

Chapitre 9

Le testing adaptatif

Table des matières

- 9.0 Introduction
- 9.1 Problèmes de précision et limites à l'administration des tests papier crayon
- 9.2 Testing adaptatif
 - 9.2.1 Déroulement d'un test adaptatif
- 9.3 Le testing adaptatif : une application fort pertinente de la théorie de la réponse à l'item
 - 9.3.1 Les stratégies quant à la règle d'arrêt
 - 9.3.2 Les stratégies quant à la règle de suite
 - 9.3.2.1 Stratégie de maximisation de l'information
 - 9.3.2.2 Stratégie de minimisation de l'espérance de l'erreur-type a posteriori
 - 9.3.2.3 Minitests
- 9.4 Considérations diverses
 - 9.4.1 Une formule de prophétie adaptée aux tests adaptatifs
 - 9.4.2 Logiciels disponibles
- 9.5 Défis et enjeux du testing adaptatif
- 9.6 Exercices
- 9.7 Corrigé des exercices nécessitant des calculs
- 9.8 Références

Liste des tableaux

- 9.1 Algorithme décrivant le déroulement normal d'un test papier crayon fixe et invariable
- 9.2 Algorithme décrivant le déroulement d'un test adaptatif basé sur la théorie de la réponse à l'item
- 9.3 Biais de l'estimateur du niveau d'habileté selon la distance entre l'estimateur a priori et le niveau d'habileté en fonction de quatre valeurs de l'erreur-type retenue pour la règle d'arrêt lorsque l'estimateur a priori du niveau d'habileté est fixé à 0,00
- 9.4 Estimateur du niveau d'habileté, erreur-type de celui-ci et réponse à l'item en testing adaptatif en fonction du nombre d'items administrés selon quatre méthodes d'estimation provisoire lorsque le niveau d'habileté est fixé à - 3,00

Liste des figures

- 9.1 Déroulement général d'un test adaptatif
- 9.2 Structure d'un test adaptatif basé sur la théorie de la réponse à l'item
- 9.3 Biais de l'estimateur du niveau d'habileté selon la distance entre l'estimateur a priori et le niveau d'habileté en fonction de quatre valeurs de l'erreur-type retenue pour la règle d'arrêt lorsque l'estimateur a priori du niveau d'habileté est fixé à 0,00
- 9.4 Exemple d'un minitest
- 9.5 Estimateur du niveau d'habileté, erreur-type de celui-ci et réponse à l'item en testing adaptatif en fonction du nombre d'items administrés selon quatre méthodes d'estimation provisoire lorsque le niveau d'habileté est fixé à - 3,00
- 9.6 Estimateur du niveau d'habileté en testing adaptatif en fonction du nombre d'items administrés selon quatre valeurs du niveau d'habileté
- 9.7 Erreur-type de l'estimateur du niveau d'habileté en testing adaptatif en fonction du nombre d'items administrés selon quatre valeurs du niveau d'habileté

Liste des équations

- 9.1 Prédiction du biais par régression cubique quant $S_{\theta} = 0,40$
- 9.2 Prédiction du biais par régression cubique quant $S_{\theta} = 0,35$
- 9.3 Prédiction du biais par régression cubique quant $S_{\theta} = 0,30$
- 9.4 Prédiction du biais par régression cubique quant $S_{\theta} = 0,20$
- 9.5 Information, au sens de Fisher, fournie par chaque item
- 9.6 Approximation du coefficient de fidélité
- 9.7 Correction du biais par la méthode de Bock et Mislevy
- 9.8 Formule de prophétie de l'erreur-type en testing adaptatif
- 9.9 Formule de prophétie du nombre d'items à administrer en testing adaptatif

Chapitre 9

Le testing adaptatif

Gilles Raïche, Faculté des sciences de l'éducation, Université de Moncton

24 février 2002

9.0 Introduction

Selon les habitudes développées au 20^e 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 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.

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. Il peut aussi être d'ordre psychomoteur et le niveau d'habileté vise ainsi un comportement moteur. Dans tous ces cas, le test ne permet d'obtenir qu'un estimateur de ce niveau d'habileté : il n'est qu'une occasion pour l'étudiant de manifester son habileté.

Au 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 et plusieurs limites à l'administration d'un tel test papier crayon fixe et invariable existent cependant. Nous décrivons ici ces problèmes ainsi que ces limites pour ensuite présenter une proposition de solution à ceux-ci, soit le testing adaptatif.

9.1 Problèmes de précision et limites à l'administration des tests papier crayon

Dans un test papier crayon fixe et invariable, le niveau de difficulté des items auxquels doit répondre l'étudiant ne correspond pas toujours au niveau d'habileté de ce dernier. L'étudiant peut faire face à certains items trop faciles ou trop difficiles pour lui. Dans le premier cas, aucun défi n'est relevé, et l'étudiant peut avoir l'impression de perdre son temps. Cela peut alors se traduire par des réponses erronées de la part de l'étudiant parce que celui-ci ne se concentre pas sur la tâche qui lui semble sans intérêt. Dans le second cas, soit lorsque les items sont trop difficiles, l'étudiant peut se décourager au point de ne pas compléter le test. 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.

De plus, pour permettre l'administration d'un test papier crayon fixe et invariable à des étudiants dont le niveau d'habileté varie beaucoup, ce test doit être constitué d'items dont le niveau de difficulté est très varié. Des items faciles ne sont donc pas nécessairement administrés à des étudiants dont le niveau d'habileté est faible, tandis que des items difficiles ne sont pas forcément administrés aux élèves dont le niveau d'habileté est plus élevé. Pour cette raison surtout, les tests papier crayon fixes et invariables ne permettent généralement pas d'obtenir un estimateur précis du niveau d'habileté dans les points extrêmes de l'échelle d'habileté où les niveaux d'habileté sont très faibles ou très élevés. Weiss (1982, p. 474) souligne ainsi que plus ce type de test permet d'estimer une large étendue de niveaux d'habileté, donc plus il est constitué d'items dont le niveau de difficulté varie de très facile à très difficile, moins la précision du test est élevée. À l'inverse, lorsque le test est composé d'items dont le niveau de difficulté varie peu, donc lorsqu'ils sont destinés à estimer un niveau d'habileté spécifique, une plus grande précision de l'estimateur du niveau d'habileté est obtenue lorsque les items administrés ne sont ni trop faciles, ni trop difficiles pour l'étudiant. C'est ce que souligne Weiss (1982, p. 474) lorsqu'il met en relief le dilemme entre la largeur de bande et la fiabilité du test.

Lors de l'administration d'un test papier crayon fixe et invariable, il est à noter que l'étudiant ne peut recevoir immédiatement son résultat au test ; il doit attendre que celui-ci soit corrigé. Ainsi, pour les tests à fonction diagnostique ou formative, qui nécessitent le plus souvent une 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. Ainsi, les séquences vidéo et les éléments auditifs sont peu employés et, lorsque c'est le cas, dans des conditions souvent inadéquates. Par exemple, les tests de classement en langue seconde comportent souvent une section visant à estimer le niveau d'habileté en compréhension orale. À cette fin, l'étudiant doit écouter un texte enregistré sur cassette et par la suite répondre à des items destinés à estimer son niveau d'habileté en compréhension auditive.

Enfin, des problèmes de sécurité peuvent se poser lors de l'administration d'un test. Ainsi, il peut y avoir plagiat au moment même de l'administration du test. Ou encore, la confidentialité des réponses peut être affectée par la circulation d'une copie du test, de la feuille réponse ou de la grille de correction.

9.2 Testing adaptatif

Pour remédier à ces problèmes de précision de l'estimateur du niveau d'habileté et à ces limites d'administration, des chercheurs ont proposé l'utilisation du testing adaptatif (TA). Le testing adaptatif est une forme de testing sur mesure (*tailored testing*) spécifiquement adaptée à la personne à qui on administre le test. Le testing adaptatif a connu de multiples transformations depuis son introduction : test à deux étapes, tests à niveaux flexibles, test pyramidal ou test stratifié. Ces diverses formes de test adaptatif étant abordées ailleurs par Auger (1989, p. 51-71), Laurier (1993b, p. 37-46) et Raïche (2000, p. 18-36), nous jugeons plus approprié de ne traiter que de la forme de test adaptatif la plus prometteuse, soit celle basée sur les propositions modernes de modélisation de la réponse à l'item. En fait, l'utilisation du testing adaptatif a été facilitée principalement depuis 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. 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.

Plusieurs programmes de testing à grande échelle (*large scale testing*) utilisent des versions adaptatives par ordinateur de leurs tests. C'est le cas, notamment, de plusieurs tests développés par l'*Educational Testing Service* (ETS) tels que le SAT (*Scholastic Assessment 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 d'application tandis que les travaux d'Auger (1989 ; Auger et Séguin, 1992) sur le testing adaptatif de maîtrise en éducation économique au secondaire, de Laurier en révision de texte (1996) et de Raïche (1994, 2000, 2001a, 2001b) et Raïche et Blais (2002a, 2002b, 2002c) sur la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing en sont des exemples de travail de recherche.

Le testing adaptatif offre plusieurs avantages par rapport aux tests papier crayons fixes et invariables. L'une des caractéristiques les plus importantes du testing adaptatif est de permettre l'administration d'items dont le niveau de difficulté correspond au niveau d'habileté de la personne passant le test. À l'opposé des tests papier crayons fixes et invariables, où tous les items du test sont administrés sans égard pour le niveau d'habileté de la personne, le testing adaptatif permet l'administration de tests sur mesure, de façon à ce que le niveau de difficulté des items à ces tests ne soit ni trop difficile, ni trop facile. Le nombre d'items administrés, tout comme la durée de l'administration, sont ainsi réduits par rapport à une version papier crayon du test, sans que la précision de l'estimateur du niveau d'habileté diminue pour autant. Le testing adaptatif devrait d'ailleurs permettre d'obtenir un estimateur plus précis du niveau d'habileté, plus spécifiquement lorsque le niveau d'habileté est faible ou élevé.

En testing adaptatif, chaque personne peut recevoir une version du test dont les items ont un niveau de difficulté adapté à son niveau d'habileté et dont la séquence des items peut varier d'une personne à une autre. Toutefois, cette caractéristique du testing adaptatif fait en sorte que le nombre de bonnes réponses au test ne permet plus de comparer les personnes entre elles puisqu'elles obtiennent toutes, selon certains auteurs (Weiss, 1985, p. 776), environ le même pourcentage de bonnes réponses aux items. Il serait alors plus approprié d'estimer le niveau d'habileté indépendamment du choix particulier des items d'une version du test.

Des propositions de modélisation 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.

9.2.1 Déroulement d'un test adaptatif

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 9.1.

