0.77	0.44	0.18	-0.03	0.05	-0.17
42	1	-0.77	-0.74	0.45	0.18	-0.02	0.05	-0.14
43	1	-0.74	-0.71	0.47	0.18	-0.01	0.04	-0.11
44	0	-0.71	-0.74	0.45	0.18	-0.02	0.04	-0.14
45	1	-0.74	-0.71	0.47	0.18	-0.01	0.04	-0.11

Tableau V.3 (suite)
Exemple de résultats de la simulation pour le sujet 3 (θ = -0,60)

Item	r	b	EAP(θ)	p	$S_{EAP(\theta)}$	$a3_{EAP(\theta)}$	$a4_{EAP(\theta)}$	ERREUR $_{EAP(\theta)}$
46	1	-0,71	-0,69	0.48	0.17	-0.01	0.04	-0.09
47	1	-0,69	-0,66	0.49	0.17	0.00	0.04	-0.06
48	0	-0,66	-0,69	0.48	0.17	-0.01	0.04	-0.09
49	0	-0,69	-0,71	0.47	0.17	-0.01	0.04	-0.11
50	0	-0,71	-0,74	0.46	0.17	-0.02	0.04	-0.13
51	1	-0,74	-0,71	0.47	0.16	-0.01	0.04	-0.11
52	0	-0,71	-0,73	0.46	0.16	-0.02	0.04	-0.13
53	0	-0,73	-0,76	0.45	0.16	-0.02	0.04	-0.15
54	1	-0,76	-0,74	0.46	0.16	-0.02	0.04	-0.13
55	1	-0,74	-0,71	0.47	0.16	-0.01	0.05	-0.11
56	0	-0,71	-0,73	0.46	0.16	-0.02	0.05	-0.13
57	1	-0,73	-0,71	0.47	0.16	-0.01	0.05	-0.11
58	1	-0,71	-0,69	0.48	0.15	0.00	0.05	-0.09
59	0	-0,69	-0,71	0.47	0.15	-0.01	0.05	-0.11
60	0	-0,71	-0,73	0.47	0.15	-0.01	0.05	-0.13

Annexe VI

Vérification du calcul des estimateurs en fonction du nombre d'items administrés et du nombre de points de quadrature

Les tableaux VI.1 à VI.5 présentent respectivement la réponse aux items, l'estimateur du niveau d'habileté ainsi que l'erreur type, l'asymétrie et la kurtose de l'estimateur du niveau d'habileté en fonction du nombre de points de quadrature dans la méthode d'estimation du niveau d'habileté par la méthode EAP, lorsque le niveau d'habileté, θ, est égal à -0,49 et que le niveau de difficulté du premier item, b_1, est de -0,01. La simulation s'arrête lorsque 80 items ont été administrés. Notons qu'il a été jugé prudent de vérifier les résultats en fonction d'un nombre d'items, soit 80, plus grand que celui qui est utilisé dans cette recherche, soit 60. Les zones ombragées soulignent les résultats obtenus au delà du 60^{e} item administré.

À l'intérieur de tous les tableaux, les valeurs en caractères gras sont celles qui se démarquent de celles obtenues lorsque 80 points de quadrature sont utilisés. Ainsi, en ce qui concerne la réponse aux items, au 49^{e} item administré, il est possible de constater que, lorsque 10 points de quadrature sont utilisés, la réponse à cet item (0) diffère de celle obtenue avec 80 points de quadrature (1). À l'intérieur de tous les autres tableaux, une valeur est jugée différente de celle obtenue à partir de 80 points de quadrature quand elle est inférieure ou supérieure d'au moins 0,05. Par exemple, déjà au 13^{e} item administré, l'estimateur du niveau d'habileté obtenu à partir de 10 points de quadrature (-0,57) est différent de celui obtenu à partir de 80 points de quadrature (-0,65). Lorsque 21 points de quadrature sont utilisés, c'est au 58^{e} item qu'on remarque une différence, soit de -0,45.

Tableau VI.1
Réponse à l'item (r) en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
1	1	1	1	1	1	1	1	1	1
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1
5	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0
9	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1
11	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0
14	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0
16	1	1	1	1	1	1	1	1	1
17	1	1	1	1	1	1	1	1	1
18	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0
20	1	1	1	1	1	1	1	1	1
21	0	0	0	0	0	0	0	0	0
22	1	1	1	1	1	1	1	1	1
23	0	0	0	0	0	0	0	0	0

Tableau VI.1 (suite)
Réponse à l'item (r) en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
24	1	1	1	1	1	1	1	1	1
25	1	1	1	1	1	1	1	1	1
26	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0
30	1	1	1	1	1	1	1	1	1
31	0	0	0	0	0	0	0	0	0
32	0	1	1	1	1	1	1	1	1
33	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0
35	0	1	1	1	1	1	1	1	1
36	1	1	1	1	1	1	1	1	1
37	1	1	1	1	1	1	1	1	1
38	1	1	1	1	1	1	1	1	1
39	0	0	0	0	0	0	0	0	0
40	1	1	1	1	1	1	1	1	1
41	1	1	1	1	1	1	1	1	1
42	0	0	0	0	0	0	0	0	0
43	1	1	1	1	1	1	1	1	1
44	0	0	0	0	0	0	0	0	0
45	0	0	0	0	0	0	0	0	0
46	0	0	0	0	0	0	0	0	0
47	1	1	1	1	1	1	1	1	1
48	1	1	1	1	1	1	1	1	1

