

**Direction des bibliothèques**

**AVIS**

Ce document a été numérisé par la Division de la gestion des documents et des archives de l'Université de Montréal.

L'auteur a autorisé l'Université de Montréal à reproduire et diffuser, en totalité ou en partie, par quelque moyen que ce soit et sur quelque support que ce soit, et exclusivement à des fins non lucratives d'enseignement et de recherche, des copies de ce mémoire ou de cette thèse.

L'auteur et les coauteurs le cas échéant conservent la propriété du droit d'auteur et des droits moraux qui protègent ce document. Ni la thèse ou le mémoire, ni des extraits substantiels de ce document, ne doivent être imprimés ou autrement reproduits sans l'autorisation de l'auteur.

Afin de se conformer à la Loi canadienne sur la protection des renseignements personnels, quelques formulaires secondaires, coordonnées ou signatures intégrées au texte ont pu être enlevés de ce document. Bien que cela ait pu affecter la pagination, il n'y a aucun contenu manquant.

**NOTICE**

This document was digitized by the Records Management & Archives Division of Université de Montréal.

The author of this thesis or dissertation has granted a nonexclusive license allowing Université de Montréal to reproduce and publish the document, in part or in whole, and in any format, solely for noncommercial educational and research purposes.

The author and co-authors if applicable retain copyright ownership and moral rights in this document. Neither the whole thesis or dissertation, nor substantial extracts from it, may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms, contact information or signatures may have been removed from the document. While this may affect the document page count, it does not represent any loss of content from the document.

Université de Montréal

# **TAARAC : Test d'Anglais Adaptatif par Raisonnement À base de Cas**

par

Zakia Lakhlili

Département d'Informatique et de Recherche Opérationnelle  
Faculté des Arts et des Sciences

Mémoire présenté à la Faculté des études supérieures  
en vue de l'obtention du grade de Maîtrise ès Sciences  
en Informatique

Décembre, 2007

© Zakia Lakhlili, 2007



Université de Montréal  
Faculté des études supérieures

Ce mémoire intitulé :

TAARAC : Test d'anglais adaptatif par raisonnement à base de cas

Présenté par :  
Zakia Lakhli

a été évalué par un jury composé des personnes suivantes :

Claude Frasson, président-rapporteur  
Esma Aïmeur, directrice de recherche  
Jian-Yun Nie, membre du jury

## Résumé

De nos jours, les tests assistés par ordinateur sont devenus une pratique courante dans les systèmes éducatifs, notamment les tests adaptatifs (*Computerized adaptive tests*). La plupart des tests de réputation internationale possèdent une version adaptative, en adaptant les questions selon le niveau d'habileté du candidat. Habituellement un test adaptatif par ordinateur est construit en se basant soit sur la théorie des réponses aux items (*item response theory*), ou sur les modélisations Bayésiennes. Nous présentons ici TAARAC, *Test Adaptatif en Anglais par Raisonnement À base de Cas*, construit à partir de deux techniques : *la théorie des réponses aux items* couplée au *raisonnement à base de cas* (*case-based reasoning*). Nous avons opté pour la première technique à cause de son pouvoir incrémental. Elle est utilisée pour adapter les questions selon le niveau d'habileté du candidat. Alors que la deuxième technique a permis le choix de la première question à présenter au candidat, celle dont la difficulté se rapproche le plus de son niveau d'habileté. Les résultats obtenus, après avoir testé TAARAC sur 102 personnes, ont montré que le raisonnement à base de cas a rendu le test : *plus court*, ainsi 67% des personnes dont la longueur du test est de 16 questions avaient pour niveau d'habileté initial celui déduit par le raisonnement à base de cas; *plus précis*, puisqu'avec un test plus court on atteint facilement la précision ou l'erreur type fixée au préalable à 0.2; et *plus efficace*, puisque le raisonnement à base de cas favorise le rapprochement entre le niveau d'habileté initial et celui estimé à la fin du test, ainsi tout au long du test les questions choisies par TAARAC sont de difficulté proche du niveau réel du candidat.

**Mots-clés** : test adaptatif par ordinateur, théorie des réponses aux items, l'espérance à postériori, estimateur du niveau d'habileté, raisonnement à base de cas, similarité des profils.

## Abstract

Nowadays, Computer-Based Tests have become a common practice in educational systems, specifically *computerized adaptive testing*. Most of the well-known tests have an adaptive version so as to adapt questions according to the candidate's skill level. Usually, a computerized adaptive test is built by using the *item response theory* or the Bayesian modelling. In this work, we present TAARAC, a case-based reasoning adaptive English test. TAARAC is based on *item response theory* coupled with *case based reasoning*. We opted for the first technique because of its incremental power: it is used to adapt the questions according to the candidate's skill level. The second technique allows a better choice of the first question presented to the candidate, whose difficulty best matches his skill level. The results obtained by testing TAARAC on 102 persons, showed that with the case-based reasoning the test becomes: *shorter*, because 67 % of the persons whose length of the test is 16 questions had the skill level initially deduced by the case based reasoning; *more precise*, because a shorter test indicates reaching the desired precision or the standard error set beforehand to 0.2; and *more effective*, because the case-based reasoning estimates the initial skill level close to that established at the end of the test, thus, throughout the test, the questions selected by TAARAC are those whose difficulty corresponds better to the real candidate's level.

**Keywords:** computerized adaptive testing, item response theory, expected a posteriori, estimation of skill level, case-based reasoning, profile similarity.

# Table des matières

<b>Chapitre 1 : Introduction.....</b>	<b>1</b>
1.1. <i>Notre approche</i> .....	3
1.2. <i>Plan du mémoire</i> .....	4
<b>Chapitre 2 : La théorie des réponses aux items (Item response theory) .....</b>	<b>6</b>
2.1. <i>Introduction</i> .....	6
2.2. <i>Généralités</i> .....	7
2.3. <i>Courbe caractéristique de l’item</i> .....	8
2.3.1. <i>Modèle logistique à un paramètre</i> .....	10
2.3.2. <i>Modèle à deux paramètres</i> .....	11
2.3.3. <i>Modèle à trois paramètres</i> .....	12
2.4. <i>Estimation du niveau d’habileté</i> .....	14
2.4.1. <i>Méthode de vraisemblance maximale</i> .....	14
2.4.2. <i>La méthode bayésienne de maximisation à posteriori (MAP)</i> .....	17
2.4.3. <i>Méthode de l’espérance à posteriori (EAP)</i> .....	18
2.5. <i>Estimation des paramètres de l’item</i> .....	20
2.5.1. <i>L’estimation de vraisemblance marginale maximale (MML)</i> .....	20
2.6. <i>Erreur type et fonction d’information d’item et de test</i> .....	21
2.7. <i>Autres modèles</i> .....	23
2.7.1. <i>Modèle polytomique</i> .....	23
2.7.2. <i>Modèle multidimensionnel</i> .....	24
2.8. <i>Conclusion</i> .....	25
<b>Chapitre 3 : Test Adaptatif par Ordinateur (TAO).....</b>	<b>26</b>
3.1. <i>Test adaptatif par ordinateur basé sur une modélisation Bayésienne</i> .....	27
3.1.1. <i>La structure d’un réseau Bayésien</i> .....	28
3.1.2. <i>Le déroulement d’un test adaptatif par ordinateur dans un réseau bayésien</i> .....	30
3.1.3. <i>Le modèle POKS</i> .....	31
3.2. <i>Test adaptatif par ordinateur basé sur la théorie des réponses aux items</i> .....	32
3.2.1. <i>Banque d’items</i> .....	32
3.2.2. <i>Déroulement d’un test adaptatif par ordinateur</i> .....	33
3.2.2.1. <i>Étape du départ</i> .....	35
3.2.2.2. <i>Étape de succession</i> .....	36
a) <i>La sélection de l’item</i> .....	37
b) <i>Contrôle du taux d’exposition</i> .....	39
c) <i>Procédures de stratification</i> .....	40
3.2.2.3. <i>Étape d’arrêt</i> .....	42
3.2.3. <i>L’estimateur final du niveau d’habileté</i> .....	43

3.3.	<i>Conclusion</i> .....	43
<b>Chapitre 4 : Raisonnement à base de cas</b> .....		<b>45</b>
4.1.	<i>Définitions</i> .....	46
4.2.	<i>Le processus d'un raisonnement à base de cas</i> .....	47
4.2.1.	La recherche .....	48
4.2.2.	L'adaptation .....	49
4.2.3.	La maintenance .....	49
4.2.4.	L'apprentissage .....	50
4.3.	<i>Les parties composant un cas</i> .....	50
4.4.	<i>Modèles de la représentation des cas</i> .....	51
4.5.	<i>Mesure de similarité dans un système de raisonnement à base de cas</i> .....	52
4.6.	<i>Domaines d'application du raisonnement à base de cas</i> .....	54
4.6.1.	L'art et le divertissement .....	54
4.6.2.	La Science .....	54
4.6.3.	L'industrie .....	55
4.6.4.	L'éducation .....	55
4.7.	<i>Conclusion</i> .....	56
<b>Chapitre 5 : Conception et méthodologie</b> .....		<b>57</b>
5.1.	<i>Objectifs</i> .....	57
5.2.	<i>Approche de TAARAC</i> .....	58
5.2.1.	Algorithme de TAARAC .....	58
5.2.2.	Fonctionnement général .....	60
5.2.3.	Raisonnement à base de cas .....	61
5.3.	<i>Méthodologie de TAARAC</i> .....	63
5.3.1.	Banque d'items .....	63
5.3.2.	Étape de départ .....	65
5.3.3.	Étape de succession .....	66
5.3.4.	Étape d'arrêt .....	70
5.4.	<i>Architecture de TAARAC</i> .....	70
5.4.1.	Le module du raisonnement à base de cas .....	72
5.4.2.	Le module du test adaptatif .....	75
5.4.3.	Base des cas .....	83
5.4.4.	Base d'items .....	84
5.5.	<i>Comparaison</i> .....	84
5.6.	<i>Conclusion</i> .....	88
<b>Chapitre 6 : Implémentation et validation</b> .....		<b>90</b>
6.1.	<i>Implémentation de TAARAC</i> .....	90
6.1.1.	Architecture trois-tiers .....	90

6.1.2. Le modèle MVC .....	91
6.2. <i>Environnement d'un candidat</i> .....	93
6.2.1. Page d'accueil de TAARAC.....	93
6.2.2. Enregistrement.....	94
6.2.3. Identification.....	96
6.2.4. Pendant le test du candidat .....	98
6.2.5. À la fin du test du candidat.....	102
6.3. <i>Validation</i> .....	106
6.3.1. Mise en contexte.....	106
6.3.2. L'adaptabilité de TAARAC.....	107
6.3.3. L'apport du raisonnement à base de cas .....	110
6.3.3.1. Un test plus court.....	110
6.3.3.2. Un test plus efficace .....	112
6.3.4. Feedback du candidat .....	113
6.3.4.1. Feedback inféré.....	114
6.3.4.2. Feedback réel.....	115
6.3.5. Conclusion.....	116
<b>Chapitre 7 : Conclusion</b> .....	<b>117</b>
<b>Bibliographie</b> .....	<b>121</b>



## Liste des tableaux

<b>Tableau 3.1</b> : Une décomposition fictive du cours d'anglais .....	29
<b>Tableau 3.2</b> : Une hiérarchie du sujet l'Anglais.....	29
<b>Tableau 5.2</b> : Les items classés selon leurs paramètres de difficulté .....	67
<b>Tableau 5.3</b> : Les blocs d'items classés selon le paramètre de discrimination .....	68
<b>Tableau 5.4</b> : Les blocs d'items formés par ceux du tableau 5.3, en regroupant les items du même niveau.....	68
<b>Tableau 5.5</b> : Les profils de Fatima, Rachid et Marc. ....	73
<b>Tableau 5.6</b> : Les distances des attributs de chaque paire de profils.....	74
<b>Tableau 5.7</b> : Les paramètres des items, 33, 21 et 39.....	78
<b>Tableau 5.8</b> : Le récapitulatif des étapes du calcul de niveau d'habileté. ....	79
<b>Tableau 5.9</b> : Le récapitulatif du test de Marc.....	82
<b>Tableau 5.10</b> : Comparaison des tests adaptatifs.....	86
<b>Tableau 6.1</b> : Résumé des résultats obtenus par un candidat de niveau moyen.....	108
<b>Tableau 6.2</b> : Résumé des résultats obtenus par un candidat de niveau assez fort.....	109
<b>Tableau 6.3</b> : Les candidats répartis selon la longueur du test et le niveau d'habileté initial.....	111
<b>Tableau 6.4</b> : Les candidats répartis selon le niveau d'habileté initial et l'exactitude du test.....	113
<b>Tableau 6.5</b> : Les candidats répartis selon la différence entre le niveau auto évalué et celui estimé à la fin du test. ....	114
<b>Tableau 6.6</b> : Les moyennes des appréciations de 42 candidats. ....	116

## Liste des figures

<b>Figure 2.1</b> : Proportions de réussite d'un item calculées à partir des scores totaux centrés et réduits d'un test de mathématique [Bertrand & Blais 2004] .....	9
<b>Figure 2.2</b> : CCI de trois items selon le modèle à un paramètre, avec $b_1=-1$ , $b_2=0$ et $b_3=1$ [URL2].....	11
<b>Figure 2.3</b> : CCI de trois items selon le modèle à deux paramètres avec $b_4=b_5=b_6=0$ , $a_4=0.5$ , $a_5=1$ et $a_6=2$ [URL2].....	12
<b>Figure 2.4</b> : CCI de trois items selon le modèle à trois paramètres ( $a=1$ , $b=0$ et $D=1.7$ ). [Raïche 2000] .....	13
<b>Figure 2.5</b> : Illustration de la méthode de Newton-Raphson. [Hambleton & Swaminathan 1985].....	15
<b>Figure 3.1</b> : La structure d'un réseau Bayésien ( $I_1, I_2, \dots, I_{11}$ sont des items).....	29
<b>Figure 3.2</b> : Organigramme du processus du test adaptatif .....	35
<b>Figure 4.1</b> : Processus d'un système de CBR.....	47
<b>Figure 5.1</b> : L'algorithme de TAARAC .....	59
<b>Figure 5.2</b> : Vue d'écran de PARAM-3PL.....	64
<b>Figure 5.3</b> : Fichier de données qui représente l'entrée de PARAM-3PL.....	64
<b>Figure 5.4</b> : Fichier des paramètres de l'item représentant la sortie de PARAM-3P ( $a\_calib$ est le paramètre de discrimination, $b\_calib$ est celui de difficulté et $c\_calib$ est celui de pseudo-chance).....	65
<b>Figure 5.5</b> : Architecture de TAARAC .....	71
<b>Figure 6.1</b> : Modèle MVC. ....	92
<b>Figure 6.2</b> : Page d'accueil de TAARAC.....	93
<b>Figure 6.3</b> : Première partie de l'enregistrement dans TAARAC. ....	94
<b>Figure 6.4</b> : Page d'erreur.....	95
<b>Figure 6.5</b> : Deuxième étape de l'enregistrement.....	95
<b>Figure 6.6</b> : Page d'identification .....	96
<b>Figure 6.7</b> : Page affichée après l'identification d'un candidat qui n'a pas passé le test. 97	
<b>Figure 6.8</b> : Page affichée après l'identification d'un candidat qui a passé le test.....	97
<b>Figure 6.9</b> : Page affichée après l'identification d'un candidat qui a suspendu le test.....	98

<b>Figure 6.10</b> : Page affichée avant de commencer le test. ....	99
<b>Figure 6.11</b> : Page représentant la question affichée au candidat.....	100
<b>Figure 6.12</b> : Page affichée en cliquant sur le bouton de l'arrêt du test. ....	101
<b>Figure 6.13</b> : Page affichée en cliquant sur le bouton de la suspension du test.....	101
<b>Figure 6.14</b> : Page affichant l'item dans les dix dernières secondes. ....	102
<b>Figure 6.15</b> : Page affichant le niveau d'habileté du candidat. ....	103
<b>Figure 6.16</b> : Avertissement affiché avant la correction.....	103
<b>Figure 6.17</b> : Page affichant la correction du test. ....	104
<b>Figure 6.18</b> : Page affichant les résultats du candidat. ....	105
<b>Figure 6.19</b> : Graphe répartition des candidats selon leur niveau auto évalué.....	106
<b>Figure 6.20</b> : Diagramme représentant le nombre des candidats selon la longueur du test et leurs niveaux d'habileté initial.....	111
<b>Figure 6.21</b> : Diagramme représentant le nombre des candidats selon les tranches de différence entre le niveau auto évalué et le niveau estimé à la fin du test.....	114
<b>Figure 6.22</b> : Le formulaire du feedback du candidat. ....	115

## Remerciements

Je voudrais en premier lieu adresser mes sincères remerciements à ma directrice de recherche, professeure Esma Aïmeur : elle a su me guider lors de l'accomplissement de mon projet, m'a facilité la tâche en fournissant de précieux conseils et a toujours été disponible pour moi, avec des encouragements permanents.

Je remercie également les membres de l'honorable jury qui ont bien voulu évaluer mon travail, les professeurs Claude Frasson et Jian-Yun Nie.

Par cet ouvrage, je tiens également à témoigner toute mon affection et mon attachement à  
Ma petite famille :

Mon cher mari Abdelouahab Assikar qui m'a toujours apporté soutien et réconfort, et sans qui ce travail n'aurait pas été possible.

Mes chers enfants Hatim et Ali qui étaient patients avec moi, qui ont compris, malgré leur jeune âge, que maman a du travail et il faut la laisser travailler.

Je vous remercie de votre immense aide et de votre présence à mes côtés.

Ma grande famille :

Ma chère mère R'kia Oummouy , mon cher père Omar Lakhli, mes frères Yassine et Hicham ainsi que ma cousine Nawal qui m'ont beaucoup soutenu et encouragé. Que Dieu vous protège et vous garde en bonne santé.

Mes amis :

Sihem Benlizidia, Narimel Bendakir et Soumaya Chaffar, qui se sont avérées d'une aide précieuse pour l'accomplissement de mon projet de maîtrise. Je tiens particulièrement à remercier Leila Arras, Karima Echihabi, Ghita Kouadri Mostafaoui et Malika Zidani qui m'ont fourni soutien et amitié jour après jour.

Les membres du laboratoire Héron :

Sébastien Gambs, Flavien Serge Mani Onana, Emmanuel Blanchard, Hicham Hage, Pierre Chalfoun, Soumaya Chaffar, Narimel Bendakir, Sihem Benlizidia , Moez Mabrouk, Khadija Benadada, Ai Thanh Ho, Jeanne d'Arc Uwatowenimana et Zbigniew Rakowski qui m'ont apporté un grand soutien et m'ont fourni un environnement de travail propice.

Je remercie particulièrement les responsables du test de classement d'anglais de la faculté de l'éducation permanente de l'université de Montréal qui m'ont fourni, données et renseignements nécessaires pour la réalisation de ce travail.

## Chapitre 1 : Introduction

Le e-learning ou l'apprentissage à distance a vu une expansion considérable dans ces dernières années. Plusieurs systèmes éducatifs ont vu le jour grâce à l'intérêt croissant qu'on leur accorde, la disponibilité de plus en plus importante du matériel informatique et le développement considérable que connaissent les technologies de l'information ont largement contribué à cette expansion. Ainsi, la plupart des écoles et des universités possèdent un ou plusieurs systèmes d'apprentissage en ligne, qui réduisent les contraintes du temps et de l'espace entre le formateur et l'apprenant.

Qui dit système éducatif, dit évaluation des compétences et des connaissances de l'apprenant qui se fait généralement à travers des tests. Dans le cas de l'apprentissage à distance, ces tests sont assistés par ordinateur. Au début ceux-ci, consistaient en une traduction et en une mise en ligne des tests papier existants. Leur avantage était l'automatisation de l'analyse des résultats, facilitant ainsi la tâche aux formateurs. Mais il reste que ces tests sont très demandeurs en matière de temps et de coût de production, puisque au fur et à mesure que l'apprenant avance dans sa formation, un nouveau test doit être conçu dépendamment du niveau atteint. *La durée* de vie du test devient ainsi très courte et sa *réutilisabilité* limitée. Pour remédier aux contraintes des tests assistés par ordinateur, s'est très vite ajoutée la capacité adaptative des séries des questions selon le niveau de l'apprenant. Ce sont les tests adaptatifs par ordinateur.

Comme son nom l'indique, un test adaptatif par ordinateur est un test spécifiquement adapté à la personne à qui on administre le test. C'est une forme de test sur mesure (*tailored testing*). Son utilisation en éducation est facilitée par l'apparition des propositions de *la théorie des réponses aux items (item response theory)* [Wainer 2000] qui sont tout à fait différentes de celles suggérées dans le contexte de *la théorie classique des tests* qui se contente de faire l'analyse des scores obtenus dans le test. Alors que la théorie des réponses aux items est fondée sur la réponse du sujet à l'item. Nous désignons ici par item, une question.

Afin de mieux comprendre l'approche d'un test adaptatif, nous allons prendre comme exemple le cas d'un enseignant qui veut déterminer oralement le niveau d'habileté d'un sujet donné. Ainsi si l'enseignant a une idée du niveau du sujet, il va lui proposer une question de difficulté proche de son niveau. Sinon, il se contentera de lui présenter une de difficulté moyenne. La réponse correcte entraînera la suggestion d'une question plus difficile. Alors qu'une réponse incorrecte amène l'enseignant à lui énoncer une plus facile. Ainsi de suite jusqu'à ce que ce dernier juge qu'il est capable de classer le sujet dans le niveau adéquat.

Le test adaptatif par ordinateur utilise la même approche citée ci-dessus, puisqu'il présente tout le temps au candidat la question dont la difficulté se rapproche du niveau d'habileté du candidat. Au début du test, le niveau d'habileté du candidat n'est pas encore connu, c'est pourquoi il peut être estimé à partir des tests passés auparavant par le candidat, ou par la valeur de la moyenne des compétences des candidats ayant déjà passé le test si aucun préalable n'est connu sur le candidat. Une question de difficulté proche du niveau d'habileté estimé initialement est présentée au candidat. Dépendamment de la réponse de ce dernier, son niveau d'habileté estimé sera soit augmenté ou diminué. Par la suite, la prochaine question présentée au candidat sera respectivement soit plus difficile ou plus facile. Ce processus se poursuit jusqu'à ce que l'erreur type (la précision) du niveau d'habileté estimé du candidat atteigne une valeur fixée au préalable, ou qu'un nombre bien défini de questions soient allouées au candidat. Ce genre de test est capable d'évaluer un apprenant tout au long de sa formation quelque soit le niveau atteint, sans que le formateur ne soit obligé de construire un nouveau test, car il est capable de classer les candidats dans une compétence selon le niveau atteint dans celle-ci. Plusieurs tests de classement de réputation internationale possèdent une version adaptative, notons par exemple : divers tests développés par l'ETS (l'Educational Testing Service) tel que GRE (Graduate Record Examination) [URL4], le TOEFL (Test Of English as a Foreign Language) [URL5], le GMAT (Graduate Management Test) [URL8], et plusieurs autres, ainsi que les tests de certification de Microsoft.

Dans la section suivante, nous allons définir la problématique du test adaptatif et présenter notre approche.

## 1.1. Notre approche

Comme nous venons de voir, dans un test adaptatif les questions présentées au candidat ont une difficulté proche du niveau d'habileté du candidat. Alors qu'au début du test, le niveau d'habileté du candidat n'est pas encore connu. C'est pourquoi la plupart des tests adaptatifs choisissent comme première question : celle de difficulté moyenne, rendant ainsi le test plus long que nécessaire. TAARAC essaye de remédier à ce problème en estimant pour tout nouveau candidat, un niveau d'habileté initial.

Afin de mieux comprendre notre approche, revenons à l'exemple cité précédemment. Supposons que notre enseignant est quelqu'un d'expérimenté dans sa matière, et a accumulé de nombreuses connaissances sur les différents sujets qu'il a vus au cours de ses multiples années d'enseignement. Ainsi, avant de commencer le test oral pour un nouveau sujet, l'enseignant va essayer de collecter le plus d'informations possibles sur ce sujet afin d'avoir une idée sur son profil. Au fil des années il a probablement fait une correspondance entre les profils des sujets qu'il a formés et leurs niveaux d'habileté. C'est pourquoi connaissant le profil du sujet, il sera capable de lui estimer un niveau d'habileté initial, et par conséquent lui présenter comme première question, celle qui sera de difficulté proche du niveau d'habileté estimé.

C'est l'approche présentée dans l'exemple ci-dessus que nous avons adoptée dans TAARAC. Le fait de faire appel aux expériences déjà vécues par l'enseignant pour trouver une solution au nouveau problème, nous a fait penser au *raisonnement à base de cas*. C'est pourquoi nous avons choisi d'intégrer cette technique dans TAARAC. Puisque ce dernier est un test d'anglais adaptatif, nous avons essayé de recueillir un maximum d'informations, en relation avec l'apprentissage de l'anglais, de notre nouveau candidat, formant ainsi son profil. À partir des candidats qui ont déjà passé le test, dégager ceux dont le profil se rapproche de celui de notre candidat. La moyenne de leurs niveaux d'habileté *estimés* par TAARAC sera considérée comme le niveau d'habileté *initial* pour notre nouveau candidat.



Un niveau d'habileté initial propre à chaque candidat fait en sorte que le premier item présenté soit différent d'un candidat à un autre, rendant ainsi le test *plus fiable*. Car sinon un risque existe que les candidats ayant déjà passé le test vont informer le nouveau candidat sur la première question. Si celui-ci répond correctement alors que normalement il n'aurait su la réponse la suite du test sera biaisée.

Raïche a montré [Raïche 2000] qu'en diminuant la longueur du test, la différence entre le niveau d'habileté *estimé* et le *vrai* niveau d'habileté du candidat augmente au fur et à mesure que la différence entre le niveau d'habileté *initial* et le *vrai* niveau d'habileté du candidat augmente. Ainsi le niveau d'habileté *initial* affecte considérablement le niveau d'habileté *estimé*, d'autant plus que la longueur du test diminue. L'objectif principal d'un test adaptatif est de faire une bonne estimation du vrai niveau d'habileté des candidats avec un nombre minimal d'items. En ajoutant le raisonnement à base de cas, nous cherchons à rendre le test plus court, plus précis et plus efficace. C'est ce que nous allons démontrer dans ce travail.

## 1.2. Plan du mémoire

Ce mémoire est organisé de la façon suivante : dans le deuxième chapitre nous décrivons la théorie des réponses aux items, en définissant la courbe caractéristique de l'item, en présentant ses différents modèles logistiques et en montrant ses différentes façons d'estimer le niveau d'habileté et les paramètres de l'item. Le troisième chapitre présente le test adaptatif à travers sa banque d'items, un algorithme qui représente son déroulement, et les différentes étapes qui le constituent. Le quatrième chapitre propose un état de l'art sur le raisonnement à base de cas, dans lequel son processus, son modèle et une description du cas sont expliqués. Au cinquième chapitre, l'architecture de TAARAC et les différents modules qui le composent, sont présentés. Le sixième chapitre décrit l'implémentation et la validation de notre système. Finalement, une conclusion et des perspectives futures sont présentées dans le dernier chapitre.

Notons que dans la suite de ce mémoire, nous allons utiliser, de façon interchangeable, le mot « item » pour désigner « question », le mot « compétence » ou

« trait latent » pour désigner « niveau d'habileté » et le mot « précision » pour désigner « l'erreur type ». De plus, l'utilisation du masculin sera privilégiée pour des fins de clarté.

## Chapitre 2 : La théorie des réponses aux items (Item response theory)

Dans ce chapitre nous allons voir les fonctions de base de la théorie des réponses aux items en se basant sur l'aspect mathématique, ainsi nous montrons sa pertinence dans un test adaptatif par ordinateur.

### 2.1. Introduction

Comme nous connaissons tous, c'est à partir des scores obtenus dans un test qu'on peut retrouver le niveau de difficulté de ce dernier, ainsi analyser un test revient à étudier les scores obtenus dans celui-ci et leurs erreurs de mesure. Cette façon d'analyser les tests s'appelle généralement *la théorie classique du test*. C'est une théorie qui a été créée par Charles Spersman en 1904, qui ne prend pas en considération ni la réponse à l'item, ni le niveau d'habileté des candidats et dont le niveau de difficulté<sup>1</sup> de l'item et celui de discrimination<sup>2</sup> dépendent directement des réponses du groupe passant le test. Ainsi dans un groupe où la plupart des candidats ont répondu correctement à un item, le niveau de difficulté de ce dernier va être supérieur à celui déduit à partir d'un groupe dont la plupart des candidats ont une réponse incorrecte à cet item [Hambleton & Swaminathan 1985]. De même, puisque l'indice de discrimination classique de l'item est calculé à partir de la corrélation bisériale entre l'item et le score total du test [URL1], alors celui-ci dépend directement des scores des candidats dans un groupe donné. Ainsi, dans un groupe homogène, c'est-à-dire dont les scores se rapprochent, l'indice de discrimination de l'item est faible, alors qu'il est élevé dans un groupe hétérogène [Bertrand & Blais 2004].

---

<sup>1</sup> Le niveau de difficulté de l'item est la proportion des candidats d'un groupe donné qui réussissent l'item, il est en fait un indice de facilité [Bertrand & Blais 2004].

<sup>2</sup> Le niveau de discrimination de l'item est sa puissance de faire la distinction entre les candidats les plus habiles et les moins habiles. Il est souvent calculé par la corrélation de Pearson entre les scores à l'item et les scores au test, appelée corrélation item-total. Si l'item est corrigé de manière dichotomique, celle-ci est appelée corrélation bisériale [Bertrand & Blais 2004].

C'est pourquoi, afin de faire une analyse adéquate au test et remédier à tous les problèmes rencontrés lors de la théorie classique de l'item, les chercheurs ont créé une nouvelle théorie appelée *la Théorie des Réponses aux Items* (TRI).

## 2.2. Généralités

La théorie des réponses aux items a vu le jour grâce aux différents travaux faits par plusieurs chercheurs au cours de ces cinq dernières décennies. Notamment Lawley, Lord, Birnbaum, Rasch, Wright et plusieurs autres [Hambleton & Swaminathan 1985].

Cette théorie est conçue afin de régler les faiblesses de *la théorie classique de l'item*. Ainsi, au lieu de s'intéresser au test au complet et aux scores obtenus dans ce dernier pour en faire une analyse, la théorie des réponses aux items prend en considération l'item et plus précisément la réponse à l'item, comme son nom l'indique.

D'après [Lord & Novick 1968], la performance d'un candidat dans un test dépend directement de son niveau d'habileté [Frederiksen *et al.* 1993]. C'est la raison pour laquelle un modèle des réponses aux items est conçu, il détermine une relation entre le niveau de compétence d'un candidat et sa réponse à l'item. C'est un modèle mathématique capable de déterminer la probabilité qu'un candidat donné réponde correctement à un item donné, soit une fonction de son niveau de compétence et les caractéristiques de l'item [Frederiksen *et al.* 1993]. Il est appelé généralement *la fonction caractéristique de l'item* ou *Courbe Caractéristique de l'Item* (CCI), sur laquelle est fondée la théorie des réponses aux items.

Cette théorie a pour avantages [Hambleton & Swaminathan 1985] :

(a) Une estimation du niveau d'habileté du candidat indépendamment du choix des items du test. (b) Une estimation des paramètres d'un item indépendamment du groupe des candidats passant le test. (c) Une précision dans l'estimation du niveau d'habileté qui diffère selon le niveau d'habileté. Ainsi, elle a fait en sorte de résoudre plusieurs problèmes dont les solutions étaient difficiles à trouver par la théorie classique de l'item. C'est pourquoi plusieurs chercheurs l'ont considérée comme un terrain fertile pour des nouvelles applications, notamment :

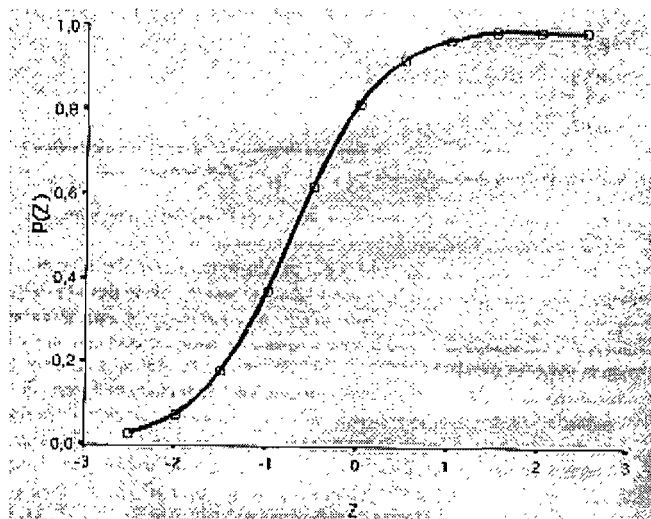
- Le *test adaptatif* [Lord 1980; Weiss 1984], en sélectionnant le meilleur item à administrer au candidat selon la quantité d'informations fournie par différents items, et en estimant son niveau d'habileté provisoire selon la réponse du candidat à l'item.
- L'évaluation à large échelle en éducation [Bock *et al.* 1982; Choppin 1976; Messick *et al.* 1983], en mesurant la compétence d'une population au lieu de celle d'une personne afin d'évaluer un programme ou les tendances d'un formateur.

Remarque : Nous ne parlons ici que du modèle dit unidimensionnel, celui avec un seul paramètre d'habileté, soit le niveau d'habileté.

### 2.3. Courbe caractéristique de l'item

C'est une fonction mathématique qui relie le niveau d'habileté d'un candidat, mesuré lors d'un test, avec la probabilité que ce candidat répond correctement à un item du test. De façon plus simple, c'est une fonction régressive non linéaire de la performance (réussite de l'item) dans différents niveaux d'habileté mesurés par le test [Hambleton & Swaminathan 1985]. Le niveau d'habileté est souvent représenté par la lettre grecque  $\theta$ . Il est supposé être distribué de façon normale au sein de la population, ainsi il est défini entre  $-\infty$  et  $+\infty$ .