Tableau 9.1

Algorithme décrivant le déroulement normal d'un test papier crayon fixe et invariable (d'après Raïche, 2000, p. 19)

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 quel item. Les questions sont présentées à tous dans le même ordre, mais le point de départ est laissé à la discrétion du répondant. Dans les faits, même si presque tous débutent avec la première question et répondent de manière séquentielle aux questions suivantes, la suite n'est pas nécessairement la même pour tous.

Dans un test adaptatif, à l'opposé, la première question proposée, les questions subséquentes, l'ordre de ces questions ainsi que la fin du test peuvent varier d'une personne à une autre selon des règles préétablies. Les règles de départ, de suite et d'arrêt permettent de présenter une première question selon des caractéristiques préalables du répondant, de déterminer quelle 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 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'examineur 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'examineur devrait tenter d'ajuster le niveau de difficulté des questions au niveau d'habileté du répondant.

La figure 9.1 illustre, de manière générale, le déroulement d'un test adaptatif. Au départ, un estimateur provisoire du niveau d'habileté du répondant est déterminé. Cet estimateur peut être obtenu en se basant sur des caractéristiques du répondant telles que son âge, des résultats antérieurs à d'autres tests ou, tout simplement, un estimateur fourni par le répondant lui-même. En l'absence d'informations préalables sur les caractéristiques du répondant, le niveau de difficulté de la première question est fréquemment fixé à un niveau moyen. À la suite de la réponse choisie par le répondant, un nouvel estimateur provisoire de son niveau d'habileté est alors calculé et une nouvelle question est administrée. Tant que la règle d'arrêt n'est pas satisfaite, de nouvelles questions, dont le niveau de difficulté est conditionnel aux réponses précédentes et à leur taux de succès, sont présentées. Cette règle d'arrêt peut être aussi simple que de cesser le test lorsqu'un nombre fixe de questions a été présenté, comme elle peut être aussi complexe que de mettre fin à l'administration du test lorsqu'un niveau prédéterminé de précision de l'estimateur du niveau d'habileté est atteint.

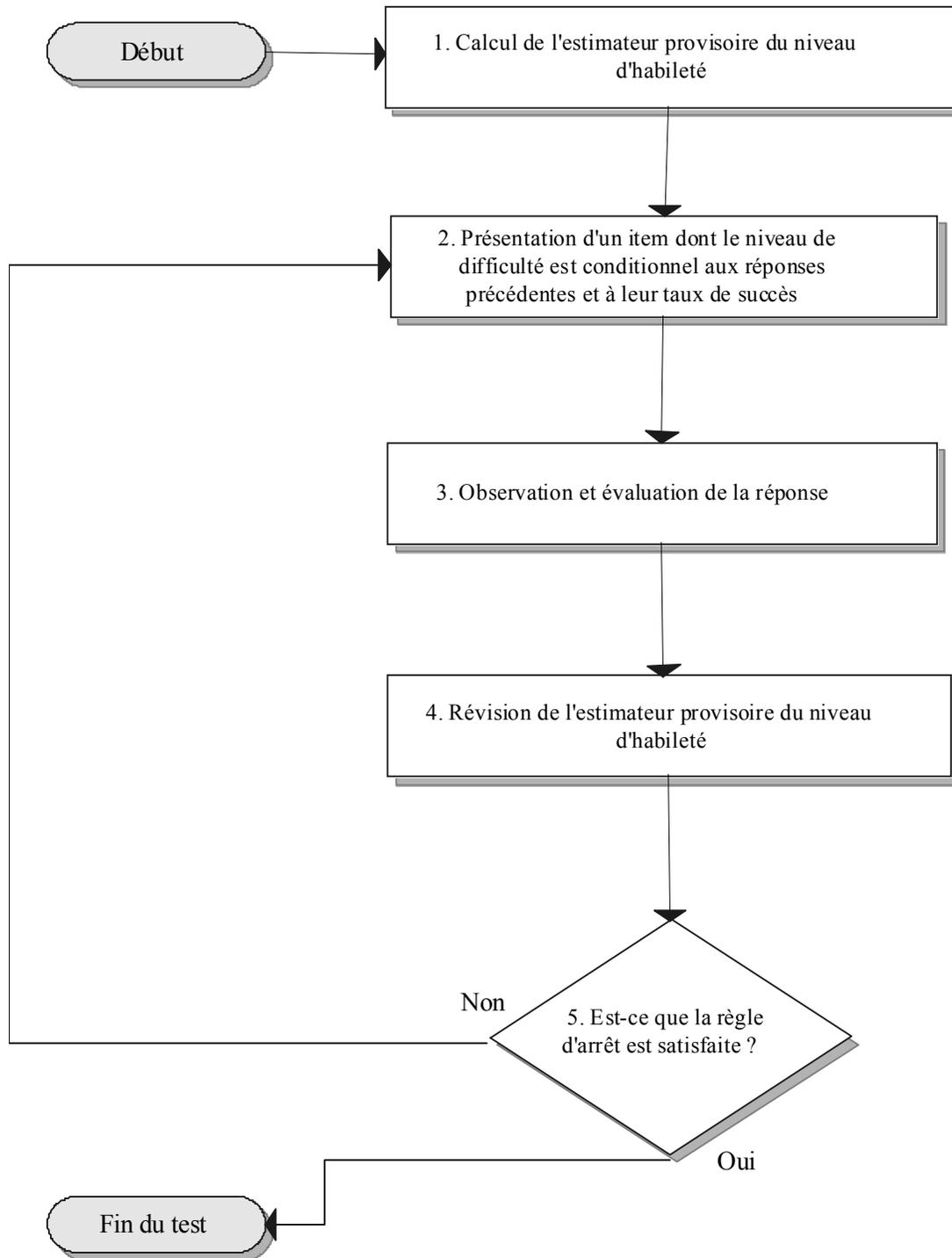


Figure 9.1 Déroulement général d'un test adaptatif (d'après Raïche, 2000, p. 22)

9.3 Le testing adaptatif : une application fort pertinente 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. En second lieu, il y a un avantage corollaire à ceci : le niveau d'habileté peut être estimé à partir de n'importe quel sous-ensemble d'items administrés. Cette caractéristique est très utile en testing adaptatif puisqu'elle permet d'administrer des items différents à des personnes différentes, tout en permettant d'obtenir des scores sur une même échelle. Les tests peuvent donc être réellement considérés sur mesure.

Troisièmement, un test adaptatif fondé sur des propositions de modélisation 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 souligné par Weiss (1982) est que la règle d'arrêt peut être basée sur la précision de l'estimateur du niveau d'habileté après chaque réponse. La règle d'arrêt peut ainsi être conditionnelle à l'atteinte d'un niveau de précision prédéterminé de l'estimateur du niveau d'habileté.

Dans un test adaptatif, où sont présentés des items dont le niveau de difficulté se rapproche le plus possible du niveau d'habileté, des décisions doivent être prises en ce qui concerne les caractéristiques du ou des premiers items administrés ; autrement dit, une règle de départ doit être établie. Par suite de la performance à un premier item ou aux premiers items, d'autres items dont le niveau de difficulté est de plus en plus près du niveau d'habileté sont proposés ; il est alors question de la règle de suite. Enfin, un ou des critères ayant pour but de décider de mettre fin à la situation de mesure doivent être adoptés ; il s'agit de la règle d'arrêt.

La figure 9.2 et le tableau 9.2 décrivent le déroulement d'un tel test. Nous présentons, pour chacune des règles considérées, soient celles de départ, de suite et d'arrêt, des stratégies proposées par la littérature.

Tableau 9.2

Algorithme décrivant le déroulement d'un test adaptatif basé sur la théorie de la réponse à l'item (d'après Raïche, 2000, p. 72)

RÈGLE	ACTION
1. Règle de départ	Administrer un item dont le niveau de difficulté est conditionnel à certaines caractéristiques du candidat
2. Règle de suite	Administrer un item dont le niveau de difficulté se rapproche de la valeur de l'estimateur provisoire du niveau d'habileté
3. Règle d'arrêt	Terminer le test après l'administration d'un nombre prédéterminé d'items, lorsqu'une erreur-type prédéterminée de l'estimateur du niveau d'habileté est obtenue ou lorsqu'il n'y a plus d'items qui puissent fournir une quantité d'information minimale au niveau d'habileté estimé

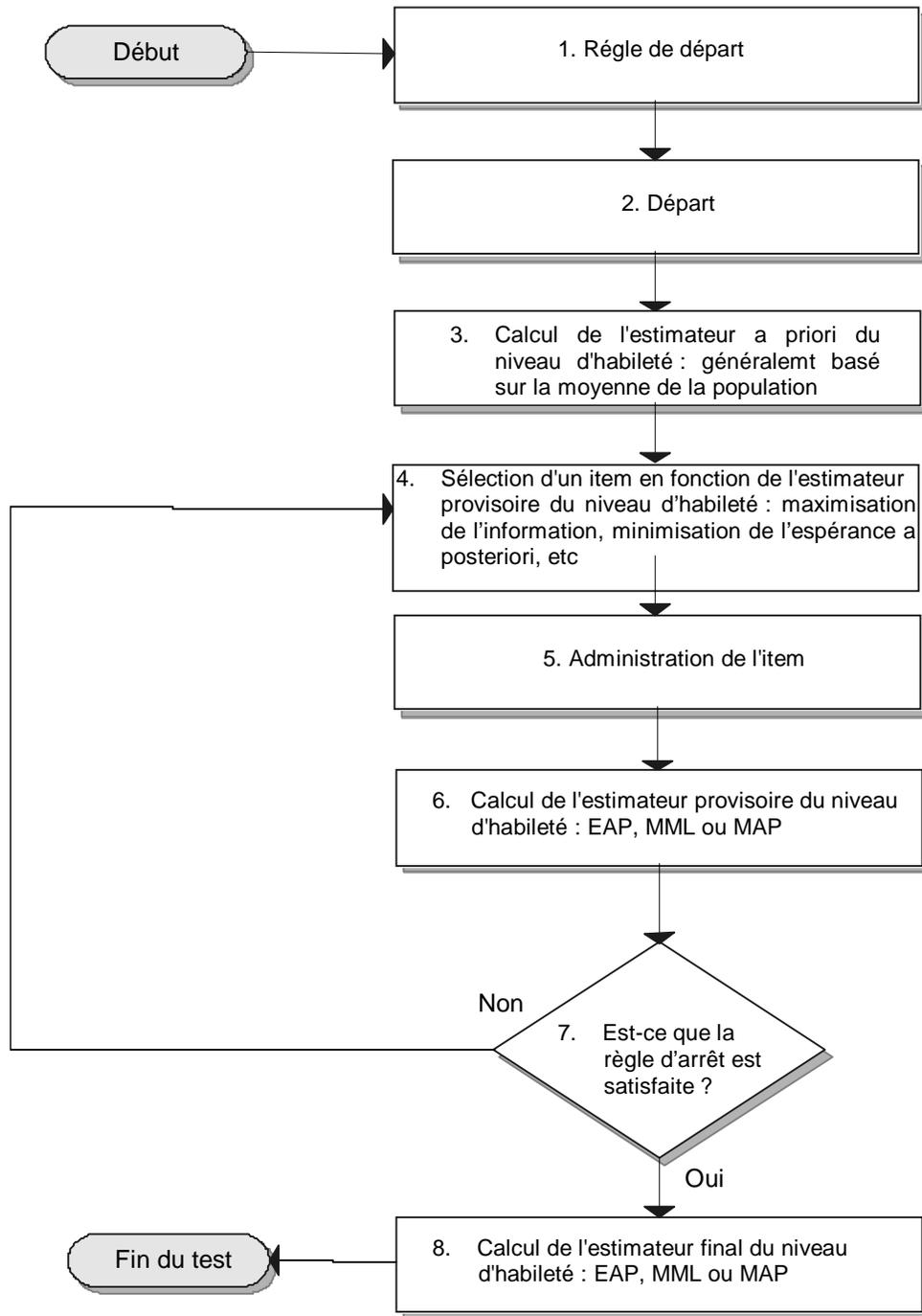


Figure 9.2 Structure d'un test adaptatif basé sur la théorie de la réponse à l'item (d'après Raïche, 2000, p. 71)