Tableau VI.1 (suite)
Réponse à l'item (r) en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
49	**0**	1	1	1	1	1	1	1	1
50	1	1	1	1	1	1	1	1	1
51	1	1	1	1	1	1	1	1	1
52	0	0	0	0	0	0	0	0	0
53	1	1	1	1	1	1	1	1	1
54	1	1	1	1	1	1	1	1	1
55	1	1	1	1	1	1	1	1	1
56	1	1	1	1	1	1	1	1	1
57	0	0	0	0	0	0	0	0	0
58	0	0	0	0	0	0	0	0	0
59	1	1	1	1	1	1	1	1	1
60	1	1	1	1	1	1	1	1	1
61	0	0	0	0	0	0	0	0	0
62	1	1	1	1	1	1	1	1	1
63	1	1	1	1	1	1	1	1	1
64	0	0	0	0	0	0	0	0	0
65	0	0	0	0	0	0	0	0	0
66	1	1	1	1	1	1	1	1	1
67	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0
69	1	1	1	1	1	1	1	1	1
70	1	1	1	1	1	1	1	1	1
71	1	1	1	1	1	1	1	1	1
72	0	0	0	0	0	0	0	0	0
73	1	1	1	1	1	1	1	1	1

Tableau VI.1 (suite)
Réponse à l'item (r) en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
74	1	1	1	1	1	1	1	1	1
75	1	1	1	1	1	1	1	1	1
76	0	0	0	0	0	0	0	0	0
77	1	1	1	1	1	1	1	1	1
78	1	1	1	1	1	1	1	1	1
79	0	0	0	0	0	0	0	0	0
80	1	1	1	1	1	1	1	1	1

Tableau VI.2
Estimateur du niveau d'habileté, EAP(θ), en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
1	0,58	0,58	0,58	0,58	0,58	0,58	0,58	0,58	0,58
2	0,17	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16
3	-0,15	-0,15	-0,15	-0,15	-0,15	-0,15	-0,15	-0,15	-0,15
4	0,12	0,11	0,11	0,11	0,11	0,11	0,11	0,11	0,11
5	-0,10	-0,10	-0,10	-0,10	-0,10	-0,10	-0,10	-0,10	-0,10
6	-0,31	-0,29	-0,29	-0,29	-0,29	-0,29	-0,29	-0,29	-0,29
7	-0,47	-0,46	-0,46	-0,46	-0,46	-0,46	-0,46	-0,46	-0,46
8	-0,60	-0,61	-0,61	-0,61	-0,61	-0,61	-0,61	-0,61	-0,61
9	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47
10	-0,39	-0,34	-0,34	-0,34	-0,34	-0,34	-0,34	-0,34	-0,34
11	-0,46	-0,45	-0,45	-0,45	-0,45	-0,45	-0,45	-0,45	-0,45
12	-0,51	-0,55	-0,55	-0,55	-0,55	-0,55	-0,53	-0,55	-0,55
13	**-0,57**	-0,65	-0,65	-0,65	-0,65	-0,65	-0,65	-0,65	-0,65
14	**-0,49**	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56
15	**-0,53**	-0,64	-0,64	-0,64	-0,64	-0,64	-0,64	-0,64	-0,64
16	**-0,48**	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56
17	-0,45	-0,49	-0,49	-0.49	-0,49	-0,49	-0,49	-0,49	-0,49
18	**-0,46**	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56
19	-0,48	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62
20	**-0,46**	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56
21	**-0,47**	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62
22	**-0,45**	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56
23	**-0,46**	-0,61	-0,61	-0,61	-0,61	-0,61	-0,61	-0,61	-0,61

Tableau VI.2 (suite)
Estimateur du niveau d'habileté, EAP(θ), en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
24	**-0,45**	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56
25	**-0,45**	-0,51	-0,51	-0,51	-0,51	-0,51	-0,51	-0,51	-0,51
26	**-0,45**	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56
27	**-0,45**	-0,60	-0,61	-0,61	-0,61	-0,61	-0,61	-0,61	-0,61
28	**-0,45**	-0,65	-0,65	-0,65	-0,65	-0,65	-0,65	-0,65	-0,65
29	**-0,46**	-0,70	-0,69	-0,69	-0,69	-0,69	-0,69	-0,69	-0,69
30	**-0,45**	-0,65	-0,65	-0,65	-0,65	-0,65	-0,65	-0,65	-0,65
31	**-0,45**	-0,70	-0,69	-0,69	-0,69	-0,69	-0,69	-0,69	-0,69
32	**-0,46**	-0,66	-0,65	-0,65	-0,65	-0,65	-0,65	-0,65	-0,65
33	**-0,46**	-0,70	-0,69	-0,69	-0,69	-0,69	-0,69	-0,69	-0,69
34	**-0,48**	-0,73	-0,73	-0,73	-0,73	-0,73	-0,73	-0,73	-0,73
35	**-0,50**	-0,70	-0,69	-0,69	-0,69	-0,69	-0,69	-0,69	-0,69
36	**-0,47**	-0,66	-0,66	-0,66	-0,66	-0,66	-0,66	-0,66	-0,66
37	**-0,45**	-0,63	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62
38	**-0,45**	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59
39	**-0,45**	-0,63	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62
40	**-0,45**	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59
41	**-0,45**	-0,55	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56
42	**-0,45**	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59
43	**-0,45**	-0,55	-0,57	-0,56	-0,56	-0,56	-0,56	-0,56	-0,56
44	**-0,45**	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59
45	**-0,45**	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62
46	**-0,45**	-0,66	-0,65	-0,64	-0,64	-0,64	-0,64	-0,64	-0,64
47	**-0,45**	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62	-0,62
48	**-0,44**	-0,59	-0,60	-0,59	-0,59	-0,59	-0,59	-0,59	-0,59