D'après [Bertrand & Blais 04], la CCI prend souvent la forme d'un « S », et cela apparaît évident dans la figure 2.1 où la représentation des proportions de réussite d'un item donné, retrouvées à partir des scores totaux centrés (moyen=0) et réduits (écart-type=1) d'un test de mathématique, a la forme d'un « S ». Ainsi on constate que, par exemple, 40% des candidats dont le score est de l'ordre de  $z = -1$  ont répondu correctement à cet item. Alors que ceux dont le score est  $z = 1$ , environ 80% d'eux ont donné une bonne réponse à ce même item.



**Figure 2.1** : Proportions de réussite d'un item calculées à partir des scores totaux centrés et réduits d'un test de mathématique [Bertrand & Blais 2004]

Ainsi Lord en 1952 a proposé un modèle qui illustre bien la forme de « S », le *modèle de l'ogive normale* dont la formule algébrique est la suivante :

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

(Formule 2.1)

$P_i(\theta)$  Est la probabilité qu'un candidat donné, avec un niveau d'habileté  $\theta$ , réponde correctement à un item  $i$ .

$a_i$  et  $b_i$  sont les paramètres qui caractérisent l'item  $i$ .

$z$  représente la déviation d'une distribution normale avec une moyenne de  $b_i$  et une déviation standard de  $1/a_i$ .

Birnbaum (1957) a simplifié cette formule par *un modèle logistique* dont la formule est la suivante :

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

(Formule 2.2)

$D$  est une constante qui est égale à 1.7 afin de se rapprocher du modèle de l'ogive normale.

$a_i$  est le paramètre de discrimination de l'item.

$b_i$  est le paramètre de difficulté de l'item.

Le choix de  $D=1.7$  est pris de sorte que la différence entre les valeurs de  $P_i(\theta)$  dans le modèle de l'ogive normale et celles obtenues par le modèle logistique, soit de moins de 1% pour tous les valeurs de  $\theta$  [Hambleton & Swaminathan, 1985].

Ce modèle logistique a donné naissance à trois modèles principaux selon le nombre de paramètres de l'item utilisés.

### 2.3.1. Modèle logistique à un paramètre

Rasch a travaillé séparément [Rasch, 1966], indépendamment des autres modèles de la théorie des réponses aux items, et il a conçu un nouveau modèle qui est considéré par la suite comme un cas particulier de celui de Birnbaum. Ce modèle prend en considération un seul paramètre de l'item, celui de difficulté, ainsi tous les items ont une puissance de discrimination égale et minimale.

L'équation d'un modèle logistique à un paramètre pour la CCI est la suivante :

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

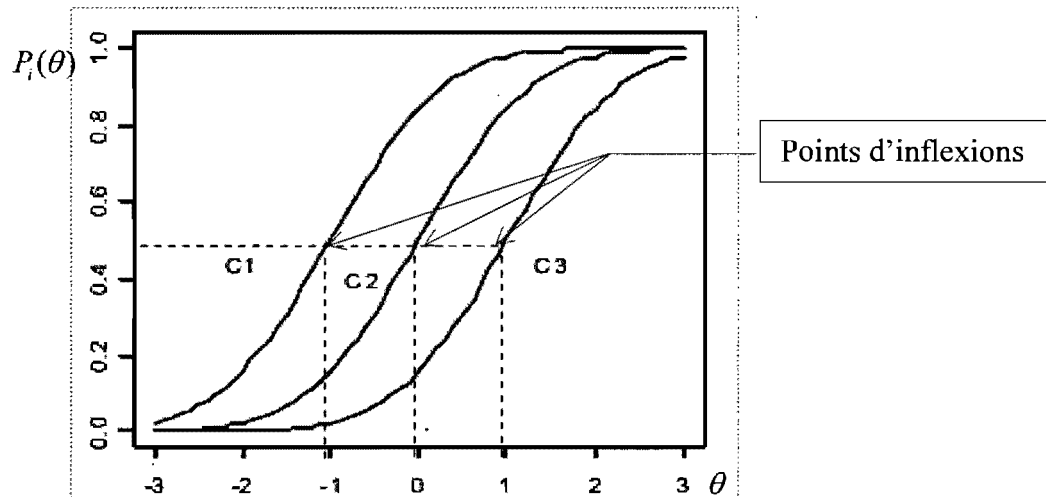
(Formule 2.3)

$D=1.7$

$b_i$  est le paramètre de difficulté de l'item.

On constate que le paramètre de discrimination de l'item  $a_i=1$ .

Il existe d'autres modèles logistiques à un paramètre dont le paramètre de discrimination est constant mais différent de 1, alors que le plus utilisé c'est celui décrit précédemment qui est appelé fréquemment modèle de Rasch.



**Figure 2.2** : CCI de trois items selon le modèle à un paramètre, avec  $b_1=-1$ ,  $b_2=0$  et  $b_3=1$   
[URL2]

Dans la CCI de ce modèle, le paramètre de difficulté est représenté par la valeur du niveau d'habileté au point d'inflexion de la courbe.

La figure 2.2 représente la CCI de trois différents items, on constate que la pente des trois courbes reste la même, et plus l'item devient difficile plus la probabilité de réussir l'item diminue. Ainsi plus  $b_i$  augmente plus la courbe évolue à droite.

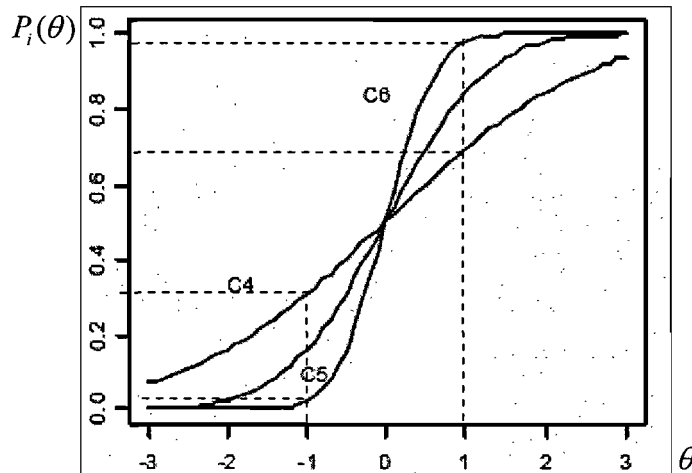
Cette courbe peut être interprétée d'une autre manière : c'est la représentation d'un ensemble d'items possédant le même paramètre de difficulté,  $P(\theta)$  est la portion d'items réussis par un candidat dont le niveau d'habileté est  $\theta$ .

Dans plusieurs applications de la TRI, le modèle à un seul paramètre reste très limité puisque il ne prend pas en considération les courbes avec différentes pentes, d'où la nécessité d'ajouter un autre paramètre, celui de discrimination [Wainer 2000].

### 2.3.2. Modèle à deux paramètres

Le modèle de Birnbaum déjà cité précédemment représente le modèle à deux paramètres. Il prend en considération le paramètre de difficulté de l'item et celui de discrimination. Une représentation graphique de ce modèle est montrée dans la figure 2.3.





**Figure 2.3 :** CCI de trois items selon le modèle à deux paramètres avec  $b_4=b_5=b_6=0$ ,  $a_4=0.5$ ,  $a_5=1$  et  $a_6=2$  [URL2].

Nous constatons que plus le paramètre de discrimination augmente, plus la pente de la CCI devient grande aux environs du point d'inflexion. Ainsi, l'item 6 est plus discriminatoire que l'item 4, puisque la différence entre  $P_6(1)$  et  $P_6(-1)$  est largement supérieure par rapport à celle de  $P_4(1)$  et  $P_4(-1)$ .

Dans ces deux derniers modèles, l'asymptote inférieure des CCI tend toujours vers 0, ce qui signifie qu'avec un niveau d'habileté faible, le candidat n'a aucune chance de réussir l'item. Ce qui n'est pas toujours le cas, car dans un questionnaire à choix multiples, le candidat a une chance de répondre correctement à l'item, même s'il a un niveau d'habileté faible. Ainsi, [Lord 1974] a ajouté un troisième paramètre qui définit la probabilité minimale avec laquelle un candidat réussit l'item.

### 2.3.3. Modèle à trois paramètres

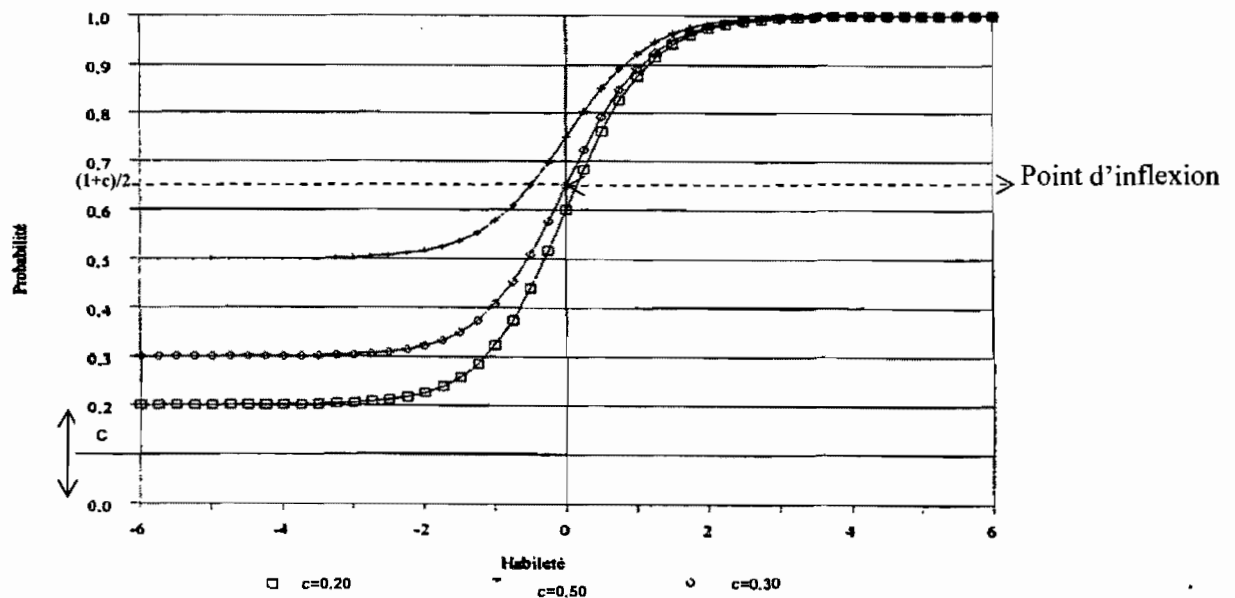
En prenant en considération le fait qu'un candidat avec un niveau d'habileté faible peut réussir l'item, un nouveau paramètre s'ajoutera aux deux premiers, le paramètre  $c$  de *pseudo chance*, ainsi l'équation de ce modèle prend la forme suivante :

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

(Formule 2.4)

 $D=1.7$  $a_i$  est le paramètre de discrimination de l'item. $b_i$  est le paramètre de difficulté de l'item. $c_i$  est le paramètre de pseudo chance de l'item.

Une représentation graphique de ce module est montrée dans la figure ci-dessous.



**Figure 2.4** : CCI de trois items selon le modèle à trois paramètres ( $a=1$ ,  $b=0$  et  $D=1.7$ ).

[Raïche 2000]

Remarque : dans le modèle à un seul paramètre comme dans le modèle à deux paramètres, le point d'inflexion de la CCI  $i$  a pour coordonnées  $[b_i, 0.5]$ . Alors que dans le modèle à trois paramètres, l'abscisse reste le même,  $b_i$  le paramètre de difficulté de l'item, et l'ordonnée devient  $(1+c_i)/2$ .

Les trois modèles de la théorie des réponses aux items représentent, via la CCI, la relation qui existe entre le niveau d'habileté d'un candidat et les paramètres de l'item, tout en supposant que ces deux facteurs sont déjà connus, alors que ce n'est pas toujours le cas. Ainsi plusieurs techniques ont été utilisées afin d'en faire une estimation plus adéquate. Nous allons citer ci-dessous celles les plus connues.

## 2.4. Estimation du niveau d'habileté

Dans ce qui suit, nous supposons que les paramètres de l'item sont déjà connus, et nous désirons estimer le niveau d'habileté pour un candidat donné. Plusieurs méthodes ont été utilisées jusqu'à présent, parmi lesquelles nous citons quelques unes.

### 2.4.1. Méthode de vraisemblance maximale

D'après [Hambleton & Swaminathan 1985] et [Bertrand & Blais 2004],  $P(U_i / \theta)$  est la probabilité qu'un candidat de niveau d'habileté  $\theta$  réponde à un item  $i$  par une réponse  $U_i$ . Dans notre cas celle-ci est binomiale, c'est-à-dire  $U_i=0$  si la réponse est fautive et  $U_i=1$  sinon.

Soit  $P(U_i / \theta) = P_i^{U_i} Q_i^{1-U_i}$  avec  $P_i = P_i(\theta) = P(U_i / \theta)$  et  $Q_i = 1 - P_i$ .

Ainsi la probabilité qu'un individu de niveau d'habileté  $\theta$  réponde à  $n$  items par  $U_1, U_2, \dots, U_n$  selon le principe de l'indépendance locale est :

$$P(U_1, U_2, \dots, U_n / \theta) = \prod_{i=1}^n P_i^{U_i} Q_i^{1-U_i}.$$

Une fois que le candidat réponde à ces items par les valeurs  $u_1, u_2, \dots, u_n$ ; une nouvelle fonction est obtenue, on la nomme, fonction de vraisemblance. Elle est définie comme suit :

$$L(u_1, u_2, \dots, u_n / \theta) = \prod_{i=1}^n P_i^{u_i} Q_i^{1-u_i}.$$

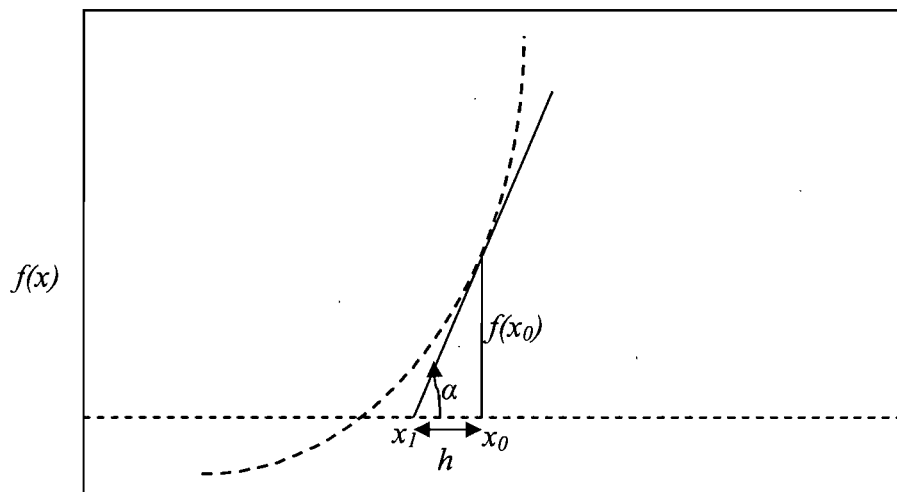
Ainsi chercher une bonne estimation de  $\theta$  revient à trouver sa valeur pour laquelle la fonction de vraisemblance est maximale. Maximiser  $L$  revient à maximiser son logarithme naturel, c'est à dire :

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

Nous cherchons alors la valeur de  $\theta$  où  $\frac{\partial \ln L}{\partial \theta} = 0$ . Afin de résoudre cette équation, nous utilisons la procédure de Newton-Raphson qui est la plus utilisée dans ce cas ci. Cette procédure est définie comme suit :

Nous supposons que nous cherchons  $x$  tel que  $f(x) = 0$  (voir la figure 2.5)<sup>3</sup>.  $x_0$  est une solution approximative.  $x_1$  est une solution plus précise tel que  $x_1 = x_0 - h$ , avec  $h = \frac{f(x_0)}{\tan \alpha}$ ,  $\tan \alpha$  est la pente de la fonction  $f(x)$  au point  $x_0$ ,  $\tan \alpha = f'(x_0)$  la dérivée de la fonction au point  $x_0$ . Ainsi  $h = \frac{f(x_0)}{f'(x_0)}$ .

De même si  $x_m$  est une solution approximative à la  $m$ -ième itération, une solution plus précise est donnée par  $x_{m+1} = x_m - \frac{f(x_m)}{f'(x_m)}$ . Ce processus est répété jusqu'à ce que  $x_{m+1} - x_m$  atteint une valeur fixe et petite,  $x_{m+1}$  est alors la solution de  $f(x) = 0$ .



**Figure 2.5 :** Illustration de la méthode de Newton-Raphson. [Hambleton & Swaminathan 1985]

<sup>3</sup> La courbe représentée dans la figure 2.5 est celle d'une fonction convexe. Notons que la méthode de Newton-Raphson peut être aussi illustrée avec une fonction concave.

Ainsi l'estimation du maximum de vraisemblance  $\hat{\theta}$  de  $\theta$  est obtenue lorsque la convergence entre  $\theta_{m+1}$  et  $\theta_m$  prend place, avec

$$\theta_{m+1} = \theta_m - \left[ \frac{d}{d\theta} \ln L(u | \theta) \right]_m / \left[ \frac{d^2}{d\theta^2} \ln L(u | \theta) \right]_m \text{ à l'itération } m.$$

Dans un modèle logistique à un paramètre cette formule devient :

$$\theta_{m+1} - \theta_m = \frac{r - \sum_{i=1}^n P_i(\theta_m)}{-D \sum_{i=1}^n P_i(\theta_m) Q_i(\theta_m)} \quad (\text{Formule 2.5})$$

$$r = \sum_{i=1}^n u_i$$

Dans un modèle logistique à deux et trois paramètres, l'équation de vraisemblance  $\frac{\partial \ln L}{\partial \theta} = 0$  est exprimée de manière générale dans la formule suivante :

$$\frac{\partial \ln L}{\partial \theta} = \sum_{i=1}^n k_i u_i - \sum_{i=1}^n k_i P_i = 0 \quad (\text{Formule 2.6})$$

La valeur de  $k_i$  pour le modèle logistique à un, deux et trois paramètres est respectivement :

$$k_i = D$$

$$k_i = Da_i$$

$$k_i = Da_i(P_i - c_i) / P_i(1 - c_i)$$

La méthode de vraisemblance maximale donne de bons résultats sur un modèle logistique à un paramètre avec un nombre d'items supérieur à vingt. Mais elle nécessite un grand nombre d'items pour donner de bon résultat, pour les modèles à deux et trois paramètres. Par exemple, dans un modèle logistique à trois paramètres avec moins de vingt items, la fonction de vraisemblance possède plusieurs maximums. De plus, si un

candidat répond correctement ou incorrectement à *tous* les items,  $\hat{\theta}$  tend vers l'infini comme montré dans ce qui suit.

Revenons à la formule 2.6 précédente. Comme le paramètre de discrimination  $a_i$  est toujours positif, et  $P_i - c_i$  est aussi positif ( $c_i$  est l'asymptote inférieure de  $P_i$ ), alors  $k_i$  est toujours positif dans un modèle logistique à trois paramètres. Ainsi, si un candidat répond incorrectement à tous les items, la formule 2.6 devient :

$$\sum_{i=1}^n k_i P_i = 0$$

Puisque  $k_i$  est positif et  $P_i$  est la probabilité d'avoir une bonne réponse à l'item  $i$ , alors cette équation n'est satisfaite que si  $\theta = -\infty$ . De même si un candidat répond correctement à tous les items, l'équation de vraisemblance devient :

$$\sum_{i=1}^n k_i = \sum_{i=1}^n k_i P_i \text{ est satisfaite seulement pour } \theta = +\infty.$$

Comme nous avons vu, la méthode de vraisemblance maximale connaît beaucoup de problèmes. Pour y remédier, la méthode *bayésienne de Maximisation À Posteriori* (MAP) a été suggérée.

#### 2.4.2. La méthode bayésienne de maximisation à posteriori (MAP)

Elle se base sur le théorème de Bayes, à partir duquel nous retrouvons la formule suivante :

$$P(\theta_a|u) = P(u|\theta_a)P(\theta_a)/P(u)$$

Avec  $\theta_a$  est le niveau d'habileté d'un candidat  $a$  et  $u$  est l'ensemble de ses réponses.

Puisque  $P(u)$  est constant, cette dernière formule sera écrite d'une autre manière, appelée formule de probabilité à posteriori :

$$f(\theta_a|u) \propto L(u|\theta_a)f(\theta_a)$$

$f(\theta_a|u)$  est la fonction de probabilité à posteriori.

$L(u|\theta_a)$  est la fonction de vraisemblance.

$f(\theta_a)$  est la fonction de probabilité à priori.

(Formule 2.7)

D'après [Hambleton & Swaminathan, 1985],  $f(\theta_a)$  est presque équivalente à  $\exp(-\frac{1}{2}\theta_a^2)$ .

De la même manière que dans l'estimation du maximum de vraisemblance, estimer  $\theta$  revient à trouver sa valeur pour un  $f(\theta_a|u)$  maximal, tout en utilisant la procédure de Newton- Raphson.

L'avantage de cette méthode est qu'elle permet l'estimation du niveau d'habileté pour un score parfait de 0 ou de 1, voir [Hambleton & Swaminathan, 1985]. Mais puisque la procédure de calcul reste la même que celle utilisée dans la méthode de maximum de vraisemblance, la plupart des limites de cette dernière sont transférées à la méthode bayésienne de maximisation à posteriori. Ainsi pour un modèle à trois paramètres, il est nécessaire d'avoir un grand nombre d'items. Si ce dernier est inférieur à 20, il est fort probable de retrouver plus qu'un maximum, la solution de l'estimation du niveau d'habileté n'est plus alors unique [Raïche 2000].

### 2.4.3. Méthode de l'espérance à posteriori (EAP)

La méthode bayésienne de maximisation à posteriori a été améliorée en utilisant l'espérance mathématique moyenne sur une distribution à priori obtenue à partir des données [Raïche 2000]. L'estimation du niveau d'habileté obtenu par la méthode de l'espérance à posteriori est :

$$E(\theta | u_1, u_2, \dots, u_n) = \frac{\sum_{j=1}^k \theta_j L(u_1, u_2, \dots, u_n | \theta_j) f(\theta_j)}{\sum_{j=1}^k L(u_1, u_2, \dots, u_n | \theta_j) f(\theta_j)} \quad (\text{Formule 2.8})$$

$\theta_j$  est un des  $k$  points de quadrature.

$f(\theta_j)$  est le poids de quadrature associé à  $\theta_j$ , c'est la densité à priori.

Et l'erreur type  $S_{EAP}$  de l'estimation du niveau d'habileté par la méthode de l'Espérance À Posteriori est le suivant :

$$S_{EAP}(\theta | u_1, u_2, \dots, u_n) = \left[ \frac{\sum_{j=1}^k (\theta_j - E(\theta))^2 L(u_1, u_2, \dots, u_n | \theta_j) f(\theta_j)}{\sum_{j=1}^k L(u_1, u_2, \dots, u_n | \theta_j) f(\theta_j)} \right]^{1/2}$$

Selon [Ayala 1995], pour retrouver  $f(\theta_j)$ , nous divisons l'intervalle de propagation de  $\theta$  (exemple -4 à 4) en  $k$  points équidistants (ces points sont les  $\theta_j$ , point de quadrature), puis nous calculons la probabilité de densité selon la loi de distribution normale pour chaque point  $\theta_j$ , et nous la multiplions par la différence entre deux points de quadrature consécutifs ( $\theta_j - \theta_{j-1}$ ), nous obtenons ainsi le poids de quadrature  $f(\theta_j)$ , tout en tenant compte du fait que  $\sum_{j=1}^k f(\theta_j) = 1$ .

Cette technique a beaucoup d'avantages, notamment le calcul est facile, puisque les itérations ne sont plus utilisées. Elle donne aussi de bons résultats même avec un nombre petit d'items et même si les réponses sont toutes correctes ou incorrectes, ainsi l'estimateur du niveau d'habileté est plus efficace. C'est la raison pour laquelle cette méthode est la plus recommandée dans un test adaptatif.



## 2.5. Estimation des paramètres de l'item

Après avoir estimé le niveau d'habileté, tout en considérant les paramètres de l'item connus. Supposons maintenant que ces derniers sont inconnus et que nous cherchons à faire une estimation de leurs valeurs. L'une des méthodes utilisée est *l'estimation de vraisemblance marginale maximale*.

### 2.5.1. L'estimation de vraisemblance marginale maximale (MML)

D'après ce qu'on a vu précédemment, la probabilité pour qu'un candidat obtienne un vecteur de réponses  $U$  est:

$$P(U/\theta) = \prod_{i=1}^n P_i^{U_i} Q_i^{1-U_i}.$$

Ainsi pour une population de candidats avec une distribution continue des niveaux d'habileté  $g(\theta)$ , cette équation devient :

$$P(U/\theta) = \int_{-\infty}^{+\infty} \prod_{i=1}^n P_i^{U_i} Q_i^{1-U_i} g(\theta) d\theta = \pi_u$$

Puisque pour  $n$  items binaires il existe  $2^n$  patrons de réponses, et si on suppose que  $r_u$  est le nombre de candidats dont le patron de réponses est  $u$ , alors la fonction de vraisemblance prend la forme suivante :

$$L \propto \prod_{u=1}^{2^n} \pi_u^{r_u} \Rightarrow \ln L = k + r_u \sum_{u=1}^{2^n} \ln \pi_u \quad k \text{ est une constante.}$$

D'où l'estimateur de vraisemblance marginale maximale des paramètres d'un item  $i$  est obtenu en résolvant l'équation suivante :

$$\frac{\partial \ln L}{\partial \phi} = 0 \quad \text{Où} \quad \phi = a_i, b_i \text{ ou } c_i$$

Avec  $a_i$  est le paramètre de discrimination,  $b_i$  est le paramètre de difficulté et  $c_i$  est celui de pseudo-chance.

Cette équation est résolue de la même manière que l'estimateur du niveau d'habileté par la méthode de vraisemblance maximale. C'est-à-dire par la procédure de Newton-Raphson. Pour en savoir plus, voir [Bock & Aikin 1981].

D'autres techniques sont utilisées pour estimer les paramètres de l'item, notamment celle du maximum de vraisemblance conditionnelle (*Conditional Maximum Likelihood*, CML), mais cette dernière n'est utilisée qu'avec un modèle à un seul paramètre. Ainsi la procédure de MML reste la plus utilisée, et la plus efficace pour un modèle à trois paramètres. On la retrouve dans plusieurs logiciels de l'estimation des paramètres de l'item comme : TESTFACT, BILOG-MG, MULTILOG et PARSCALE développés par SSI (Scientific Software International) [URL3], des outils d'analyse pour différent type de tests et d'items, ainsi que PRAM-3PL (voir la section 5.3.1).

Dans un test adaptatif, en plus de connaître les paramètres de l'item, il est primordial de savoir comment choisir l'item adéquat à présenter au candidat. De même, il faut connaître la précision avec laquelle le niveau d'habileté est estimé. Ainsi il est nécessaire d'ajouter une nouvelle notion : *la fonction d'information*.

## 2.6. Erreur type et fonction d'information d'item et de test

Cette fonction est particulièrement utile pour la construction et la comparaison des tests, ainsi que la sélection d'un item à présenter au candidat lors d'un test adaptatif. Elle est déduite à partir de l'estimateur du maximum de vraisemblance  $\hat{\theta}$  (le niveau d'habileté estimé du candidat) de  $\theta$  (le niveau d'habileté réel du candidat) qu'on a vu précédemment (voir la section 2.4.1). La distribution de l'estimateur est supposée normale asymptotiquement, avec une moyenne de  $\theta$  et un écart type de  $\sigma(\hat{\theta}|\theta)$  appelé aussi *erreur type* ou *précision*. Ainsi, la fonction d'information pour le niveau d'habileté  $\theta$  est définie par :

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

Autrement, cette fonction est maximale au voisinage d'un niveau d'habileté  $\theta$  si et seulement si la différence entre ce dernier et celui estimé  $\hat{\theta}$  est minimale, soit l'erreur type est petite et l'habileté  $\theta$  est connue de manière plus précise.

Cette formule est traduite par Lord (1980), pour un item  $i$ , avec un niveau d'habileté fixe  $\theta$ , en une fonction de la pente  $P'_i(\theta)$  de la courbe caractéristique de l'item (CCI) à  $\theta$ :

$$I_i(\theta) = \frac{P'_i(\theta)^2}{P_i(\theta)Q_i(\theta)}$$

$P'_i(\theta)$  est la dérivée première de  $P_i(\theta)$ .

Dans un modèle à trois paramètres cette dernière équation devient :

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

(Formule 2.9)

$\theta$  est le niveau d'habileté estimé.

$a_i$  est le paramètre de discrimination de l'item.

$c_i$  est le paramètre de pseudo chance de l'item.

$P_i(\theta)$  est la probabilité de réussite de l'item, indiqué dans la formule 2.4

Ainsi plus l'item est discriminatoire, plus sa fonction d'information augmente. Et plus la valeur de son paramètre de pseudo chance diminue, elle devient de plus en plus grande.

Pour un test à  $n$  items, avec  $\theta$  toujours fixe, la fonction d'information prend la forme suivante :

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

Or nous savons que dans la CCI, la pente est maximale au point d'inflexion dont les coordonnées sont :  $(b_i, (1+c_i)/2)$ , par conséquent la fonction d'information d'un item est maximale au voisinage de  $b_i$ , c'est-à-dire au point :

$$\theta_{\max} = b_i + \frac{1}{Da_i} \ln \left[ \frac{1}{2} + \frac{1}{2} \sqrt{1 + 8c_i} \right]$$

De plus (Lord, 1980) a montré que le maximum de la fonction de l'information d'un item  $i$  dans un modèle à trois paramètres est le suivant :

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

Cette dernière devient  $I_i(\theta)_{\max} = \frac{D^2 a_i^2}{4}$  pour un modèle à deux paramètres, et une constante de valeur  $I_i(\theta)_{\max} = \frac{D^2}{4}$  pour un modèle à un seul paramètre [Bertrand & Blais 2004].

Cette fonction est beaucoup utilisée dans un test adaptatif, notamment lors de la procédure de la sélection de l'item (nous l'expliquons en détail dans le chapitre 3).

## 2.7. Autres modèles

Dans tout ce qui est cité auparavant, nous ne tenons compte que des modèles dichotomiques (deux sorte de réponse, 1 pour une bonne réponse et 0 pour une mauvaise réponse) et unidimensionnels (un seul niveau d'habileté). Alors qu'ils en existent d'autres, notamment les *modèles polytomiques* qui prennent en considération les différentes options de réponses des items comme les tests à choix multiple<sup>4</sup>, et Les *modèles multidimensionnels* en optant pour plusieurs niveaux d'habileté.

### 2.7.1. Modèle polytomique

Dans les modèles à un, deux et trois paramètres que nous avons vu précédemment; la réponse à l'item prend deux sortes de valeurs fixes : soit 1 pour une bonne réponse, soit 0 pour une mauvaise réponse. Alors que parfois il est nécessaire d'étudier le cas d'un test dont les réponses aux items prennent plusieurs options. C'est ce qu'ont fait plusieurs chercheurs, en créant ainsi différents modèles polytomiques. Citons par exemple :

- *Le modèle nominal* de Bock (1972) qui prend en considération la réponse à choix multiples dont les différentes options constituent une échelle nominale. Dans ce modèle, une courbe caractéristique est produite pour chaque option,

---

<sup>4</sup> Les tests à choix multiple sont des tests dont les items possèdent plusieurs types de réponses dépendamment de la performance du candidat à l'item.

permettant ainsi d'analyser le comportement de chaque choix de réponse [Bertrand & Blais 2004].

- *Le modèle gradué* de Samejima (1969) : c'est une généralisation du modèle dichotomique à deux paramètres (section 2.3), tout en prenant en considération pour un item donné, un seul paramètre de discrimination, et plusieurs paramètres de difficulté selon le nombre de choix de réponses [Bertrand & Blais 2004].
- *Le modèle d'échelle d'évaluation (rating scale)* d'Andrich (1978), idéal pour traiter les échelles de type Likert<sup>5</sup>. il peut être considéré comme un cas général du modèle dichotomique à un seul paramètre (voir la section 2.3), où chaque item possède plusieurs paramètres de difficulté, variant selon le nombre des options de ses réponses [Andrich 1978].
- *Modèle à crédit partiel* de Masters (1982) : il est aussi membre de la famille des modèles dichotomiques à un seul paramètre, comme celui cité ci-dessus. Sauf qu'ici les réponses de l'item ont des valeurs qui croissent selon leur exactitude [Bertrand & Blais 2004]. Une extension de ce dernier est faite par Muraki (1992), c'est le modèle à *crédit partiel généralisé* [Van Rijn *et al.* 2002].

### 2.7.2. Modèle multidimensionnel

Parfois il est nécessaire d'évaluer plusieurs niveaux d'habileté d'un candidat. Par exemple dans le cas d'un test de mathématique dont les items sont des problèmes sous forme d'un récit écrit, le candidat est évalué dans sa compétence en mathématique et son niveau vocabulaire. D'où l'utilité de considérer un modèle avec différentes compétences. C'est le modèle multidimensionnel. Il est défini, par Reckase (1997), comme une généralisation du modèle unidimensionnel à trois paramètres, tout en considérant le paramètre de difficulté et celui de pseudo-chance comme ceux de l'item. Alors que celui de discrimination est propre à chaque compétence [Marvelde *et al.* 2006].

---

<sup>5</sup> Échelle de Likert est un questionnaire psychologique en 5 ou 7 niveaux pour exprimer le degré de satisfaction.

Un autre modèle multidimensionnel est proposé par Embreston (1985), qui est considéré comme une généralisation du modèle unidimensionnel à un seul paramètre, appelé modèle de trait latent à plusieurs composantes (*multicomponent latent trait model*) MLTM. Dans un tel modèle on suppose que la réussite d'un candidat à une tâche donnée (item total) dépend de la difficulté des différentes composantes de cette tâche ainsi que de l'habileté du candidat à résoudre chacune des composantes de la tâche. Un exemple est l'item à deux composantes suivant :

Chat : Tigre : Chien : \_\_\_\_\_

a) Lion b) Loup c) Aboiement d) Chiot e) Cheval

La première composante est d'indiquer la règle sous-jacente à l'appariement des animaux cités. La deuxième composante est de compléter l'item d'analogie, une fois la règle connue. Chacune de ces deux composantes devrait normalement être réussies pour que l'item total le soit.