9.3.1 Les stratégies quant à la règle de départ

Un test adaptatif débute généralement par l'administration d'un item dont le niveau de difficulté est conditionnel à l'information disponible a priori : moyenne de groupe, âge ou même appréciation subjective de la part de l'étudiant évalué. Une règle de départ basée sur l'information a priori disponible à propos du niveau d'habileté doit être adoptée. Selon Thissen et Mislevy (2000, p. 107), la moyenne de la population d'où provient l'individu en situation de testing, estimée préalablement selon une modélisation issue de la théorie 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.

Laurier (1993b, p. 146-148), quant à lui, utilise des informations qu'il obtient directement auprès du répondant en cours d'administration d'un test de classement en langue seconde. Ainsi, avant d'administrer les items du test, le répondant doit répondre à quelques questions qui permettent d'obtenir des renseignements sur son habileté perçue et sur son expérience antérieure dans la langue seconde. Des questions du type : *À quelle année remonte le dernier cours suivi dans la langue seconde ?* ou encore, *Quel est ton degré d'aisance dans la langue seconde ?* Le niveau de difficulté du premier item administré est tributaire du niveau d'habileté établi en fonction des réponses à ces questions préalables.

On pourrait aussi imaginer une stratégie très simple pour obtenir de l'information a priori sur le niveau d'habileté d'un étudiant lors de l'administration d'un test de classement en anglais langue seconde, un test, non adaptatif pour le moment, qui est d'ailleurs utilisé actuellement dans la plupart des collèges et cégep du Québec, soit le TCALS II (Laurier, Froio, Paero et Fournier, 1999). Il s'agirait d'utiliser les résultats en anglais obtenus en secondaire V et IV. Selon des résultats non publiés obtenus chez les étudiants inscrits au Collège de l'Outaouais (n=1715), le coefficient de détermination entre les notes en secondaire V et le résultat au TCALS II est d'ailleurs de 0,64 ; une valeur assez importante pour justifier l'utilisation du résultat en secondaire V comme estimateur du niveau d'habileté a priori dans une éventuelle version adaptative du TCALS II.

Il faut aussi signaler que la détermination de l'estimateur a priori du niveau d'habileté, $\hat{\theta}_{\text{a priori}}$, peut affecter l'estimateur final du niveau d'habileté, $\hat{\theta}$, lorsque trop peu d'items sont administrés. C'est pourquoi les auteurs (Thissen et Mislevy, 1989, p. 110) suggèrent d'utiliser la même valeur de l'estimateur a priori du niveau d'habileté pour toutes les personnes à qui est administré un test adaptatif. Raïche (2000, p. 188-189) a réalisé une modélisation de l'impact de la détermination de l'estimateur a priori sur la valeur obtenue du biais de l'estimateur du niveau d'habileté en fonction de quatre valeurs courantes de la règle d'arrêt selon l'erreur-type : 0,40, 0,35, 0,30 et 0,20. À titre de rappel, le biais de l'estimateur du niveau d'habileté correspond à la valeur moyenne de la différence entre l'estimateur du niveau d'habileté et le niveau d'habileté, soit

$\widehat{\theta} - \theta$. Les équations de régression cubique (équations 9.1 à 9.4) permettent de calculer le biais de l'estimateur du niveau d'habileté lorsque l'estimateur a priori du niveau d'habileté a été fixé à 0,00. Dans ces équations $\text{Biais}_{0,40}$, $\text{Biais}_{0,35}$, $\text{Biais}_{0,30}$ et $\text{Biais}_{0,20}$ représentent le biais de l'estimateur du niveau d'habileté selon les quatre valeurs retenues de la règle d'arrêt selon l'erreur-type. Dans le but de généraliser à toutes les valeurs possibles de l'estimateur a priori du niveau d'habileté, on peut, bien sûr, remplacer dans ces équations $\widehat{\theta}_{\text{a priori}}$ par $(\theta - \widehat{\theta}_{\text{a priori}})$.

$$\text{Biais}_{0,40} = 0,00206 - 0,15132 \theta + 0,00040 \theta^2 - 0,00078 \theta^3 \quad (9.1)$$

$$\text{Biais}_{0,35} = 0,00419 - 0,11620 \theta + 0,00127 \theta^2 - 0,00088 \theta^3 \quad (9.2)$$

$$\text{Biais}_{0,30} = 0,00962 - 0,08577 \theta + 0,00593 \theta^2 - 0,00007 \theta^3 \quad (9.3)$$

$$\text{Biais}_{0,20} = -0,01069 - 0,04019 \theta + 0,00096 \theta^2 - 0,00041 \theta^3 \quad (9.4)$$

Le tableau 9.3, ainsi que la figure 9.3, présentent les valeurs du biais de l'estimateur du niveau d'habileté prédites par ces fonctions. Il y est très clair que plus l'erreur-type retenue pour la règle d'arrêt est élevée, situation où moins d'items sont administrés, plus le biais de l'estimateur du niveau d'habileté est important. On remarque d'ailleurs que la valeur du biais de l'estimateur du niveau d'habileté peut s'approcher de la valeur de l'erreur-type retenue, ou même la surpasser, lorsque le niveau d'habileté s'éloigne considérablement de l'estimateur a priori et que l'erreur-type retenue pour la règle d'arrêt est inférieure à 0,20. À titre d'exemple, lorsque l'erreur-type retenue pour la règle d'arrêt est égale à 0,40 et que le niveau d'habileté est égal à -3,00, le biais de l'estimateur du niveau d'habileté atteint 0,48. Le tableau 9.3 et la figure 9.3 nous permettent aussi de constater que le biais de l'estimateur du niveau d'habileté est peu important, quelle que soit la valeur du niveau d'habileté, lorsque le niveau d'habileté ne dépasse pas 1,00 en valeur absolue : dans ces conditions, il est au plus de |0,15|. Ces résultats ne sont pas surprenants et sont en concordance avec ceux obtenus par la communauté scientifique. Nous devons donc faire preuve de prudence lorsque l'estimateur du niveau d'habileté s'éloigne considérablement de la valeur de l'estimateur a priori et que l'erreur-type retenue pour la règle d'arrêt est élevée. Ces considérations nous amènent à recommander, contrairement à ce que suggèrent Thissen et Mislevy, l'utilisation de valeurs adaptées au sujet comme estimateur a priori du niveau d'habileté en testing adaptatif. Une règle de départ qui permet l'utilisation de valeurs adaptées au sujet comme estimateur a priori du niveau d'habileté offre aussi l'avantage de minimiser l'exposition des mêmes items à différents sujets puisque, pour des raisons de sécurité, il faut s'assurer que le premier item administré puisse varier d'un répondant à un autre. Sinon, il serait risqué que les

répondants se transmettent l'information et soient ainsi informés à l'avance du contenu du premier item administré.

Tableau 9.3

Biais de l'estimateur du niveau d'habileté selon la distance entre l'estimateur a priori et le niveau d'habileté en fonction de quatre valeurs de l'erreur-type retenue pour la règle d'arrêt (S_{θ}) lorsque l'estimateur a priori du niveau d'habileté est fixé à 0,00

Niveau d'habileté θ	S_{θ}			
	0,40	0,35	0,30	0,20
-3,00	0,48	0,39	0,30	0,13
-2,00	0,31	0,25	0,19	0,08
-1,00	0,15	0,12	0,08	0,03
0,00	0,00	0,00	-0,01	-0,01
1,00	-0,15	-0,11	-0,09	-0,05
2,00	-0,31	-0,23	-0,16	-0,09
3,00	-0,47	-0,36	-0,22	-0,13

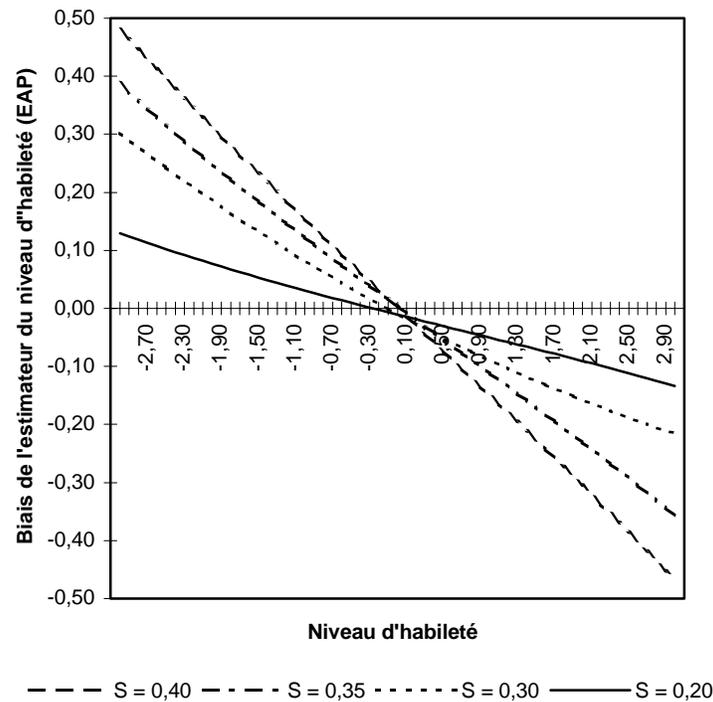


Figure 9.3 Biais de l'estimateur du niveau d'habileté en testing adaptatif selon quatre valeurs de l'erreur-type (S_{θ}) retenue pour la règle d'arrêt lorsque l'estimateur a priori du niveau d'habileté est fixé à 0,00

9.3.2 Les stratégies quant à la règle de suite

Selon la performance au premier item ou aux items précédents, un item optimal, dont le niveau de difficulté se rapproche de l'estimateur provisoire du niveau d'habileté, doit être sélectionné 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. Wainer et Kiely (1987) proposent une stratégie de sélection des items permettant d'exercer un meilleur contrôle sur l'équilibre du contenu des items, celle des minitests (*testlets*). Enfin, de nouvelles stratégies ont été récemment proposées pour tenir compte de contraintes spécialisées dans la sélection des items (Hetter et Sympson, 1997 ; van der Linden et Pashley, 2000). Nous présentons maintenant ces diverses stratégies quant à la règle de suite.