Tableau VI.2 (suite)
Estimateur du niveau d'habileté, EAP(θ), en fonction du nombre de points de quadrature ($\theta = -0,49$, $b_1 = -0,01$)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
49	**-0,44**	-0,55	-0,57	-0,57	-0,57	-0,57	-0,57	-0,57	-0,57
50	**-0,44**	-0,52	-0,55	-0,54	-0,54	-0,54	-0,54	-0,54	-0,54
51	**-0,44**	-0,49	-0,52	-0,52	-0,52	-0,52	-0,52	-0,52	-0,52
52	**-0,44**	-0,52	-0,55	-0,54	-0,54	-0,54	-0,54	-0,54	-0,54
53	**-0,44**	-0,49	-0,52	-0,52	-0,52	-0,52	-0,52	-0,52	-0,52
54	-0,44	-0,46	-0,50	-0,50	-0,50	-0,50	-0,50	-0,50	-0,50
55	-0,44	-0,44	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47
56	-0,44	-0,43	-0,45	-0,45	-0,45	-0,45	-0,45	-0,45	-0,45
57	-0,44	-0,44	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47
58	**-0,44**	**-0,45**	-0,50	-0,49	-0,50	-0,50	-0,50	-0,50	-0,50
59	-0,44	-0,44	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47
60	-0,44	-0,42	-0,45	-0,45	-0,45	-0,45	-0,45	-0,45	-0,45
61	-0,44	-0,43	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47
62	-0,44	-0,42	-0,44	-0,45	-0,46	-0,46	-0,46	-0,46	-0,46
63	-0,44	-0,41	-0,42	-0,43	-0,44	-0,44	-0,44	-0,44	-0,44
64	-0,44	-0,42	-0,44	-0,45	-0,46	-0,45	-0,45	-0,45	-0,45
65	-0,44	-0,43	-0,46	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47
66	-0,44	-0,42	-0,44	-0,45	-0,46	-0,46	-0,46	-0,46	-0,46
67	-0,44	-0,43	-0,46	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47
68	**-0,44**	**-0,44**	-0,49	-0,49	-0,49	-0,49	-0,49	-0,49	-0,49
69	-0,44	**-0,42**	-0,46	-0,47	-0,47	-0,47	-0,47	-0,47	-0,47
70	-0,44	**-0,41**	-0,44	-0,45	-0,46	-0,46	-0,46	-0,46	-0,46
71	-0,44	-0,41	-0,42	-0,44	-0,44	-0,44	-0,44	-0,44	-0,44
72	-0,44	**-0,41**	-0,44	-0,45	-0,46	-0,46	-0,46	-0,46	-0,46
73	-0,44	-0,41	-0,42	-0,44	-0,44	-0,44	-0,44	-0,44	-0,44

Tableau VI.2 (suite)
Estimateur du niveau d'habileté, EAP(θ), en fonction du nombre de points de quadrature ($\theta = -0,49$, $b_1 = -0,01$)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
74	-0,44	-0,40	-0,40	-0,42	-0,42	-0,42	-0,42	-0,42	-0,42
75	-0,44	-0,40	-0,38	-0,41	-0,41	-0,41	-0,41	-0,41	-0,41
76	-0,44	-0,40	-0,39	-0,42	-0,42	-0,42	-0,42	-0,42	-0,42
77	-0,44	-0,40	-0,38	-0,41	-0,41	-0,41	-0,41	-0,41	-0,41
78	**-0,44**	-0,40	-0,37	-0,39	-0,39	-0,39	-0,39	-0,39	-0,39
79	-0,44	-0,40	-0,38	-0,41	-0,41	-0,41	-0,41	-0,41	-0,41
80	**-0,44**	-0,40	-0,36	-0,40	-0,39	-0,39	-0,39	-0,39	-0,39