Contrairement au modèle précédent, chaque item possède autant de paramètres de difficulté que ses composantes [Bertrand & Blais 2004].

L'application d'un tel modèle dans les tests est souvent limitée par des problèmes statistiques rencontrés lors de son analyse, et les difficultés de l'interprétation de ses paramètres [Raymond *et al.* 1997].

## 2.8. Conclusion

Les modèles cités précédemment sont tous capables de nous élaborer un test adaptatif. Mais nous nous sommes surtout attardés sur le modèle dichotomique unidimensionnel car c'est celui qui nous intéresse ici. Ainsi, nous constatons que le point fort de cette théorie est la fonction caractéristique de l'item. Cette dernière définit une relation entre la probabilité qu'un candidat réussisse un item et ses paramètres. Elle est la base de toutes autres formules de la théorie des réponses aux items, notamment celle qui donne une estimation du niveau d'habileté du candidat, de son erreur type ainsi que la fonction d'information de l'item. Ces trois notions, comme nous verrons plus tard, sont la base même d'un test adaptatif par ordinateur utilisant la théorie des réponses aux items.

## Chapitre 3 : Test Adaptatif par Ordinateur (TAO)

Comme son nom le décrit, un test adaptatif est un test adapté au candidat selon son niveau d'habileté. D'après [Salcedo *et al.* 2005] « l'idée de base d'un test adaptatif c'est d'imiter de façon judicieuse le comportement d'un examinateur ». Ainsi, une fois qu'un candidat répond correctement à un item (une question), une autre plus difficile lui sera présentée. Si le candidat échoue, c'est celle qui est la moins difficile qui lui sera présentée. Par conséquent, un candidat ne peut être évalué, comme dans un test ordinaire, par le nombre des réponses correctes; mais c'est une estimation de son niveau d'habileté qui va être retenue à la fin du test.

Dans un test papier-crayon traditionnel, tous les candidats doivent répondre aux même nombre d'items. Alors que dans au test adaptatif, chacun d'eux reçoit un test différent, de longueur inégale avec un nombre d'items différent. Ainsi le test devient plus court, avec une estimation efficace du niveau d'habileté du candidat, et un résultat immédiat [Meijer & Nering 1999].

Les premiers principes du test adaptatif sont utilisés dans un test qui calcule l'âge mental de l'enfant en 1990, appelé le test d'intelligence de Binet [Raîche 2000]. Dans ce dernier, le rôle de l'ordinateur est joué par un administrateur qui choisit les items à présenter à l'enfant dépendamment de sa réponse. Avec l'apparition de la TRI (Théorie des Réponses aux Item) vers les années 70, qui est devenue la base mathématique des tests adaptatifs, et l'apparition de l'ordinateur qui a facilité les calculs; plusieurs travaux et applications ont été développés dans ce domaine. Il est même resté à nos jours, un terrain fertile pour d'autres recherches.

Comme nous avons vu précédemment, la TRI était le fondement même d'un TAO et la seule technique qui était utilisée dans ce domaine. Avec l'apparition des systèmes tutoriels intelligents, dont la modélisation de l'étudiant est considérée comme une de ses principales composantes, plusieurs approches ont vu le jour, notamment, *l'approche de modélisation Bayésienne* qui est devenue de plus en plus utilisée pour modéliser l'habileté d'un candidat en se basant sur les réponses de ce dernier aux items dans un test. Cette méthode est fondée spécialement sur les réseaux Bayésiens, puisque

un graphe Bayésien est construit afin de modéliser le niveau d'habileté du candidat. Plusieurs chercheurs ont développé cette technique et ils y ont trouvé de bons résultats. Notamment Millàn [Millàn & Pérez-de-la-Cruz 2002], Vomlel [Vomlel 2004] et Desmarais [Desmarais & Pu 2005] avec son modèle POKS.

Dans ce chapitre, nous allons voir ces deux différentes techniques, en insistant sur le TAO avec la TRI, ainsi nous expliquons toutes les étapes nécessaires pour une construction efficace de ce dernier. Notamment, la banque d'items, les règles de calculs basées sur la TRI et finalement son processus de déroulement.

### **3.1. Test adaptatif par ordinateur basé sur une modélisation Bayésienne.**

Avant de retourner au TAO avec la TRI, nous décrivons ici l'apport d'une modélisation Bayésienne dans un test adaptatif par ordinateur.

Au début du 21ème siècle, plusieurs chercheurs ont introduit l'approche Bayésienne dans le domaine de la modélisation et de l'évaluation de l'étudiant. Ainsi un TAO est créé à partir d'une fonction de probabilité calculée pour un événement donné, en utilisant des informations tirées des occurrences d'un ou plusieurs tests passés précédemment. C'est la probabilité conditionnelle qu'un candidat réponde correctement à un item étant donné les réponses aux items précédents, en utilisant des échantillons contenant ces mêmes patrons de réponses. Or dans le cas où les réponses sont dichotomiques, retrouver cette probabilité exige le calcul de  $2^k$  probabilités conditionnelles, où  $k$  est le nombre des items posés précédemment. Ainsi avec le nombre des items qui augmente au cours du test, les probabilités conditionnelles augmentent de manière exponentielle. Afin d'éviter ce problème, seules les probabilités conditionnelles pertinentes sont retenues.

Plusieurs chercheurs ont proposé d'utiliser ce modèle dans le test adaptatif. Notamment [Collins *et al.* 1996] qui a montré son utilité pour propager une connaissance le long d'une hiérarchie granulée faite sur des cours, et construire ainsi un



test adaptatif par ordinateur capable de mesurer plusieurs traits latents. La structure du réseau Bayésien est décrite ci-dessous.

### 3.1.1. La structure d'un réseau Bayésien

Un réseau Bayésien est un graphe hiérarchique où les nœuds représentent des variables et les arcs sont des probabilités de dépendances entre ces variables. Ainsi si  $\{X_1, X_2, \dots, X_n\}$  est l'ensemble de variables du réseau et  $pa(X_i)$  représente l'ensemble des parents de  $X_i$  pour chaque  $i$ ,  $\{P(X_i / pa(X_i)), i = 1, \dots, n\}$  est la distribution des probabilités conditionnelles pour chaque variable étant donné la valeur de ses parents. La probabilité de l'ensemble du réseau en supposant que chaque variable (nœud), étant donné ses parents, est indépendante des autres variables sauf de ses descendants. Elle est définie par :

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i / pa(X_i)) \quad (\text{Formule 3.1})$$

Dans un modèle de l'étudiant, les variables peuvent être des règles, concepts, problèmes, habiletés, etc. indépendamment du domaine. Alors que les relations ont la forme de «*partie de*» ou «*prés-requis de*» etc. Pour bien comprendre ce concept, nous prenons ici comme exemple le modèle Bayésien de l'étudiant proposé par Millan [Millan 2002]. Dans ce dernier les variables sont soit des items, des concepts, des thèmes ou un sujet dont nous cherchons à évaluer la connaissance pour un étudiant donné; ainsi le sujet est composé d'un ensemble de thèmes avec leurs poids respectivement, et de même chaque thème est composé à son tour de plusieurs concepts et de leurs poids. Un poids mesure l'importance du concept (thème) dans le thème (sujet). Les items font partie des concepts et parfois même de plusieurs concepts. Les items et les concepts dans un sujet donné sont déterminés par l'expert du sujet.

Un cas concret de ce modèle est montré dans l'exemple d'un cours d'anglais, qui est considéré comme sujet d'apprentissage. Le poids d'un concept (thème) est mesuré par le pourcentage du temps accordé à son apprentissage par rapport au temps global

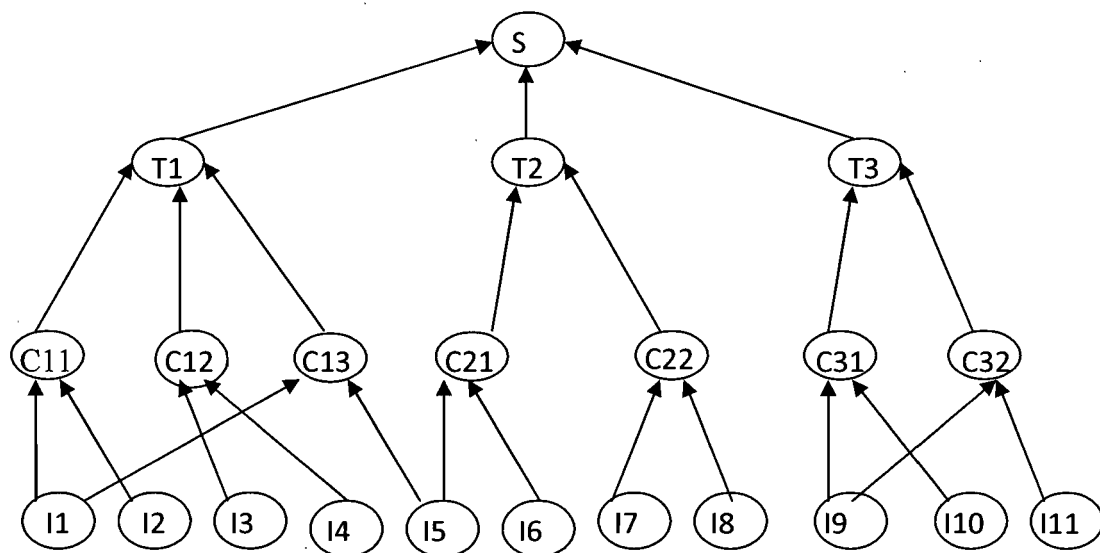
alloué à son thème (sujet). Dans le tableau 3.1 nous trouvons une décomposition de ce sujet avec le temps attribué pour chaque partie, dans le tableau 3.2 la hiérarchie ainsi construite et dans la figure 3.1 le graphe qui en résulte.

**Tableau 3.1** : Une décomposition fictive du cours d'anglais

Sujet	Temps (par mois)	Thèmes	Temps (par mois)	Concepts	Temps (par jours)
Anglais	12	Grammaire	5	Pronoms et possessifs	50
				Les adjectifs	70
				Les auxiliaires modaux	30
		Vocabulaire	4	Prononciation	40
				Verbe irrégulier	80
		Compréhension orale et écrite	3	Compréhension orale et écrite	Compréhension orale
Compréhension écrite	45				

**Tableau 3.2** : Une hiérarchie du sujet l'Anglais

Sujet	Thèmes	Poids	Concepts	Poids
S : Anglais	T1 : Grammaire	0.41	C11 : Pronoms et possessifs	0.33
			C12 : Les adjectifs	0.46
			C13 : Les auxiliaires modaux	0.2
	T2 : Vocabulaire	0.33	C21 : Prononciation	0.33
			C22 : Verbe irrégulier	0.66
	T3 : Compréhension orale et écrite	0.25	C31 : Compréhension orale	0.5
			C32 : Compréhension écrite	0.5



**Figure 3.1** : La structure d'un réseau Bayésien (I1, I2, ..., I11 sont des items)

L'évaluation d'un étudiant dans ce modèle-ci, se déroule de la manière suivante : à partir des réponses de l'étudiant aux items d'un ou de plusieurs concepts, on détermine son niveau de connaissance dans ces derniers. À l'aide des résultats trouvés, des probabilités seront propagées afin de déterminer le niveau atteint par l'étudiant dans chaque thème et par conséquent dans le sujet.

### **3.1.2. Le déroulement d'un test adaptatif par ordinateur dans un réseau bayésien**

Ce qui nous intéresse ici c'est la partie inférieure de la hiérarchie (les concepts et les items) puisqu'on cherche à déterminer le taux de connaissance d'un étudiant pour un ou plusieurs concepts.

Une fois le modèle construit, une estimation initiale des taux de connaissance de l'étudiant pour les concepts est faite. Cette estimation est soit déduite à partir de l'historique de l'étudiant, ou elle a une valeur de 0.5 pour chaque concept. Ainsi un item est sélectionné de l'hiérarchie afin de le présenter à l'étudiant. Cette sélection est faite selon différents critères, une description bien détaillée de cette section est faite dans l'article de [Van Der Linden 1998] et celui de Millan [Millan 2002]. Une fois la réponse de l'étudiant à cet item connue, une mise à jour des probabilités des concepts est faite, et un autre item est sélectionné et présenté à l'étudiant. Ainsi de suite jusqu'à ce qu'un nombre fixe des items soit présenté à l'étudiant ou que tous les concepts sont évalués (un concept est évalué si il est diagnostiqué comme étant connu ou non connu).

L'inconvénient avec cette méthode est qu'elle n'est pas tout à fait automatique, puisque elle nécessite tout le temps l'expertise d'un expert pour la construction du modèle. La rendre totalement automatique est très demandant coté données, puisque l'élaboration, par exemple, d'un petit modèle de dix nœuds exige l'analyse des données empiriques de milliers de cas [Desmarais et Pu 2005].

Un des modèles les plus connus est celui de [Vomlel 2004], ce dernier est construit à partir des données en utilisant un algorithme d'induction des réseaux Bayésiens appelé *Hugin*, ainsi que l'expertise de l'expert en catégorisant les items dans des concepts. Desmarais (1996) [Desmarais & Maluf 1996] a aussi créé un nouveau

modèle automatiquement construit à partir des données appelé POKS (*Partial Order Knowledge Structure*). Ce modèle est décrit ci-dessous.

### 3.1.3. Le modèle POKS

Il est basé sur *la théorie de la structure des connaissances* un formalisme de représentation de l'ordre dans lequel les unités de connaissance sont apprises [Desmarais & Maluf 1996]. Dans ce cas-ci les unités de connaissance sont des items, les liens existant entre les unités des connaissances sont définis par une relation  $A \Rightarrow B$  signifiant que si un étudiant connaît correctement  $A$ , il connaîtra nécessairement  $B$ . Idem s'il ne connaît pas  $B$  il ne connaîtra pas  $A$ . Cette relation est déduite à partir des analyses faites sur des tests antérieurs tout en leur appliquant les trois conditions suivantes :

$$P(B/A) \geq p_c \quad (1)$$

$$P(\overline{A}/\overline{B}) \geq p_c \quad (2)$$

$$P(B/A) \neq P(B) \quad (3)$$

Avec  $p_c$  est la probabilité conditionnelle minimale dans (1) et (2).

La condition (1) correspond à la probabilité conditionnelle de réussir  $B$ , étant donné le succès de  $A$ , qui doit être supérieure à  $p_c$ . De même (2) correspond à la probabilité conditionnelle de ne pas réussir  $B$ , étant donné l'échec dans  $A$ , qui doit être aussi supérieure à  $p_c$ . Alors que la condition (3) montre qu'il y a une différence entre la probabilité conditionnelle et celle non conditionnelle.

Une des caractéristiques de POKS est que les parents  $(X_{p1}, \dots, X_{pn})$  de  $X$  sont indépendants entre eux :

$$P(X/X_{p1}, \dots, X_{pn}) = \prod_{i=1}^n P(X/X_{pi})$$

Cette technique permet l'induction des structures de connaissance à partir d'un petit nombre des cas de données empiriques, puisque l'analyse sera limitée à des relations binaires [Desmarais et Pu 2005].

Une comparaison faite entre cette technique et celle utilisant les réseaux Bayésiens par Desmarais dans [Desmarais & al. 2006] montre que ces deux approches

sont efficaces, mais que POKS reste la meilleure puisqu'elle affiche une forte puissance de prédiction, et un faible temps de réponse. Alors que la comparaison faite par le même chercheur [Desmarais & Pu 2005] entre POKS et le modèle à deux paramètres de la théorie des réponses aux items dans un TAO montre que les deux approches classent efficacement et rapidement les étudiants dans un sujet donné.

### **3.2. Test adaptatif par ordinateur basé sur la théorie des réponses aux items.**

Tous les modèles que nous venons de voir ne sont pas incrémentaux, c'est-à-dire si nous voulons ajouter de nouveaux items il faut reconstruire le modèle. Alors que dans le TAO à base de la TRI, il suffit d'ajouter les items et ses paramètres dans la base d'items pour qu'ils soient pris en charge par le programme, une des raisons essentielles qui nous a poussés à le choisir dans notre test d'anglais adaptatif.

Comme nous avons dit précédemment, c'est la théorie des réponses aux items qui était à l'origine d'un test adaptatif, elle reste à nos jours la plus efficace et la plus utilisée. Afin de mieux comprendre ce processus, nous décrivons ici les étapes nécessaires pour la construction et le déroulement d'un TAO, notamment le choix des items qui composent *la banque d'items*, l'utilisation *des règles de calculs* à base de la TRI afin *d'estimer le niveau d'habileté* du candidat et la détermination de l'algorithme de *sélection de l'item* le plus efficace possible.

#### **3.2.1. Banque d'items**

La première étape à faire dans un TAO est de construire une banque d'items adéquate, puisque la qualité des items rend meilleure la tâche de l'algorithme du test adaptatif. Ainsi un bon programme adaptatif ne fonctionne pas de manière efficace si la banque d'items n'est pas bien construite, telle celle avec un nombre très petit des items ou ceux de mauvaise qualité [Wainer 2000].

Une Banque d'items contient en général, le texte de l'item, ses différentes réponses, correctes et incorrectes, ainsi que les paramètres de l'item. Ces derniers sont

choisis selon le modèle utilisé, ainsi dans un modèle à trois paramètres, il faut intégrer le paramètre de difficulté, celui de discrimination et celui de pseudo-chance de chaque item dans la banque. Afin de construire une base ou banque d'items, il faut passer par les étapes suivantes :

1. Choisir un ensemble d'items reliés à un sujet ou un thème bien défini, où le candidat va être évalué.
2. Faire passer ce test, qui va être au départ un test papier-crayon, à un ensemble très grand de candidats, avec différents niveaux d'habileté. Les résultats obtenus seront collectés pour être ensuite analysés.
3. Analyser les résultats, en utilisant la théorie des réponses aux items, et extraire les paramètres de difficulté, de discrimination et de pseudo-chance de chaque item. Comme nous avons vu précédemment, plusieurs logiciels font cette analyse automatiquement, nous avons utilisé dans ce travail PARAM-3PL (voir la section 5.3.1).
4. Nettoyer la banque d'items en éliminant ceux qui sont peu discriminatoires ou qui ont un facteur de pseudo chance très grand. Ainsi d'après [Wainer 2000], pour un modèle à trois paramètres, une bonne banque d'items doit être composée par des items avec des valeurs des paramètres de discrimination élevées (supérieures à 1), par ceux avec des paramètres de pseudo-chance très petits (inférieures à 0.2) et avec à peu près le même nombre d'items pour chaque niveau de difficulté.

### **3.2.2. Déroulement d'un test adaptatif par ordinateur**

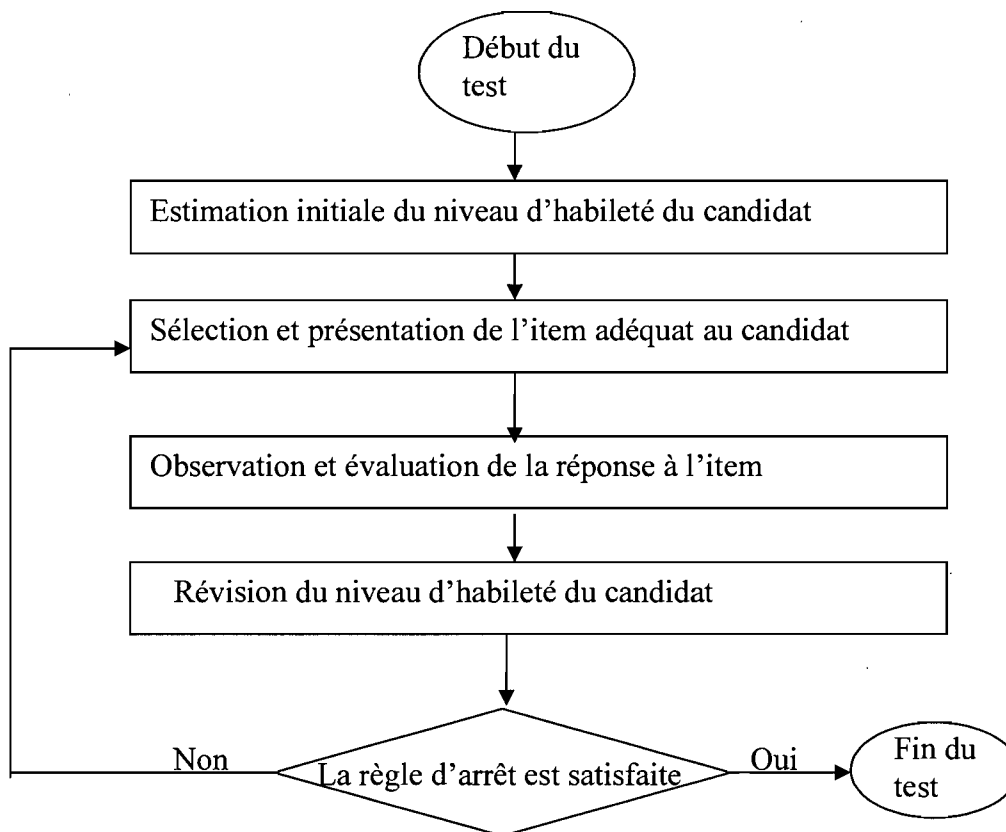
La façon la plus claire et la plus évidente pour décrire le déroulement d'un test adaptatif c'est à travers un algorithme. En général même un test papier-crayon dont le déroulement est simple peut être décrit par un algorithme qui n'apparaît pas dans le processus du test :

1. Répondre à la première question.
2. Répondre à la question suivante.
3. Répéter l'étape 2 jusqu'à ce qu'une réponse à la dernière question est donnée ou que le temps du test est écoulé.

Or nous savons que dans un test adaptatif, nous cherchons à estimer le niveau d'habileté d'un candidat, tout en utilisant les réponses aux items précédentes. Et cela en lui présentant la question de difficulté proche de son niveau d'habileté. Ainsi l'algorithme du test adaptatif peut être décrit de la manière suivante :

1. Une estimation initiale du niveau d'habileté du candidat précisera le premier item à présenter. C'est l'étape du départ.
2. Une fois la réponse du candidat connue, une estimation de son niveau d'habileté est calculée. L'item dont la difficulté est au voisinage de ce niveau d'habileté est par la suite sélectionné de la banque d'items et présenté au candidat. C'est l'étape de la succession.
3. L'étape 2 est répétée jusqu'à ce qu'un nombre bien défini des items est présenté au candidat, ou que la précision souhaitée du niveau d'habileté du candidat soit atteinte. C'est l'étape d'arrêt ou la règle d'arrêt.

Cet algorithme est représenté schématiquement par l'organigramme de la figure 3.2 ci-dessous.



**Figure 3.2 :** Organigramme du processus du test adaptatif

Afin de bien comprendre ce processus, une description bien détaillée de ces trois étapes s'en suit.

### 3.2.2.1. Étape du départ.

La règle de base d'un test adaptatif est l'estimation du niveau d'habileté du candidat tout en utilisant les réponses de ce dernier aux items qui lui sont présentés. Le problème qui se pose ici est comment faire pour avoir une estimation initiale du niveau d'habileté du candidat au début du test, c'est-à-dire au moment où aucun item n'est encore présenté à celui-ci et par conséquent aucune réponse n'est encore connue. Plusieurs solutions ont été adoptées jusqu'ici :



Généralement c'est à partir de l'information à priori des niveaux d'habileté qu'une estimation initiale de celui-ci, pour un candidat donné, est déduite. Ainsi selon Thissen et Mislevy [Wainer 2000] cette dernière peut être, soit la moyenne des estimations des niveaux d'habileté dans une population à qui appartient le candidat, soit retrouvée à partir de l'historique du candidat tel les tests qu'il a déjà passés dans le même domaine, soit inspirée du profil du candidat. Enfin, La manière la plus simple d'estimer initialement la compétence d'un candidat, est de lui affecter la valeur de la moyenne des compétences des candidats ayant déjà passé le test, généralement c'est 0, ou parfois lui administrer un pré-test à partir duquel, son niveau d'habileté initiale est déduit.

D'après Raïche [Bertrand & Blais 2004] la détermination de l'estimateur du niveau d'habileté préliminaire peut affecter l'estimateur final du niveau d'habileté si le nombre des items administrés est petit. Ainsi, un bon choix est primordial à cette étape, c'est-à-dire une solution qui donne une meilleure estimation initiale à la compétence du candidat et par la suite l'estimation finale de cette dernière sera rapide, précise et avec un petit nombre d'items. C'est pourquoi nous avons intégré dans notre travail *le raisonnement à base de cas* à cette étape. Ainsi, en se basant sur le profil du candidat, une recherche est faite dans la base des cas (constituée par les profils des candidats ayant déjà passé le test avec leurs niveaux d'habileté estimés) des profils qui se rapprochent le plus à celui du candidat, et la moyenne des compétences estimées pour ces profils est par la suite considérée comme une estimation initiale du niveau d'habileté du candidat.

### **3.2.2.2. Étape de succession**

Une fois l'estimateur du niveau d'habileté connu, soit initial dans le cas du début du test ou calculé par la TRI dans le cas contraire, ainsi que la réponse du candidat; une sélection de l'item dont la difficulté est proche du niveau d'habileté du candidat est faite. Or dans la section 2.6, nous avons vu que la fonction d'information est maximal pour un item au voisinage de son paramètre de difficulté. Ainsi l'item sélectionné sera celui le plus informatif au voisinage du niveau d'habileté estimé, et dont le taux d'exposition<sup>6</sup>

---

<sup>6</sup> Le taux d'exposition est le nombre de fois qu'un item est administré sur le nombre total des candidats.

est minimal. Cet item sera par la suite l'item suivant à être présenté au candidat. Une explication bien détaillée de chaque partie de ce processus est étalée dans ce qui suit.

Nous avons déjà expliqué dans le chapitre de la TRI, les différentes méthodes pour l'estimation du niveau d'habileté, donc ce qu'il nous reste à faire est de donner les différentes façons de la sélection de l'item et expliquer c'est quoi un taux d'exposition et la manière dont nous pouvons le minimiser.

#### a) La sélection de l'item

La sélection de l'item dans un TAO est une stratégie qui permet de sélectionner un item de manière adaptative étant donné le niveau d'habileté du candidat. Les deux approches les plus utilisées jusqu'à présent sont : l'approche de maximisation de l'information (*maximum information*) et la stratégie Bayésienne de sélection de l'item (*Bayesian item selection*) [Meijer & Nering 1999].

##### ❖ Stratégie de maximisation de l'information.

Théoriquement comme nous cherchons à estimer le niveau d'habileté  $\hat{\theta}$  d'un candidat, il faut sélectionner l'item qui minimise l'espérance de l'erreur-type de l'estimateur du niveau d'habileté, ce qui revient à choisir l'item qui maximise le mieux la fonction de l'information à  $\hat{\theta}$ , compte tenu du fait que cette dernière est l'inverse de l'erreur-type, comme nous avons mentionné dans la définition de la fonction de l'information de l'item dans le chapitre précédent (voir la section 2.6). Ainsi il faut

choisir l'item  $i$  qui donne un  $I_i(\hat{\theta})$  maximale, avec 
$$I_i(\hat{\theta}) = \frac{[P_i(\hat{\theta})]^2}{P_i(\hat{\theta})(1 - P_i(\hat{\theta}))}$$

Où  $P_i(\hat{\theta})$  est la probabilité qu'un candidat de niveau d'habileté  $\hat{\theta}$  répond correctement à l'item  $i$ ,  $P_i'(\hat{\theta})$  est la dérivée première de la fonction de probabilité.

Selon Thissen et Mislevy [Wainer 2000], la façon la plus efficace pour implémenter cette approche est de construire au préalable une table (table d'informations) où pour chaque niveau d'habileté est associé un nombre déterminé d'items triés selon l'information qu'ils fournissent à ce niveau d'habileté. Cette

approche ne demande pas beaucoup de calcul, en outre elle permet l'obtention de meilleurs résultats. Cette méthode peut être simplifiée en choisissant comme prochain item celui dont le paramètre de difficulté se rapproche le plus de l'estimateur du niveau d'habileté [Bertarand & Blais 2004].

❖ La stratégie Bayésienne de sélection de l'item

Cette stratégie est souvent connue sous le nom de la méthode Bayésienne de Owen, en se référant à son fondateur Owen(1969). Ce dernier a développé une formule qui calcule approximativement la moyenne à postériori de  $\theta$  et sa variance après l'administration de  $n$  items. À partir de cette formule une fonction de sélection de l'item est développée qui permet de choisir l'item qui réduit maximalelement la variance à postériori. Selon Raïche [Bertarand & Blais 2004], dans cette méthode une approximation de la loi normale par une loi logistique est fréquemment appliquée. Pour avoir plus de détails sur cette méthode voir [Van Der Linden 1998].

D'après Thissen et Mislevy [Wainer 2000], même si cette méthode possède des équations complexes, elle a l'avantage de diminuer le temps de calcul puisqu'on n'a plus recours aux itérations. Son inconvénient c'est que l'approximation de  $\theta$  dépend de l'ordre dans lequel les items son présentés, ce qui est contradictoire avec les principes de la TRI où l'estimation de  $\theta$  est indépendante de la manière dont les items sont administrés au candidat. C'est pour cette raison que la méthode Bayésienne de Owen est de moins en moins utilisée.

Ces deux dernières méthodes font en sorte que le meilleur item soit présenté au candidat pour chaque étape du déroulement du test, ce qui revient à choisir souvent les items possédant des paramètres de discriminations élevés. En outre, un certain nombre d'items sont fréquemment sélectionnés alors que d'autres ne le seront jamais. Le test sera ainsi non fiable, puisque la connaissance d'un de ces items et par conséquent sa réponse, cause une mauvaise estimation du niveau d'habileté du candidat. C'est pourquoi l'intégration d'un mécanisme qui contrôle le taux d'exposition de l'item est primordiale dans cette étape.

## b) Contrôle du taux d'exposition

Plusieurs procédures ont été créées jusqu'ici pour satisfaire cette condition. Selon [Boyd 2003], elles peuvent être classées en deux catégories : des *procédures de sélection aléatoire*, et des *procédures de sélection conditionnelle*. Dans la première catégorie, après avoir sélectionné un nombre bien défini d'items avoisinants le niveau d'habileté estimé en maximisant la fonction d'information, un de ces items est choisi aléatoirement pour être administré par la suite au candidat. Malgré qu'elles soient simples à intégrer, ces procédures n'arrivent pas à atteindre un taux d'exposition maximal.

Par contre, les procédures de sélection conditionnelle permettent d'atteindre un taux d'exposition maximal prédéfini auparavant à l'aide des paramètres de contrôle d'exposition alloués initialement aux items en utilisant des simulations; or ces dernières nécessitent des procédures et un temps de calcul respectivement complexes et long. Une des procédures de sélection conditionnelle la plus connue est celle de Sympson-Hetter (1985) et d'autres s'en suivent.

La méthode de Sympson-Hetter utilise des paramètres de contrôle d'exposition de l'item pour contrôler, à l'aide des probabilités, les fréquences avec lesquelles les items sont administrés. Ces paramètres de contrôle sont obtenus à partir d'une série de simulations de tests adaptatifs administrés à une population donnée.

Plusieurs chercheurs ont amélioré ces dernières techniques en utilisant les tests fantômes (*shadow tests*), notamment [Van der Linden & Reese 1998], [Li & Schafer 2005] et plusieurs autres. Avec ces tests fantômes, en plus d'atteindre le maximum du taux d'exposition, d'autres contraintes plus complexes sont satisfaites, elles peuvent être spécifiques aux caractéristiques de l'item (nombre de mots, nombre de réponses,...etc.), ou conditionnées par les caractéristiques de la personne passant le test (sa culture, sa langue, son sexe,...etc.) Raïche [Bertarand & Blais 2004].

En plus de ces deux catégories, d'autres procédures appelées de stratification ont été créées par [Chang & Ying 1999] où les items avec des petits paramètres de discrimination sont les premiers à être administrés, alors que ceux avec des paramètres de discrimination élevés sont les derniers à être présentés, là où l'exactitude du niveau

d'habileté est plus forte. Ainsi ce n'est pas juste les items les plus discriminatoires qui sont administrés, comme c'est le cas dans la plupart des tests adaptatifs, mais même les items avec des petits paramètres de discrimination auront la chance d'être alloués. Nous nous attardons un peu plus sur cette dernière catégorie car c'est celle-ci qui nous intéresse dans ce travail.

### c) Procédures de stratification

La sélection de l'item par une maximisation de la fonction de l'information, fait en sorte que c'est les items avec des paramètres de discrimination élevés qui sont les plus sélectionnés. Chang [Chang & Ying, 1999] a proposé une méthode appelée *test adaptatif par ordinateur a-stratifié*. Dans cette approche les items sont stratifiés selon leur paramètre de discrimination  $a$ . Ainsi au début du test, se sont ceux avec de faibles  $a$ s qui sont sélectionnés, et au fur et à mesure que le test avance c'est les items avec  $a$ s de plus en plus élevés qui sont choisis. L'algorithme de cette méthode est donné ci-dessous.

#### ❖ Algorithme :

1. Partitionner la banque d'items en  $K$  niveaux selon le paramètre de discrimination  $a$ .
2. Partitionner le test dont la longueur maximale est toujours prédéterminée, en  $K$  étapes.
3. À la  $k$ -ème étape, sélectionner  $n_k$  items du niveau  $k$  en se basant sur la similarité de  $b$  (paramètre de difficulté) avec  $\hat{\theta}$  (notant que  $n_1 + n_2 + \dots + n_K =$  longueur du test)
4. Recommencer l'étape 3 pour  $k = 1, 2, \dots, K$ .

Pour déterminer  $K$ , le nombre des niveaux, plusieurs facteurs doivent être pris en considération : le premier facteur est la variation des valeurs de  $a$  dans le niveau. Ainsi, dans une banque d'items composée par des questions dont les valeurs de  $a$  sont presque égales, l'effet de stratification est minimisé, et par la suite le nombre des niveaux exigés est petit. Alors que dans une banque d'items composée par des questions dont les valeurs de  $a$  sont très différentes, le nombre des niveaux exigés est grand. Le second

facteur à considérer est la richesse de la banque d'items, en se référant au paramètre de difficulté  $b$  dont la valeur doit être similaire à la valeur de  $\hat{\theta}$  estimée dans le test. Ainsi dans chaque niveau, les valeurs de  $b$  doivent suffisamment couvrir une large gamme de valeur. Le troisième et le quatrième facteur sont respectivement la longueur du test et la taille de la banque d'items. Si celle-ci est suffisamment large,  $K$  peut être choisi dans le même ordre que la longueur du test.