9.3.2.1 Stratégies de maximisation de l'information

La première de ces stratégies de sélection du prochain item à administrer consiste à choisir l'item pour lequel l'information est maximale. Plusieurs méthodes peuvent être utilisées pour maximiser l'information : par information maximale sans contrainte, par une table des valeurs de l'information pour chaque item ou par la méthode de Urry.

La méthode de sélection par information maximale sans contrainte (*unconstrained maximum information selection*) permet de choisir un item pour lequel l'information, au sens de Fisher (1922), évaluée au niveau d'habileté estimé provisoirement après l'administration de l'item i est maximale (Lord, 1980b, p. 199). C'est le concept d'information qui permet d'obtenir une mesure de la précision de l'estimateur du niveau d'habileté lorsque celui-ci est obtenu par la méthode de vraisemblance maximale (Baker, 1992, p. 79-81). L'information fournie par l'item i au niveau d'habileté θ étant évaluée en conformité avec l'équation 9.5.

$$I_i(\theta) = \frac{[P'_i(r_i | \theta)]^2}{P_i(r_i | \theta)Q_i(r_i | \theta)} \quad (9.5)$$

Où $P_i(r_i | \theta)$ correspond à la probabilité d'obtention d'une bonne réponse à l'item i calculée selon une des modélisations issues de la théorie de la réponse à l'item sachant que le niveau d'habileté

est égal à θ . $Q_i(r_i|\theta)$ correspond à la probabilité d'obtenir une mauvaise réponse à l'item i et $P_i'(r_i|\theta)$ est égale à la dérivée première de la fonction de probabilité. L'information offre l'avantage d'être additive de façon telle que l'information totale à un test à un niveau d'habileté fixé est égale à la somme de l'information fournie par chacun des items administrés.

Il est cependant possible d'obtenir, avec une précision satisfaisante, une approximation de l'information fournie au prochain item en recourant à une table de valeurs où l'information, calculée au préalable, apportée par chacun des items d'une banque d'items disponibles est indiquée pour différentes valeurs du niveau d'habileté. La procédure de sélection consiste alors à choisir l'item qui fournit le plus d'information à une valeur rapprochée du niveau d'habileté. Selon Thissen et Mislevy (1990, p. 111), cette méthode a l'avantage d'être moins exigeante en temps de calcul tout en permettant d'obtenir une approximation généralement satisfaisante.

Urry (1970, p. 82) propose une méthode 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.

9.3.2.2 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 basée 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 ne sont pas présentées ici. Selon Thissen et Mislevy (1990, p. 112), ces équations, quoique complexes, permettent de diminuer le temps de calcul de façon significative puisqu'elles ne reposent pas sur des calculs itératifs, comme c'est le cas dans la méthode de sélection par information maximale sans contrainte. 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).

9.3.2.3 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.

Un exemple de minitest appliqué au domaine de la statistique est offert à la figure 9.4. On peut remarquer que les réponses aux items 2, 3 et 4 ne sont pas indépendantes de la réponse fournie à l'item 1 puisque la valeur de la moyenne est nécessaire à la réussite de ces items. De plus, la réussite des items 3 et 4 nécessite la connaissance de la valeur de l'écart-type calculé à l'item 2. Il est alors possible d'attribuer, soit un score de succès ou d'échec au minitest, soit un score variant entre 0 et 4 représentant un échec, un succès partiel ou un succès total au minitest.

Soit la série de valeurs suivante :		
10, 25, 5, 15, 30, 2, 24, 18 et 30		
1. Quelle est la valeur de la moyenne arithmétique ?	a) 17,67 b) 21,32 c) 25,01 d) 10,00	1 point
2. Quelle est la valeur de l'écart-type ?	a) 2,32 b) 45,10 c) 5,51 d) 10,43	1 point
3. Quelle est la valeur du coefficient d'asymétrie ?	a) 4,55 b) -0,28 c) -8,76 d) 0,12	1 point
4. Quelle est la valeur du coefficient de kurtose ?	a) -1,39 b) 56,34 c) -8,02 d) 1,39	1 point
TOTAL :		4 points

Figure 9.4 Exemple d'un minitest

Prometteuse, selon Wainer et *al.* (1990, p. 253-254), cette stratégie exige toutefois des modélisations de la réponse à l'item plus sophistiquées que celles qui sont utilisées dans les modèles habituels. Les modèles à réponses nominales ou ordonnées se montrent alors

intéressants. Dans cette veine, Thissen (1993) propose d'ailleurs certains modèles spécifiques à une démarche de testing adaptatif par minitests, principalement la modélisation de la réponse à l'item par crédit partiel (*partial credit model*) de Masters (1982). La recherche sur l'utilisation des minitests en testing adaptatif est très active actuellement. Plus récemment, des auteurs tels que Wainer, Bradlow et Du (2000), Glas, Wainer et Bradlow (2000) ainsi que Vos et Glas (2000) proposent diverses stratégies pour soutenir la mise en oeuvre des minitests lors de tests adaptatifs. Wainer, Bradlow et Wu (2000) explorent l'utilisation d'une généralisation multidimensionnelle de la modélisation logistique à trois paramètres. Glas, Wainer et Bradlow (2000) explorent l'utilisation des méthodes d'estimations basées sur les chaînes de Markov Monte Carlo (*Monte Carlo Markov Chain*) tandis que Vos et Glas appliquent la stratégie des minitests aux tests de maîtrise.

9.3.2.4 Nouvelles stratégies de sélection des items

Enfin, de nouvelles stratégies ont été récemment proposées pour tenir compte de contraintes spécialisées dans la sélection des items. Ces stratégies reçoivent actuellement beaucoup d'attention de la part des chercheurs et il est fort probablement trop tôt pour vraiment juger de leur supériorité sur les stratégies présentées plus haut.

Dans la plupart des cas, il s'agit de contraintes qui visent à minimiser la probabilité d'exposition de chacun des items qui composent la banque d'items disponibles. Par exemple, Hetter et Sympson (1997) ont développé une stratégie de sélection du prochain item visant à réduire les séquences d'items prédictibles et, ainsi, la surexposition éventuelle des items qui fournissent le plus d'information au sens de Fisher. Van der Linden (2000), ainsi que van der Linden et Pashley (2000), pour leur part, suggèrent l'utilisation de tests fantômes (*shadow tests*), soit, plus ou moins des super minitests satisfaisant à des contraintes plus complexes que celle qui vise le simple contrôle de la surexposition des items. Les tests fantômes, ainsi construits, sont ceux qui fournissent le plus d'information tout en répondant aux contraintes spécifiées. Ces contraintes peuvent s'adresser aux caractéristiques des items, telles que le nombre de mots, le nombre de choix de réponse. Elles peuvent aussi dicter l'administration d'items différents selon les personnes à qui sont administrés les tests : par exemple, selon la langue, le sexe ou la culture.

9.3.3 Stratégies d'estimation provisoire du niveau d'habileté

Selon Thissen et Mislevy (1990, p. 113), les méthodes d'estimation provisoire du niveau d'habileté les plus utilisées après l'administration de j items sont celles basées sur les fonctions de vraisemblance telle que la méthode de vraisemblance maximale (*maximum likelihood, ML*) (section 6.1). Les méthodes bayésiennes d'estimation du niveau d'habileté sont aussi utilisées, soient la méthode bayésienne de maximisation a posteriori (*maximization a posteriori, MAP*) (section 6.2) et la méthode de l'espérance a posteriori (*expected a posteriori, EAP*) (section 6.3). Wainer et Thissen (1987, p. 353) ont comparé différentes méthodes d'estimation du niveau d'habileté et en arrivent à la conclusion que les estimateurs du niveau d'habileté obtenus par la méthode de l'espérance a posteriori sont ceux dont l'erreur-type est généralement la plus petite.

Selon Thissen et Mislevy (1990, p. 113), la méthode 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.

Les figures 9.5 et 9.6 représentent respectivement les valeurs de l'estimateur provisoire du niveau d'habileté et de l'erreur-type associés à chacun des 60 items obtenues par une simulation de quatre tests adaptatifs. La modélisation logistique à un paramètre et la méthode de l'estimateur a posteriori (EAP) sont utilisées. Les quatre tests adaptatifs diffèrent par la valeur du niveau d'habileté simulé pour quatre sujets : -3,00, -2,00, -1,00 et 0,00. On peut remarquer à la figure 9.5 que la convergence de l'estimateur du niveau d'habileté est plus rapide quand le niveau d'habileté est fixé à 0,00. En fait, lorsque peu d'items sont administrés et que le niveau d'habileté s'éloigne substantiellement de 0,00, le biais de l'estimateur du niveau d'habileté est assez important.

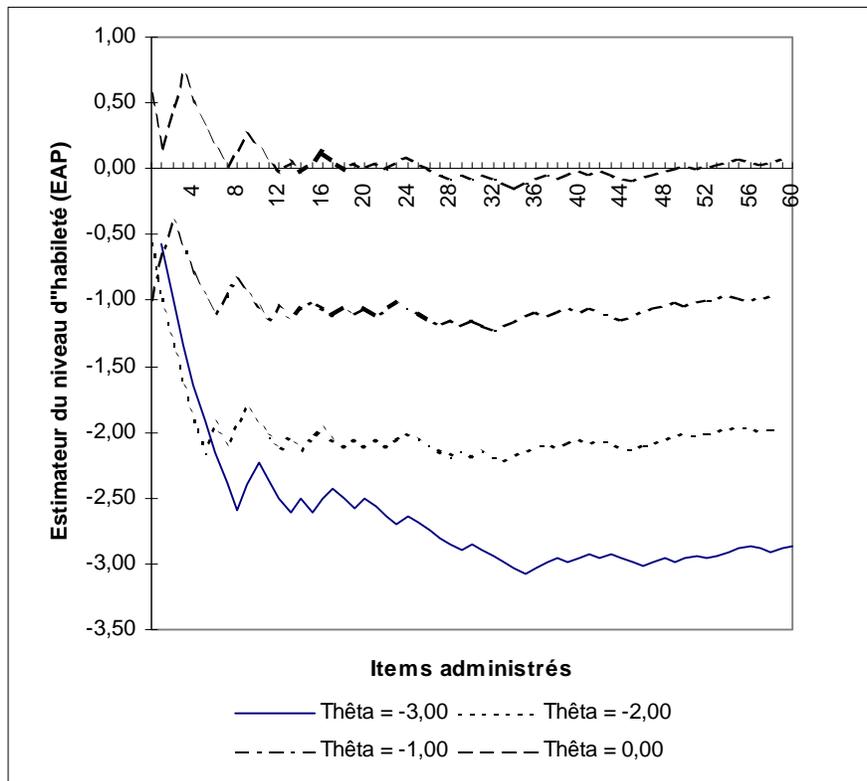


Figure 9.5 Estimateur du niveau d'habileté en testing adaptatif en fonction du nombre d'items administrés selon quatre valeurs du niveau d'habileté

Selon la figure 9.6, plus le niveau d'habileté s'éloigne de 0,00, plus l'erreur-type de l'estimateur du niveau d'habileté est importante. Ce phénomène devient toutefois de moins en moins important avec l'augmentation du nombre d'item administrés. À partir du 12^e item administré, l'erreur-type est d'environ 0,40, tandis qu'elle n'est que de 0,20 autour du 40^e item.