Tableau VI.3
Erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
1	0,82	0,82	0,82	0,82	0,82	0,82	0,82	0,82	0,82
2	0,69	0,69	0,69	0,69	0,69	0,69	0,69	0,69	0,69
3	0,63	0,62	0,62	0,62	0,62	0,62	0,62	0,62	0,62
4	0,55	0,55	0,55	0,55	0,55	0,55	0,55	0,55	0,55
5	0,52	0,50	0,50	0,50	0,50	0,50	0,50	0,50	0,50
6	0,47	0,48	0,48	0,48	0,48	0,48	0,48	0,48	0,48
7	0,42	0,46	0,46	0,46	0,46	0,46	0,46	0,46	0,46
8	0,42	0,44	0,44	0,44	0,44	0,44	0,44	0,44	0,44
9	**0,34**	0,40	0,40	0,40	0,40	0,40	0,40	0,40	0,40
10	**0,31**	0,38	0,38	0,38	0,38	0,38	0,38	0,38	0,38
11	**0,27**	0,36	0,36	0,36	0,36	0,36	0,36	0,36	0,36
12	**0,28**	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35
13	0,32	0,35	0,35	0,35	0,35	0,35	0,35	0,35	0,35
14	**0,23**	0,33	0,33	0,33	0,33	0,33	0,33	0,33	0,33
15	**0,27**	0,32	0,32	0,32	0,32	0,32	0,32	0,32	0,32
16	**0,18**	0,30	0,30	0,30	0,30	0,30	0,30	0,30	0,30
17	**0,13**	0,29	0,29	0,29	0,29	0,29	0,29	0,29	0,29
18	**0,14**	0,29	0,29	0,29	0,29	0,29	0,29	0,29	0,29
19	**0,18**	0,28	0,28	0,28	0,28	0,28	0,28	0,28	0,28
20	**0,11**	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27
21	**0,14**	0,27	0,27	0,27	0,27	0,27	0,27	0,27	0,27
22	**0,09**	0,26	0,26	0,26	0,26	0,26	0,26	0,26	0,26
23	**0,11**	0,26	0,25	0,25	0,25	0,25	0,25	0,25	0,25

Tableau VI.3 (suite)
Erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
24	**0,07**	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,25
25	**0,05**	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,24
26	**0,05**	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,24
27	**0,07**	0,24	0,23	0,23	0,23	0,23	0,23	0,23	0,23
28	**0,08**	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
29	**0,11**	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
30	**0,06**	0,23	0,22	0,22	0,22	0,22	0,22	0,22	0,22
31	**0,08**	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,22
32	**0,10**	0,22	0,21	0,21	0,21	0,21	0,21	0,21	0,21
33	**0,13**	0,21	0,21	0,21	0,21	0,21	0,21	0,21	0,21
34	**0,17**	0,21	0,21	0,21	0,21	0,21	0,21	0,21	0,21
35	**0,21**	0,21	0,20	0,20	0,20	0,20	0,20	0,20	0,20
36	**0,14**	0,21	0,20	0,20	0,20	0,20	0,20	0,20	0,20
37	**0,08**	0,21	0,20	0,20	0,20	0,20	0,20	0,20	0,20
38	**0,05**	0,21	0,19	0,19	0,19	0,19	0,19	0,19	0,19
39	**0,06**	0,21	0,19	0,19	0,19	0,19	0,19	0,19	0,19
40	**0,04**	0,21	0,19	0,19	0,19	0,19	0,19	0,19	0,19
41	**0,02**	0,20	0,19	0,19	0,19	0,19	0,19	0,19	0,19
42	**0,03**	0,21	0,18	0,18	0,18	0,18	0,18	0,18	0,18
43	**0,02**	0,20	0,18	0,18	0,18	0,18	0,18	0,18	0,18
44	**0,02**	0,21	0,18	0,18	0,18	0,18	0,18	0,18	0,18
45	**0,03**	0,20	0,17	0,18	0,18	0,18	0,18	0,18	0,18
46	**0,04**	0,20	0,17	0,18	0,18	0,18	0,18	0,18	0,18
47	**0,02**	0,20	0,17	0,17	0,17	0,17	0,17	0,17	0,17
48	**0,01**	0,20	0,17	0,17	0,17	0,17	0,17	0,17	0,17

Tableau VI.3 (suite)
Erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
49	**0,02**	0,20	0,17	0,17	0,17	0,17	0,17	0,17	0,17
50	**0,01**	0,19	0,18	0,17	0,17	0,17	0,17	0,17	0,17
51	**0,01**	0,18	0,18	0,17	0,17	0,17	0,17	0,17	0,17
52	**0,01**	0,19	0,17	0,17	0,17	0,17	0,17	0,17	0,17
53	**0,00**	0,17	0,18	0,17	0,16	0,16	0,16	0,16	0,16
54	**0,00**	0,16	0,18	0,16	0,16	0,16	0,16	0,16	0,16
55	**0,00**	0,14	0,17	0,16	0,16	0,16	0,16	0,16	0,16
56	**0,00**	0,13	0,17	0,16	0,16	0,16	0,16	0,16	0,16
57	**0,00**	0,13	0,17	0,16	0,16	0,16	0,16	0,16	0,16
58	**0,00**	0,15	0,17	0,16	0,16	0,16	0,16	0,16	0,16
59	**0,00**	0,13	0,17	0,15	0,15	0,15	0,15	0,15	0,15
60	**0,00**	0,11	0,17	0,15	0,15	0,15	0,15	0,15	0,15
61	**0,00**	0,12	0,17	0,15	0,15	0,15	0,15	0,15	0,15
62	**0,00**	0,11	0,17	0,15	0,15	0,15	0,15	0,15	0,15
63	**0,00**	0,10	0,16	0,15	0,15	0,15	0,15	0,15	0,15
64	**0,00**	0,10	0,16	0,15	0,15	0,15	0,15	0,15	0,15
65	**0,00**	0,11	0,17	0,15	0,15	0,15	0,15	0,15	0,15
66	**0,00**	0,10	0,16	0,14	0,15	0,15	0,15	0,15	0,15
67	**0,00**	0,11	0,17	0,14	0,14	0,14	0,14	0,14	0,15
68	**0,00**	0,12	0,17	0,15	0,14	0,14	0,14	0,14	0,14
69	**0,00**	0,10	0,17	0,14	0,14	0,14	0,14	0,14	0,14
70	**0,00**	**0,09**	0,16	0,14	0,14	0,14	0,14	0,14	0,14
71	**0,00**	**0,08**	0,15	0,13	0,14	0,14	0,14	0,14	0,14
72	**0,00**	**0,09**	0,16	0,13	0,14	0,14	0,14	0,14	0,14
73	**0,00**	**0,08**	0,15	0,13	0,14	0,14	0,14	0,14	0,14