Nous constatons que la méthode de la maximisation de la fonction de l'information de l'item n'est pas utilisée ici dans la sélection de l'item suivant, mais plutôt la similarité entre le niveau de difficulté et l'estimateur du niveau d'habileté.

Selon [Chang & Ying 1999], cette méthode donne un taux d'exposition équilibré dans certaines banques d'items. La moyenne des taux d'exposition de tous les items est très inférieure comparativement à celle récupérée lors des autres méthodes de sélection de l'item vues précédemment. En plus l'exactitude de l'estimateur du niveau d'habileté est maintenue.

Le même chercheur a amélioré cette technique dans son article [Chang *et al.* 2001] en essayant d'avoir plus d'exactitude dans l'estimateur du niveau d'habileté, il l'a appelé *algorithme a-stratifié avec blocage de b* décrit de la manière suivante :

1. Partitionner la banque d'items en  $M$  blocs selon la valeur du paramètre de difficulté  $b$ .
2. Partitionner chacun des  $M$  blocs en  $K$  niveaux selon le paramètre de discrimination  $a$ .
3. Pour  $k=1, 2, \dots, K$  recombinaison les  $k$ -ième niveaux de chacun des  $M$  blocs pour former une nouvelle strate.
4. Partitionner le test en  $K$  étapes.
5. À la  $k$ -ième étape, sélectionner  $n_k$  items du niveau  $k$  en se basant sur la similarité de  $b$  avec  $\hat{\theta}$ .
6. Recommencer l'étape 5 pour  $k = 1, 2, \dots, K$ .

C'est cette dernière technique que nous utilisons dans notre travail, tout en lui ajoutant la contrainte de maximisation de la fonction d'information. Ainsi nous utilisons

les deux contraintes ensemble : la minimisation du taux d'exposition et la maximisation de la fonction d'information.

### 3.2.2.3. Étape d'arrêt

Comme mentionné auparavant, dans l'algorithme du test adaptatif, il faut choisir un ou plusieurs critères qui font en sorte que le test s'arrête à un moment donné, et bien sûr avec un estimateur du niveau d'habileté le plus exact possible. Ces critères sont soit l'administration d'un nombre bien déterminé d'items, soit l'atteinte d'une certaine précision dans l'estimateur du niveau d'habileté, ou que le temps alloué au test soit écoulé.

Si par exemple nous fixons la longueur du test, l'erreur type entre le niveau d'habileté réel et l'estimateur du niveau d'habileté variera d'un candidat à un autre, par conséquent pour certains candidats le test ne sera pas efficace. Par contre, d'après Raïche [Bertrand & Blais 2004], si c'est le niveau de précision qui est fixé sans tenir compte de la longueur du test, l'atteinte de ce niveau devient parfois difficile, en provoquant ainsi un test très long.

Chang et Ying ont essayé d'atteindre un niveau de précision minimale prédéterminée [Chang & Ying 2003], tout en variant la longueur du test par candidat et cela en utilisant des simulations de tests. Les résultats sont satisfaisants, c'est-à-dire que la précision voulue est presque atteinte avec différents niveaux d'habileté. Mais il s'avère que cette technique nécessite une banque d'item très grande, ce qui n'est pas toujours le cas.

Une combinaison des deux critères précédents est préférée, ainsi le test se termine une fois qu'une précision prédéterminée du niveau d'habileté soit atteinte ou qu'un nombre bien précis d'items soit administré au candidat. Éventuellement il est nécessaire de fixer ces deux valeurs, l'erreur type de l'estimateur du niveau d'habileté et la longueur du test, pour avoir un test adaptatif le plus efficace possible.

Selon Raïche dans [Bertrand & Blais 2004], lorsque l'erreur type de l'estimateur du niveau d'habileté exigé par le test n'est pas assez petite, l'administration de 13 items

donnera une bonne estimation de la compétence du candidat. Si la précision (erreur type) exigée est de l'ordre de 0.2 ou moins, la longueur du test doit être égale à 40 au moins. En plus, comme mentionné dans l'article de [De Ayala *et al.* 1995], après l'administration de 20 items, l'erreur type de l'estimateur du niveau d'habileté reste virtuellement inchangeable. Donc d'après ce qui précède, l'administration de 20 items donne probablement de bons résultats.

De même [Raïche 2000] a montré que la précision de l'estimateur du niveau d'habileté ne peut pas dépasser 0.40, si un modèle logistique à un paramètre et une méthode de l'espérance à posteriori sont appliqués.

Finalement, il ne faut pas oublier d'ajouter la notion du temps parmi les critères d'arrêt du test, surtout pour des raisons administratives, et bien sûr sans tenir compte ni de la longueur du test ni de l'erreur type de l'estimateur final du niveau d'habileté.

### **3.2.3. L'estimateur final du niveau d'habileté**

Naturellement, toutes les méthodes utilisées pour l'estimateur intermédiaire du niveau d'habileté, restent aussi efficaces pour l'estimateur du niveau d'habileté final. Toutefois, selon Thissen et Mislevy dans [Wainer 2000], arrivé à la phase finale du test, l'estimateur de la maximisation à posteriori et celui de l'espérance à postérieure donnent les mêmes résultats après 20 items. En plus, selon [Raïche 2000], une très grande précision est obtenue en utilisant ces deux dernières méthodes, comparativement à la méthode de vraisemblance maximale. Ainsi, l'utilisation de la méthode de l'espérance à postérieure donne de bons résultats depuis le début du test jusqu'à la phase finale. C'est pourquoi nous avons opté pour cette méthode dans notre travail.

## **3.3. Conclusion.**

Comme nous avons vu précédemment, le test adaptatif par ordinateur a vu le jour depuis les années 80. Plusieurs recherches ont été faites dans chacune des étapes de son algorithme afin d'avoir un estimateur du niveau d'habileté le plus efficace possible avec un test le plus court possible. De nos jours, il reste un terrain fertile pour plusieurs autres



recherches. Plusieurs solutions sont de plus en plus utilisées, notamment la méthode Bayésienne et Poks, mais la théorie des réponses aux items reste la plus efficace.

Avec la puissance des ordinateurs, les calculs complexes ne présentent plus aucun obstacle pour l'élaboration d'un bon test adaptatif. Reste son coût, qui est toujours élevé par rapport à celui du test papier-crayon. En plus, afin d'avoir une bonne estimation des paramètres des items, un échantillonnage d'une grande taille est exigible.

Malgré ces contraintes plusieurs applications du test adaptatif ont vu le jour, notamment aux États-Unis où plusieurs tests possèdent une version adaptative par ordinateur opérationnelle. Tels que: GRE (*Gradate Record Examining*) [URL4] et TOEFL (*Test Of English as a Foreign Language*) [URL5] des tests exigés pour étudier dans des grandes écoles ou universités anglophones faites par l'ETS (*Educational Testing Service*), NCLEX (*National Concil Licensure EXamination*) [URL6] un examen national des infirmiers développé par NCSBN (*National Concil of State Boards of Nursing*), ASVAB (*the Armed Services Vocational Aptitude Battery*) [Legree *et al.* 1998] un test adaptatif de classement pour le recrutement de nouveaux militaires développé par l'institut de recherche de l'armée américain. En plus, Microsoft utilise le test adaptatif par ordinateur dans plusieurs de ses examens de certification.

En Espagne, [Conejo *et al.* 2004] ont développé un outil de test adaptatif sur le Web appelé SIETTE (Spanish translation of Intelligent Evaluation system using Tests for TeleEducation). Au Canada, un test adaptatif de la langue française a vu le jour en 1999, développé par Laurier [Laurier 1999], FrenchCapt (*French Computerized Adaptive Placement Test*). Les exemples cités jusqu'ici ne sont pas les seuls, il en existe plusieurs autres.

Ainsi l'utilisation du test adaptatif n'est pas une technique récente, plusieurs chercheurs l'ont utilisé avant. Mais l'apport du raisonnement à base de cas n'est pas étudié jusqu'à présent, selon nos connaissances, qui représente notre contribution dans ce travail.

## Chapitre 4 : Raisonnement à base de cas

Le raisonnement de l'être humain est basé sur la mémorisation, chaque fois qu'il est confronté à un problème, il déduit sa solution à partir des solutions trouvées pour des problèmes semblables dans le passé. Prenons par exemple un médecin qui cherche à diagnostiquer le cas d'un malade. En étudiant les symptômes de ce dernier tout en ajoutant d'autres examens, si c'est nécessaire, le médecin aura assez de données afin de reconnaître la maladie, et par la suite, prescrire un remède au malade. Le médecin a adopté ici *un raisonnement à base de cas*, puisqu'un malade est considéré comme un nouveau cas. Puis, en se basant sur les caractéristiques de ce nouveau cas, il a essayé de trouver la maladie correspondante la plus à ce patient, tout en faisant une recherche soit dans ses connaissances, ou parfois dans ses livres. La maladie ainsi décelée, un traitement adéquat sera prescrit au malade. Par conséquent, on constate que le médecin s'est inspiré des vieux problèmes (cas) pour trouver une solution pour le nouveau.

C'est un raisonnement qui nous paraît facile et évident, mais très difficile pour une machine. Des chercheurs ont essayé de l'adapter pour celle-ci, et cela en rassemblant les expériences (les cas) dans une base des cas; ainsi devant chaque nouveau problème, la machine cherche l'expérience qui se rapproche le plus au problème présent pour en extraire la solution.

Les premières contributions du raisonnement à base de cas étaient de Schank et Abelson en 1977 [Kolodner 1993], en proposant de décoder notre connaissance générale des situations sous forme de scénarios qui nous permettent de fonder des espérances sur ce que nous venons d'entendre et inférer des rapports entre ce que nous avons entendu.

Plusieurs applications ont vu le jour dans cette période notamment :

CHEF : proposé par Hammond à l'université du Chicago [Hammond 1989], un planificateur à base de cas qui génère des recettes : à partir de caractéristiques recherchées dans un plat par un utilisateur (ingrédients et goûts), une nouvelle recette est construite satisfaisant les besoins de l'utilisateur.

CASEY : Un diagnostiqueur à base de cas, proposé par Koton à l'institut technologique du Massachusetts [Koton 1988]. Sa fonction consiste à chercher à partir des signes et des symptômes d'un nouveau patient, un diagnostique qui se rapproche le plus au cas de ce patient dans la base de cas.

HYPO : un interpréteur du raisonnement utilisé dans le domaine de la loi, proposé par Ashley à l'université du Massachusetts [Ashley & Rissland 1988], et capable de créer des arguments pour le client, pouvant aider soit la défense soit le plaignant, à partir des situations légales.

Nous nous limitons à ces cas-ci, mais il en existe plusieurs autres.

Dans ce chapitre, nous allons voir une explication détaillée du raisonnement à base de cas, tout en passant par chacune de ses étapes.

## 4.1. Définitions

D'après Riesbeck [Riesbeck & Schank 1989] un raisonnement à base de cas est le fait de résoudre de nouveaux problèmes en adaptant les solutions des anciens problèmes.

Selon Kolodner [Kolodner 1993], un raisonnement à base de cas est l'adaptation des anciennes solutions pour satisfaire les nouvelles demandes. Des anciens cas sont utilisés pour expliquer, critiquer ou interpréter de nouvelles situations, ou même créer une solution équitable pour un nouveau problème.

De la même façon, Leake [Leake 1996] a défini le raisonnement à base de cas comme une approche qui résout les problèmes, en retrouvant dans la mémoire, les cas les plus pertinents et les adapter par la suite aux nouvelles situations.

Après avoir défini un raisonnement à base de cas, une question est soulevée qui demande une réponse : Qu'est-ce qui est alors un cas?

D'après Kolodner [Kolodner 1993], un cas est « *a contextualized piece of knowledge representing an experience that teaches lesson fundamental to achieving the goals of the reasoner* ». D'une autre façon, un cas est défini comme une partie

contextuelle d'une connaissance représentant une expérience qui aide à résoudre de manière facile de nouveaux problèmes dans le futur.

En général un cas est composé d'un « problème » et de sa « solution ». Il est parfois même construit à partir des conséquences déduites lors de l'application de la solution (succès ou échec).

## 4.2. Le processus d'un raisonnement à base de cas

Un raisonnement à base de cas (Case Based Reasoning « CBR ») est un processus qui peut être résumé en « *se souvenir et adapter* » ou « *se souvenir et comparer* ». Ainsi à partir d'un nouveau cas, on cherche dans la base de cas celui qui se rapproche le plus de ce dernier en utilisant les techniques de similarité. La solution du cas retrouvée sera par la suite adaptée pour le nouveau cas. Le processus d'un CBR est montré dans la figure ci-dessous, adaptée de [Marling & Cheetham 2006].

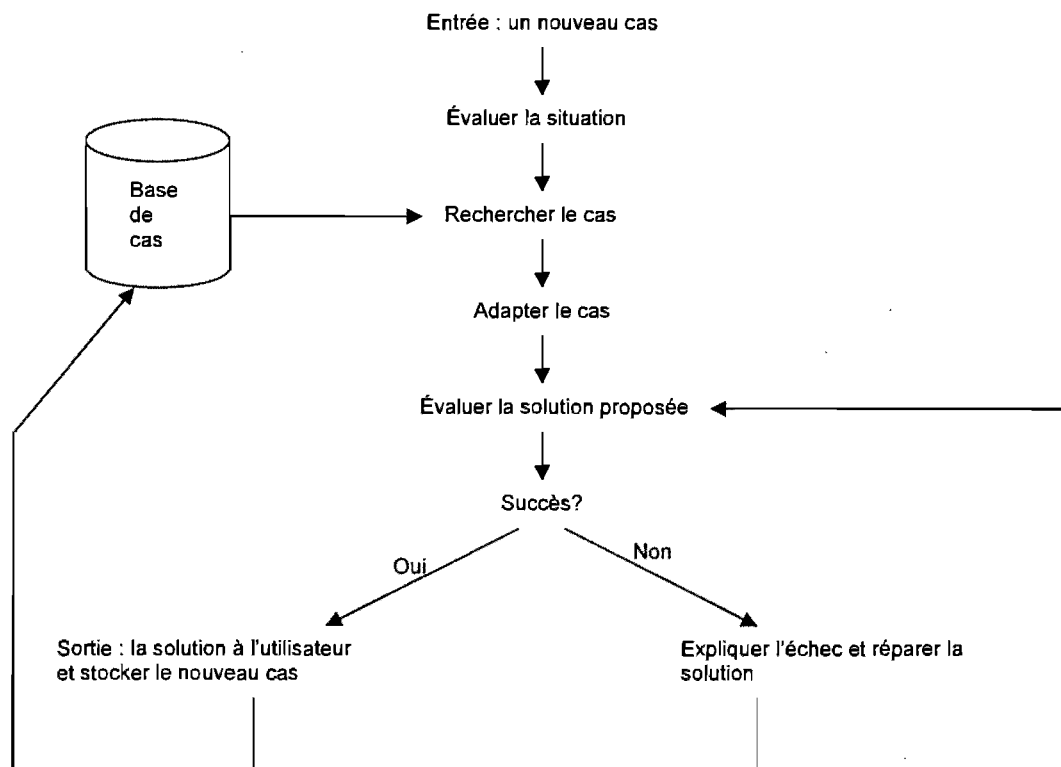


Figure 4.1 : Processus d'un système de CBR

Ce schéma représente les quatre étapes du cycle de vie d'un CBR, qui sont, *rechercher*, *adapter*, *maintenir* et *apprendre* (retrieve, reuse, revise, retain) appelé souvent le modèle à « 4R ». Alors que Finnie et Sun [Finnie & Sun 2003] ont proposé un autre modèle à cinq étapes appelé « 5R » (repartition, retrieve, reuse, revise, retain), mais le modèle « 4R » reste le plus connu et le plus utilisé. Nous expliquons chacune de ses étapes dans ce qui suit.

#### 4.2.1. La recherche

Une fois un problème à résoudre ou une situation à interpréter est présentée au système de raisonnement à base de cas, une évaluation des caractéristiques du problème et de ses attributs est faite afin d'extraire ceux qui sont pertinents pour une meilleure recherche. Cette dernière qui se déroulera dans la base de cas (qui contient l'ensemble des problèmes précédents ainsi que leurs solutions) afin de sélectionner le problème qui se rapproche le plus de celui initial et cela en utilisant des métriques de similarité.

Selon [Azuaje *et al.* 2000], trois méthodes sont généralement utilisées afin de retrouver le cas le plus pertinent, soit *l'approche calculatoire* en se basant sur les mesures de similarité, *l'approche représentative* en se basant sur les structures de l'indexation, ou *l'approches hybride* qui combine les deux.

Dans la première approche la similarité entre les cas est calculée en mesurant les distances entre leurs attributs. D'après [Lopez *et al.* 2005], habituellement les cas retrouvés sont les  $k$  les plus similaires au nouveau problème, appelés *k plus proche voisin* (*k Nearest Neighbor k-NN*). C'est une procédure qui demande beaucoup de temps si la base des cas est large. Ainsi pour contourner ce problème, un arbre binaire est utilisé, en partitionnant les cas en groupes, chacun d'eux contient les cas qui sont similaires entre eux, c'est ce qu'on appelle *l'arbre k-d* (*k-d tree*).

Dans la deuxième approche, la recherche est basée sur une similarité structurelle entre les cas, qui retourne en général les cas les plus pertinents. Selon [Lopez de Mantaras *et al.* 2006] la similarité structurelle est la structure de graphe la plus spécifique qui existe en commun entre le problème cible et le cas stocké, et un ensemble de règles de transformation utilisées pour déterminer cette structure en commun. Ces

stratégies de recherche sont basées généralement sur les arbres de décision, le regroupement (*clustering*) et autres méthodes de l'apprentissage machine (*machine Learning*).

Alors que dans la dernière approche, les structures d'indexation et les mesures de distances sont combinées entre elles pour performer le processus de la recherche des cas les plus pertinents.

#### **4.2.2. L'adaptation**

Après avoir retrouvé le cas (solution et problème) recherché, sa solution ne sera pas automatiquement appliquée à notre nouveau problème, mais une adaptation est souvent nécessaire afin de générer une nouvelle solution. Cela se fait de trois façons différentes, *substitutionnelle* en utilisant quelques parties de la solution récupérée, *transformationnelle* en transformant la structure de la solution afin de satisfaire notre nouveau cas, ou *dérivationnelle* en regardant la façon dont le problème est résolu dans le cas retrouvé et déduire ainsi une nouvelle solution.

Normalement en retrouvant le cas le plus similaire au nouveau problème, sa solution sera systématiquement appliquée au nouveau cas. Mais parfois la différence est assez évidente entre ces deux cas qu'il est nécessaire de faire une adaptation de la solution retrouvée afin qu'elle satisfasse notre nouveau cas. C'est dans le domaine de la médecine que l'adaptation est généralement utilisée. Puisque il n'y a pas deux malades qui se ressemblent parfaitement, il est toujours primordial d'adapter le diagnostique d'un cas semblable à un nouveau malade.

#### **4.2.3. La maintenance**

La solution retrouvée est par la suite testée et évaluée afin de voir si elle est satisfaisante ou non. Si un échec se produit, les raisons qui ont entraînées cet échec sont apprises afin de les éviter plus tard, en plus les explications fournis sur l'échec sont utilisées pour réparer cette solution de sorte que l'insuccès ne se reproduit plus. La solution réparée est par la suite testée une autre fois de la même manière que précédemment.

#### 4.2.4. L'apprentissage

Si le succès est atteint, ce nouveau cas sera ajouté à la base des cas, en plus d'une reconfiguration de cette base qui est parfois nécessaire, puisque l'ajout consécutif de nouveaux cas rend sa taille énorme, et par la suite la phase de recherche devient lourde et celle d'adaptation réduite puisque les cas les plus semblables à notre nouveau cas deviennent de plus en plus fréquents [Lopez de Mantaras *et al.* 2006].

Afin de réduire la taille de la base des cas, des techniques d'apprentissage machine sont utilisées pour éliminer les cas redondants de la mémoire, différents sortes de filtres sont ainsi utilisés pour supprimer la connaissance nocive dans différentes étapes du cycle du raisonnement à base de cas, et la notion de compétence du cas est ajoutée afin de catégoriser les cas selon leurs compétences, ce qui va aider à leur élimination ou leur sélection et ainsi de suite.

Dans cette dernière étape, ce n'est pas simplement le fait de savoir quels sont les cas appris qui sera traité, mais aussi quand et comment le faire. Ainsi [Reinartz *et al.* 2001] a proposé une modification du modèle 4R du raisonnement à base de cas en un modèle à six étapes, ainsi la dernière étape est décomposée en trois sous étapes : *retention (retain)*, *la révision (review)* en évaluant et supervisant la qualité du système de connaissance, puis *la restauration (restore)* en modifiant le contenu de la base de cas selon les recommandations résultant de la sous étape précédente.

### 4.3. Les parties composant un cas

Selon [Aamodt & Plaza 1994], un raisonnement à base de cas dépend étroitement de la structure et le contenu de la collection des cas. Ainsi, une description bien détaillée d'un cas est primordiale. Comme nous avons mentionné auparavant, un cas est composé principalement d'un problème (description de la situation), d'une solution (la solution au problème déjà décrit ou la réaction de sa situation) et d'un résultat (obtenu après l'application de la solution ou la réaction de la situation) [Kolodner 1993].

Un problème est composé des buts à atteindre une fois résolu, ces buts sont soit abstraits ou concrets avec une résolution totale ou approximative du problème; des

contraintes posées sur ces buts; et des attributs qui donnent une description plus ou moins détaillée sur la situation la plus pertinente pour atteindre les buts du problème [Kolodne 1993]. Ainsi un problème contient généralement des conditions considérées explicitement lors de la résolution d'un problème spécifique, en plus des informations utilisées normalement par les gens pour décrire ce type de problème [Marling & Cheetham 2006].

Une solution représente en général les concepts et les objectifs que l'ensemble des buts du problème essaye d'atteindre, en tenant compte des contraintes déjà spécifiées et des attributs contextuels. Dépendamment du type du problème, la solution peut être soit une conception, un plan, une classification ou une explication ...etc.

En plus de la solution elle-même, d'autres composantes sont intégrées au système du raisonnement à base de cas pour avoir une meilleur adaptation, notamment, l'ensemble des étapes du raisonnement utilisées pour résoudre le problème, les explications faites sur la solution, les solutions acceptables qui ne sont pas choisies, les solutions non acceptables qui sont éliminées et les résultats espérés après l'application de la solution [Kolodner 1993].

Le résultat du cas spécifie ce qui va arriver une fois la solution est appliquée, soit le feedback ou l'interprétation du feedback de la solution. Plusieurs composantes peuvent lui être ajoutées, comme le fait que le résultat est un succès ou un échec, ou que ce dernier a satisfait ou non les espérances. Dans le cas de l'échec du résultat, ce dernier peut inclure, l'explication sur la cause de cet échec, une stratégie de réparation, ce qu'il faut faire pour éviter le problème et un pointeur sur une meilleure solution [Marling & Cheetham 2006].

#### **4.4. Modèles de la représentation des cas**

À travers les formalismes de la représentation des connaissances les plus connus dans l'intelligence artificielle, un cas peut être représenté de plusieurs manières, cas à vecteurs d'attributs (*propositionnel*), cas structurés (*relationnel*) ou cas textuels (*semi-structuré*) [Bergmann *et al*, 2006].



Le modèle propositionnel : dans ce modèle un cas est représenté par un vecteur des paires <attribut, valeur> comme celui utilisé dans l'apprentissage machine qui supporte le  $k$ -plus proche voisin. Un attribut est une composante caractérisant le cas.

Le modèle relationnel : où chaque cas est représenté par un ensemble d'objets structurés selon l'ordre logique du déroulement du cas, de la même manière qu'une représentation relationnelle est faite dans l'apprentissage machine. Ainsi le cas est formé par un ensemble de groupe de relations entres des objets élémentaires qui constituent le cas. Il ressemble à une représentation orienté-objet de cas. Dans cette dernière représentation le cas est une collection d'objets, chacun d'eux est décrit par un ensemble de paires <attribut, valeur>. La structure de l'objet est décrite par un objet classe.

Le modèle textuel : où le cas qui est sous forme de texte, est décomposé en un ensemble d'entités d'information. Ces dernières sont soit des mots ou des phrases pertinentes contenues dans le texte, de cette façon le cas sera facilement automatisé. Cette décomposition est faite de la même manière que dans la recherche d'information (*information retrieval*) en utilisant par exemple des algorithmes de lemmatisation. Cependant cette structure permet l'introduction de beaucoup plus de sémantiques afin d'avoir une meilleure interprétation pour la réutilisation du cas.

D'autres représentations des cas ont été utilisées pour des domaines particuliers. Notons par exemple le domaine de conception ou de planification. Dans ce dernier les cas sont représentés sous forme de plans, où le problème est décrit par un état initial et un but final, et la solution est totalement ou partiellement ordonnée par une séquence d'actions.

## **4.5. Mesure de similarité dans un système de raisonnement à base de cas**

Comme nous avons vu auparavant, la recherche des cas similaires est la première étape dans un système à base de cas. En plus elle joue un rôle très important dans la recherche de cas. C'est pourquoi il est nécessaire de décrire cette notion et voir les différentes façons de la calculer.

Souvent, dans un système de raisonnement à base de cas, la similarité est calculée soit à partir de l'inverse de la distance euclidienne normalisée dont la formule est la suivante :

$$SIM(X,Y)=1-D(X,Y)=1-\sqrt{\sum_{i=1}^n w_i^2 dist^2(x_i, \tilde{y}_i)}$$

(Formule 4.1)

$w_i$  est le poids normalisé montrant l'importance de l'attribut  $i$ .

$n$  est le nombre des attributs dans le cas.

Soit par l'inverse de la distance de Hamming dont la formule est la suivante :

$$SIM(X,Y)=1-D(X,Y)=1-\sum_{i=1}^n w_i dist(x_i, y_i)$$

(Formule 4.2)

$w_i$  est le poids normalisé montrant l'importance de l'attribut  $i$ .

$n$  est le nombre des attributs dans le cas.

La distance normalisée  $dist(x_i, y_i)$  est souvent calculée de la manière suivante :

$$dist(x_i, y_i) = |x_i - y_i| / |\max_i - \min_i|$$

Pour les attributs numériques,  $\max_i$  et  $\min_i$  sont respectivement la valeur maximale et minimale du  $i$ ème attribut. Alors que pour les attributs symbolique  $dist(x_i, y_i) = 0$  si  $x_i = y_i$  et  $dist(x_i, y_i) = 1$  si  $x_i \neq y_i$ . Ainsi la similarité entre  $X$  et  $Y$  est maximale si les deux cas sont identiques. Par contre elle est égale à 0 si les deux cas sont complètement différents.

En plus de ces deux équations, d'autres façons ont été utilisées pour calculer la similarité entre deux cas dans un système de raisonnement à base de cas. Cela est bien décrit dans l'article de Liao [Liao *et al.* 1998].

## **4.6. Domaines d'application du raisonnement à base de cas**

Le raisonnement à base de cas est appliqué dans plusieurs domaines : l'industrie, la médecine, les sciences, l'art et le divertissement, l'éducation, la recherche d'information, la gastronomie et plusieurs autres domaines. Nous verrons ici quelques exemples de ces applications.

### **4.6.1. L'art et le divertissement**

Le premier travail qui a appliqué le raisonnement à base de cas dans la musique est *SaxEx* proposé par Acros [Acros & Lòpez de Màntaras 2001]. Ce système est capable de générer des mélodies de grandes qualités sous forme de balades de jazz à partir des cas représentant les performances humaines. C'est un travail qui lui a permis l'octroi du prix du meilleur papier dans la conférence internationale de la musique sur ordinateur en 1997.

Dernièrement les jeux sur ordinateur utilisent de plus en plus le raisonnement à base de cas. Dans la conférence internationale du raisonnement à base de cas (*the International Conference on Case-Based Reasoning ICCBR*) à Chicago en 2005, les jeux sur ordinateur est le domaine qui a attiré plus d'attention, avec plusieurs ateliers réussis. Cependant, le premier prix du meilleur papier revient à celui de [Aha *et al.* 2005] sur la planification à base de cas dans les stratégies des jeux sur ordinateur.

### **4.6.2. La Science**

Le raisonnement à base de cas est appliqué aussi dans les sciences notamment le domaine de la biologie moléculaire, en essayant de résoudre plusieurs problèmes notamment ceux de l'analyse des séquences génomiques et la détermination de la structure de la protéine. Cette dernière se fait en comparant l'index de précipitation d'une nouvelle protéine avec les indexes de précipitation de toutes les protéines stockées dans la base de cas. Les plans de cristallisation des protéines dont les index de précipitation sont les plus similaires sont pris en considération lors de la planification de

la cristallisation d'une nouvelle protéine. Pour plus d'informations, voir [Jurisica & Glasgow 2004].

En plus cette technique est utilisée aussi dans la géographie où une combinaison est faite entre le raisonnement à base de cas et les systèmes d'information géographique afin de résoudre des problèmes de raisonnement spatiaux. Le travail de [Holt & Benwell 1999] est considéré comme le premier à utiliser cette notion.

### **4.6.3. L'industrie**

CLAVER est un système fondé sur le raisonnement à base de cas, il est capable de déterminer de manière efficace les composantes d'un matériel utilisé dans le domaine aéronautique. Celui-ci sera traité par la suite dans un grand four appelé « Autoclave ». Ce système utilise le raisonnement à base de cas afin de comparer la liste des composantes à traiter, avec une librairie des composantes ayant donné des meilleurs résultats auparavant, et proposer ainsi une charge efficace dans le grand four [Hinkle & Toomey 1994].

De même GE plastics a employé le raisonnement à base de cas pour déterminer la formule de couleurs qui satisfait la requête d'un client donné. Cet outil appelé FormTool a sauvé plusieurs heures de travail et par conséquent plusieurs millions de Dollars nécessaires pour retrouver la formule de couleurs appropriée. [Cheetham 2005].

### **4.6.4. L'éducation**

Un raisonnement à base de cas dans le domaine de l'éducation aide surtout à superviser et à ordonner les activités d'apprentissage manuelles faites dans la classe tout en incluant le rôle du professeur. Plusieurs outils ont été développés afin d'améliorer la façon de raisonner et de réfléchir de l'étudiant. L'apprentissage par conception, par exemple, utilise les librairies de cas où l'apprenti peut retrouver des exemples annotés qui pourront l'aider au cours de son raisonnement. En plus, dernièrement des systèmes d'apprentissage à base de jeux sont de plus en plus développés afin de satisfaire ce besoin [Kolodner *et al.* 2006]. Notant par exemple JV<sup>2</sup>M une application développée par [Gòmez-Martìn *et al.* 2005] qui représente une simulation d'une machine virtuelle de

Java où les étudiants peuvent apprendre le langage de compilation Java et améliorer leur performance dans la programmation orienté-objet. Dans ce système, le raisonnement à base de cas est utilisé de différentes façons selon la méthodologie d'apprentissage utilisée.

## 4.7. Conclusion

Dans ce chapitre nous avons décrit le processus du raisonnement à base de cas, puis les différentes parties qui composent un cas et ensuite nous avons vu les trois modèles représentant un cas puis quelques différentes façons de mesurer la similarité dans un système de raisonnement à base de cas et enfin quelques domaines où ce dernier est appliqué. Ainsi le raisonnement à base de cas est utilisé dans presque tous les domaines où les solutions des problèmes précédents sont apprises afin de retrouver une solution pour un nouveau problème.

Nous avons opté pour cet outil parce que dans notre travail nous avons besoin de savoir le niveau d'habileté initial d'un candidat, étant donné son profil. Ce que nous pouvons retrouver si nous avons des cas de profils similaires à notre candidat, ainsi la moyenne de leurs niveaux d'habileté estimés précédemment sera par conséquent le niveau d'habileté initial pour notre candidat. Or, on constate que ce raisonnement est le même que celui d'un raisonnement à base de cas, tout en considérant notre problème comme le profil du candidat et la solution comme le niveau d'habileté estimé par le test adaptatif. Ainsi la base de cas sera l'ensemble des profils avec leurs niveaux d'habileté estimés.

## Chapitre 5 : Conception et méthodologie

Dans ce chapitre, nous présentons les objectifs de notre système TAARAC [URL11], faire une présentation de notre approche, décrire l'architecture et l'algorithme de notre système et expliquer la méthodologie adoptée pour atteindre nos objectifs.

### 5.1. Objectifs

La plupart des gens essayent de prendre un ou plusieurs cours dans les langues secondes pour différentes raisons : soit à cause des exigences de leur travail, soit pour une passion, ou encore pour leurs intérêts personnels. Et comme nous connaissons tous, avant de prendre un cours dans une langue seconde, il est nécessaire de passer un test afin de se voir classer dans le niveau approprié. Ce test est souvent sous forme papier-crayon. Le même test, avec le même nombre d'items, dans une durée bien déterminée, est proposé à tous les candidats sans tenir compte de leurs compétences. La correction est souvent, soit manuelle faite par un expert, soit semi automatique en lisant les feuilles des candidats sur une lecture optique. À partir des scores obtenus, le niveau d'habileté du candidat est déduit. Le test papier crayon consomme beaucoup de temps et croît avec l'augmentation du nombre des candidats. Ainsi le temps de correction est toujours long.

Afin de remédier aux problèmes cités ci-dessus, une version sur ordinateur, adaptée selon le niveau d'habileté du candidat, est créée pour quelques uns de ces tests. Comme nous avons vu dans le chapitre 3, il existe plusieurs façons de construire un test adaptatif par ordinateur. Il reste à connaître celle qui donne une meilleure estimation pour le niveau d'habileté du candidat et cela avec un petit nombre d'items.

Dans TAARAC, nous avons transformé le test de classement en anglais de la faculté de l'éducation permanente de l'université de Montréal, qui est sous forme papier crayon, en un test adaptatif par ordinateur. Notre objectif n'est pas simplement de construire un test adaptatif en anglais, mais d'avoir le meilleur test possible. C'est pourquoi nous avons combiné entre deux techniques totalement différentes : *la théorie des réponses aux items* et *le raisonnement à base de cas*. La première technique nous a permis de construire un test propre à chaque candidat, selon son niveau d'habileté : un

test sur mesure. Le nombre des items administrés au candidat et la durée du test sont alors réduits. Grâce à la deuxième technique une meilleure estimation du niveau d'habileté initial est obtenue. Par la suite le nombre des items administrés au candidat et le temps de réalisation du test sont encore plus réduits, et l'estimateur final est plus précis. Nous expliquons ceci plus tard dans ce chapitre.