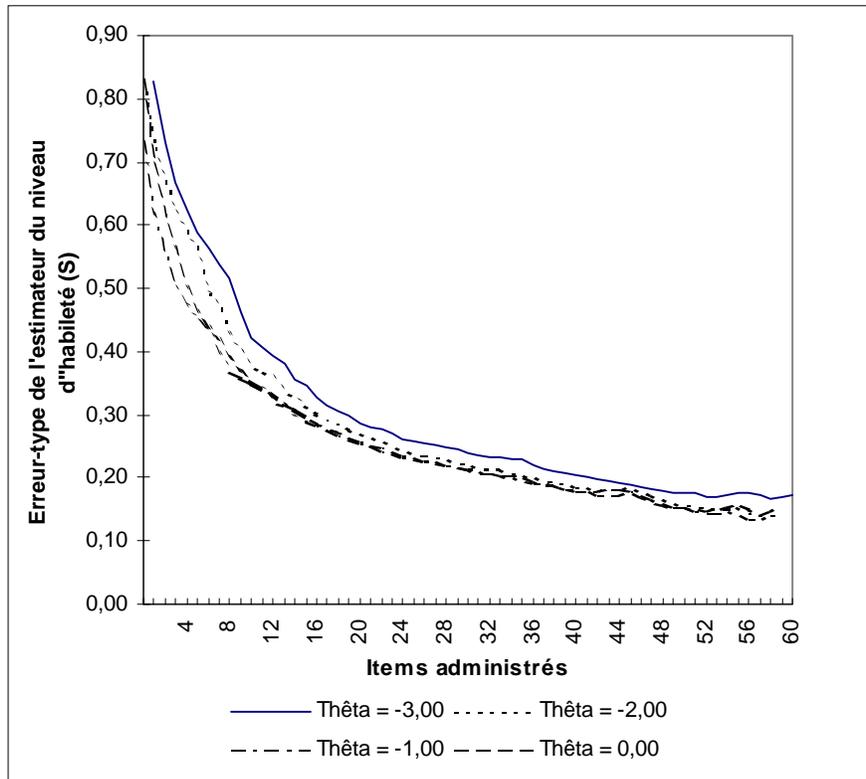


Figure 9.6 Erreur-type de l'estimateur du niveau d'habileté en testing adaptatif en fonction du nombre d'items administrés selon quatre valeurs du niveau d'habileté

Raïche et Blais (2002b) expérimentent actuellement des méthodes d'estimation qui sont elles-mêmes adaptatives : soient l'estimation par intervalle d'intégration adaptatif, l'estimation par estimateur a priori adaptatif et l'estimation avec correction adaptative pour biais. Toutes ces stratégies visent à centrer l'estimation provisoire θ_i du niveau d'habileté autour de l'estimateur provisoire précédent θ_{i-1} du niveau d'habileté.

La méthode d'estimation par intervalle d'intégration adaptatif est utilisée pour permettre d'ajuster l'intervalle d'intégration de la méthode de l'espérance a posteriori (EAP) à la valeur de l'estimateur du niveau d'habileté obtenue au cycle d'estimation précédent. Dans la méthode de l'estimation de l'espérance a posteriori, les limites d'intégration sont généralement fixées au plus à $\pm 4,00$ autour d'un point milieu égal à 0,00. La méthode d'estimation par intervalle adaptatif fait en sorte que les limites d'intégrations varient de $\pm 4,00$ autour de la valeur de l'estimateur provisoire précédent θ_{i-1} .

La méthode d'estimation par estimateur a priori adaptatif peut aussi bien s'appliquer à la méthode de maximisation a posteriori qu'à la méthode de l'espérance a posteriori où l'estimateur a priori du niveau d'habileté est fixe tout au long de l'administration du test adaptatif. Dans la

méthode d'estimation par estimateur a priori adaptatif, l'estimateur a priori varie donc en fonction de la valeur de l'estimateur provisoire θ_{i-1} obtenu après l'administration de l'item précédent.

Enfin, la méthode d'estimation avec correction adaptative pour biais applique l'ajustement proposé par Bock et Mislevy (1982, p. 439-442) pour diminuer l'importance du biais de l'estimateur du niveau d'habileté. Bock et Mislevy effectuent cette correction en divisant l'estimateur du niveau d'habileté par une approximation du coefficient de fidélité, r_{tt} :

$$r_{tt} = 1 - S_{\theta}^2 \quad (9.6)$$

Où S_{θ} correspond à l'erreur-type associée à l'estimateur du niveau d'habileté. L'estimateur corrigé du niveau d'habileté $\hat{\theta}_c$ devient alors égal à :

$$\hat{\theta}_c = \frac{\hat{\theta}}{1 - S_{\theta}^2} \quad (9.7)$$

Généralement, la correction de Bock et Mislevy n'est appliquée qu'à l'estimateur final du niveau d'habileté. Dans la méthode d'estimation avec correction adaptative pour biais, elle est effectuée à chaque estimation provisoire du niveau d'habileté.

Le tableau 9.4 illustre la simulation des résultats obtenus après l'administration de chaque item à quatre tests adaptatifs qui se terminent au 15^e item chez une personne dont le niveau d'habileté est égal à -3,00. Le premier test adaptatif utilise la méthode de l'espérance a posteriori (EAP), tandis que des stratégies d'estimation adaptative sont utilisées dans les trois autres tests adaptatifs. La méthode usuelle de l'espérance a posteriori et la méthode avec intervalle d'intégration adaptatif sont celles qui présentent la valeur la plus importante du biais de l'estimateur du niveau d'habileté, soit 0,40 (-2,60 - (-3,00) = 0,40). Toutefois, la méthode avec intervalle d'intégration adaptative permet d'obtenir une meilleure précision de l'estimateur du niveau d'habileté puisque l'erreur-type est inférieure (0,31) à celle obtenue par la méthode de l'espérance a posteriori (0,35). Cependant, c'est la méthode avec estimateur a priori adaptatif qui semble la plus intéressante, car elle permet d'obtenir l'estimateur du niveau d'habileté le moins biaisé, soit seulement -0,07 et que l'erreur-type associée est parmi les plus petites.

Tableau 9.4

Estimateur du niveau d'habileté, erreur-type de celui-ci et réponse à l'item en testing adaptatif en fonction du nombre d'items administrés selon quatre méthodes d'estimation lorsque le niveau d'habileté est égal à -3,00

Item	r*	EAP		Correction de Bock et Mislevy adaptative			Estimateur a priori adaptatif			Intervalle d'intégration adaptatif		
		$\hat{\theta}$	$S_{\hat{\theta}}$	r	$\hat{\theta}$	$S_{\hat{\theta}}$	r	$\hat{\theta}$	$S_{\hat{\theta}}$	r	$\hat{\theta}$	$S_{\hat{\theta}}$
1	0	-0,57	0,83	0	-1,81	0,83	0	-0,57	0,83	0	-0,57	0,83
2	0	-0,99	0,73	0	-3,05	0,73	0	-1,30	0,73	0	-0,99	0,69
3	0	-1,34	0,67	1	-2,56	0,67	0	-2,05	0,62	0	-1,34	0,61
4	0	-1,64	0,62	1	-2,11	0,62	0	-2,74	0,56	0	-1,64	0,56
5	0	-1,91	0,59	0	-2,48	0,59	0	-3,25	0,50	0	-1,91	0,50
6	0	-2,15	0,56	0	-2,74	0,56	0	-3,56	0,47	0	-2,16	0,46
7	0	-2,38	0,54	0	-2,96	0,49	1	-3,49	0,45	0	-2,38	0,44
8	0	-2,59	0,51	0	-3,16	0,47	1	-3,34	0,43	0	-2,60	0,42
9	1	-2,39	0,46	1	-2,99	0,43	1	-3,16	0,40	1	-2,40	0,39
10	1	-2,24	0,42	1	-2,84	0,40	1	-2,97	0,37	1	-2,24	0,37
11	0	-2,37	0,41	0	-2,97	0,38	0	-3,09	0,36	0	-2,37	0,35
12	0	-2,50	0,39	0	-3,09	0,37	1	-2,98	0,35	0	-2,50	0,34
13	0	-2,62	0,38	0	-3,21	0,36	0	-3,08	0,34	0	-2,62	0,33
14	1	-2,50	0,36	1	-3,10	0,34	1	-2,99	0,32	1	-2,51	0,32
15	0	-2,60	0,35	0	-3,20	0,33	0	-3,07	0,31	0	-2,60	0,31

* r est égal à 1, lors d'une bonne réponse à l'item et à 0, lors d'une mauvaise réponse.

La figure 9.7 permet d'obtenir une représentation visuelle des valeurs disponibles au tableau 9.4. La plus rapide convergence vers la valeur du niveau d'habileté des méthodes de l'estimateur a priori adaptatif et de la correction pour biais de Bock et Mislevy est très nette.