Tableau VI.3 (suite)
Erreur type de l'estimateur du niveau d'habileté, $S_{EAP(\theta)}$, en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
74	**0,00**	**0,07**	0,14	0,13	0,14	0,14	0,14	0,14	0,14
75	**0,00**	**0,07**	0,13	0,13	0,14	0,14	0,14	0,14	0,14
76	**0,00**	**0,07**	0,14	0,13	0,14	0,14	0,14	0,14	0,14
77	**0,00**	**0,06**	0,13	0,12	0,14	0,14	0,14	0,14	0,14
78	**0,00**	**0,07**	0,12	0,12	0,14	0,14	0,13	0,13	0,13
79	**0,00**	**0,06**	0,12	0,12	0,13	0,13	0,13	0,13	0,13
80	**0,00**	**0,06**	0,11	0,12	0,13	0,13	0,13	0,13	0,13

Tableau VI.4
Asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
1	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10
2	-0,06	-0,05	-0,05	-0,05	-0,05	-0,05	-0,05	-0,05	-0,05
3	-0,14	-0,15	-0,15	-0,15	-0,15	-0,15	-0,15	-0,15	-0,15
4	**-0,10**	-0,02	-0,02	-0,02	-0,02	-0,02	-0,02	-0,02	-0,02
5	**0,00**	-0,11	-0,11	-0,11	-0,11	-0,11	-0,11	-0,11	-0,11
6	**0,15**	-0,16	-0,16	-0,16	-0,16	-0,16	-0,16	-0,16	-0,16
7	-0,19	-0,20	-0,20	-0,20	-0,20	-0,20	-0,20	-0,20	-0,20
8	**-0,76**	-0,23	-0,23	-0,23	-0,23	-0,23	-0,23	-0,23	-0,23
9	**-0,35**	-0,14	-0,14	-0,14	-0,14	-0,14	-0,14	-0,14	-0,14
10	**-0,89**	-0,08	-0,08	-0,08	-0,08	-0,08	-0,08	-0,08	-0,08
11	**-0,48**	-0,11	-0,11	-0,11	-0,11	-0,11	-0,11	-0,11	-0,11
12	**-1,81**	-0,13	-0,13	-0,13	-0,13	-0,13	-0,13	-0,13	-0,13
13	**-1,81**	-0,15	-0,15	-0,15	-0,15	-0,15	-0,15	-0,15	-0,15
14	**-2,59**	-0,11	-0,11	-0,11	-0,11	-0,11	-0,11	-0,11	-0,11
15	**-2,48**	-0,12	-0,12	-0,12	-0,12	-0,12	-0,12	-0,12	-0,12
16	**-3,54**	-0,09	-0,09	-0,09	-0,09	-0,09	-0,09	-0,09	-0,09
17	**-2,40**	-0,06	-0,05	-0,05	-0,05	-0,05	-0,05	-0,05	-0,05
18	**-4,73**	-0,08	-0,07	-0,07	-0,07	-0,07	-0,07	-0,07	-0,07
19	**-4,43**	-0,09	-0,09	-0,09	-0,09	-0,09	-0,09	-0,09	-0,09
20	**-6,25**	-0,08	-0,06	-0,06	-0,06	-0,06	-0,06	-0,06	-0,06
21	**-5,81**	-0,08	-0,07	-0,07	-0,07	-0,07	-0,07	-0,07	-0,07
22	**-8,21**	-0,09	-0,05	-0,05	-0,05	-0,05	-0,05	-0,05	-0,05
23	**-7,60**	-0,07	-0,06	-0,06	-0,06	-0,06	-0,06	-0,06	-0,06