## **5.2. Approche de TAARAC**

Dans cette partie nous montrons le fonctionnement général de TAARAC, ainsi que les différentes étapes du déroulement du test avant que le niveau d'habileté du candidat lui soit affiché.

### **5.2.1. Algorithme de TAARAC**

De manière générale, le fonctionnement global de TAARAC est présenté ici à l'aide de l'algorithme ci-dessous. Au fur et à mesure que nous avançons dans ce chapitre, les détails s'en suivent.

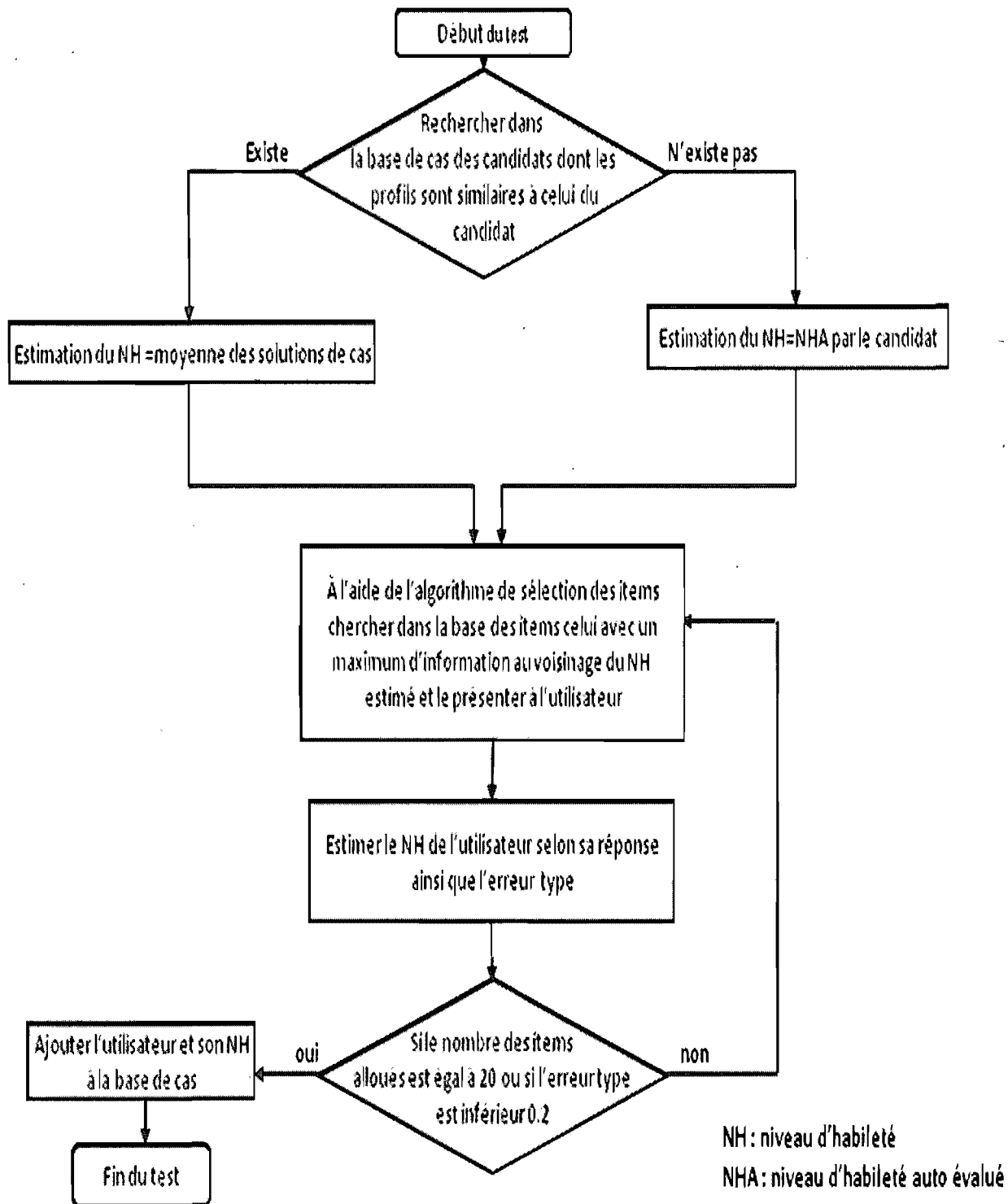


Figure 5.1 : L'algorithme de TAARAC



## 5.2.2. Fonctionnement général

Avant de passer le test, le candidat doit obligatoirement s'enregistrer, et cela en passant par les étapes suivantes. Premièrement, il est invité d'enter son nom, son prénom, son adresse électronique ainsi que son mot de passe. Puis, le candidat doit fournir quelques informations, par exemple, son âge, son pays d'origine, son niveau d'étude complété, sa langue maternelle, celle maîtrisée, et celle apprise. Le candidat doit aussi préciser le nombre d'année qu'il a passées dans l'apprentissage de l'anglais, s'il connaît cette langue depuis son enfance, s'il l'utilise dans son milieu de travail, ainsi que la raison qui l'a poussé à l'apprendre. Son niveau d'apprentissage doit aussi être déterminé et finalement il est invité à donner une estimation de son niveau d'anglais. Toutes ces informations représentent le profil du candidat.

Un candidat qui s'est déjà enregistré doit s'identifier pour que le système le reconnaisse. S'il a déjà passé le test, il aura le choix entre refaire le test ou revoir ses résultats. Sinon, il est invité à le faire ou à le finir si le système a constaté que le candidat avait commencé le test sans le terminer.

Le profil du candidat est comparé avec les profils de ceux qui ont déjà passé le test en appliquant l'algorithme des *k-plus proches voisin* (section 4.2.1). La moyenne des compétences des candidats dont les profils se rapprochent à celui de notre candidat, sera son niveau d'habileté initial. Si aucun profil ne satisfait cette condition, le niveau d'habileté initial sera celui auto évalué par le candidat.

À l'aide de la théorie des réponses aux items, et plus précisément *l'algorithme de sélection de l'item*, nous sélectionnons l'item approprié et nous le présentons au candidat. Ce dernier a un délai bien déterminé dans lequel il doit répondre à l'item. S'il ne le fait pas dans le temps accordé, sa réponse est considérée comme fausse. Une fois la réponse connue, le nouveau niveau d'habileté du candidat et son erreur type<sup>7</sup> sont calculés à l'aide de *la méthode de l'espérance à postériori* (section 2.4.3).

---

<sup>7</sup> L'erreur type d'un niveau d'habileté d'un candidat estimé est la différence entre celui-ci et le niveau d'habileté réel du candidat, autrement c'est la précision avec laquelle celui-ci est estimé (section 2.6).

De la même manière que précédemment, un autre item est affecté au candidat. Et par la suite, un nouveau niveau d'habileté est calculé ainsi que son erreur type. Si le nombre des items présenté au candidat a atteint 20, ou que l'erreur type est inférieure ou égale à 0.2 le test sera arrêté automatiquement. Enfin, le dernier niveau d'habileté calculé sera affiché au candidat. La correction du test et un histogramme qui représente son niveau par rapport à ceux des autres qui ont déjà passé le test seront aussi affichés.

### 5.2.3. Raisonnement à base de cas

Dans la première étape du test adaptatif, nous cherchons à déterminer le niveau d'habileté initial. Puisque nous n'avons comme données que le profil du candidat et les niveaux d'habileté des candidats qui ont déjà passés le test, nous avons pensé à utiliser le raisonnement à base de cas (ce choix est justifié dans la section 3.2.2.1). Nous avons choisi comme *problème*, le profil du candidat dont on cherche à estimer la compétence en anglais. Et comme *solution* le niveau d'habileté en anglais estimé. Ainsi nous passons par toutes les étapes du raisonnement à base de cas à savoir *la recherche, l'adaptation, la maintenance et l'apprentissage* :

*La recherche* : La première étape consiste à choisir les attributs qui déterminent le problème (le profil du candidat). L'expertise d'un professionnel de l'université de Montréal nous a aidés à choisir ces attributs et leurs poids respectifs. Le résultat de notre collaboration est montré dans le tableau 5.1. Ainsi chaque profil candidat est représenté par ces attributs. Un nouveau candidat est considéré comme un nouveau problème dont on cherche la solution qui est son niveau d'habileté en anglais. Celui-ci est obtenu à partir des profils des candidats qui ont déjà passé le test. Ainsi une recherche est faite dans la base des profils de ceux qui se rapprochent le plus à celui du nouveau candidat, et cela en appliquant les métriques de similarité à savoir *l'inverse de la distance euclidienne normalisée* (section 4.5).

**Tableau 5.1 : Les attributs du profil candidat avec leurs poids**

Attributs	Poids
Age	$\sqrt{1/24}$
Pays d'origine	$\sqrt{1/24}$
Profession	$\sqrt{1/24}$
Niveau d'étude complété	$\sqrt{1/24}$
Langue maternelle	$\sqrt{2/24}$
Langue la plus maîtrisée	$\sqrt{3/24}$
Langue apprise (environnement académique)	$\sqrt{3/24}$
Nombre d'années passées dans l'apprentissage d'anglais	$\sqrt{1/24}$
Connaissance de l'anglais dès l'enfance	$\sqrt{2/24}$
Utilisation de l'anglais en milieu de travail	$\sqrt{4/24}$
Autoévaluation du niveau d'anglais	$\sqrt{1/24}$
Motivation (travail, intérêt personnel, passion)	$\sqrt{3/24}$
Aptitude d'apprentissage	$\sqrt{1/24}$

*L'adaptation* : les niveaux d'habileté en anglais des candidats dont les profils les plus proches sont adaptés pour former une solution au nouveau problème. Ainsi la moyenne de ces compétences sera le niveau d'habileté initial en anglais pour le nouveau candidat.

*La maintenance* : le candidat passe le test. Après chaque réponse, le niveau d'habileté du candidat est recalculé ainsi que son erreur type. Une fois le niveau le plus précis est atteint, le test s'arrête et la compétence finale est considérée comme solution au nouveau problème.

*L'apprentissage* : dans l'étape de la recherche, si au moins une des similarités calculées est égale à 1, alors le nouveau profil n'est pas stocké dans la base des cas,

limitant ainsi l'extension de cette dernière. Dans le cas échéant, nous nous contentons de stocker le profil du nouveau candidat ainsi que son niveau d'habileté final sans y faire aucune modification.

### 5.3. Méthodologie de TAARAC

Comme nous avons mentionné auparavant, TAARAC est un test adapté en anglais, avec un raisonnement à base de cas, capable de retrouver le niveau d'habileté du candidat en anglais, après avoir passé un test adapté à son niveau. Afin de réaliser ce système, nous avons passé par les étapes suivantes :

#### 5.3.1. Banque d'items

Dans un test adaptatif, il faut bien choisir nos items, comme mentionné dans la section 3.2.1. En même temps il faut être capable de trouver ses paramètres. Cette dernière procédure nécessite un nombre très large de données.

C'est à partir du test de classement en anglais de la faculté de l'éducation permanente à l'université de Montréal que nous avons extrait nos items. Ces derniers sont indépendants entre eux comme l'exige les techniques du test adaptatif. Le test est sous forme papier crayon que l'on fait passer à tous les candidats qui veulent suivre un cours d'anglais dans cette faculté. Une fois les items choisis, il faut estimer leurs paramètres, notamment le *paramètre de difficulté*, celui de *discrimination* et celui de *pseudo-chance* (section 2.3). Nous avons vu dans la section 2.5.1 qu'il existe plusieurs logiciels qui peuvent faire l'affaire. Nous avons opté pour PARAM-3PL [URL7] car il est gratuit et il est nouveau, selon nos connaissances. C'est une application faite par Lawrence M. Rudner de l'université de Maryland, *Département de mesures statistiques et évaluation* en 2005 et modifiée en 2007. Elle prend comme entrée un fichier de réponses à tous les items pour l'ensemble des candidats (voir figure 5.2), et a pour sortie un fichier des valeurs des paramètres de chaque item (voir figure 5.3). L'estimation y est faite par *la méthode de l'estimation de vraisemblance marginale maximale* (section 2.5.1) en utilisant *la méthode de Newton-Raphson* (section 2.4.1). Une bonne estimation de ces paramètres exige un très grand nombre de candidats. Ainsi à partir d'un

échantillon de 1029 candidats qui ont passé le test en automne 2006 et hiver 2007, comprenant les réponses de chaque candidat pour chacun des 50 items du test, et à l'aide de PARAM-3PL, nous avons obtenus les valeurs des paramètres de difficulté, de discrimination et de pseudo-chance de chacun de ces items.

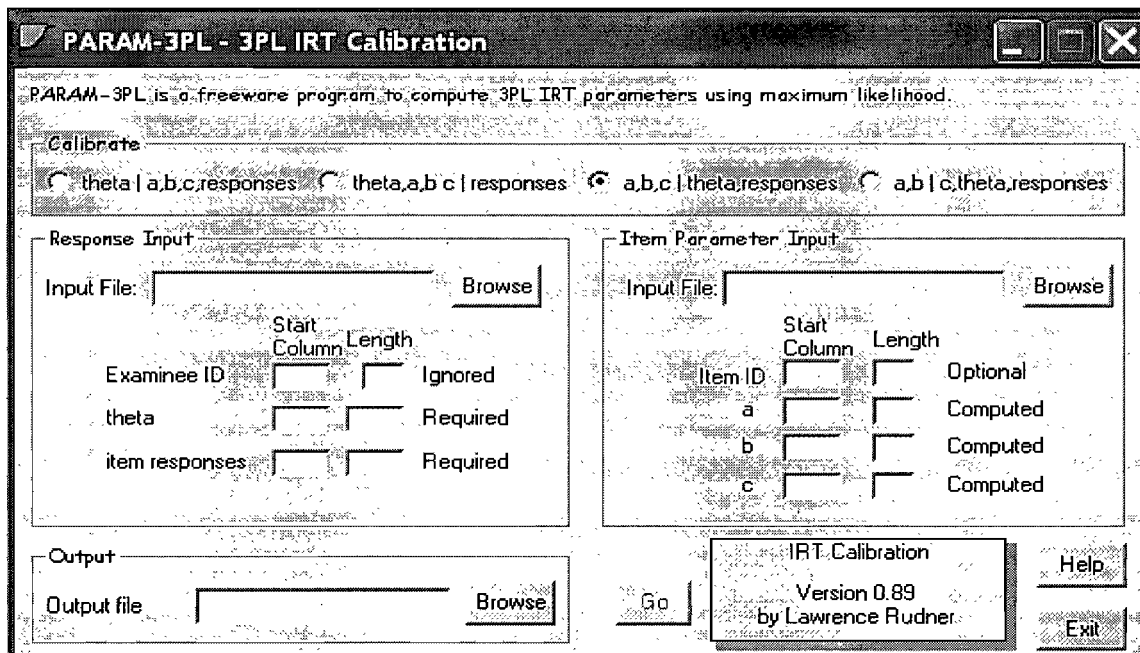


Figure 5.2 : Vue d'écran de PARAM-3PL

Numéro du candidat    réponse du candidat pour chaque item (1 pour une bonne réponse, 0 pour une mauvaise réponse)

testdata - Bloc-notes		
Fichier	Edition	Format Affichage ?
EXAM0001	-1.5841	10001101000001000 00010000010001001001011101
EXAM0002	-1.6890	011001000100000010000101101100100100011000001
EXAM0003	0.4940	1111111111101101111100111111101001110011111100
EXAM0004	-0.9810	100000011000 100000 0010000000101111111 0100
EXAM0005	1.4221	1111111011111111111101111111110111101111111
EXAM0006	0.0353	0110110011100000110100 1011010100100011101111
EXAM0007	0.7333	111011101110011111011010111110101111101011101
EXAM0008	-0.2385	01 01110101010011000000011111010011101111100
EXAM0009	-0.5911	10001110011001011100110000100000010101000000
EXAM0010	-0.6697	101000101110111101000001000110000110100111000
EXAM0011	0.0707	1100111010101011011110000101011010011111101101
EXAM0012	-0.3552	011111101110100011000011100110001110101 00111
EXAM0013	1.0341	01111010110111111110 111111111111111111101
EXAM0014	-0.7209	011110001000010001111000010011100110011101100
EXAM0015	-1.1992	110010101010100001000001010110000010101101010
EXAM0016	0.2684	11111100101110111111001011001110011111101111
EXAM0017	0.6881	0111111011101111100001011101010111011111111
EXAM0018	-0.6200	010100101110100011010001001010100011001101001
EXAM0019	-0.2659	011000101000100001001010010000100101101100100
EXAM0020	0.9098	011111101010111111100111111 100111001111110
EXAM0021	-0.8168	010110001010100010000010001110000011001000100
EXAM0022	-1.5232	000000001001000000110001100000000000101100001

Figure 5.3 : Fichier de données qui représente l'entrée de PARAM-3PL

SeqItem	AccNo	Nobs	a_calib	b_calib	c_calib	seea	seeb	seec
1	,	1983	0.900	0.796	0.197	0.065	0.052	0.014
2	,	1981	1.066	-0.603	0.142	0.055	0.040	0.021
3	,	1984	1.571	-0.360	0.221	0.094	0.033	0.019
4	,	1977	0.840	0.044	0.206	0.054	0.049	0.018
5	,	1981	0.980	-0.878	0.205	0.051	0.046	0.025
6	,	1971	1.296	-0.184	0.192	0.077	0.036	0.017
7	,	1984	1.164	-0.026	0.204	0.071	0.039	0.017
8	,	1976	1.032	2.005	0.135	0.088	0.070	0.009
9	,	1983	1.327	-1.409	0.169	0.070	0.043	0.033
10	,	1983	1.053	0.213	0.318	0.077	0.048	0.017
11	,	1975	1.427	-0.664	0.148	0.076	0.034	0.020
12	,	1978	1.631	1.249	0.150	0.137	0.040	0.010
13	,	1982	0.709	-1.327	0.181	0.035	0.059	0.032
14	,	1980	1.299	-0.036	0.147	0.075	0.034	0.015
15	,	1979	1.077	0.632	0.203	0.077	0.045	0.014
16	,	1981	0.713	0.099	0.259	0.051	0.059	0.018
17	,	1975	1.248	-0.465	0.211	0.071	0.038	0.020
18	,	1985	1.218	-0.899	0.132	0.061	0.038	0.024
19	,	1970	1.760	0.553	0.237	0.133	0.034	0.013
20	,	1978	1.486	0.642	0.209	0.109	0.037	0.013
21	,	1979	0.913	1.125	0.133	0.062	0.052	0.011

**Figure 5.4** : Fichier des paramètres de l'item représentant la sortie de PARAM-3P  
(*a\_calib* est le paramètre de discrimination, *b\_calib* est celui de difficulté et *c\_calib* est celui de pseudo-chance)

Après l'analyse des données obtenues (section 3.2.1), nous avons construit une banque d'items, composée de 46 questions d'anglais, les différentes réponses possibles, les réponses correctes parmi celles-ci ainsi que leurs paramètres de difficulté, de discrimination et de pseudo-chance.

### 5.3.2. Étape de départ

La première des choses à faire dans un test adaptatif est de trouver le niveau d'habileté initial. Comme nous avons mentionné dans la section 5.3.2, nous avons utilisé le raisonnement à base de cas dans cette partie. Ainsi une fois le nouveau candidat entre ses données, qui constituent son profil, un calcul de similarité est fait entre ce dernier et chacun des autres profils dans la base. Ce calcul est fait à l'aide de *l'inverse de la distance euclidienne* de la formule 4.1. À partir des trois profils qui ont donné une plus grande similarité, tout en restant supérieur à un certain seuil prédéterminé, nous

retrouvons le niveau d'habileté initial, qui est la moyenne des niveaux d'habileté finaux des trois profils similaires. Si nous ne trouvons que deux candidats qui sont les plus proches alors le niveau d'habileté initial est la moyenne des niveaux d'habileté finaux des deux candidats. Si seul un candidat est trouvé, alors le niveau d'habileté initial est le niveau d'habileté final du candidat le plus proche. Et si aucun candidat n'est proche au profil du nouveau candidat, alors le niveau d'habileté initial sera la valeur auto évaluée par celui-ci. Cette valeur est ajustée pour qu'elle soit entre -4 et 4. Puisque théoriquement, le niveau d'habileté dans la théorie des réponses aux items est toujours compris entre  $-\infty$  et  $+\infty$  avec une moyenne de 0 (voir la section 2.3). Pratiquement, il est très rare qu'il soit inférieur à -4 ou supérieur à 4 [Bertrand & Blai 2004]. C'est pourquoi nous avons limité le niveau d'habileté dans l'intervalle de [-4,4].

### 5.3.3. Étape de succession

Une fois le niveau d'habileté du candidat initial connu, il faut sélectionner l'item adéquat à présenter au candidat, celui qui maximise la fonction d'information au voisinage du niveau d'habileté initial et minimise le taux d'exposition. Pour réussir cela, nous avons opté pour l'*algorithme a-stratifié avec blocage de b* qui donne un taux d'exposition minimal, cité dans la partie *c* de la section 3.2.2.2. Pour ajouter la notion de maximisation de la fonction de l'information, nous sélectionnons alors l'item qui maximise la fonction de l'information au voisinage du niveau d'habileté initial du candidat, au lieu de choisir l'item dont le paramètre de difficulté est proche de sa compétence initiale. L'algorithme ainsi modifié est présenté ci-dessous :

1. Partitionner la banque d'items en  $M$  blocs selon la valeur du paramètre de difficulté  $b$ .
2. Partitionner chacun des  $M$  blocs en  $K$  niveaux selon le paramètre de discrimination  $a$ .
3. Pour  $k=1, 2, \dots, K$  recombinaison des  $k$ -ième niveaux de chacun des  $M$  blocs pour former une nouvelle strate.
4. Partitionner le test en  $K$  étapes.
5. À la  $k$ -ième étape, sélectionner  $n_k$  items du niveau  $k$  en se basant sur la maximisation de la fonction de l'information au voisinage du niveau d'habileté estimé.
6. Recommencer l'étape 5 pour  $k = 1, 2, \dots, K$ .

❖ *Exemple*

Dans ce cas, nous avons choisi le nombre de bloc  $M=5$  et le nombre de niveaux  $K=5$ . Ainsi après avoir classé les items de la base selon leurs paramètres de difficulté (tableau 5.2). Les 46 items de la base sont partagés en 5 blocs. Quatre des blocs contiennent 9 items et les autres 10. Puis les items de chaque bloc sont classés selon leurs paramètres de discrimination, comme montré dans le tableau 5.3. Chaque bloc est partagé en 5 niveaux. Dans chaque niveau, il y a 2 items, sauf pour les quatre premiers blocs, où le dernier niveau contient un seul item. Et enfin, à chaque niveau du bloc, les items sont combinés pour former 5 nouveaux blocs de 10 items, sauf bien sûr le dernier où il n'y a que 6 (tableau 5.4). Chaque bloc contient à peu près le même nombre d'items car à chaque étape du test, le même nombre d'items est sélectionné du bloc correspondant pour être administré au candidat.

**Tableau 5.2 :** Les items classés selon leurs paramètres de difficulté

Numéro du bloc	Numéro de l'item	P. de discrimination	P. de difficulté
1	9	0,439	-2,419
1	25	0,729	-2,401
1	15	0,586	-1,809
1	27	0,56	-1,788
1	16	0,98	-1,24
1	17	0,361	-1,207
1	6	1,2	-0,988
1	32	1,2	-0,988
1	34	1,661	-0,964
2	7	1,438	-0,936
2	20	1,438	-0,866
2	39	0,908	-0,658
2	46	1,191	-0,567
2	13	1,482	-0,564
2	12	1,441	-0,486
2	30	1,308	-0,469
2	47	1,496	-0,418
2	33	1,281	-0,403
3	26	1,182	-0,353
3	29	0,877	-0,35



**Tableau 5.3 :** Les blocs d'items classés selon le paramètre de discrimination

Numéro du bloc	Numéro de l'item	P. de discrimination	P. de difficulté
1	17	0,361	-1,207
1	9	0,439	-2,419
1	27	0,56	-1,788
1	15	0,586	-1,809
1	25	0,729	-2,401
1	16	0,98	-1,24
1	6	1,2	-0,988
1	32	1,2	-0,988
1	34	1,661	-0,964
2	39	0,908	-0,658
2	46	1,191	-0,567
2	33	1,281	-0,403
2	30	1,308	-0,469
2	7	1,438	-0,936
2	20	1,438	-0,866
2	12	1,441	-0,486
2	13	1,482	-0,564
2	47	1,496	-0,418
3	49	0,687	0,027
3	29	0,877	-0,35

**Tableau 5.4 :** Les blocs d'items formés par ceux du tableau 5.3, en regroupant les items du même niveau

Numéro du bloc	Numéro de l'item	P. de discrimination	P. de difficulté
1	17	0,361	-1,207
1	9	0,439	-2,419
1	39	0,908	-0,658
1	46	1,191	-0,567
1	49	0,687	0,027
1	29	0,877	-0,35
1	18	0,643	0,333
1	21	0,738	0,487
1	41	1,2	1,716
2	3	1,286	0,719
2	27	0,56	-1,788
2	15	0,586	-1,809
2	33	1,281	-0,403
2	30	1,308	-0,469
2	19	0,962	0,112
2	1	1,053	-0,133
2	35	1,043	0,337
2	14	1,335	0,494
3	2	1,287	0,967
3	8	1,496	0,87
3	25	0,729	-2,401
3	16	0,98	-1,24

Les 5 blocs ainsi construits, font en sorte que le test est décomposé en 5 niveaux. À chaque niveau, 4 items sont sélectionnés, puisque le test est en général de longueur 20. D'après le tableau 5.4, nous constatons que dans chaque bloc, les items sont repartis de façon presque égale entre une difficulté minimale et celle maximale. Ainsi quel que soit le niveau d'habileté estimé, l'item approprié sera retrouvé dans chaque bloc. En plus, dans ce dernier, le paramètre de discrimination est presque pareil pour tous les items qui le composent. Ce paramètre augmente d'un niveau à l'autre. Ainsi les items qui sont sélectionnés au début du test ont un paramètre de discrimination faible. Et à la fin du test, ce sont les items dont le paramètre de discrimination est élevé qui sont sélectionnés.

À chaque niveau du test, nous calculons la fonction d'information de tous les items du bloc correspondant, au voisinage du niveau d'habileté estimé. L'item qui maximise cette fonction sera sélectionné et présenté au candidat. Le calcul est fait à l'aide de la formule 2.9 de la section 2.6.

Une fois la réponse du candidat connue, le nouveau niveau d'habileté ainsi que son erreur type sont calculés à l'aide de *la méthode de l'espérance à postériori* de la formule 2.8. Afin de diminuer l'effet de l'erreur type sur les calculs, le niveau d'habileté est ajusté à l'aide de la méthode *d'estimation avec correction adaptative pour biais* [Raïche & Blais 2002] en utilisant la formule de correction du niveau d'habileté suivante :

$$\hat{\theta}_c = \frac{\hat{\theta}}{1 - S_{\hat{\theta}}^2}$$

(Formule 5.1)

$\hat{\theta}$  est le niveau d'habileté estimé.

$\hat{\theta}_c$  est le niveau d'habileté corrigé.

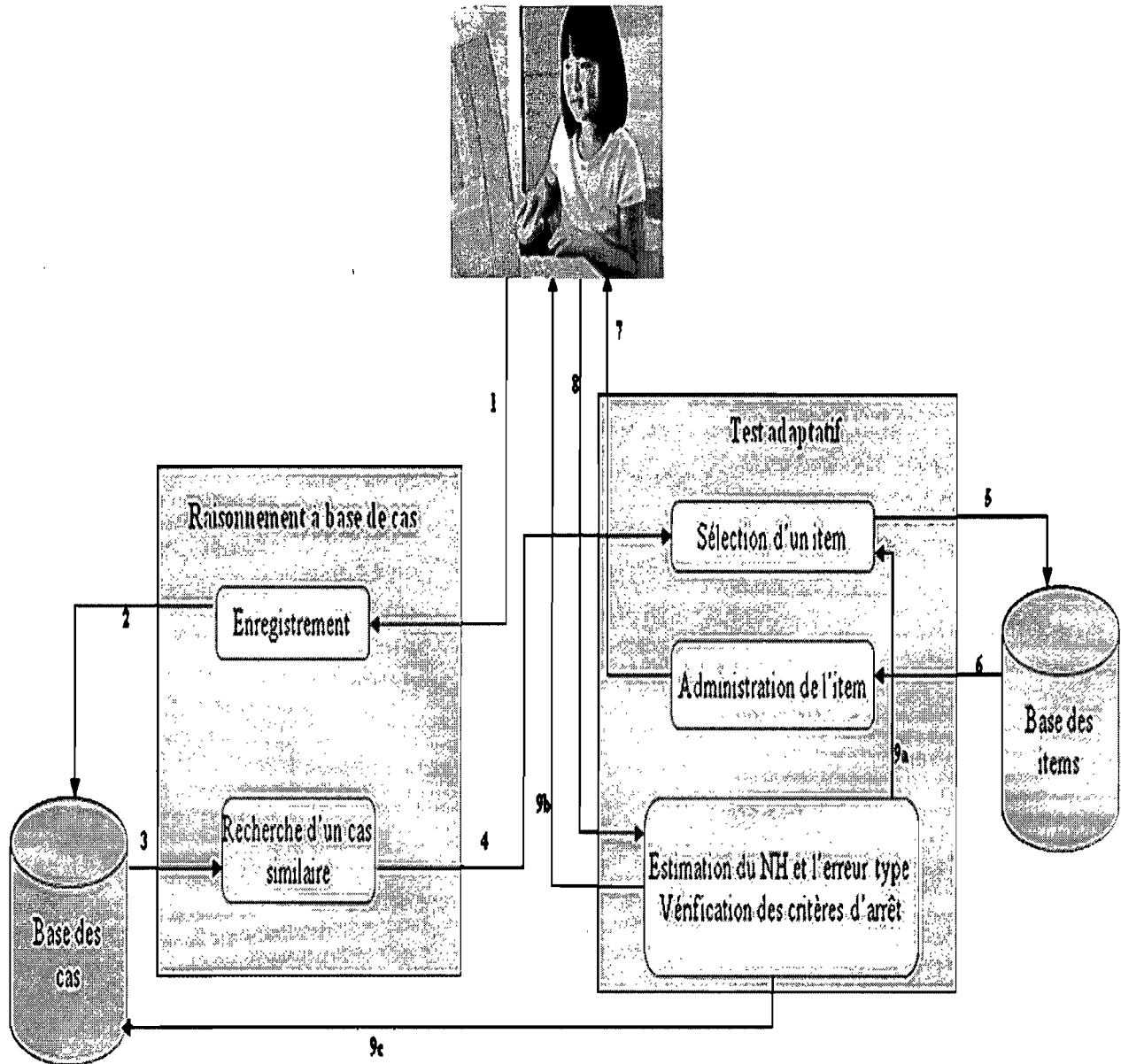
$S_{\hat{\theta}}$  est l'erreur type au niveau de  $\hat{\theta}$ .

#### **5.3.4. Étape d'arrêt**

Dans l'étape précédente, le niveau d'habileté et son erreur type sont calculés. Si ce dernier est plus grand que 0.2 et que nous n'avons pas encore présenté les 20 questions au candidat, un autre item est sélectionné et le niveau d'habileté est recalculé de la même manière que dans la section précédente. Nous avons vu dans la section 3.2.2.3 que dans un test adaptatif, l'administration de 20 items donne une bonne précision dans le calcul du niveau d'habileté, qui pourrait atteindre 0.2. C'est pourquoi nous avons choisi d'arrêter le test, une fois le nombre d'item administré est égale à 20, ou que l'erreur type est de l'ordre de 0.2.

#### **5.4. Architecture de TAARAC**

Dans cette section, nous allons voir l'architecture de TAARAC, et les différents modules qui la composent, notamment *le test adaptatif* et *le raisonnement à base de cas*, ainsi que les différentes bases de données : celle des cas et celle des items. Cette architecture est représentée dans la figure 5.3.



- 1- Données utilisateur
- 2- Profil utilisateur
- 3- Profil similaire
- 4- Estimateur du niveau d'habileté initial
- 5- Code de l'item possédant le maximum d'information

- 6- L'item possédant le maximum d'information
- 7- Affichage de l'item à l'utilisateur
- 8- Réponse de l'utilisateur
- 9a- Estimateur du niveau d'habileté provisoire
- 9b, 9c- Estimateur du niveau d'habileté final

**Figure 5.5 :** Architecture de TAARAC

### 5.4.1. Le module du raisonnement à base de cas

Au début TAARAC invite le candidat à s'enregistrer en entrant ses données personnelles. *L'enregistrement* se fait une seule fois, et en deux étapes. Dans la première, le candidat entre son nom, son prénom, son adresse courriel et un mot de passe. Dans la deuxième étape, il est invité à entrer d'autres informations personnelles, citées dans la section 5.2.2, qui constituent son profil. Si l'enregistrement est déjà fait, le candidat doit s'identifier en entrant son courriel et son mot de passe. Le système le reconnaît, et charge son profil de la base des cas. Son profil est utilisé par le module *recherche d'un cas similaire* afin de faire une recherche, dans la base, des profils qui s'en rapprochent le plus, en utilisant les métriques de similarité. À partir des profils les plus proches, nous retrouvons le niveau d'habileté initial. Le pseudo-code suivant montre de manière explicite, la façon dont on retrouve ce niveau d'habileté initial.