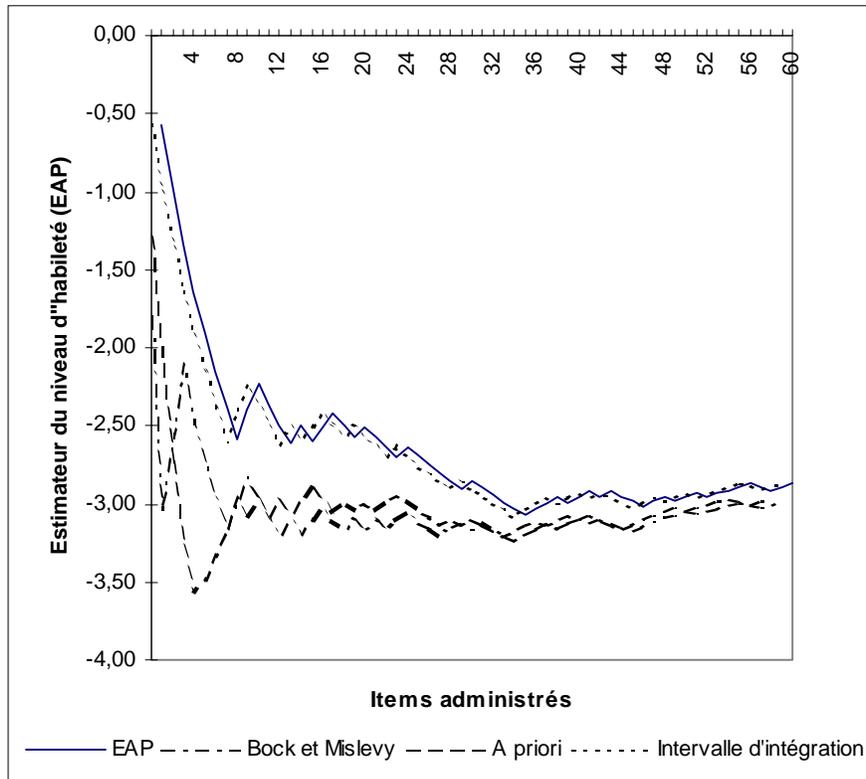


Figure 9.7 Estimateur du niveau d'habileté en testing adaptatif en fonction du nombre d'items administrés selon quatre méthodes d'estimation provisoire du niveau d'habileté lorsque le niveau d'habileté est égal à -3,00

9.3.4 Stratégie quant à la règle d'arrêt

Deux règles sont généralement utilisées dans le but de mettre fin au test. La première consiste à arrêter le test après l'administration d'un nombre fixe et prédéterminé d'items. Aucun critère absolu n'a été arrêté quant à ce nombre d'items. Selon Thissen et Mislevy (2000, p. 113), l'administration d'un nombre minimal de 20 items permet d'obtenir un estimateur du niveau d'habileté presque identique, que l'on utilise la méthode d'estimation 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. Raîche (2000), ainsi que Raîche et Blais (2002a, 2000c), arrivent à la conclusion que l'administration d'aussi peu que 13 items est suffisante lorsque la situation n'exige pas que l'erreur-type de l'estimateur du niveau d'habileté soit très petite. Ainsi, lorsque la modélisation logistique à un paramètre et que la méthode de l'espérance a posteriori sont utilisées, l'erreur-type de l'estimateur du niveau d'habileté est égale à 0,40 avec seulement 12 items administrés. Toutefois, en dehors de l'intervalle $[-1,50, 1,50]$ le biais est important. Quand la précision exigée est plus importante, l'administration d'au moins 40 items est nécessaire. L'erreur-type est alors égale à 0,20 ou moins.

Une seconde règle d'arrêt consiste à terminer l'administration du test lorsqu'une erreur-type prédéterminée de l'estimateur du niveau d'habileté est obtenue. En pratique, 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. Selon Raîche (2000), ainsi que Raîche et Blais (2002a, 2000c), pour que le biais de l'estimateur du niveau d'habileté ne soit pas trop important aux valeurs extrêmes du niveau d'habileté, l'erreur-type de l'estimateur du niveau d'habileté doit être d'au plus 0,40 lorsque la modélisation logistique à un paramètre et que la méthode de l'espérance a posteriori sont utilisées. Toutefois, si on désire obtenir l'homogénéité de l'erreur-type de l'estimateur du niveau d'habileté sur un intervalle important du niveau d'habileté, l'erreur-type retenue pour la règle d'arrêt doit être d'au plus 0,20. L'homogénéité de l'erreur-type de l'estimateur du niveau d'habileté est importante quand vient le moment d'appliquer certaines procédures de comparaisons de moyennes telles que des tests t de Student ou des analyses de la variance : ces procédures reposent d'ailleurs sur le postulat de l'homogénéité de la variance.

Dans certaines situations spécifiques, d'autres règles d'arrêt peuvent être utilisées. Ainsi, Dodd (1990) et Dodd, Koch et de Ayala (1993), à l'intérieur d'études de certaines règles d'arrêt, ont utilisé une règle basée sur l'information minimale de l'item (*minimum item information*). Selon cette stratégie, l'administration du test se termine lorsqu'il n'y a plus d'items dans la banque d'items qui puissent fournir une quantité d'information minimale prédéterminée au niveau d'habileté estimé. Dans une autre situation, lorsque certains tests sont destinés à mesurer l'exactitude des réponses dans un test où le fait de répondre avec rapidité est important (*accuracy at speed*), 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.

9.3.5 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. Toutefois, les travaux de Raïche et Blais (2002b) sur les méthodes d'estimation provisoires adaptatives et les commentaires formulés à la section traitant des stratégies quant à la règle de départ, ne nous permettent pas d'endosser le point de vue de Thissen et Mislevy. Leur position serait réaliste seulement dans des situations irréalistes, où trop peu d'items seraient administrés.

Bock et Mislevy (1982) suggèrent d'utiliser la méthode de l'espérance a posteriori pour calculer l'estimateur final du niveau d'habileté. Leurs études indiquent que l'estimateur final du niveau d'habileté obtenu par cette méthode affiche, en général, une valeur plus petite de son erreur-type. Comme nous l'avons signalé plus haut, à la section qui traite de l'estimateur provisoire du niveau d'habileté, ils suggèrent aussi d'utiliser une correction du biais en divisant l'estimateur final du niveau d'habileté par une approximation du coefficient de fidélité. Selon Raïche (2000, p. 191-194), la correction proposée par Bock et Mislevy est recommandée lorsque le nombre d'items administrés est inférieur à 40 ou que l'erreur-type de l'estimateur du niveau d'habileté retenue pour la règle d'arrêt est supérieur à 0,20.

Certains auteurs ont proposé d'utiliser des méthodes d'estimation du niveau d'habileté qui seraient moins affectées par des patrons de réponses atypiques ou par l'effet potentiel du petit nombre d'items sur le comportement des estimateurs. Selon eux, l'utilisation d'estimateurs robustes (Hojtink et Boomsma, 1995, p. 54 ; Mislevy et Bock, 1982 ; Thissen et Mislevy, 1990, p. 115; Wainer, 1983, p. 71) pourrait être plus appropriée. En ce sens, Mislevy et Bock (1982) suggèrent l'utilisation d'une méthode d'estimation à double pondération (*biweight*) tandis que Wainer et Thissen (1987, p. 344-345) ainsi que Wainer et Wright (1980) explorent une méthode reposant sur une technique de rééchantillonnage sans remise (*jackknife*), soit la méthode AMJACK. Dans la pratique, toutefois, ces méthodes semblent peu employées car les situations pour lesquelles elles ont été proposées au départ sont extrêmes et laissent peu de crédibilité quant leur capacité d'estimer le niveau d'habileté.

9.4 Considérations diverses

9.4.1 Une formule de prophétie adaptée aux tests adaptatifs

Dans la théorie classique des tests, la formule de prophétie de Spearman-Brown (Laveault et Grégoire, 1997, p. 154 ; Wainer et Thissen, 2001, p. 31) permet de prédire le nombre d'items qu'il est nécessaire d'administrer $n_{\text{prédit}}$ pour obtenir un niveau de fidélité désiré connaissant le niveau de fidélité r_{tt} observé à partir d'un test de longueur n . Par extension, la formule de prophétie permet aussi de prédire le niveau de fidélité qu'afficherait un test n fois plus long, ou encore n fois plus court, que le test qui a servi à calculer la fidélité.

Raïche et Blais (2000a) ont élaboré des formules de prophétie spécifiques à un test adaptatif construit autour d'une modélisation logistique à un paramètre. Ainsi, l'équation 9.8 permet de prédire l'erreur-type de l'estimateur du niveau d'habileté $S_{\text{prédite}}$ lorsque n fois plus d'items sont administrés. Selon Raïche et Blais, une différence d'au plus 0,04 est obtenue quant à l'erreur-type prédite de l'estimateur du niveau d'habileté.

$$S_{\text{prédite}} = \sqrt{1 - \frac{n * (1 - S_{\theta}^2)}{1 + (n - 1) * (1 - S_{\theta}^2)}} \quad (9.8)$$

Par exemple, à partir des données disponibles au tableau 9.4, si nous tentons de prédire l'erreur-type obtenue suite à l'administration du quatrième item à partir de l'erreur-type obtenue après l'administration du premier item (0,83), nous obtiendrons :

$$S_{\text{prédite}} = 0,60 = \sqrt{1 - \frac{4 * (1 - 0,83^2)}{1 + (4 - 1) * (1 - 0,83^2)}}$$

La différence entre l'erreur-type de l'estimateur du niveau d'habileté obtenue au quatrième item, selon la méthode EAP, et l'erreur-type prédite de l'estimateur du niveau d'habileté (0,62 - 0,60) n'est que de 0,02.

À partir de l'équation 9.9, nous pouvons aussi prédire le nombre d'items qu'il est nécessaire d'administrer $n_{\text{prédit}}$ pour obtenir une valeur prédéterminée de l'erreur-type de l'estimateur du niveau d'habileté.

$$n_{\text{prédit}} = n * \left(\frac{(1 - S_{\text{prédite}}^2) * [1 - (1 - S_{\theta}^2)]}{(1 - S_{\theta}^2) * [1 - (1 - S_{\text{prédite}}^2)]} \right) \quad (9.9)$$

Où n correspond au nombre d'items administrés au moment où nous obtenons une erreur-type de l'estimateur du niveau d'habileté.

Toujours en se basant sur les données fournies au tableau 9.4, connaissant l'erreur-type obtenue suite à l'administration du second item (0,73) selon la méthode EAP, nous désirons prédire le nombre d'items nécessaires pour obtenir une erreur-type égale à environ 0,39. En insérant les valeurs requises à l'intérieur de l'équation 9.9, nous obtenons :

$$n_{\text{prédit}} = 12,72 = 2 * \left(\frac{(1 - 0,39^2) * [1 - (1 - 0,73^2)]}{(1 - 0,73^2) * [1 - (1 - 0,39^2)]} \right)$$

Au tableau 9.4, un nombre minimal de 12 items était nécessaire pour obtenir une erreur-type de l'estimateur du niveau d'habileté égale à 0,39. La différence entre la valeur prédite et la valeur obtenue n'est alors que de 0,72 (12,00 - 12,72).

Il est intéressant de constater la précision des valeurs obtenues à partir de ces équations avec si peu d'items administrés.

9.4.2 Logiciels disponibles

Les premiers tests adaptatifs basés sur les modélisations issues de la théorie de la réponse à l'item ont été développés à partir du début des années 70 pour les besoins de la Marine américaine en sélection de personnel, principalement autour des travaux de McBride, Urry, et Weiss (voir McBride et Martin, 1983, p. 223-236). Au début, ces tests difficilement disponibles au grand public, nécessitaient l'utilisation d'ordinateurs centraux puissants. Avec l'arrivée des micro-ordinateurs, il a été possible de développer des logiciels plus accessibles et dont les coûts étaient plus abordables. L'un de ces premiers logiciels de testing adaptatif a été MicroCat (Assessment systems corporation, 1984). MicroCat a été conçu pour être utilisé sous le système d'exploitation DOS et supporte des fichiers graphiques au plus au format CGA. Toujours sur le marché actuellement, on lui préfère généralement une version plus contemporaine fonctionnant sous Windows et qui permet la gestion des fichiers multimédias accessibles par ce système d'opération.