Tableau VI.4 (suite)
Asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
24	-10,72	**-0,10**	-0,04	-0,04	-0,05	-0,05	-0,05	-0,05	-0,05
25	-8,09	**-0,13**	-0,03	-0,03	-0,03	-0,03	-0,03	-0,03	-0,03
26	-13,99	**-0,12**	-0,03	-0,04	-0,04	-0,04	-0,04	-0,04	-0,04
27	-12,93	-0,06	-0,03	-0,05	-0,05	-0,05	-0,05	-0,05	-0,05
28	-10,44	**0,03**	-0,05	-0,06	-0,06	-0,06	-0,06	-0,06	-0,06
29	-8,16	**0,10**	-0,07	-0,07	-0,07	-0,07	-0,07	-0,07	-0,07
30	-13,55	**0,08**	-0,04	-0,05	-0,05	-0,05	-0,05	-0,05	-0,05
31	-10,62	**0,17**	-0,07	-0,06	-0,06	-0,06	-0,06	-0,06	-0,06
32	-8,25	**0,13**	-0,03	-0,05	-0,05	-0,05	-0,05	-0,05	-0,05
33	-6,36	**0,24**	-0,07	-0,05	-0,06	-0,06	-0,06	-0,06	-0,06
34	-4,86	**0,27**	**-0,11**	-0,06	-0,06	-0,06	-0,06	-0,06	-0,06
35	-3,66	**0,32**	-0,07	-0,05	-0,05	-0,05	-0,05	-0,05	-0,05
36	-6,27	**0,26**	-0,01	-0,03	-0,04	-0,04	-0,04	-0,04	-0,04
37	-10,53	**0,11**	**0,06**	-0,02	-0,03	-0,03	-0,03	-0,03	-0,03
38	-17,58	**-0,09**	**0,11**	-0,01	-0,02	-0,02	-0,02	-0,02	-0,02
39	-13,70	**0,13**	**0,09**	-0,01	-0,03	-0,03	-0,03	-0,03	-0,03
40	-22,84	**-0,09**	**0,15**	-0,01	-0,02	-0,02	-0,02	-0,02	-0,02
41	-37,99	**-0,34**	**0,16**	-0,01	-0,01	-0,01	-0,01	-0,01	-0,01
42	-29,70	**-0,11**	**0,19**	0,00	-0,02	-0,02	-0,02	-0,02	-0,02
43	-49,38	**-0,37**	**0,20**	-0,01	-0,01	-0,01	-0,01	-0,01	-0,01
44	-38,62	**-0,13**	**0,24**	0,01	-0,02	-0,02	-0,02	-0,02	-0,02
45	-30,16	**0,15**	**0,22**	0,02	-0,02	-0,02	-0,02	-0,02	-0,02
46	-23,53	**0,43**	**0,13**	**0,02**	-0,02	-0,03	-0,03	-0,03	-0,03
47	-39,19	**0,17**	**0,27**	**0,03**	-0,02	-0,02	-0,02	-0,02	-0,02
48	-65,22	**-0,11**	**0,35**	0,03	-0,01	-0,02	-0,01	-0,01	-0,01

Tableau VI.4 (suite)
Asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
49	-50,95	**-0,41**	**0,35**	0,01	-0,01	-0,01	-0,01	-0,01	-0,01
50	-84,78	**-0,72**	**0,28**	-0,02	-0,01	0,00	0,00	0,00	0,00
51	-140,91	**-1,02**	**0,15**	**-0,06**	-0,01	0,00	0,00	0,00	0,00
52	-110,22	**-0,79**	**0,31**	-0,03	-0,02	0,00	0,00	0,00	0,00
53	-183,19	**-1,11**	**0,17**	**-0,07**	-0,01	0,00	0,00	0,00	0,00
54	-302,88	**-1,39**	-0,01	**-0,11**	-0,01	0,01	0.01	0.01	0,01
55	-489,13	**-1,56**	**-0,20**	**-0,12**	0,01	0,01	0.01	0.01	0,01
56	-710,63	**-1,51**	**-0,38**	**-0,10**	0,02	0,02	0,02	0,02	0,02
57	-635,88	**-1,71**	**-0,22**	**-0,14**	0,01	0,01	0,01	0,01	0,01
58	-511,81	**-1,65**	**-0,03**	**-0,15**	-0,01	0,01	0,01	0,01	0,01
59	-826,63	**-1,86**	**-0,24**	**-0,17**	0,00	0,01	0,01	0,01	0,01
60	-1201,00	**-1,82**	**-0,45**	**-0,15**	0,03	0,02	0,01	0,01	0,01
61	-1074,50	**-2,01**	**-0,27**	**-0,20**	0,00	0,01	0,01	0,01	0,01
62	-1561,25	**-1,98**	**-0,49**	**-0,18**	0,03	0,02	0,01	0,01	0,01
63	-1392,75	**-1,52**	**-0,69**	**-0,11**	0,05	0,02	0,02	0,02	0,02
64	-2029,75	**-2,14**	**-0,53**	**-0,21**	0,03	0,02	0,01	0,01	0,01
65	-1816,25	**-2,34**	**-0,33**	**-0,27**	0,00	0,02	0,01	0,01	0,01
66	-2638,50	**-2,32**	**-0,57**	**-0,24**	0,03	0,02	0,01	0,01	0,01
67	-2361,00	**-2,52**	**-0,36**	**-0,31**	0,00	0,02	0,01	0,01	0,01
68	-1900,50	**-2,41**	**-0,13**	**-0,32**	-0,04	0,01	0,01	0,00	0,00
69	-3069,50	**-2,70**	**-0,38**	**-0,34**	0,00	0,02	0,01	0,01	0,01
70	-4459,00	**-2,70**	**-0,63**	**-0,32**	0,04	0,03	0,01	0,01	0,01
71	-3978,00	**-2,16**	**-0,87**	**-0,22**	**0,08**	0,03	0,01	0,01	0,01
72	-5797,00	**-2,90**	**-0,67**	**-0,35**	0,05	0,03	0,01	0,01	0,01
73	-5171,00	**-2,33**	**-0,93**	**-0,25**	**0,09**	0,03	0,01	0,01	0,01