#### ❖ Pseudo-code de l'algorithme de l'étape de départ

Dans le pseudo-code suivant,  $s$  est le seuil que nous avons choisi pour considéré que deux profils sont similaires. Nous avons opté pour  $s=0.7$  car c'est celui qui nous a donné de bons résultats.  $k$  est le nombre des profils proches au candidat. C'est le  $k$  de l'algorithme *k-plus proches voisins*. Nous avons pris  $k=3$ , mais il est à noter qu'il est modifiable en tout temps.

```

Si la base des profils n'est pas vide
| Calculer la similarité entre le profil du candidat  $c$  et chacun des
| profils de la base (voir l'exemple suivant).
| Tant que la similarité est supérieure au seuil  $s$  et le nombre des
| profils similaires n'a pas encore atteint la valeur  $k$ 
|   extraire le niveau d'habileté estimé à partir du profil
|   similaire.
| Fin tant que
| Calculer la moyenne des niveaux d'habileté estimés.
| Le Sauvegarder comme étant le niveau d'habileté initial de  $c$ .
Sinon
| Le niveau d'habileté initial de  $c$  sera son niveau auto-évalué ajusté
| (voir la section 5.3.4).
Fin si

```

Dans l'exemple ci-dessous, nous montrons comment calculer la similarité entre deux profils.

❖ *Exemple 1*

Soient Fatima, Rachid et Marc des candidats dont les profils sont représentés dans le tableau 5.5. Calculons la similarité entre chaque paire de ces trois candidats.

**Tableau 5.5** : Les profils de Fatima, Rachid et Marc.

Profil	Fatima	Rachid	Marc
Age	25	24	30
Pays d'origine	Maroc	Maroc	Canada
Profession	Étudiante	Étudiant	Employé
Niveau d'étude complété	maîtrise	maîtrise	bac
Langue maternelle	Arabe	Arabe	Français
Langue la plus maîtrisée	Arabe	Arabe	Français
Langue apprise (environnement académique)	Français	Anglais	Anglais
Nombre d'années passées dans l'apprentissage d'anglais	4	6	10
Connaissance de l'anglais dès l'enfance	non	non	non
Utilisation de l'anglais en milieu du travail	oui	oui	oui
Autoévaluation du niveau d'anglais	5	6	8
Motivation (travail, intérêt personnel, passion)	travail	travail	passion
Aptitude d'apprentissage	moyenne	moyenne	Très forte

Le calcul de la similarité se fait à l'aide de la formule 4.1. Les poids sont ceux du tableau 5.1. Pour calculer la similarité entre deux candidats, il faut calculer la distance qui existe entre chacun de leurs attributs. Le tableau 5.6 représente les distances entre les attributs de chaque paire des candidats.

**Tableau 5.6 : Les distances des attributs de chaque paire de profils.**

Profil	Dist <sub>Fati, Rachid</sub>	Dist <sub>Rachid, Marc</sub>	Dist <sub>Fatima, Marc</sub>
Age	0.16	1	0.83
Pays d'origine	0	1	1
Profession	0	1	1
Niveau d'étude complété	0	1	1
Langue maternelle	0	1	1
Langue la plus maîtrisée	0	1	1
Langue apprise (environnement académique)	0	0	0
Nombre d'années passées dans l'apprentissage d'anglais	0.33	0.66	1
Connaissance de l'anglais dès l'enfance	0	0	0
Utilisation de l'anglais en milieu du travail	0	0	0
Autoévaluation du niveau d'anglais	0.5	0.5	1
Motivation	0	1	1
Aptitude d'apprentissage	0	1	1

La distance entre deux profils, pour un attribut donné, se calcule comme dans la section 4.5. Ainsi :

$$Dist_{age}(Fatima, Marc) = |Age_{Fatima} - Age_{Marc}| / |Max_{Age} - Min_{Age}| = |25 - 30| / |30 - 24| = 5/6 = 0.83$$

$$Dist_{niveau\ d'\acute{e}tudes}(Fatima, Rachid) = 1 \text{ car } Niveau\ d'\acute{e}tudes_{Fatima} \neq Niveau\ d'\acute{e}tudes_{Rachid}$$

$$Dist_{niveau\ d'\acute{e}tudes}(Rachid, Marc) = 0 \text{ car } Niveau\ d'\acute{e}tudes_{Rachid} = Niveau\ d'\acute{e}tudes_{Marc}$$

$Dist_{langue\ apprise}(Fatima, Rachid) = 0$  même si la langue apprise par Fatima est le français, et celle apprise par Rachid est l'anglais. Puisque ces deux langues s'écrivent par des lettres latines, nous les considérons comme égales. D'après l'expert, ce qui nous

intéresse ici, à propos d'une langue, est de savoir si celle-ci est latine ou non. Car en connaissant une ou plusieurs langues latines, on peut apprendre facilement l'anglais.

Puisque  $SIM(Fatima, Rachid) = 1 - D(Fatima, Rachid)$ , alors calculons d'abord  $D(Fatima, Rachid)$ .

$$D(Fatima, Rachid) = \sqrt{\sum_{i=1}^{13} w_i^2 dist^2(x_i, y_i)} =$$

$$\sqrt{\frac{1}{24} \times 0.16^2 + \frac{1}{24} \times 0 + \frac{1}{24} \times 0 + \frac{1}{24} \times 0 + \frac{2}{24} \times 0 + \frac{3}{24} \times 0 + \frac{3}{24} \times 0 + \frac{1}{24} \times 0.33^2 + \frac{2}{24} \times 0 + \frac{4}{24} \times 0 + \frac{1}{24} \times 0.5^2 + \frac{3}{24} \times 0 + \frac{1}{24} \times 0 =}$$

$$D(Fatima, Rachid) = \sqrt{\frac{0.37}{24}} = 0.12 \quad \text{Ainsi :}$$

$SIM(Fatima, Rachid) = 1 - 0.12 = 0.88$ . De même  $SIM(Rachid, Marc) = 0.28$  et  $SIM(Fatima, Marc) = 0.25$ .

Nous constatons alors que c'est le profil de Fatima et celui de Rachid qui donnent la plus grande similarité, ainsi se sont eux qui sont les plus proches parmi les trois.

Une fois le niveau d'habileté initial trouvé, il sera transmis au module suivant, celui du *test adaptatif*. En fait, le test peut commencer.

#### 5.4.2. Le module du test adaptatif

Dans ce module, nous montrons, comment choisir la question adéquate, (*sélection de l'item*), comment la présenter au candidat (l'administration de l'item), comment calculer le niveau d'habileté et son erreur type, et à quel moment le test doit s'arrêter.

Avant de passer à *la sélection de l'item*, comme nous avons expliqué dans la section 5.3.3, il faut décomposer la banque d'items en  $M$  blocs, afin de minimiser le taux d'exposition de l'item. Le pseudo-code suivant décrit le processus utilisé pour réaliser cette décomposition.



❖ Pseudo-code de la décomposition de la banque d'items

Début

Trier les items selon leurs paramètres de difficulté.  
 Décomposer ces items en  $M$  blocs.  
 Trier chaque bloc selon le paramètre de discrimination.  
 Décomposer chaque bloc en  $K$  niveaux.  
 Combiner chaque niveau pour former de nouveaux blocs.

Fin

$M$  et  $K$  sont respectivement, le nombre de blocs et le nombre de niveaux. Nous avons choisi de donner à ces deux paramètres la valeur 5. De ce fait, chaque bloc contient le maximum d'items possibles, et par conséquent un large choix d'item.

Les  $K$  blocs des items ainsi construits, le test est partagé, par conséquent, en  $K$  niveau. À chaque niveau,  $20/K$  items seront sélectionnés. *La sélection de l'item* le plus approprié, à présenter au candidat, est montré de manière explicative dans le pseudo code suivant.

❖ Pseudo-code de la procédure de sélection de l'item

Si le nombre des items alloués au candidat jusqu'à présent est inférieur à 20 et l'erreur type est supérieur à 0.2.  
 Trouver le niveau courant  $N$ , étant donné le nombre des items présentés au candidat.  
 Calculer la fonction d'information pour tous les items du niveau  $N$ , au voisinage du niveau d'habileté initial, à l'aide de la formule 2.9.  
 Présenter au candidat l'item dont la fonction d'information au voisinage de son niveau d'habileté estimé est maximale.  
 Supprimer cet item du tableau des blocs des items.

Sinon

Arrêter le test.

Fin si

La fonction d'information se calcule à l'aide de la formule 2.9. L'exemple 2 donne une idée sur la démarche suivie afin de la calculer.

❖ Exemple 2

Soit un candidat dont le niveau d'habileté initial est de l'ordre de 50%, soit 0 sur l'échelle de -4 et 4. On cherche à calculer la fonction d'information de l'item 9, par

exemple, au voisinage de 0. Cet item a pour paramètres, de difficulté -2.419, de discrimination 0.439 et de pseudo chance 0 (voir la section 5.3.1).

Dans un modèle à trois paramètres, la fonction d'information de la formule 2.9, déjà vu dans le chapitre 2 a la forme suivante :  $I_i(\theta) = D^2 a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[ \frac{P_i(\theta) - c_i}{(1 - c_i)} \right]^2$

$$\text{Avec } P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}}, \quad Q_i(\theta) = 1 - P_i(\theta), \quad D = 1.7$$

$a_i$ ,  $b_i$ , et  $c_i$  sont respectivement les paramètres de discrimination, de difficulté et de pseudo-chance.

En substituant ces variables par les valeurs citées précédemment, les calculs auront l'allure suivante :  $P_9(0) = 0 + \frac{1 - 0}{1 + e^{-1.7 \times 0.439(0 + 2.419)}} = 0.858$  puis  $Q_9(0) = 1 - 0.858 = 0.142$

$$\text{Par conséquent } I_9(0) = 1.7^2 \times 0.439^2 \times \frac{0.142}{0.858} \times \left[ \frac{0.858 - 0}{(1 - 0)} \right]^2 = 0.067$$

Après avoir sélectionné l'item approprié, on le présente au candidat. C'est la phase de *l'administration de l'item*. La question et ses différentes réponses sont ainsi affichées au candidat. Ce dernier a un délai de 60 secondes pour répondre, sinon la réponse sera considérée fautive. La notion du temps est ajoutée ici afin d'éviter des tests trop longs. La position des différentes réponses est modifiée, avant chaque affichage, afin de minimiser les chances à un candidat, de se rappeler des réponses à partir de leurs positions.

Une fois la réponse du candidat connue, il faut calculer le niveau d'habileté, ainsi que l'erreur type. Ces derniers sont calculés à l'aide de la méthode de *l'espérance à posteriori*, selon la formule 2.8. L'exemple 3 montre de manière explicative, comment ces deux données sont calculées.

❖ Exemple 3

Supposons que nous ayons présenté jusqu'ici, trois questions au candidat, les items, 33, 21 et 39 dont les paramètres sont représentés dans le tableau 5.7, et que les réponses du candidat sont respectivement, 1, 0 et 1.

**Tableau 5.7 :** Les paramètres des items, 33, 21 et 39.

		Item 33	Item 21	Item 39
Paramètre de difficulté	de	-0.403	0.487	-0.658
Paramètre de discrimination	de	1.281	0.738	0.908
Paramètre de pseudo-chance	de	0.026	0.113	0.084

D'après la formule 2.8, le niveau d'habileté estimé et l'erreur type sont respectivement :

$$E(\theta|1,0,1) = \frac{\sum_{j=1}^{30} \theta_j L(1,0,1|\theta_j) f(\theta_j)}{\sum_{j=1}^{30} L(1,0,1|\theta_j) f(\theta_j)}, \quad S_{EAP}(\theta|1,0,1) = \left[ \frac{\sum_{j=1}^{30} (\theta_j - E(\theta))^2 L(1,0,1|\theta_j) f(\theta_j)}{\sum_{j=1}^{30} L(1,0,1|\theta_j) f(\theta_j)} \right]^{1/2}$$

avec  $\theta_j$  est un des 30 points de quadrature et  $f(\theta_j)$  est le poids associé à ce point, soit la densité de probabilité normale calculée à ce point.

$L(1,0,1|\theta_j) = P_{33}(\theta_j)^1 Q_{33}(\theta_j)^0 \times P_{21}(\theta_j)^0 Q_{21}(\theta_j)^1 \times P_{39}(\theta_j)^1 Q_{39}(\theta_j)^0$  est la fonction de vraisemblance.

Soit  $L(1,0,1|\theta_j) = P_{33}(\theta_j) \times Q_{21}(\theta_j) \times P_{39}(\theta_j)$ , avec  $P_i(\theta_j)$  est la probabilité de réussite de l'item  $i$  avec le niveau d'habileté  $\theta_j$  (voir formule 2.4) et  $Q_i(\theta_j) = 1 - P_i(\theta_j)$ .

Le détail des calculs est montré dans le tableau 5.8.

**Tableau 5.8 :** Le récapitulatif des étapes du calcul de niveau d'habileté.

	Les points de quadratures ( $\theta_j$ )	Densité $f(\theta_j)$	$P_{33}(\theta_j)$	$P_{33}(\theta_j) \times Q_{21}(\theta_j)$	$P_{33}(\theta_j) \times Q_{21}(\theta_j) \times P_{33}(\theta_j)$	$\theta_j L(1,0,1 \theta_j) f(\theta_j)$	$L(1,0,1 \theta_j) f(\theta_j)$	$(\theta_j - E(\theta))^2 L(1,0,1 \theta_j) f(\theta_j)$
	-4,000	0,000	0,026	0,023	0,004	0,000	0,000	0,000
	-3,724	0,000	0,027	0,024	0,004	0,000	0,000	0,000
	-3,448	0,001	0,027	0,024	0,004	0,000	0,000	0,000
	-3,172	0,001	0,028	0,025	0,004	0,000	0,000	0,000
	-2,897	0,002	0,030	0,026	0,005	0,000	0,000	0,000
	-2,621	0,004	0,034	0,029	0,005	0,000	0,000	0,000
	-2,345	0,007	0,040	0,034	0,006	0,000	0,000	0,000
	-2,069	0,013	0,051	0,044	0,008	0,000	0,000	0,001
	-1,793	0,021	0,071	0,060	0,011	0,000	0,000	0,001
	-1,517	0,033	0,105	0,086	0,017	-0,001	0,001	0,002
	-1,241	0,050	0,161	0,128	0,029	-0,002	0,001	0,004
	-0,966	0,067	0,247	0,189	0,052	-0,003	0,003	0,006
	-0,690	0,086	0,366	0,264	0,097	-0,006	0,008	0,009
	-0,414	0,101	0,507	0,340	0,173	-0,007	0,017	0,010
	-0,138	0,109	0,650	0,396	0,267	-0,004	0,029	0,007
	0,138	0,109	0,771	0,415	0,339	0,005	0,037	0,002
	0,414	0,102	0,859	0,399	0,361	0,015	0,037	0,000
	0,690	0,088	0,918	0,355	0,340	0,021	0,030	0,003
	0,966	0,071	0,953	0,299	0,293	0,020	0,021	0,008
	1,241	0,053	0,974	0,242	0,239	0,016	0,013	0,010
	1,517	0,036	0,985	0,188	0,188	0,010	0,007	0,009
	1,793	0,022	0,992	0,143	0,143	0,006	0,003	0,007
	2,069	0,012	0,996	0,107	0,107	0,003	0,001	0,004
	2,345	0,006	0,998	0,078	0,078	0,001	0,000	0,002
	2,621	0,003	0,999	0,057	0,057	0,000	0,000	0,001
	2,897	0,001	0,999	0,041	0,041	0,000	0,000	0,000
	3,172	0,001	1,000	0,030	0,030	0,000	0,000	0,000
	3,448	0,000	1,000	0,021	0,021	0,000	0,000	0,000
	3,724	0,000	1,000	0,015	0,015	0,000	0,000	0,000
	4,000	0,000	1,000	0,011	0,011	0,000	0,000	0,000
Somme:						0,073	0,210	0,086

$$\text{Donc } E(\theta|1,0,1) = \frac{0.073}{0.210} = 0.347 \text{ et } S_{EAP}(\theta|1,0,1) = \left[ \frac{0.086}{0.210} \right]^{1/2} = 0.639$$

Si l'erreur type obtenue est supérieure à 0.2 et le test n'a pas encore atteint une longueur de 20, on passera encore une autre fois à la *sélection de l'item* tout en substituant dans son pseudo code, le niveau d'habileté initial par celui calculé. Alors que

si le nombre des items présentés est 20 ou la précision du niveau d'habileté est inférieure à 0,2, le test s'arrête et le niveau d'habileté du candidat est affiché sur l'écran. Après l'avoir ajusté, en le passant de l'échelle -4 à 4, à l'échelle 0 à 100 à l'aide de la formule 5.2 suivante.

$$NA = 12,5 \times \theta + 50$$

$\theta$  est le niveau d'habileté estimé dans l'échelle de -4 à 4. (Formule 5.2)

$NA$  est le niveau d'habileté ajusté dans une échelle de 0 à 100.

En plus du niveau d'habileté affiché au candidat, ce dernier peut voir la correction du test qu'il vient de passer, ainsi qu'un histogramme qui représente sa compétence en anglais par rapport aux celles des autres candidats qui ont déjà passé le test.

Afin de mieux comprendre le déroulement d'un test adaptatif, nous présentons dans ce qui suit, un vrai scénario d'un test adaptatif passé par un candidat appelé Marc.

#### ❖ Scénario

Marc est un candidat qui a passé notre test. Nous montrons ici la procédure utilisée, afin de trouver son niveau d'habileté en Anglais.

Marc possède un niveau d'habileté initial en Anglais égale à 0, soit 50%. Après la décomposition de la banque d'items en 5 blocs, comme montré précédemment, le test est décomposé aussi en 5 niveaux. À chaque niveau nous sélectionnons 4 items.

–Le 1<sup>er</sup> niveau qui correspond au 1<sup>er</sup> bloc est constitué des items dont les numéros sont les suivants : 9, 17, 39, 33, 29, 49, 18, 21, 2, 23.

- Ces items sont classés selon leur fonction d'information au voisinage du niveau d'habileté initial, soit 0. On obtient ainsi, en ordre décroissant, les items : 33, 39, 49, 21, 29, 18, 17, 9, 2, 23.
- L'item 33 possède le maximum de la fonction d'information au voisinage de 0. Ainsi, c'est celui-ci qui est présenté à Marc ( $I_1=33$ ).

- Marc répond correctement à cet item ( $u_1=1$ ).
  - À l'aide de la méthode de l'espérance à posteriori, comme montré dans l'exemple 3, le niveau d'habileté  $\theta_1$  et son erreur type  $S_1$  sont calculés, ainsi que le niveau d'habileté corrigé  $\theta_{c1}$  en utilisant la formule 5.1. Soit :  $\theta_1=0.445$ ,  $S_1=0.807$  et  $\theta_{c1}=1.30$ .
  - Après suppression de l'item 33, les autres qui restent dans le 1<sup>er</sup> bloc sont eux aussi classés selon leur fonction d'information au voisinage de 1.30 en ordre décroissant : 2, 21, 49, 18, 39, 17, 29, 9, 23.
  - De même, l'item 2 sera présenté à Marc ( $I_2=2$ ).
  - Sa réponse est encore correcte, soit  $u_2=1$ .
  - Après le calcul, comme montré précédemment,  $\theta_2=0.799$ ,  $S_2=0.807$  et  $\theta_{c2}=2.29$ .
  - L'item 2 est supprimé aussi du 1<sup>er</sup> bloc. Les items qui restent sont classés selon leur fonction d'information au voisinage de 2.29. Soit 21, 49, 18, 17, 39, 9, 29, 23.
  - L'item 21 sera présenté alors à Marc ( $I_3=21$ ).
  - Cette fois ci, sa réponse est incorrecte ( $u_3=0$ ).
  - Une fois les calculs sont faits,  $\theta_3=0.407$ ,  $S_3=0.749$  et  $\theta_{c3}=0.929$ .
  - Les items restants après suppression du numéro 21, sont classés de la même manière que précédemment. Nous avons ainsi 49, 39, 18, 17, 29, 9, 23.
  - L'item 49 est présenté à Marc ( $I_4=49$ ).
  - La réponse est correcte, donc  $u_4=1$ .
  - Les calculs sont faits,  $\theta_4=0.975$ ,  $S_4=0.755$  et  $\theta_{c4}=2.269$ .
- Passons maintenant au 2<sup>ème</sup> niveau. Le bloc correspondant contient les items suivants : 27, 6, 46, 30, 26, 19, 35, 3, 8, 41.

- De même, ces items sont classés selon leur fonction d'information au voisinage de 2.269, soit 41, 3, 8, 27, 19, 35, 26, 46, 6, 30.
- L'item 41 est présenté à Marc ( $I_5=41$ ).
- Sa réponse est correcte ( $u_5=1$ ).
- Les calculs nous donnent,  $\theta_5=1.278$ ,  $S_5=0.761$  et  $\theta_{c5}=3.039$ .
- Ainsi de suite pour les autres items et les autres niveaux jusqu'à la fin du test.

–Le test s'arrête lorsque le nombre des items atteint 20, puisque l'erreur type est supérieure à 0.2. Le niveau d'habileté de Marc en Anglais a ainsi la valeur de  $\theta_{c20}=0.825$ , soit **60.8%** d'après la formule 5.2, avec une erreur type  $S_5=0.256$ .

De manière plus claire, un résumé de ce scénario est montré dans le tableau 5.9.

**Tableau 5.9 : Le récapitulatif du test de Marc.**

Niveau	N° item	P. Discrimination	P. difficulté	P. pseudo-chance	Réponse	EAP	Erreur Type
1	33	1,281	-0,403	0,026	1	1,307	0,807
1	2	1,287	0,967	0,222	1	2,295	0,807
1	21	0,738	0,487	0,113	0	0,929	0,749
1	49	0,687	0,027	0,111	1	2,269	0,755
2	41	1,2	1,716	0,211	1	3,039	0,761
2	3	1,286	0,719	0,113	1	2,668	0,66
2	8	1,496	0,87	0,23	0	1,354	0,564
2	26	1,182	-0,353	0,181	0	0,581	0,588
3	14	1,335	0,494	0,085	1	0,863	0,509
3	10	1,617	1,465	0,083	0	0,549	0,467
3	24	1,5	0,816	0,229	1	0,717	0,444
3	5	1,537	0,256	0,289	1	0,794	0,407
4	45	2	0,992	0,213	0	0,632	0,38
4	42	1,88	0,781	0,272	0	0,486	0,367
4	22	2	0,381	0,368	1	0,561	0,34
4	20	1,438	-0,866	0	1	0,573	0,325
5	4	2	0,683	0,18	1	0,68	0,298
5	43	1,947	0,872	0,162	1	0,778	0,278
5	50	2	0,471	0,244	1	0,823	0,258
5	16	0,98	-1,24	0	1	0,827	0,256

Après avoir expliqué le déroulement dans chacun de nos modules, nous présentons, dans ce qui suit, nos bases de données, notamment, celle des cas, et celle des items.

### 5.4.3. Base des cas

Comme nous avons vu précédemment, un cas est composé d'un ensemble d'attributs que le candidat entre une seule fois, lors de son enregistrement, ainsi que le *niveau d'habileté calculé* à la fin de chaque test. En plus des attributs qui composent le cas, d'autres données sont stockées dans la base pour assurer le bon fonctionnement du test. Notamment :

*Le niveau d'habileté initial*, qui change au fur et à mesure que le candidat repasse le test. Puisque avant de commencer le test, le système fait une nouvelle recherche dans la base des cas pour y extraire cette donnée ci.

*Les données du test*, nous en aurons besoin pour afficher la correction au candidat, et aussi pour construire l'histogramme de comparaison entre la compétence du candidat qui vient de passer le test et les autres qui l'ont déjà fait. Les attributs du test nécessaires sont :

- *idcandid* : le numéro du candidat qui passe le test.
- *iditem* : le numéro de l'item présenté au candidat courant.
- *candresp* : la réponse du candidat courant à l'item.
- *hablevel* : le niveau d'habileté du candidat courant calculé après sa réponse à l'item.
- *errortyp* : l'erreur type correspondant à ce niveau d'habileté.
- *numtest* : le nombre de fois que le candidat courant passe le test.
- *timeresp* : le temps écoulé avant la réponse du candidat courant à l'item. Il est toujours inférieur ou égale à 60 secondes, sinon la réponse est considérée comme fausse.

Qui dit test, dit un ensemble de questions. Ainsi il nous faut une base où nous stockons les items du test, appelée, base d'items.



#### 5.4.4. Base d'items

La base d'items est composée d'un ensemble d'attributs, nécessaires pour le déroulement du test et dans le calcul du niveau d'habileté. Ces attributs sont définis comme ce qui suit :

- *iditem* : le numéro de l'item.
- *textitem* : le texte de l'item affiché au candidat.
- *Respons<sub>i</sub>* : les différentes réponses de l'item que nous affichons au candidat, puisque toutes les questions sont à choix multiple,  $i=1...4$ .
- *trueresp* : La réponse correcte parmi les réponses affichées.
- *param\_a* : paramètre de discrimination de l'item.
- *param\_b* : paramètre de difficulté de l'item.
- *param\_c* : paramètre de pseudo chance de l'item.

Jusqu'à présent, nous avons 46 questions d'anglais dans notre base d'items. Mais, cette dernière peut être enrichie par d'autres items d'anglais, à condition de les avoir préparés avant, comme mentionné dans la section 3.2.1.

Pour des raisons de confidentialité, l'Université de Montréal ne nous a pas permis de divulguer les questions, c'est pourquoi nous n'avons pas donné des exemples des items se trouvant dans la base.

Cette base peut être chargée par les items d'un autre concept, mais il faut changer la deuxième partie de *l'enregistrement*. C'est-à-dire supprimer les attributs liés à l'anglais et les remplacer par ceux qui dépendent du nouveau concept.

### 5.5. Comparaison

Dans cette section, nous allons faire une comparaison entre TAARAC et quelques systèmes qui utilisent le test adaptatif sur ordinateur. Notant par exemple GMAT (Graduate Management Admission Test) [URL8], un test utilisé à travers le monde pour évaluer les compétences verbales, mathématiques et de rédaction analytique. Ces aptitudes qui sont essentielles pour la réussite des études dans un

programme de gestion, tel le MBA (Master of Business Administration). Depuis sa création en 1953, Ce test est développé et administré par l'ETS (Educational Testing Service) jusqu'à la fin de 2005. À partir du 1<sup>er</sup> janvier 2006, ces deux tâches sont partagées entre ACT Inc. et Person Vue [URL9]. Nous avons choisi dans cette comparaison, la partie mathématique du test qui est adaptative.

Le TOEFL (the Test Of English as a Foreign Language) [URL5] est un outil d'évaluation de l'anglais universitaire pour les candidats dont l'anglais n'est pas la langue maternelle. C'est un test qui est reconnu à l'échelle internationale, créé en 1964 par l'ETS. Actuellement, il est devenu un *IBT test* (Internet based test), c'est à dire accessible par Internet. Il n'est plus adaptatif, car l'évaluation des sections qui le composent ne se fait plus de manière séparée, et les questions posées regroupent souvent deux à trois sections. Dans cette comparaison, Nous avons choisi le TOEFL comme il était auparavant, *CBT test* (Computer Based Test), et plus précisément, sa section de structure qui est adaptative. Et finalement, MCSE (Microsoft Certified Systems Engineer) une certification des spécialistes de Windows 2000 et Windows server 2003 de Microsoft [URL10].

Comme TAARAC, ces tests utilisent tous la théorie des réponses aux items, avec un modèle logistique à trois paramètres. La comparaison est faite sur ces thèmes-ci :

- *Le concept utilisé* : quels sont les concepts à évaluer par le test.
- *L'estimation du niveau d'habileté initial* : la procédure utilisée pour estimer le niveau d'habileté initial.
- *Le premier item alloué* : quelles sont les caractéristiques du premier item présenté.
- *Type des questions* : quel est le type des items constituant le test.
- *La sélection de l'item* : quelles sont les techniques utilisées au cours de la sélection de l'item.
- *Les critères d'arrêt* : quels sont les critères choisis afin d'arrêter le test.
- *La longueur de test* : le nombre maximal des items alloués au candidat.

- *Le test peut il être suspendu?* : est ce que le test peut être arrêté et repris là où il est suspendu?
- *Le feedback* : après le test, quelles sont les différentes fonctions présentées par le système.

**Tableau 5.10** : Comparaison des tests adaptatifs.

	GMAT	TOEFL	MCSE	TAARAC
Le concept utilisé	Mathématiques	Anglais	Informatique	Anglais
Type des questions	À choix multiple (5 choix)	À choix multiple (4 choix)	Glisser et déposer (drag and drop)	À choix multiple (4 choix)
L'estimation du niveau d'habileté initial	N'existe pas	N'existe pas	N'existe pas	Un raisonnement à base de cas
Le premier item alloué	De difficulté moyenne	De difficulté moyenne	De difficulté faible ou moyenne	Selon l'algorithme de sélection de l'item
La sélection de l'item	Maximum de la fonction d'information et minimum du taux d'exposition.	Maximum de la fonction d'information et minimum du taux d'exposition.	ND	Maximum de la fonction d'information et minimum du taux d'exposition.
Les critères d'arrêt	Nombre maximal d'items est atteint.	Nombre maximal d'items est atteint.	Nombre maximal d'items est atteint ou une certaine précision est atteinte.	Nombre maximal d'items est atteint ou une certaine précision est atteinte.
La longueur maximale du test	37	20 à 25	35	20
Le test peut il être suspendu?	Non	Non	Non	Oui
Le feedback	Le score du candidat. Pas de correction et pas de comparaison.	Le score du candidat. Pas de correction et pas de comparaison	Le score du candidat et s'il a passé ou non. la correction et pas de comparaison.	Le score du candidat, sa comparaison avec ceux des autres candidats qui ont déjà passé le test et la correction du test.

❖ *Interprétation du tableau*

Même si deux des trois tests choisis pour cette comparaison traitent un sujet différent que celui de TAARAC, ce sont tous des tests adaptés selon la compétence du candidat. Ainsi les items constituant leurs banques des items sont différents dépendamment du thème traité. Mais il reste que leurs types est le même. C'est à dire qu'ils sont tous à choix multiple. Sauf ceux de MCSE qui sont de type « glisser et déposer ».

Commè montré dans le tableau 5.10, la grande différence entre TAARAC et les autres tests réside dans le fait que TAARAC est capable d'estimer un niveau d'habileté initial au candidat en utilisant un raisonnement à base de cas (voir la section 5.2.3). Et par conséquence, le premier item est sélectionné de la même manière que les autres, c'est-à-dire, à l'aide de l'algorithme de sélection de l'item. Ainsi le premier item sera celui dont la difficulté est proche à la compétence du candidat. Alors que les autres tests choisissent toujours comme premier item celui avec une difficulté moyenne ou faible. Un choix qui allongera le test, puisque il est nécessaire d'ajouter d'autres items au début du test avant de tomber sur celui dont la difficulté se rapproche de la compétence du candidat.

La sélection de l'item alloué au candidat se fait dans TAARAC, de la même manière que dans tous les autres tests internationaux, en choisissant celui qui maximise la fonction d'information au voisinage du niveau d'habileté du candidat tout en minimisant son taux d'exposition des items (le nombre de fois qu'un item est administré sur le nombre total des candidats). Chacun de ces tests utilise une technique différente que celle utilisée dans TAARAC pour réaliser ceci.

Le test s'arrête une fois un nombre fixe des items est présenté au candidat, qui est 37 pour GMAT et 20 à 25 pour la section de structure du TOEFL. Alors que dans MCSE et TAARAC, un autre critère est ajouté pour que le test s'arrête, c'est l'obtention d'une certaine précision dans le calcul du niveau d'habileté du candidat. Ainsi, si ce critère est atteint, le test s'arrête, même si le nombre maximal des items n'est pas encore présenté au candidat. Ce nombre est de l'ordre de 35 pour MCSE et 20 pour TAARAC. Comme

nous le constatons, c'est seulement avec TAARAC que nous pouvons estimer un niveau d'habileté avec un test très court.

Un autre point fort de TAARAC, est que ce dernier peut être suspendu à tout moment par le candidat. Lors de sa prochaine connexion, ce candidat est invité à continuer le test là où il l'a arrêté, ou le reprendre depuis le début. Ce qu'on ne peut pas faire dans les autres tests.

À la fin du test, TAARAC affiche le score du candidat. Ce dernier peut même voir la correction de son test, en plus d'un histogramme qui représente une comparaison entre son ou ses scores et ceux des candidats ayant déjà passé le test. Alors que dans les autres tests, le candidat n'est capable de voir que son score, sauf MCSE qui peut en plus, annoncer au candidat sa réussite ou son échec et lui afficher une correction sur demande.

Comparé aux trois tests, TAARAC reste faible coté nombre de données, puisque sa banque d'items ne contient que 46 items, alors que les autres tests possèdent des banques avec un nombre très large d'items, ce nombre croît au fil des années. Or la base d'items peut être enrichie en tout moment. Nous ne l'avons pas fait dans ce travail car, comme nous avons vu dans la section 5.3.1, une bonne estimation des paramètres de l'item exige la réponse de plusieurs candidats à cet item dans un test papier-crayon. Des données qui ne sont pas faciles à avoir puisque la plupart des responsables des tests ne sauvegardent que les scores totaux obtenus dans tout le test. En ce qui concerne les données collectées pour construire notre base, il nous fallait attendre deux sessions afin de les rassembler au fur et à mesure que les candidats passent le test papier-crayon. Autrement elles seront détruites dès l'obtention des scores.

## **5.6. Conclusion**

Nous avons expliqué dans ce chapitre, la méthodologie utilisée pour la conception de TAARAC, l'algorithme de son déroulement, l'apport du raisonnement à base de cas pour notre système et les différentes étapes de notre test adaptatif. Nous avons vu aussi son architecture en expliquant en détail les différents modules qui la composent. Ainsi TAARAC est un test adaptatif en anglais, capable de trouver le niveau d'habileté d'un candidat avec un nombre minimal de questions. Afin d'atteindre cet

objectif, nous avons utilisé *le raisonnement à base de cas* ainsi que *la théorie des réponses aux items*. À la fin de ce chapitre nous avons fait une comparaison entre TAARAC et trois des tests les plus connus au niveau international, tout en insistant sur les points forts de TAARAC.

Dans le chapitre suivant nous allons montrer les différentes technologies utilisées dans TAARAC, et expliquer en détail son implémentation.

## Chapitre 6 : Implémentation et validation

Dans ce chapitre nous allons voir les techniques utilisées dans l'implémentation et le développement de TAARAC, les outils et les langages de programmation choisis afin de réaliser ce système. Une vue bien détaillée de TAARAC est montrée à travers des captures d'écran.