Au Canada, Laurier a développé des versions de tests adaptatifs spécifiquement destinés au classement en langue seconde (Laurier, 1999a, 1999b). FrenchCapt (*French Computerized Adaptive Placement Test*), l'un de ces logiciels, est actuellement utilisé à l'intérieur de certaines universités québécoises. Un simulateur de tests adaptatifs, SIMCAT, utilisant une modélisation logistique à un paramètre et développé en langage SAS, est aussi proposé par Raïche (2000, p. xxx-xxxii).

Actuellement, les logiciels développés au Canada ou au Québec ne permettent pas, à notre connaissance, à la fois de modifier aisément la banque d'items et de présenter les items sous divers formats : audio, vidéo ou autres. Seuls MicroCat et MicroTest permettent ces opérations. Toutefois, leur coût élevé et la non disponibilité de banques d'items préalablement calibrées dans une langue autre que l'anglais en limitent l'usage à l'intérieur de la communauté francophone. C'est pourquoi, à notre avis, le développement de tels logiciels et la mise en marché de banques d'items spécialisées correspond actuellement à un besoin important dans la communauté francophone.

Pour ceux et celles qui désireraient se lancer dans cette aventure, Linacre (2000) rend disponible sur Internet le code source en langage Microsoft BASIC d'un test adaptatif, UCAT. Ce logiciel rend relativement facile la gestion de la banque d'items au domaine de notre choix et permet de plus, caractéristique non négligeable, la recalibration en ligne des items qui composent la banque d'items.

9.5 Défis et enjeux du testing adaptatif

D'un point de vue technologique, nous sommes maintenant prêts à mettre en application des instruments de mesure sous la forme de tests adaptatifs. Des ordinateurs suffisamment puissants et à coût abordables sont disponibles. Les algorithmes de calcul sont relativement bien développés. Les expériences d'administration de tests adaptatifs à grande échelle sont chose courante et une industrie de l'administration des tests adaptatifs est en émergence aux Etats-Unis. Cependant, comme le soulignent Wainer et Eignor (2000), les tests adaptatifs ne sont pas près de remplacer les tests papier crayon. Les coûts associés à l'administration des tests adaptatifs sont encore, pour le moment, plus élevés que ceux associés aux tests papier crayon. Par exemple, Wainer et Eignor (p. 285) considèrent que les frais d'administration du TOEFL aux États-Unis dans sa version papier crayon seraient encore de seulement 35 \$ à 40 \$ par individu, plutôt qu'actuellement d'environ 100 \$ pour la version adaptative par ordinateur. Le rapport coûts/bénéfices n'est donc pas encore justifié ; ce qui ne veut pas dire que nous ne devons pas nous préparer pour le futur. Un futur qui, d'ailleurs, ne s'avère peut-être pas si lointain.

Outre ces considérations financières, les tests adaptatifs continuent à poser de nouveaux défis à la modélisation de la réponse aux items. La plupart de ces défis sont apparus en dehors du contexte des tests adaptatifs. En fait, ils étaient déjà l'objet de recherches sur l'application des modélisations de la réponse à l'item aux tests papier crayon. Toutefois, les problèmes associés à l'administration des tests adaptatifs ont fait ressortir de façon plus aiguë l'importance de ces enjeux. Sans rechercher l'exhaustivité, voici quelques-uns de ces enjeux et ces défis.

Selon nous, le défi le plus important au cours des prochaines années sera de proposer des algorithmes de sélection du prochain item qui permettent d'exercer un contrôle quant à l'exposition des items de la banque d'items. Actuellement, il s'agit d'un domaine de recherche très actif. Les algorithmes traditionnels de sélection du prochain item, qui maximisent l'information ou qui minimisent l'erreur-type de l'estimateur du niveau d'habileté, favorisent

l'administration des mêmes items aux individus de même niveau d'habileté. Les auteurs étudient actuellement diverses stratégies dont le but est d'éviter que les mêmes items soient administrés trop fréquemment aux personnes qui affichent un niveau d'habileté similaire. La sécurité des tests adaptatifs et, conséquemment, la crédibilité de ces tests est en jeu.

Dans les tests adaptatifs, comme dans les tests papier crayon, le format de réponse aux items est presque toujours de type réponse à choix multiples. Les modélisations de la réponse à l'item utilisées pour ce type d'items reposent sur un postulat selon lequel la réponse à un item est indépendant de celle fournie à l'item précédent. Cependant, en éducation les tests sont fréquemment composés d'items dont la réponse est affectée par la réponse à l'item précédent sous la forme de minitests. De nouvelles modélisations de la réponse à l'item, ici le minitest, sont alors nécessaires.

Les tests adaptatifs nécessitent la disponibilité d'un nombre important d'items pour constituer des banques d'items qui justifient l'utilisation d'un test adaptatif. Les paramètres des items qui composent ces banques doivent de plus être préalablement estimés auprès d'échantillons dont la taille est suffisante. En dehors des évaluations à grande échelle, les efforts et les coûts associés à ces opérations sont souvent trop importants pour être réellement pratiques. Par exemple, il est actuellement difficile d'administrer un test adaptatif pour soutenir, soit l'évaluation formative, soit l'évaluation diagnostique, en salle de classe. Il ne serait pas justifiable d'investir des efforts humains et financiers considérables dans la création d'une banque d'items pour les besoins d'un seul groupe cours. Les auteurs (Bejar, 1993 ; Embretson, 1999) s'intéressent actuellement à des solutions pour résoudre ce problème. Principalement autour de la modélisation du test logistique linéaire (*linear logistic test model*) proposée par Fischer (1995), ces auteurs proposent des solutions pour déterminer le paramètre de difficulté d'un item à partir de ses caractéristiques (*model-based item generation*). Ces solutions devraient permettre de produire des items en cours d'administration d'un test adaptatif en fonction d'attributs divers sans être astreint à calibrer préalablement les paramètres des items d'une banque. Il est à noter que cette solution à la génération d'items en ligne présente aussi une solution alternative aux problèmes de sur exposition des items.

Les tests de personnalité, de valeurs ou d'intérêts devraient être tout désignés pour profiter des avantages de la technologie du testing adaptatif. Ces tests comportent généralement un nombre important de questions et le temps d'administration est assez long. Ils gagneraient à être élaborés sous forme adaptative pour permettre de diminuer de manière substantielle leur longueur et leur temps d'administration. Puisque le construit mesuré n'est plus unique, mais plutôt multidimensionnel, ces tests nécessitent toutefois des modélisations de la réponse à l'item plus complexes que celles qui ont été traditionnellement abordées en testing adaptatif. Les règles d'arrêt, de suite et de fin usuelles d'un test adaptatif doivent être aussi repensées pour soutenir ces modélisations multidimensionnelles de la réponse à l'item. Pour le moment, encore beaucoup de travail est à faire quant à la modélisation multidimensionnelle de la réponse à l'item et, conséquemment, peu de versions de tests adaptatifs permettent de mesurer des construits multidimensionnels.

Nous pourrions aborder plus en détails bien d'autres enjeux et défis associés aux tests adaptatifs, tels que l'analyse des items à réponse construites (Bennett et Ward, 1993), les mesures

d'ajustement des patrons de réponses (*person fit*) ou le fonctionnement différentiel d'item (*differential item functioning*). Le terrain de jeu est vaste et nous pouvons prédire, sans que la probabilité de se tromper soit trop importante, que le testing adaptatif sera un sujet d'intérêt en recherche, et pas seulement en éducation, pour plusieurs années encore.

9.6 Exercices

1. Décrivez un test papier-crayon conventionnel à partir des trois règles présentées à l'intérieur de ce chapitre.
2. Proposez un algorithme qui permet de décrire le déroulement d'une entrevue de sélection d'emploi à partir des trois règles présentées à l'intérieur de ce chapitre.
3. Suggérez une stratégie pour la règle de début dans un test adaptatif qui assure une protection contre la transmission de l'information quant au contenu du premier item administré.
4. Proposez un mini-test qui permet de mesurer l'habileté à développer un test adaptatif.
5. L'erreur-type retenue pour la règle d'arrêt étant égale à 0,20, estimez le biais de l'estimateur du niveau d'habileté lorsque le niveau d'habileté réel est égal à -2,50, -1,76, -0,82, 0,05, 1,13 et 4,00.
6. Appliquez la correction de Bock et Mislevy à l'estimateur final du niveau d'habileté après l'administration de chaque item lorsque la méthode de l'espérance a posteriori est utilisée. Comparez les nouvelles valeurs que vous avez calculées avec les valeurs de l'estimateur du niveau d'habileté qui proviennent des méthodes d'estimation adaptatives.
7. Quel impact a la correction pour biais de Bock et Mislevy sur l'erreur-type de l'estimateur du niveau d'habileté ? Plus précisément, l'erreur-type de la valeur corrigée de l'estimateur du niveau d'habileté devrait-elle augmenter, diminuer ou rester stable ? Justifiez votre réponse par des exemples tirés des résultats obtenus à la question 6.
8. À partir des estimateurs du niveau d'habileté du tableau 9.4, calculer la matrice des corrélations entre les diverses méthodes d'estimations du niveau d'habileté. Quelles sont les méthodes qui affichent les liens les plus marquées entre elles ?
9. Prédire l'erreur-type de l'estimateur du niveau d'habileté après l'administration du 15^e item, sachant que l'erreur-type de l'estimateur du niveau d'habileté est égale à 0,54 suite à l'administration du septième item.
10. Prédire le nombre d'items à administrer pour obtenir une erreur-type de l'estimateur du niveau d'habileté égale à 0,20 sachant que l'erreur-type obtenue après l'administration de 10 items est égale à 0,42.

9.7 Corrigé des exercices nécessitant des calculs

5. Selon l'équation 9.4

Réponse :

NIVEAU D'HABILITÉ	BIAIS ESTIMÉ
-2,50	0,10
-1,76	0,07
-0,82	0,02
0,05	-0,01
1,13	-0,06
4,00	-0,18

6. Selon l'équation 9.7

Réponse :

Item	$\hat{\theta}$	$S_{\hat{\theta}}$	Correction de Bock et Mislevy $\hat{\theta}_c$
1	-0,57	0,83	-1,83
2	-0,99	0,73	-2,12
3	-1,34	0,67	-2,43
4	-1,64	0,62	-2,66
5	-1,91	0,59	-2,93
6	-2,15	0,56	-3,14
7	-2,38	0,54	-3,36
8	-2,59	0,51	-3,50
9	-2,39	0,46	-3,04
10	-2,24	0,42	-2,71
11	-2,37	0,41	-2,85
12	-2,50	0,39	-2,94
13	-2,62	0,38	-3,06
14	-2,50	0,36	-2,88
15	-2,60	0,35	-2,97

8. À partir des valeurs de l'estimateur du niveau d'habileté obtenues selon les quatre méthodes d'estimation

Réponse :

	EAP	Correction de Bock et Mislevy adaptative	Estimateur a priori adaptatif	Intervalle d'intégration adaptatif
EAP		0,74	0,90	1,00
Correction de Bock et Mislevy adaptative			0,52	0,74
Estimateur a priori adaptatif				0,90
Intervalle d'intégration adaptatif				

9. À partir de l'équation 9.8

Réponse : 0,40

10. À partir de l'équation 9.9

Réponse : 51,40, soit 52 items.