Tableau VI.4 (suite)
Asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, $a3_{EAP(\theta)}$, en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
74	**2227,50**	**-1,04**	**-1,16**	**-0,09**	**0,12**	0,03	0,01	0,02	0,02
75	**8510,00**	**-0,73**	**-1,33**	**0,10**	**0,14**	0,02	0,01	0,02	0,02
76	**2896,00**	**-1,11**	**-1,23**	**-0,10**	**0,14**	0,03	0,01	0,01	0,02
77	**11064,00**	**0,77**	**-1,42**	**0,11**	**0,16**	0,02	0,01	0,02	0,02
78	**11600,00**	**2,50**	**-1,49**	**0,32**	**0,15**	0,01	0,02	0,02	0,02
79	**14382,00**	**0,84**	**-1,51**	**0,12**	**0,18**	0,02	0,01	0,02	0,02
80	**15080,00**	**2,67**	**-1,59**	**0,35**	**0,17**	0,00	0,01	0,02	0,02

Tableau VI.5
Kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
1	0,15	0,14	0,13	0,13	0,13	0,13	0,13	0,13	0,13
2	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,22	0,22
3	**0,13**	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,24
4	**-0,02**	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
5	**-0,31**	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
6	**0,55**	0,23	0,23	0,23	0,23	0,23	0,23	0,23	0,23
7	**1,87**	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,24
8	**1,25**	0,24	0,24	0,24	0,24	0,24	0,24	0,24	0,24
9	**4,00**	0,18	0,18	0,18	0,18	0,18	0,18	0,18	0,18
10	**4,55**	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15
11	**7,98**	0,15	0,15	0,15	0,15	0,15	0,15	0,15	0,15
12	**6,03**	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16
13	**2,37**	0,16	0,16	0,16	0,16	0,16	0,16	0,16	0,16
14	**10,67**	0,13	0,13	0,13	0,13	0,13	0,13	0,13	0,13
15	**5,30**	0,12	0,13	0,13	0,13	0,13	0,13	0,13	0,13
16	**18,63**	0,10	0,11	0,11	0,11	0,11	0,11	0,11	0,11
17	**42,87**	0,12	0,10	0,10	0,10	0,10	0,10	0,10	0,10
18	**32,99**	0,07	0,10	0,10	0,10	0,10	0,10	0,10	0,10
19	**19,80**	**0,03**	0,10	0,10	0,10	0,10	0,10	0,10	0,10
20	**56,48**	**0,02**	0,08	0,09	0,09	0,09	0,09	0,09	0,09
21	**35,06**	**-0,04**	0,09	0,09	0,09	0,09	0,09	0,09	0,09
22	**96,20**	**-0,05**	0,07	0,08	0,08	0,08	0,08	0,08	0,08
23	**60,87**	**-0,14**	0,09	0,08	0,08	0,08	0,08	0,08	0,08

Tableau VI.5 (suite)
Kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
24	**163,31**	**-0,14**	0,05	0,07	0,07	0,07	0,07	0,07	0,07
25	**356,06**	**0,02**	0,03	0,07	0,07	0,07	0,07	0,07	0,07
26	**277,63**	**-0,24**	0,04	0,07	0,07	0,07	0,07	0,07	0,07
27	**178,88**	**-0,40**	0,08	0,07	0,07	0,07	0,07	0,07	0,07
28	**108,92**	**-0,34**	**0,14**	0,07	0,07	0,07	0,07	0,07	0,07
29	**64,83**	**-0,07**	**0,17**	0,07	0,07	0,07	0,07	0,07	0,07
30	**184,69**	**-0,45**	**0,18**	0,07	0,06	0,06	0,06	0,06	0,06
31	**111,20**	**-0,10**	**0,21**	0,07	0,07	0,07	0,07	0,07	0,07
32	**66,05**	**-0,55**	**0,23**	0,06	0,06	0,06	0,06	0,06	0,06
33	**38,47**	**-0,13**	**0,27**	0,08	0,06	0,06	0,06	0,06	0,06
34	**21,65**	**-0,50**	**0,21**	0,08	0,06	0,06	0,06	0,06	0,06
35	**11,41**	**-0,12**	**0,35**	0,08	0,06	0,06	0,06	0,06	0,06
36	**37,27**	**-0,70**	**0,40**	0,07	0,05	0,05	0,05	0,05	0,05
37	**108,94**	**-1,09**	**0,31**	0,04	0,05	0,05	0,05	0,05	0,05
38	**307,44**	**-1,22**	0,09	0,01	0,05	0,05	0,05	0,05	0,05
39	**185,75**	**-1,22**	**0,38**	0,04	0,04	0,05	0,05	0,05	0,05
40	**520,13**	**-1,34**	**0,12**	**-0,01**	0,04	0,05	0,05	0,05	0,05
41	**1445,75**	**-1,17**	**-0,21**	**-0,05**	0,04	0,04	0,04	0,04	0,04
42	**880,75**	**-1,44**	**0,13**	**-0,03**	0,04	0,04	0,04	0,04	0,04
43	**2444,00**	**-1,25**	**-0,25**	**-0,08**	0,03	0,04	0,04	0,04	0,04
44	**1490,25**	**-1,53**	**0,15**	**-0,06**	0,03	0,04	0,04	0,04	0,04
45	**907,50**	**-1,54**	**0,63**	0,01	0,03	0,04	0,04	0,04	0,04
46	**551,88**	**-1,26**	**1,06**	**0,11**	0,03	0,04	0,04	0,04	0,04
47	**1534,25**	**-1,61**	**0,76**	0,01	0,03	0,04	0,04	0,04	0,04
48	**4255,00**	**-1,68**	**0,27**	**-0,11**	0,02	0,04	0,04	0,04	0,04