### 6.1. Implémentation de TAARAC

TAARAC a été développé par PHP (*Hypertext Preprocessor*) version 5.0.4, un langage de développement d'applications web, facile à intégrer dans les pages HTML. Sous le système d'exploitation Windows XP et le serveur web Apache Tomcat version 2.0.54. Le développement est fait sous l'environnement Macromedia Dreamweaver 8 version 8.0. La base de données a été conçue à l'aide du système de gestion de base de données MySQL version 4.1.20.

TAARAC est une application web avec une architecture client-serveur. C'est-à-dire que le client envoie une requête à travers le réseau vers le serveur web et ce dernier la traite et envoie la réponse au client. Le serveur peut être scindé en un serveur application et un serveur base de données. C'est ce qu'on appelle l'architecture trois-tiers.

#### 6.1.1. Architecture trois-tiers

Une application développée sous l'architecture trois-tiers vise à être séparée logiquement et de manière très nette en trois couches ou niveaux, *présentation*, *traitement* ou *métier* et *accès aux données*.

La première couche est la partie de l'application qui est visible et interactive avec l'utilisateur, elle affiche les données, prend la requête de l'utilisateur et lui présente les résultats. Elle est souvent représentée en HTML pour être interprétée par un navigateur web.

La deuxième couche correspond à la partie fonctionnelle de l'application, c'est elle qui traite la requête de l'utilisateur et envoie les résultats à la couche présentation. Elle décrit les opérations faites sur les données et met en œuvre les différentes règles de gestion et de contrôle du système.

La troisième couche est la partie correspondante à la gestion de l'accès aux données du système. Ces données peuvent être propres au système ou gérées par une autre application. Ce sont ces données qui seront traitées par la couche métier.

Dans notre système, nous avons utilisé l'architecture MVC (Modèle-Vue-Contrôleur) qui est une architecture trois-tiers ou trois niveaux. Nous expliquons dans ce qui suit, le fonctionnement général de ce modèle.

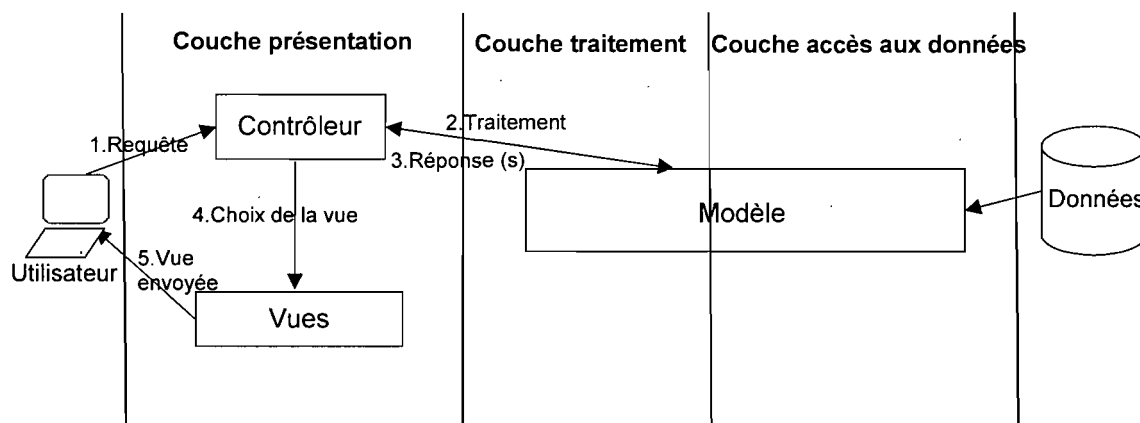
### **6.1.2. Le modèle MVC**

Dans une application web basée sur le modèle MVC (Modèle-Vue-Contrôleur), il y a une nette distinction entre les couches de présentation, traitement et accès aux données. Ainsi cette application est développée de manière à être partagée en trois parties fondamentales : le modèle, la vue et le contrôleur.

- *Le modèle* : représente le comportement de l'application, il fait le traitement et la description des données manipulées par l'application, comme il décrit les méthodes d'accès.
- *La vue* : c'est l'interface utilisateur. Elle présente les résultats renvoyés par le modèle, elle ne fait aucun traitement, comme elle permet à l'utilisateur d'interagir avec le système.
- *Le contrôleur* : il fait la gestion et la synchronisation des événements dans l'application. Il ne fait aucun traitement ni présentation. Son travail consiste à analyser la requête de l'utilisateur, appeler le modèle adéquat et renvoyer la vue correspondante.

Une application qui respecte le modèle MVC aura l'architecture présentée dans la figure 6.1 ci-dessous.





**Figure 6.1 :** Modèle MVC.

Dans un modèle MVC, Comme illustré dans la figure 6.1, l'utilisateur fait sa demande au contrôleur. Ce dernier fait appelle au modèle qui traite la demande et accède aux données, s'il y a lieu. Puis dépendamment de la demande, une ou plusieurs réponses sont expédiées au contrôleur. Celui-ci choisit parmi les vues de l'application, celle qui sera envoyée à l'utilisateur.

Cette architecture a pour avantage la clarté de la programmation, puisqu'il est facile de faire un changement dans l'application, ça suffit de connaître la partie désignée par le changement et y faire les modifications voulues sans affecter les autres parties.

Dans le développement de TAARAC, nous avons suit l'architecture MVC. Ainsi on peut y distinguer :

- Un bloc composé d'un programme principal (main.php) qui est la porte d'entrée de notre application, et d'un ensemble de scripts qui exécutent les demandes de l'utilisateur, c'est le *Contrôleur*.
- Un bloc composé d'un ensemble de modules qui assurent la logique de l'application, indépendamment du client (exemple : c-calcul-habilete.php qui a pour paramètres le numéro de l'item présenté au candidat et sa réponse. Il calcule le niveau d'habileté du candidat indépendamment de celui-ci). Ce bloc est aussi composé d'un ensemble de modules qui rapportent les données nécessaires au contrôleur (exemple : c-taarac-data.php qui s'occupe de tous ce qui fait appelle à la base de données), c'est le *Modèle*.

- Finalement un bloc qui rassemble l'ensemble des vues envoyées au client (exemple : v-accueil.php qui est une page HTML qui contient le format de la page d'accueil telle qu'elle est présentée au candidat) c'est la *Vue*

Dans ce qui suit nous allons se rapprocher encore plus de TAARAC en montrant toutes les étapes franchies par un candidat au cours de son passage du test, tout en utilisant des captures d'écran.

## 6.2. Environnement d'un candidat

Dans cette partie nous allons voir toutes les étapes passées par un candidat avant de commencer le test.

### 6.2.1. Page d'accueil de TAARAC

La page d'accueil, comme son nom le montre, accueille le candidat en lui présentant TAARAC et expliquant de manière générale son fonctionnement et toutes les tâches qu'il a à faire avant et pendant le déroulement du test (voir la figure 6.2).

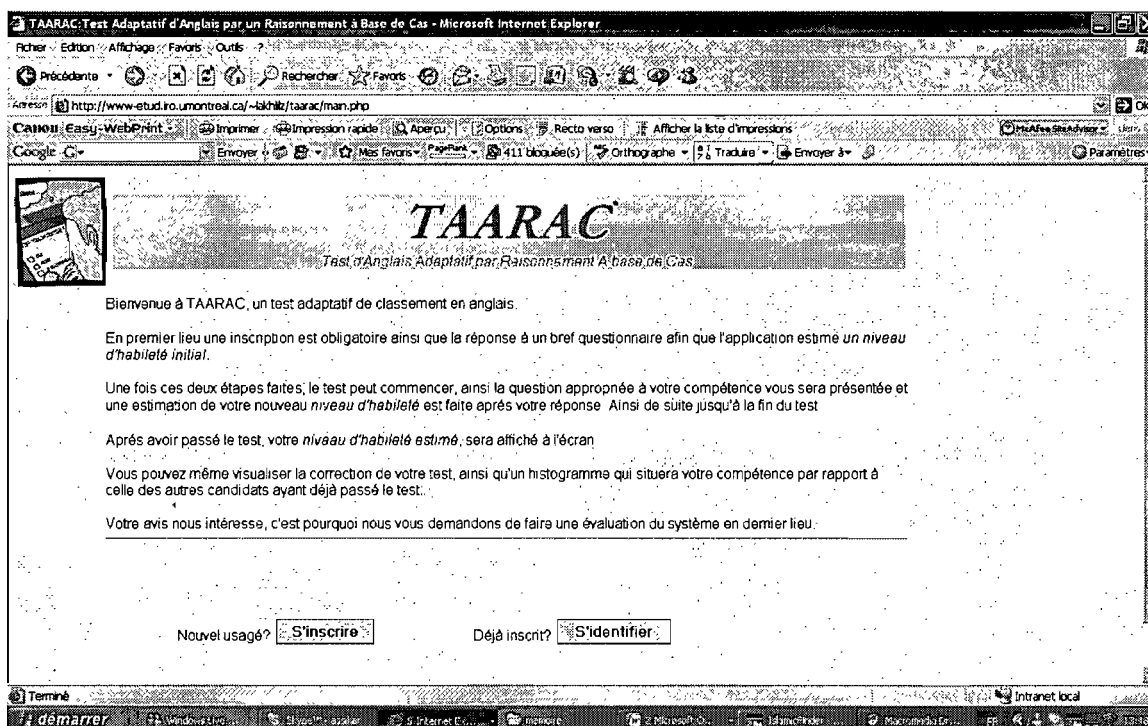


Figure 6.2 : Page d'accueil de TAARAC.

En plus le candidat est invité à s'inscrire ou s'enregistrer si c'est la première fois qu'il visite TAARAC, ou à s'identifier dans le cas contraire.

## 6.2.2. Enregistrement

Un nouveau candidat doit s'enregistrer avant de passer le test. L'enregistrement se fait en deux étapes.

Dans la première, le candidat est invité à remplir un formulaire simple, constituant son nom, son prénom, son adresse courriel et un mot de passe (figure 6.3).

Figure 6.3 : Première partie de l'enregistrement dans TAARAC.

À partir d'un lien qui se trouve dans la partie gauche de la page, le candidat peut vider le formulaire en tout moment.

Un candidat qui a omis de bien écrire son adresse mail, verra afficher un écran mettant en évidence l'erreur commise (figure 6.4)

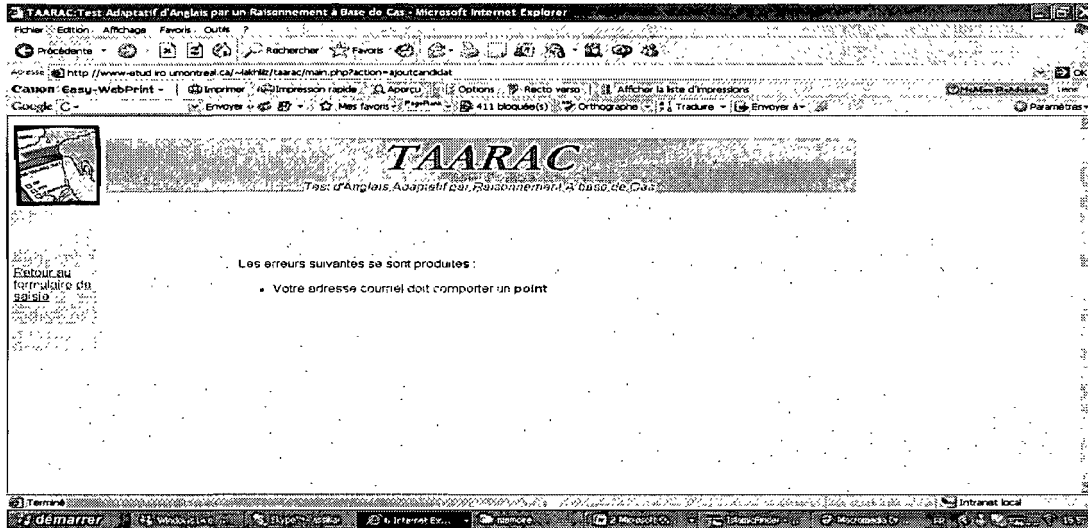


Figure 6.4 : Page d'erreur.

Une fois les données de l'enregistrement sont correctement saisies, Le candidat passe à la deuxième étape de l'enregistrement (figure 6.5), où ce dernier est invité à saisir un ensemble d'informations personnelles qui constitueront par la suite son profil.

The screenshot shows the second step of the registration process. The page title is 'CHOUKRI'. The user is welcomed and asked to provide personal information. The form includes the following fields:

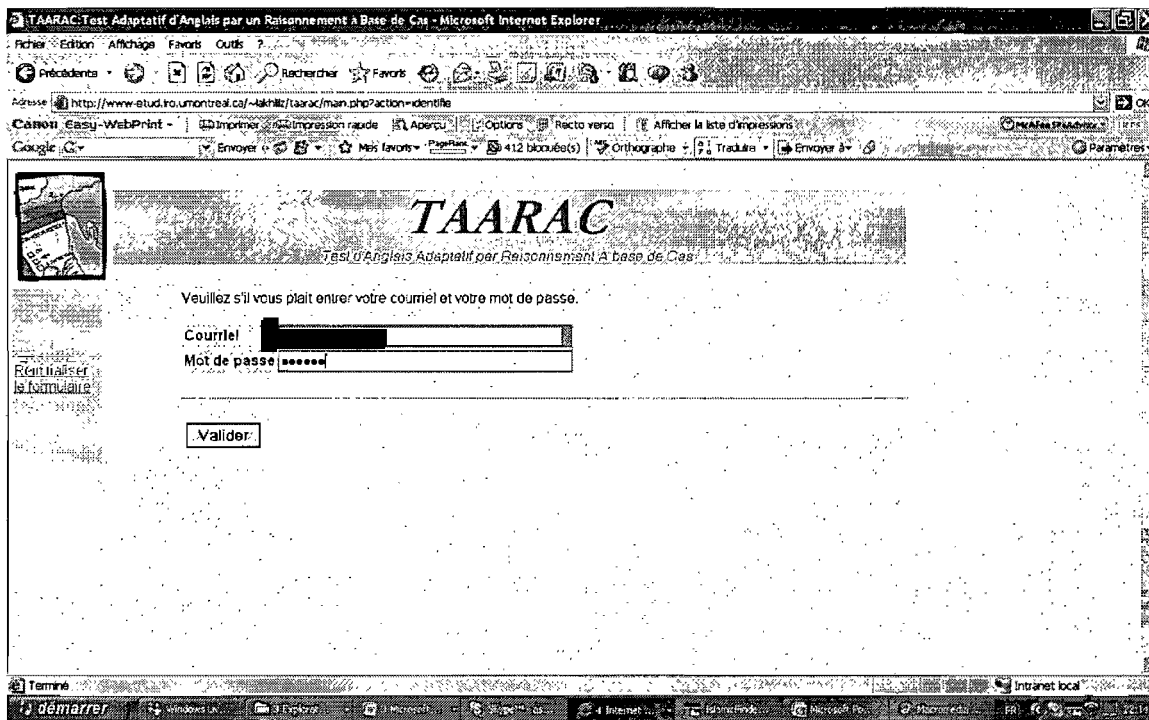
Age	6
Pays d'origine	Afghanistan
Profession	Agriculteur
Niveau d'étude complète	Collégial général
Langue maternelle	allemand
Langue la plus maîtrisée	allemand
Langue apprise (environnement académique)	allemand
Nombre d'années passées dans l'apprentissage d'anglais	0
Connaissez vous l'anglais pendant votre enfance	Non
Utilisez vous l'anglais dans votre milieu de travail	Non
Motivation pour apprendre l'anglais	Intérêt personnel
Aptitude d'apprentissage	Faible
Comment estimez vous votre niveau d'anglais (sur 10)	1

On the left side of the form, there is a link: [Reinitialiser le formulaire](#). At the bottom, there is a **Valider** button.

Figure 6.5 : Deuxième étape de l'enregistrement.

### 6.2.3. Identification

Un candidat qui est déjà enregistré, peut simplement s'identifier à l'aide de son adresse courriel et son mot de passe (figure 6.6).



**Figure 6.6 :** Page d'identification

TAARAC reconnaît le candidat. Dépendamment de l'état dans lequel ce dernier à quitter le test lors de sa dernière connexion, l'écran correspondant s'affiche.

Pour quelqu'un qui a quitté le système sans passer le test, ou après avoir arrêté le test avant de l'avoir terminé, verra afficher l'écran de la figure 6.7.

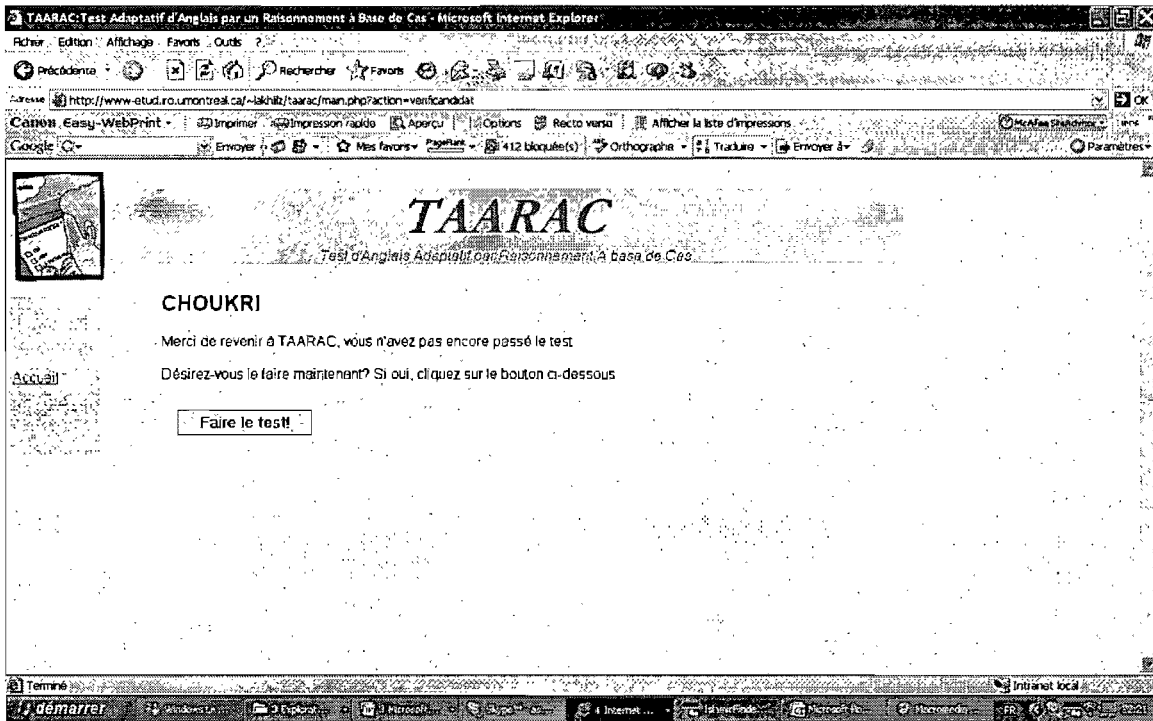


Figure 6.7 : Page affichée après l'identification d'un candidat qui n'a pas passé le test.

Si le candidat identifié a déjà passé le test, le système affiche un écran qui lui donne le choix entre refaire le test ou revoir les résultats (figure 6.8)

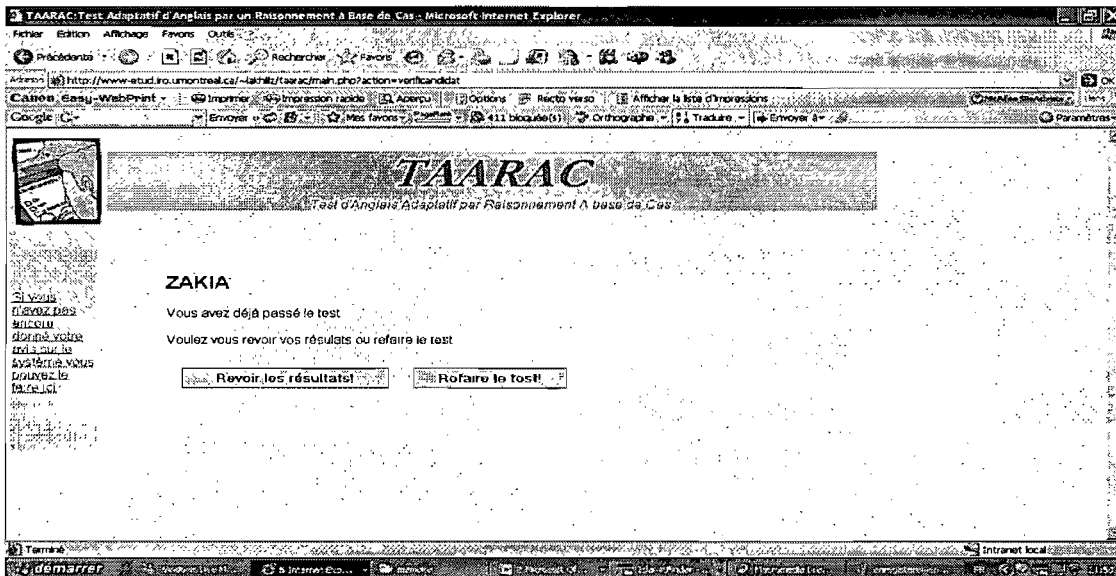


Figure 6.8 : Page affichée après l'identification d'un candidat qui a passé le test.

Si après l'identification du candidat le système s'aperçoit que ce dernier a suspendu le test avant de quitter TAARAC, il va l'inviter dans ce cas à reprendre le test là où il l'a arrêté (figure 6.9).

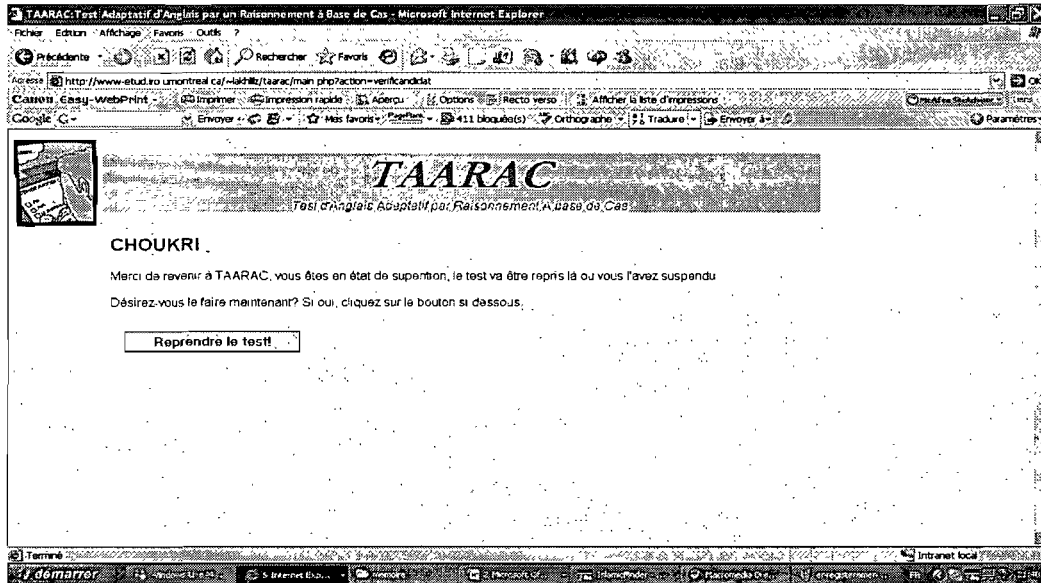
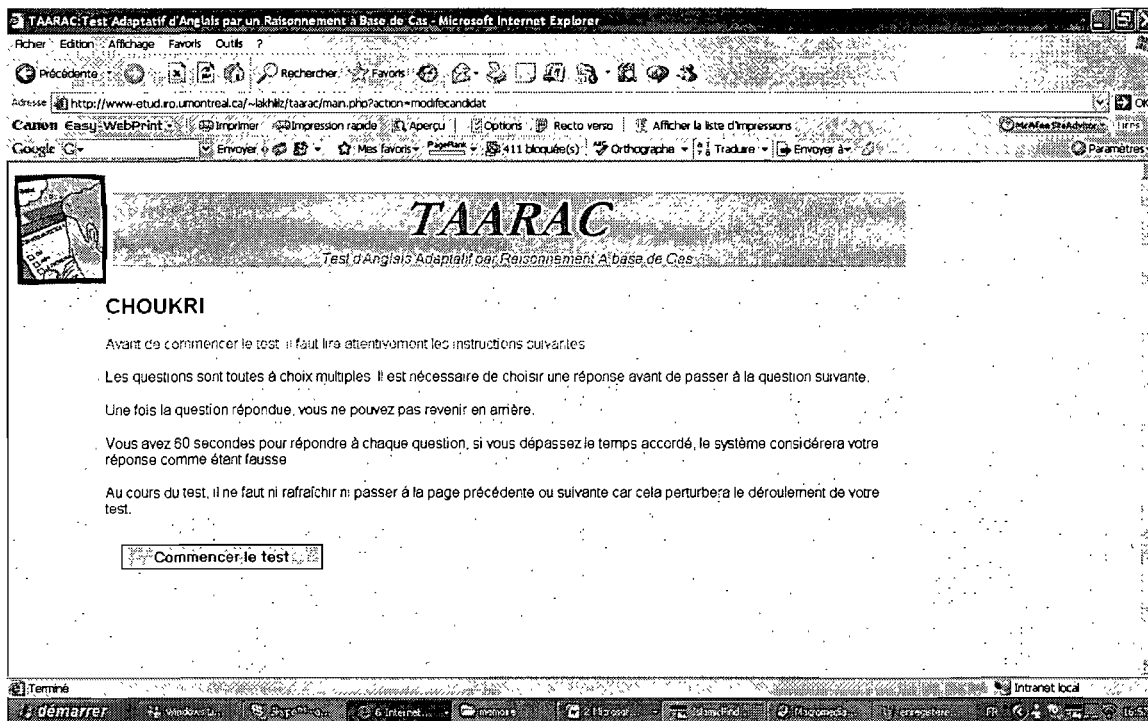


Figure 6.9 : Page affichée après l'identification d'un candidat qui a suspendu le test.

#### 6.2.4. Pendant le test du candidat

Le candidat ainsi enregistré ou identifié, voulant débiter le test, un écran expliquant les démarches à suivre pendant le test lui sera affiché (figure 6.10).



**Figure 6.10 :** Page affichée avant de commencer le test.

Ainsi le système signale au candidat qu'il faut répondre à la question avant de passer à la suivante, qu'après la confirmation de la réponse il ne peut plus retourner en arrière, et que le temps accordé pour répondre est 60 secondes. Si le candidat les dépasse, la réponse est considérée comme fausse. Il ne faut surtout pas toucher aux boutons de navigation dans le fureteur au cours du test.

En appuyant sur le bouton « Commencer le test », le système commence par chercher les profils similaires au profil courant, c'est-à-dire celui du candidat. Puis il estime son niveau d'habileté initial, comme expliqué dans la section 5.4.1. Il sélectionne par la suite l'item approprié pour l'afficher au candidat, voir la section 5.4.2. La manière dont cet item est présenté au candidat est montrée dans la figure 6.11.



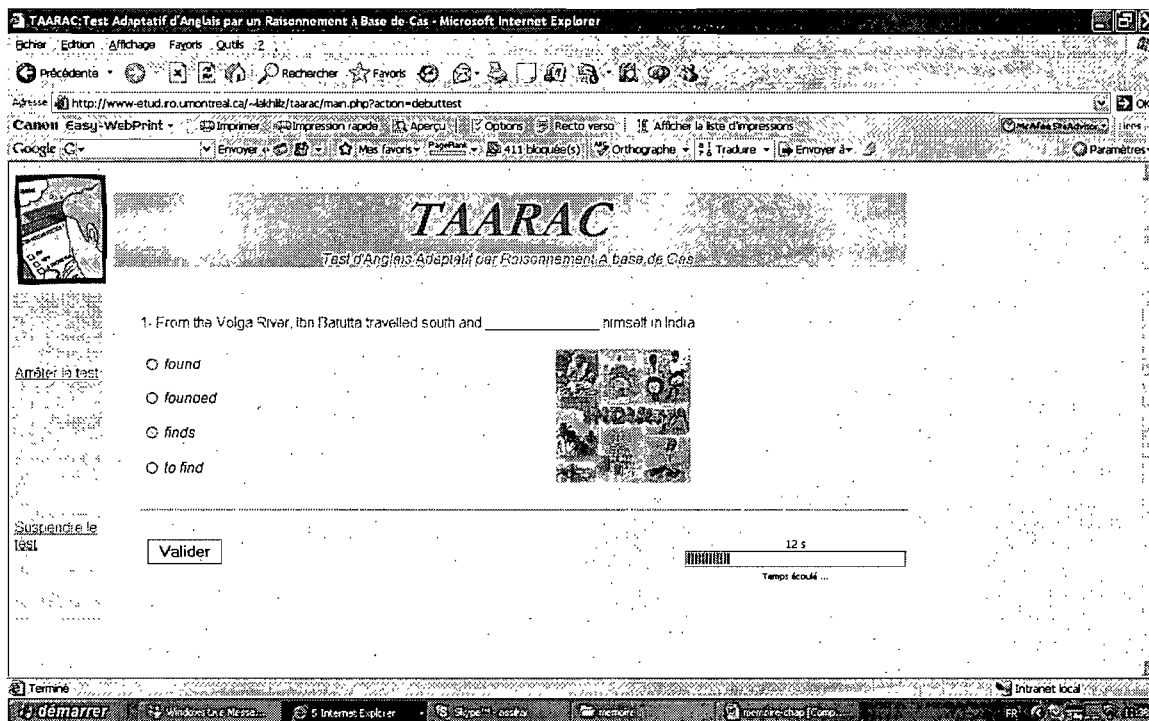


Figure 6.11 : Page représentant la question affichée au candidat.

Ainsi l'ordre de l'item, son texte et les choix de réponses sont affichés au candidat. Ces choix changent d'emplacement d'un appel à l'autre, renforçant ainsi la sécurité du test. Puisque les candidats qui sont capables de se rappeler des questions à partir de leurs emplacements n'auront plus aucune chance de le faire. Dans la partie gauche de la page, deux liens permettent au candidat d'arrêter le test ou de le suspendre s'il le désire. En cliquant sur l'un ou l'autre, l'écran de la figure 6.12 ou de la figure 6.13 respectivement sera affiché au candidat l'invitant à confirmer son choix. Une fois le test arrêté, toutes les informations sauvegardées jusqu'à présent sur le test disparaîtront. Lors de la prochaine connexion, le candidat doit refaire le test depuis le début. Alors que s'il est suspendu, le système retiendra toutes les étapes passées dans le test et le candidat pourra continuer le test là où il l'a suspendu.

Dans le côté droit de l'écran de la figure 6.11, une image donnant une idée sur le thème de la question est exposée. De même qu'une barre montrant le temps qui s'écoule avant de répondre à la question. Dans les dix dernières secondes, un autre message apparaîtra au dessus de la barre mettant en valeur le temps qui reste (figure 6.14).

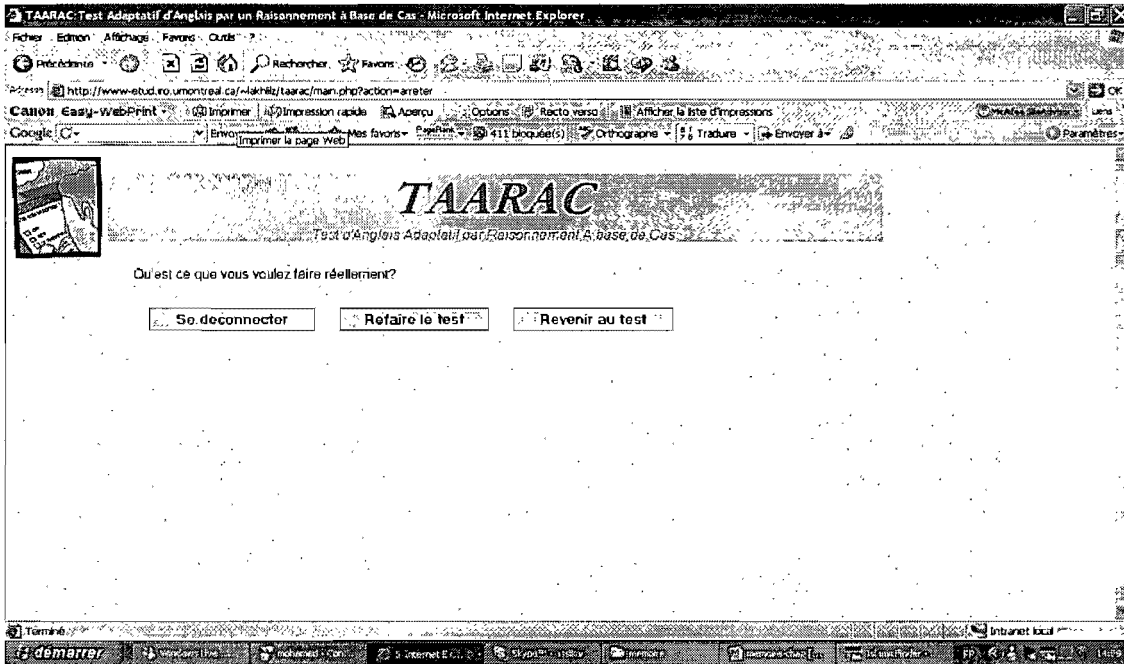


Figure 6.12 : Page affichée en cliquant sur le bouton de l'arrêt du test.