9.8 Références

- Assessment system corporation (1984). *User's manual for the MicroCat testing system*. St. Paul, Minnesota : Assessment system corporation.
- Auger, R. (1989). Étude de praticabilité du testing adaptatif de maîtrise des apprentissages scolaires au Québec : une expérimentation en éducation économique secondaire 5. Thèse de doctorat non publiée. Montréal : Université du Québec à Montréal.
- Auger, R. et Séguin, S.P. (1992). Le testing adaptatif avec interprétation critérielle, une expérience de praticabilité du TAM pour l'évaluation sommative des apprentissages au Québec. *Mesure et évaluation en éducation*, 15-1 et 2, 103-145.
- Bejar, I.I. (1993). A generative approach to psychological and educational measurement. Dans N. Frederiksen, R.J. Mislevy et I.I. Bejar (Éds) : *Test theory for a new generation of tests*. Hillsdale : Lawrence Erlbaum Associates.
- Bennett, R.E. et Ward, W.C. (1993). *Construction versus choice in cognitive measurement: Issues in constructed responses, performance testing, and port-folio assessment*. Hillsdale : Lawrence Erlbaum Associates.
- Bock, R.D. et Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a micro computer environment. *Applied psychological measurement*, 6-4, 431-444.
- Chang, H.H. et Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, 58-1, 37-52.
- Chen, S.K., Hou, L., Fitzpatrick, S.J. et Dodd, B.G. (1997). The effect of population distribution and method of theta estimation on computerized adaptive testing (CAT) using the rating scale model. *Educational and psychological measurement*, 57-3, 422-439.
- Davey, T., Godwin, J. et Mittelholtz, D. (1997). Developing and scoring an innovative computerized writing assessment. *Journal of educational measurement*, 34-1, 21-41.
- 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., Koch, W.R. et de Ayala, R.J. (1993). Computerized adaptive testing using the partial credit model effects of item pool characteristics and different stopping rules. *Educational and psychological measurement*, 53-1, 61-77.
- Embretson, S.E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika*, 64(4), 407-433.

- Fischer, G.H. (1995). The linear logistic test model. Dans 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.
- Glas, C.A.W., Wainer, H. et Bradlow, E.T. (2000). MML and EAP estimation in testlet-based adaptive testing. Dans W.J. van der Linden et C.A.W. Glas (Éds) : *Computerized adaptive testing - Theory and practice*. Dordrecht : Kluwer.
- Goldstein, H. et Wood, R. (1989). Five decades of item response modelling. *British journal of mathematical and statistical psychology*, 42, 139-167.
- Hambleton, R.K. et Zaal, J.N., Pieters, J.M.P. (1991). Computerized adaptive testing : theory, applications, and standards. Dans R.K. Hambleton et J.N. Zaal (Éds) : *Advances in educational and psychological testing - Theory and applications*. Boston : Kluwer.
- Hetter, R.D. et Sympon, J.B. (1997). Item exposure control in CAT-ASVAB. Dans W.A. Sands, B.K. Waters et J.R. McBride (Éds) : *Computerized adaptive testing - From inquiry to application*. Washington, DN : American psychological association, APA.
- Hojtink, H. et Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. Dans G.H. Fischer et I.W. Molenaar (Éds) : *Rasch models - Foundations, recent developments, and applications*. New York : Springer-Verlag.
- Jensema, C.J. (1974). An application of latent trait mental test theory. *British journal of mathematical and statistical psychology*, 27, 29-48.
- Jensema, C.J. (1977). Bayesian tailored testing and the influence of item bank characteristics. *Applied psychological measurement*, 1-1, 111-120.
- Kingsbury, G.G. et Weiss, D.J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. Dans D.J. Weiss (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 16^e session d'étude de l'ADMÉE à Laval. Montréal : Association pour le développement de la mesure et de l'évaluation en éducation.
- Laurier, M. (1993b). *L'informatisation d'un test de classement en langue seconde*. Québec : Université Laval, Faculté des lettres.

- Laurier, M. (1993c). Un test adaptatif en langue seconde : la perception des apprenants. Dans R. Hivon (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. Dans 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.
- Laurier, M., Froio, L., Paero, C. et Fournier, M. (1999). *L'élaboration d'un test provincial pour le classement des étudiants en anglais langue seconde, au collégial*. Québec : ministère de l'Éducation, Direction générale de l'enseignement collégial.
- Laveault, D. et Grégoire, J. (1997). *Introduction aux théories des tests en sciences humaines*. Paris : De Boeck.
- Linacre, J.M. (2000). *Computer-adaptive testing : a methodology whose time has come*. MESA memorandum no 69. Chicago : MESA psychometric laboratory, University of Chicago.
- Lord, F.M. (1952). A theory of test scores. *Psychometric monographs*, no 7.
- McBride, J.R. et Martin, J.T. (1983). Reliability and validity of adaptive tests in a military setting. Dans D.J. Weiss (Éd.) : *New Horizons in testing : latent trait test theory and computerized adaptive testing*. New York : Academic Press.
- Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Microsoft (2000, 27 août). Adaptive testing. Disponibilité : http://www.windowsgalore.com/cert/adaptive_testing/index.htm.
- Mislevy, R.J. et Bock, R.D. (1982). Biweight estimates of latent ability. *Educational and psychological measurement*, 42-2, 725-737.
- 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.

- Raïche, G. (1994). *La simulation de modèle sur ordinateur en tant que méthode de recherche : le cas concret de l'étude de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt*. Actes du 6^e colloque de l'Association pour la recherche au collégial. Montréal : Association pour la recherche au collégial, ARC.
- Raïche, G. (2000). *La distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt : selon l'erreur-type et selon le nombre d'items administrés*. Thèse de doctorat inédite. Montréal : Université de Montréal.
- Raïche, G. (2001a). *Principes et enjeux du testing adaptatif : de la loi des petits nombres à la loi des grands nombres*. Communication présentée à l'intérieur du 69^e congrès de l'Association canadienne française pour l'avancement de la science, Acfas. Sherbrooke : Acfas.
- Raïche, G. (2001b). *Pour une évaluation sur mesure des étudiants : défis et enjeux du testing adaptatif*. Communication présentée à l'intérieur de la 23^e session d'études de l'Association pour le développement de la mesure et de l'évaluation en éducation, ADMÉEÉ. Québec : ADMÉEÉ.
- Raïche, G. et Blais, J.G. (2002a, accepté). Étude de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt dans le contexte de l'application du modèle de Rasch. *Mesure et évaluation en éducation*
- Raïche, G. et Blais, J.G. (2002b). *Practical considerations about expected a posteriori estimation in adaptive testing: Adaptive a priori, adaptive correction for bias, and adaptive integration interval*. Communication proposée au 11^e Biannual International objective measurement workshop. New-Orleans : IOMW.
- Raïche, G. et Blais, J.G. (2002c). *Some features of the estimated sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules*. Communication proposée au 11^e Biannual International objective measurement workshop. New-Orleans : IOMW.
- Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. Dans N. Frederiksen, R.J. Mislevy et I.I. Bejar (Éds) : *Test theory for a new generation of tests*. Hillsdale : Lawrence Erlbaum Associates.
- Thissen, D. et Mislevy, R.J. (2000). Testing algorithms. Dans H. Wainer, D. Eignor, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (Éds) : *Computerized adaptive testing - A primer*. Hillsdale : Lawrence Erlbaum Associates.
- Thissen, D. et 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.
- van der Linden, W.J. (1999). Empirical initialization of the trait estimator in adaptive testing. *Applied psychological measurement*, 23-1, 21-29.
- Van der Linden, W.J. (2000). Constrained adaptive testing with shadow tests. Dans W.J. van der Linden et C.A.W. Glas (Éds) : *Computerized adaptive testing - Theory and practice*. Dordrecht : Kluwer.
- van der Linden, W.J. et Pashley, P.J. (2000). Item selection and ability estimation in adaptive testing. Dans W.J. van der Linden et C.A.W. Glas (Éds) : *Computerized adaptive testing - Theory and practice*. Dordrecht, Pays-Bas : Kluwer.
- Vos, H.J. et Glas, C.A.W. (2000). Testlet-based adaptive mastery testing. Dans W.J. van der Linden et C.A.W. Glas (Éds) : *Computerized adaptive testing - Theory and practice*. Dordrecht : Kluwer.
- Wainer, H. (1983). Are we correcting for guessing in the wrong direction ? Dans D.J. Weiss (Éd.) : *New horizons in testing - Latent trait test theory and computerized adaptive testing*. New York : Academic Press.
- Wainer, H., Bradlow, E.T. et Du, Z. (2000). Testlet response theory : an analog for the 3PL model useful in testlet-based adaptive testing. Dans W.J. van der Linden et C.A.W. Glas (Éds) : *Computerized adaptive testing - Theory and practice*. Dordrecht : Kluwer.
- Wainer, H., Dorans, N.J., Green, B.F., Mislevy, R.J., Steinberg, L. et Thissen, D. (1990). Future challenges. Dans H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (Éds) : *Computerized adaptive testing - A primer*. Hillsdale : Lawrence Erlbaum Associates.
- Wainer, H. et Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. Dans H. Wainer, D. Eignor, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg et D. Thissen (Éds) : *Computerized adaptive testing - A primer*. Hillsdale : Lawrence Erlbaum Associates.
- Wainer, H. et Kiely, G.L. (1987). Item clusters and computerized adaptive testing : a case for testlets. *Journal of educational measurement*, 24-3, 185-201.
- Wainer, H., Thissen, D. (1987). Estimating ability with the wrong model. *Journal of educational statistics*, 12-4, 339-368.
- Wainer, H., Thissen, D. (2001). True score theory : the traditional method. Dans D. Thissen. et H. Wainer (Éds) : *Test scoring* : Mahwah, NJ : Lawrence Erlbaum.

- Wainer, H. et Wright, B. (1980). Robust estimation of ability in the Rash model. *Psychometrika*, 45, 370-390.
- 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. (1985). Adaptive testing by computer. *Journal of consulting and clinical psychology*, 53-6, 774-789.