Tableau VI.5 (suite)
Kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
49	**2594,50**	**-1,49**	**-0,26**	**-0,21**	0,01	0,04	0,04	0,04	0,04
50	**7192,00**	**-0,98**	**-0,72**	**-0,24**	0,02	0,04	0,04	0,04	0,04
51	**19912**	**-0,08**	**-1,03**	**-0,20**	0,05	0,04	0,04	0,04	0,04
52	**12156**	**-0,93**	**-0,79**	**-0,30**	0,02	0,04	0,04	0,04	0,04
53	**33656**	**0,03**	**-1,11**	**-0,25**	0,05	0,04	0,03	0,03	0,03
54	**92976**	**1,50**	**-1,24**	**-0,10**	**0,09**	0,04	0,03	0,03	0,03
55	**254496**	**3,56**	**-1,17**	**0,11**	**0,11**	0,04	0,03	0,03	0,03
56	**668928**	**6,16**	**-0,88**	**0,35**	**0,12**	0,03	0,03	0,03	0,03
57	**430144**	**4,19**	**-1,24**	**0,13**	**0,13**	0,04	0,03	0,03	0,03
58	**265600**	**2,35**	**-1,41**	**-0,13**	**0,12**	0,04	0,03	0,03	0,03
59	**726912**	**4,84**	**-1,31**	**0,15**	**0,16**	0,04	0,03	0,03	0,03
60	**1910528**	**8,00**	**-0,99**	**0,45**	**0,16**	0,03	0,03	0,03	0,03
61	**1228544**	**5,63**	**-1,36**	**0,17**	**0,19**	0,04	0,03	0,03	0,03
62	**3229184**	**9,13**	**-1,03**	**0,52**	**0,19**	0,03	0,03	0,03	0,03
63	**7260160**	**12,90**	**-0,46**	**0,82**	**0,15**	0,01	0,03	0,03	0,03
64	**5456896**	**10,45**	**-1,04**	**0,59**	**0,22**	0,03	0,03	0,03	0,03
65	**3508736**	**7,59**	**-1,43**	**0,25**	**0,26**	0,04	0,03	0,03	0,03
66	**9222144**	**11,88**	**-1,06**	**0,67**	**0,26**	0,02	0,02	0,03	0,03
67	**5929984**	**8,72**	**-1,46**	**0,30**	**0,30**	0,05	0,02	0,03	0,03
68	**3661824**	**5,84**	**-1,67**	**-0,12**	**0,28**	0,07	0,03	0,03	0,03
69	**10020864**	**9,94**	**-1,50**	**0,33**	**0,34**	0,05	0,02	0,03	0,03
70	**26341376**	**15,21**	**-1,10**	**0,82**	**0,35**	0,02	0,02	0,03	0,03
71	**59219968**	**21,00**	**-0,46**	**1,27**	**0,28**	-0,01	0,02	0,03	0,03
72	**44515328**	**17,28**	**-1,09**	**0,93**	**0,40**	0,02	0,01	0,03	0,03
73	**100089856**	**23,77**	**-0,42**	**1,42**	**0,32**	**-0,02**	0,01	0,03	0,03

Tableau VI.5 (suite)
Kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, $a4_{EAP(\theta)}$, en fonction du nombre de points de quadrature (θ = -0,49, b_1 = -0,01)

	NOMBRE DE POINTS DE QUADRATURE								
ITEM	10	21	25	30	35	40	45	50	80
74	149454848	29,20	0,56	1,76	0,18	-0,05	0,02	0,03	0,03
75	133873664	31,45	1,88	1,86	0,00	-0,07	0,02	0,03	0,03
76	252575744	32,96	0,72	1,95	0,19	-0,07	0,01	0,03	0,03
77	226230272	35,48	2,12	2,06	0,00	-0,09	0,02	0,03	0,02
78	153911296	33,54	3,89	1,88	-0,21	-0,09	0,03	0,03	0,02
79	382337024	39,98	2,42	2,27	-0,01	-0,11	0,02	0,03	0,02
80	260112384	37,83	4,31	2,07	-0,23	-0,10	0,04	0,03	0,02

Remerciements

Pour tout le support qu'il m'a offert, je désire témoigner ma reconnaissance à M. Jean-Guy Blais, vice-doyen à la gestion et au développement à la Faculté des sciences de l'éducation de l'Université de Montréal. Ses conseils se sont toujours avérés judicieux et sa grande disponibilité a été fort appréciée. Mais plus que tout, il a su faire preuve d'un grand respect des conditions de réalisation de cette recherche et a été constamment une source de motivation à la réalisation de celle-ci.

Je remercie également M. Dany Laveault, vice-doyen à la recherche à la Faculté d'éducation de l'Université d'Ottawa, pour les fructueux et cordiaux moments de discussion qui m'ont permis de partager mes réflexions avec un spécialiste du sujet non impliqué directement dans le projet.

Un grand merci aussi à Suzanne qui a supporté, dans tous les sens du mot, mon travail et qui s'est montrée disponible à mes multiples moments de non disponibilité. Merci, de plus, à mes amis de leur encouragements constants.

Enfin, la rédaction de la thèse n'aurait pas eu être faite avec autant de qualité sans la contribution d'Aline Côté, éditrice aux Éditions Berger, à la révision linguistique du texte. Son travail, habile et rapide, a permis d'assurer la conformité avec les pratiques de l'écriture en langue française.