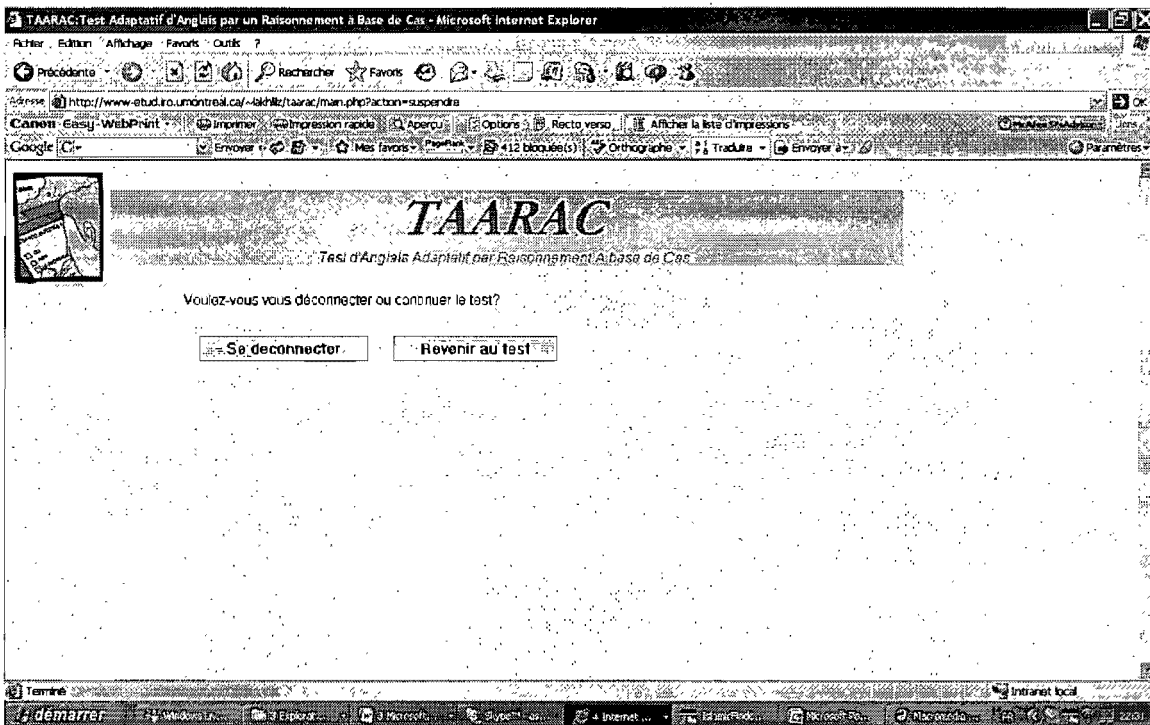


Figure 6.13 : Page affichée en cliquant sur le bouton de la suspension du test.

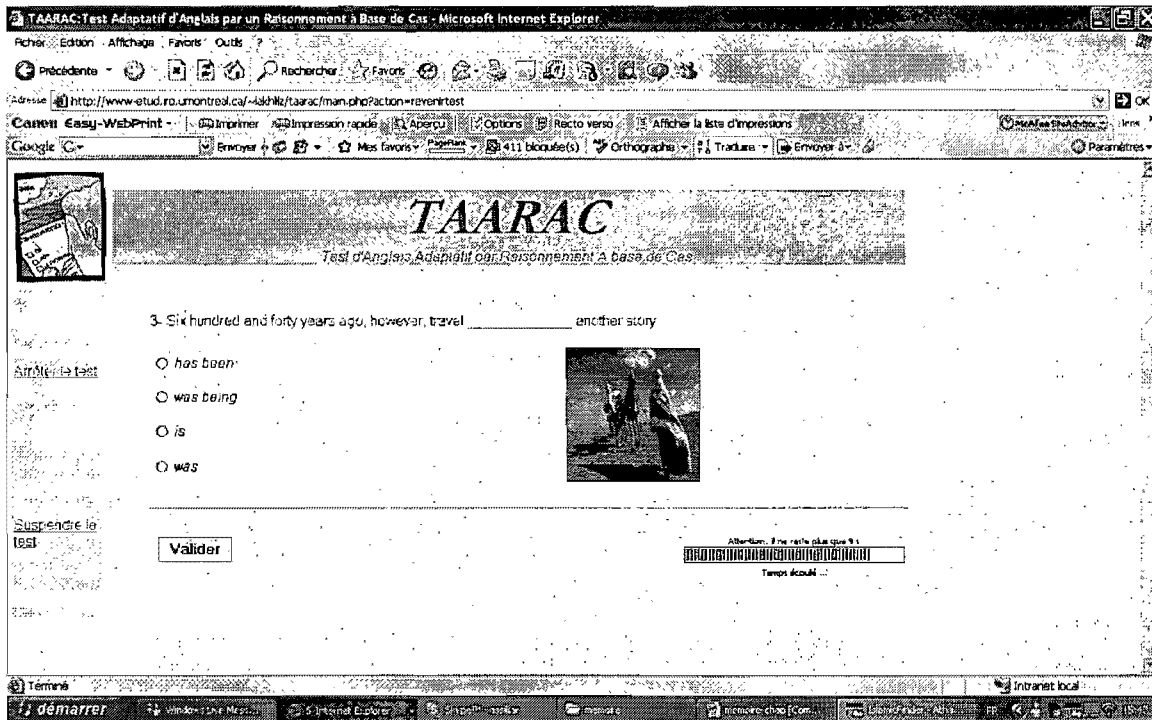


Figure 6.14 : Page affichant l'item dans les dix dernières secondes.

Le même écran est affiché pendant tout le test, les seules différences sont le numéro, le texte, l'image et les choix des réponses qui changent au fur et à mesure que la question change. Le choix de l'item à présenter au candidat est expliqué dans la partie 5.4.2.

### 6.2.5. À la fin du test du candidat

Une fois le test est arrêté, un écran montrant le niveau d'habileté du candidat (figure 6.15) est présenté à ce dernier, tout en lui rappelant son niveau auto évalué avant le test. Deux liens dans la partie gauche de l'écran donnent le choix au candidat entre voir la correction du test, ou voir ses résultats qui sont accumulés au fur et à mesure qu'il repasse le test, en même temps qu'une comparaison avec les niveaux des candidats qui ont déjà passé le test, représentée sous forme d'un histogramme.

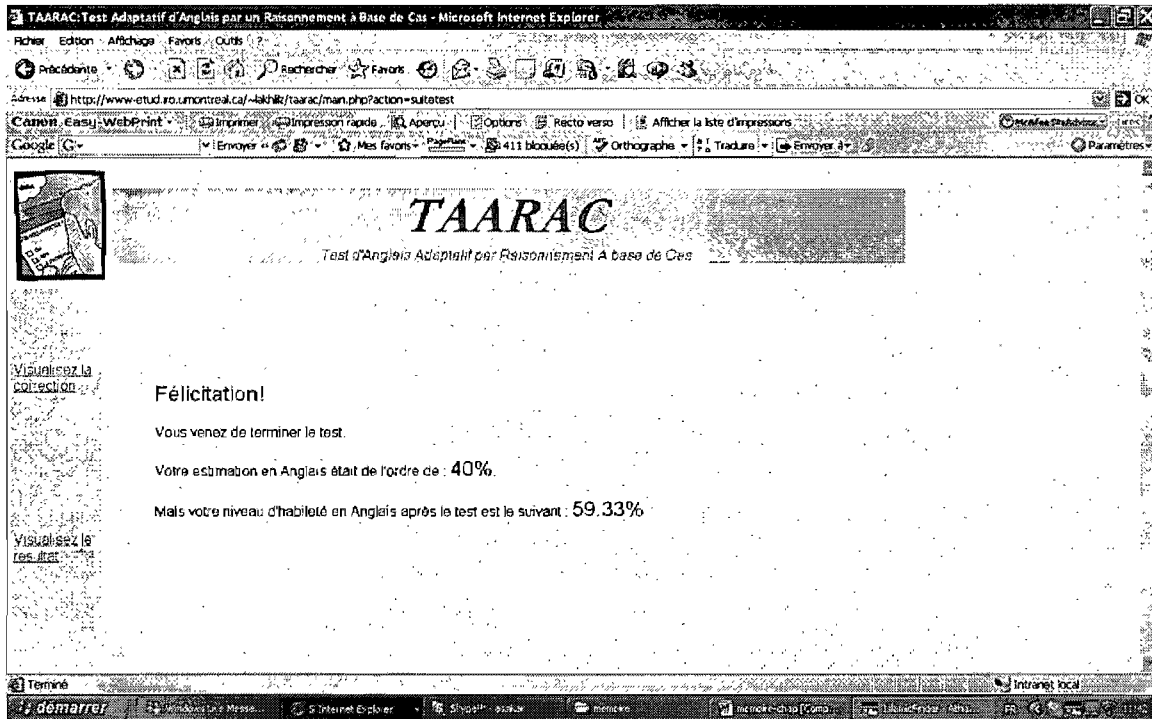


Figure 6.15 : Page affichant le niveau d'habileté du candidat.

Si le candidat choisit de visionner la correction du test, une infobulle sous forme d'avertissement est affichée au candidat. L'infobulle rappelle au candidat qu'il vient de passer un test adaptatif, et par conséquent son niveau en anglais ne peut pas être interprété à partir du nombre des questions correctes (figure 6.16).

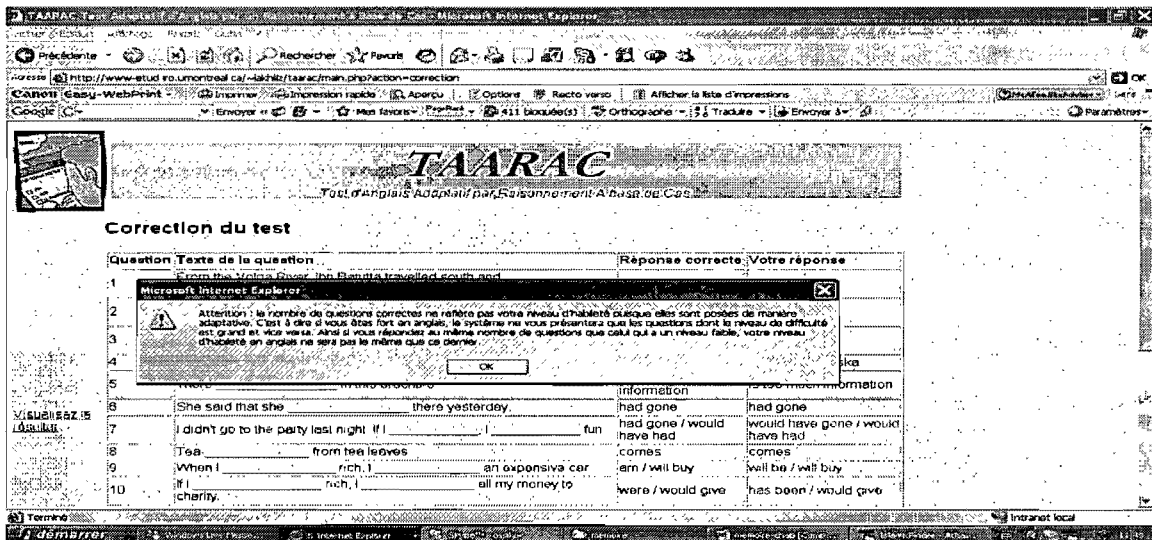


Figure 6.16 : Avertissement affiché avant la correction.

La correction est affichée par la suite sous forme d'un tableau (voir figure 6.17) rassemblant les numéros des questions posées dans le test, leurs textes, les réponses correctes et celles du candidat. Les réponses fausses sont mises en évidence par leurs couleurs rouges.



**Correction du test**

Question	Texte de la question	Réponse correcte	Votre réponse
1	From the Volga River, Ibn Batutta travelled south and himself in India.	found	found
2	Ibn Batutta's travels took him three times the distance that Marco Polo	went	went
3	That girl lives _____ Montreal _____ Sherbrooke street.	in / on	in / at
4	The forty-ninth state in _____	the U.S. is Alaska	the U.S. is Alaska
5	There _____ in this brochure.	is too much information	is too much information
6	She said that she _____ there yesterday.	had gone	had gone
7	I didn't go to the party last night. If I _____, I _____ fun.	had gone / would have had	would have gone / would have had
8	Tea _____ from tea leaves.	comes	comes
9	When I _____ rich, I _____ an expensive car.	am / will buy	will be / will buy
10	If I _____ rich, I _____ all my money to charity.	were / would give	has been / would give
11	I don't understand. Please _____	explain that to me	explain that to me
12	Today we say the world, _____	has shrunk	has shrunk
13	_____ from Constantinople, Ibn Batutta entered the steppes of central Asia.	Having crossed	Crossening
14	I _____ here _____ two years.	have lived / for	have lived / for
15	Ibn Batutta travelled across the Sahara and through Europe. He _____ with the city of Cairo.	was delighted	has delighted
16	I enjoy _____ ice cream.	eating	eating
17	Twenty-nine years after Ibn Batutta set out, he arrived home to discover that his mother _____	had died	had died

■ : Réponses incorrectes.

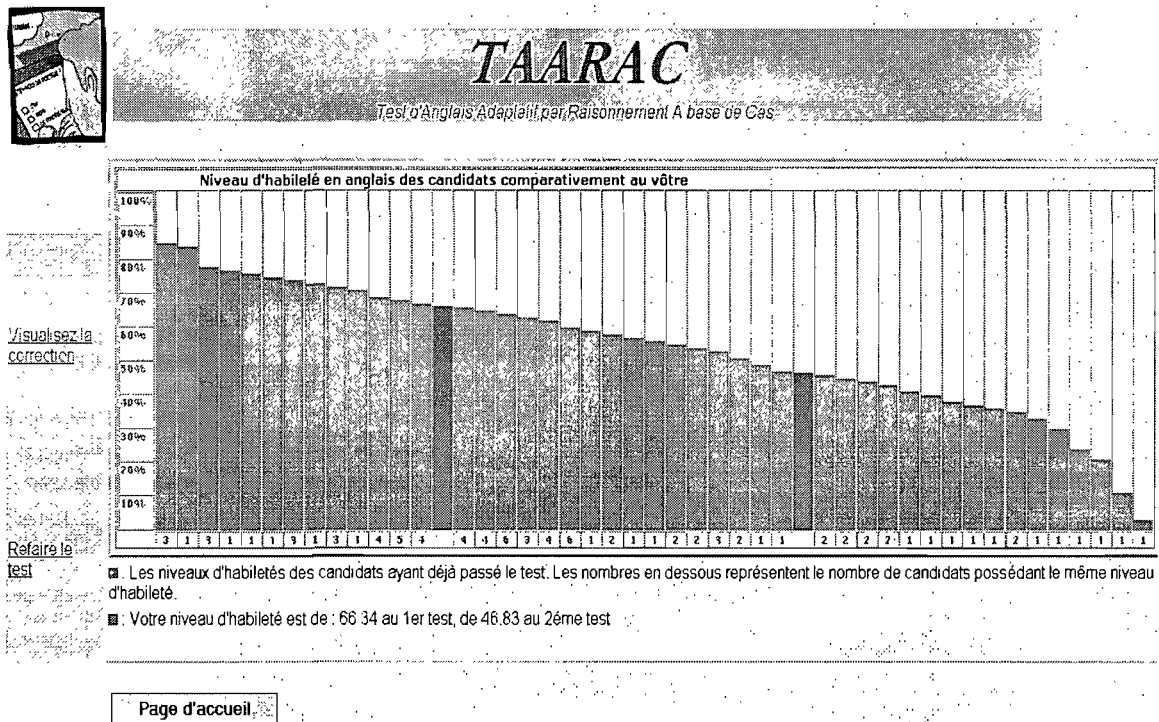
Page d'accueil

**Figure 6.17** : Page affichant la correction du test.

Nous constatons qu'à partir de cette page le candidat peut voir ses résultats, refaire le test ou se déconnecter complètement en passant à la page d'accueil.

La page résultat, comme nous avons mentionné auparavant, est une page récapitulative des résultats obtenus par le candidat lors de ses différents passages au test,

en plus d'un histogramme comparant ces résultats avec les niveaux d'habileté des candidats qui ont déjà passé le test (figure 6.18).



**Figure 6.18 :** Page affichant les résultats du candidat.

Ainsi la page résultat comprend un histogramme représentant les niveaux d'habileté en anglais de tous les candidats qui ont passé le test. Les niveaux sont classés du plus grand au plus petit. Les colonnes vertes (gris foncé) représentent les compétences en anglais du candidat courant. Alors que les colonnes orange (gris clair) représentent les niveaux d'habileté en anglais obtenus au dernier test pour les autres candidats. Les numéros en bas de l'histogramme désignent le nombre des candidats ayant le même niveau d'habileté en anglais. Grâce à cet histogramme, le candidat sera capable de se positionner par rapport aux autres candidats.

À partir de cette page, le candidat peut revoir la correction du dernier test passé, refaire le test s'il en a envi, ou même se déconnecter et revenir à la page d'accueil.

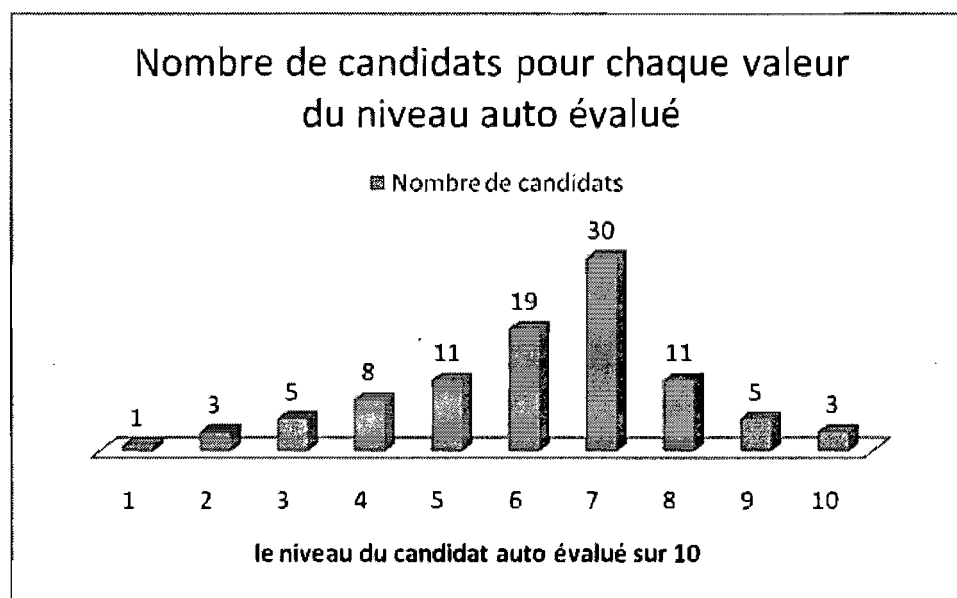
## 6.3. Validation

Dans cette section nous présentons la validation de notre système ainsi que les différents résultats obtenus.

### 6.3.1. Mise en contexte

TAARAC est un test de classement en anglais. Étant donné le domaine de l'application, qui est l'anglais, nous avons fait subir le test à un nombre très hétérogène de personnes ayant des profils différents, et par conséquent des niveaux d'habileté en anglais bien différents.

102 personnes ont testé TAARAC avec un âge variant entre 18 et 53 ans. Leurs origines sont diverses : Algérie, Argentine, Brésil, Cameroun, Canada, Chine, Liban, Maroc, Mexique, Pologne, Roumanie, Rwanda, Tunisie et Viêt Nam. Leur champ de profession varie entre étudiant, ingénieur, employé, professeur et autres. Alors que leur niveau d'étude passe du collégial, baccalauréat, maîtrise au doctorat. Leur niveau en anglais varie entre faible et plus fort, selon leur auto évaluation. Cette fluctuation est montrée dans le graphe de la figure 6.19.



**Figure 6.19 :** Graphe répartitif des candidats selon leur niveau auto évalué.

Comme nous avons déjà vu dans le chapitre précédent, notre base d'items contient 46 items, leurs paramètres de difficulté obtenus par PARAM-PL3 (voir la section 5.3.1) varient entre -2.4 et 1.7. TAARAC est un test adaptatif c'est-à-dire qu'il essaye de présenter au candidat des items dont la difficulté est proche de son niveau d'habileté. Et puisque ce dernier varie entre -4 et 4, alors les candidats qui sont très forts, verront que leur niveau n'est pas bien estimé. Car TAARAC va chercher à leur présenter des items plus difficiles, mais il n'en trouvera pas dans la base d'items. Ainsi TAARAC n'est pas efficace pour estimer le niveau d'habileté pour des personnes de niveau très fort. D'après les résultats obtenus, les personnes très fortes ont un niveau de l'ordre de 2.7 soit 84.5%, avec une erreur type égale à 0.45, une valeur qui montre parfaitement que le niveau trouvé est biaisé. Cela ne change en rien notre validation, puisque ce que nous voulons prouver ici, c'est que TAARAC est bien un système capable d'adapter le test selon le niveau d'habileté du candidat, et que l'apport du raisonnement à base de cas est bien évident. Ainsi nous allons nous baser dans cette étude sur les résultats obtenus avec des candidats de niveau moyen et faible, soit 96 candidats.

### **6.3.2. L'adaptabilité de TAARAC**

Nous mettons en évidence le premier objectif recherché par TAARAC, l'adaptabilité du test selon le niveau d'habileté du candidat. De manière plus précise, montrer que si le candidat répond correctement à un item, il aura par la suite un item plus difficile, sinon un plus facile. Nous avons choisi deux candidats de niveau différents, sur qui nous avons testé TAARAC. Le premier a un niveau plutôt moyen puisque il a obtenu 0.27 dans une échelle de -4 et 4 soit 53.3%. Alors que le deuxième possède un niveau supérieur au premier, il est estimé à 1.84 soit 73%. Les résultats obtenus sont résumés dans le tableau 6.1 pour le premier candidat et 6.2 pour le deuxième.



Tableau 6.1 : Résumé des résultats obtenus par un candidat de niveau moyen.

Numéro de l'item	P. de difficulté	P. de discrimination	P. de pseudo-chance	Réponse du candidat	Niveau d'habileté estimé	Erreur type
46	-0.56	1.19	1.33	0	-2.27	0.79
9	-2.41	0.43	0	1	-1.7	0.75
39	-0.65	0.90	0.08	1	-0.81	0.69
49	0.02	0.68	0.11	1	-0.36	0.67
33	-0.4	1.28	0.02	1	0.12	0.59
3	0.71	1.28	0.11	0	-0.1	0.52
30	-0.46	1.30	0.29	1	0.02	0.49
27	-1.78	0.56	0	1	0.07	0.48
14	0.49	1.33	0.08	0	-0.09	0.44
20	-0.86	1.43	0	1	-0.01	0.41
26	-0.35	1.18	0.18	1	0.08	0.39
16	-1.24	0.98	0	1	0.11	0.39
12	-0.48	1.44	0.11	1	0.18	0.37
32	-0.98	1.2	0	1	0.21	0.36
4	0.68	2	0.18	1	0.39	0.38
43	0.87	1.94	0.16	0	0.29	0.35
50	0.47	2	0.24	0	0.15	0.32
34	-0.96	1.66	0	1	0.17	0.31
31	-0.26	2	0.29	1	0.21	0.29
36	0	1.85	0.33	1	0.27	0.28

L'erreur type est calculée de la même manière que dans l'exemple 3 de la section 5.4.2.

**Tableau 6.2 : Résumé des résultats obtenus par un candidat de niveau assez fort.**

Numéro de l'item	P. de difficulté	P. de discrimination	P. de pseudo-chance	Réponse du candidat	Niveau d'habileté estimé	Erreur type
41	1.71	1.2	0.21	0	-0.93	0.92
46	-0.56	1.19	0.13	1	0.51	0.79
21	0.48	0.73	0.11	1	1.02	0.73
49	0.02	0.68	0.11	1	1.21	0.68
3	0.71	1.28	0.11	1	1.49	0.6
8	0.87	1.49	0.23	1	1.64	0.55
2	0.96	1.28	0.22	0	1.14	0.51
33	-0.4	1.28	0.02	1	1.15	0.47
10	1.46	1.61	0.08	1	1.45	0.45
14	0.49	1.33	0.08	1	1.49	0.41
24	0.81	1.5	0.22	1	1.55	0.39
37	0.63	1.46	0.3	1	1.58	0.38
43	0.87	1.94	0.16	1	1.63	0.35
32	0.68	2	0.18	1	1.65	0.34
4	0.78	1.88	0.27	1	1.67	0.33
42	0.25	1.53	0.28	1	1.68	0.33
5	1.62	2	0.19	1	1.8	0.33
44	0.99	2	0.21	1	1.83	0.32
45	0.47	2	0.24	1	1.83	0.32
22	0.38	2	0.36	1	1.84	0.31

❖ Interprétation des tableaux 6.1 et 6.2.

Le premier candidat avait pour niveau d'habileté initial -0.8, celui auto évalué. La question dont la difficulté est proche de cette valeur est l'item 46. Le candidat ne répond pas correctement à cette question, d'où la diminution de son niveau -2.27. Le système cherche alors l'item de difficulté proche de cette valeur, c'est la question numéro 9. Le candidat répond cette fois ci correctement, et son niveau augmente par conséquent et devient -1.7. L'item qui existe dans la base ou plus précisément dans le premier bloc, (section 5.3.3) dont la difficulté est proche de cette valeur est le 39. La réponse du candidat étant correcte, le niveau augmente et devient -0.81 et ainsi de suite

jusqu'au 20<sup>ième</sup> item dont la réponse était correcte, ainsi le test termine par un niveau égale à 0.27 avec une erreur type de l'ordre de 0.28.

Le deuxième candidat possédait un niveau d'habileté initial égale à 2.4. Cette fois-ci l'item choisi par TAARAC est le 41. Mais puisque le candidat n'a pas bien répondu, alors son niveau diminue et devient -0.93. L'item dont la difficulté est proche de cette valeur est l'item 46. Une réponse correcte à ce dernier augmente le niveau d'habileté du candidat, soit 0.51. Ainsi de suite jusqu'au dernier item dont la réponse était correcte, ce qui a donné un niveau de l'ordre de 1.84 avec une erreur type égale à 0.31.

Dans le premier tableau comme dans le deuxième, nous remarquons que TAARAC choisi souvent les items les moins discriminants en premier, et laisse les plus discriminatoires jusqu'à la fin du test, là où le niveau d'habileté doit être le plus précis possible. Cette précision nous la constatons au fur et à mesure que le test s'avance, puisque l'erreur type devient plus petite à la fin du test.

D'après ce qu'on a vu, L'adaptation est belle et bien évidente dans TAARAC, car dépendamment du niveau d'habileté du candidat, le système choisi toujours l'item dont la difficulté est proche du niveau de ce dernier. Ainsi, si le candidat ne répond pas correctement à la question courante, la prochaine sera moins difficile et vice versa.

### **6.3.3. L'apport du raisonnement à base de cas**

Dans cette section nous verrons comment le raisonnement à base de cas a amélioré notre test, en le rendant plus court et plus efficace.

#### **6.3.3.1. Un test plus court**

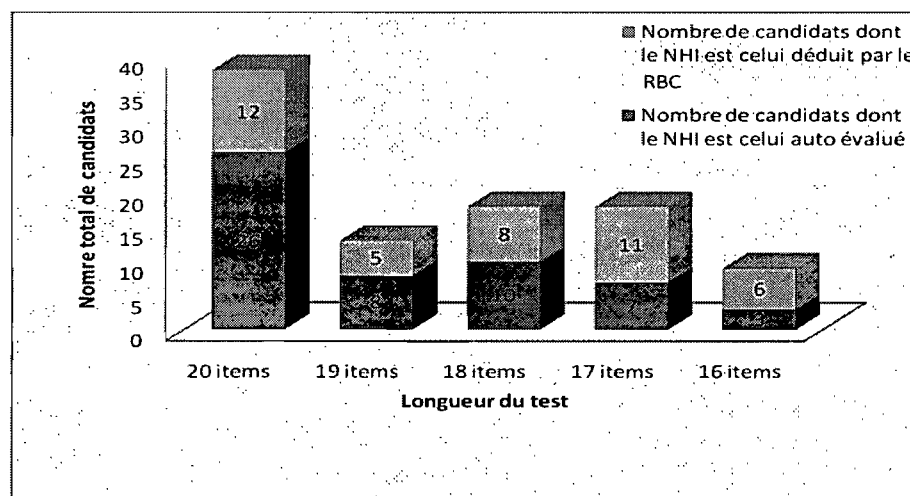
Comme nous avons mentionné précédemment, les résultats se basent sur 96 candidats qui ont testé TAARAC dont le niveau varie entre assez fort et faible. Pour montrer l'effet du raisonnement à base de cas sur le test, nous allons comparer la longueur du test entre les candidats dont le niveau d'habileté initial est celui auto évalué, avec ceux dont ce dernier est déduit par le RBC (le Raisonnement à Base de Cas), soit

55 candidats. Or nous avons constaté que 13 parmi ces derniers ont un niveau d'habileté initial (NHI) proche de celui auto évalué avec une différence de 6% près. Cette marge n'a aucun effet sur le choix du premier item à présenter au candidat (voir section 5.4.1). Ainsi, soit que le niveau d'habileté initial est déduit par le RBC ou que c'est celui auto évalué, le test reste le même. C'est pourquoi nous avons décidé de n'utiliser que les candidats dont le niveau d'habileté initial n'est pas très proche de celui auto évalué, soit 42 candidats. Le tableau 6.3 montre la répartition de ces candidats selon la longueur du test, ainsi que le pourcentage de ceux dont le NHI est celui déduit par le RBC parmi eux.

**Tableau 6.3 :** Les candidats répartis selon la longueur du test et le niveau d'habileté initial.

Longueur du test	Nombre total de candidats	Nombre de candidats dont le NHI est celui déduit par le RBC	Leur pourcentage par rapport au nombre total de candidats
20 items	38	12	32%
19 items	13	5	38%
18 items	18	8	44%
17 items	18	11	61%
16 items	9	6	67%
Total	96	42	44%

D'une manière plus représentative, un graphe est construit pour mettre en valeur les résultats trouvés, dans la figure 6.20.



**Figure 6.20 :** Diagramme représentant le nombre des candidats selon la longueur du test et leurs niveaux d'habileté initial.

### ❖ Interprétation du tableau 6.3

Comme nous avons vu précédemment, l'un des grands avantages du test adaptatif est que le test devient plus court qu'un test papier crayon. Et puisque dans TAARAC les critères d'arrêts sont la longueur du test ou la précision de l'estimateur du niveau d'habileté. Il est évident que la longueur du test variera d'un candidat à l'autre. C'est ce que nous résume la deuxième colonne du tableau. Dans la troisième colonne, nous avons placé les candidats dont le niveau d'habileté initial est déduit à partir du raisonnement à base de cas. Et dans la quatrième, le pourcentage de ces derniers pour chaque longueur du test. Nous constatant alors qu'au fur et à mesure que la longueur du test diminue, le pourcentage des candidats dont le NHI est déduit par le RBC augmente. Autrement, avec l'application du RBC, le test a la tendance d'être de plus en plus court.

Nous constatons aussi que les candidats qui ont atteint rapidement (avec un test de longueur égale à 16 ou 17 items) l'erreur type de l'ordre de 0.2, plus de 60% avaient un NHI déduit par le RBC. Ainsi avec un RBC le test atteint rapidement la précision minimal, le test est par conséquent plus précis.

#### **6.3.3.2. Un test plus efficace**

Un test efficace pour un candidat donné, est un test dont le niveau d'habileté initial est proche à celui estimé à la fin du test. Nous allons faire alors une comparaison entre les candidats dont le NHI est déduit à partir du RBC et ceux avec un NHI auto évalué, tout en se basant sur la différence entre le NHI et celui estimé à la fin du test.

Compte tenu du nombre des candidats dont le NHI est celui estimé par le RBC, soit 55 candidats, regardons le pourcentage de ceux dont le NHI est proche à celui obtenu à la fin du test et comparons le avec celui obtenu à partir des candidats dont le NHI est celui auto évalué. Ceci est résumé dans le tableau 6.4. Nous supposons que deux niveaux d'habileté sont proches, si leur différence est de l'ordre de 0.5 dans l'intervalle  $[-4,4]$ , soit 6,5%.

**Tableau 6.4 :** Les candidats répartis selon le niveau d'habileté initial et l'exactitude du test.

	Nombre total de candidats	Nombre de candidat dont le NHI est proche à celui obtenu à la fin du test	Pourcentage des candidats dont le NHI est proche à celui obtenu à la fin du test
Nombre de candidats dont le NHI est celui auto évalué	41	11	26%
Nombre de candidats dont le NHI est celui déduit du RBC	55	21	38%
Total	96	32	33%

D'après le tableau 6.4, nous constatons que le pourcentage des candidats dont le NHI est proche à celui obtenue à la fin du test, est plus grand pour ceux dont le NHI est déduit par RBC par rapport aux candidats dont le NHI est celui auto évalué. Ainsi, avec un raisonnement à base de cas le candidat a plus de chance d'avoir un niveau d'habileté à la fin du test proche de celui qu'il avait au début. Si le NHI est différent du niveau d'habileté final, les items présentés au candidat au début du test seront de difficulté très différente de son vrai niveau d'habileté, et c'est vers la fin du test qu'il aura des items de difficulté proche de son niveau. Alors que si le NHI est proche du niveau d'habileté final, du début jusqu'à la fin du test, les items présentés au candidat seront de difficulté très proche de son vrai niveau, soit un test plus efficace.

#### **6.3.4. Feedback du candidat**

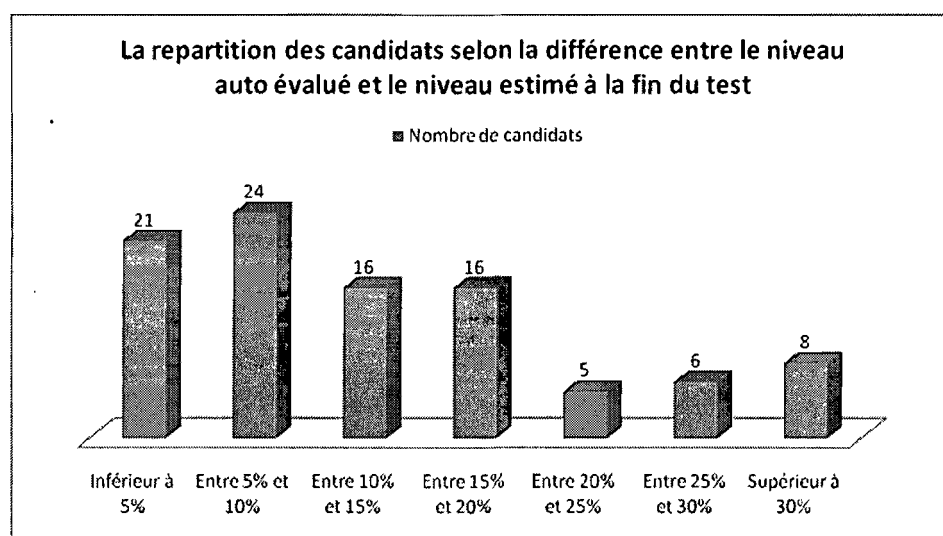
Nous allons voir dans cette section le feedback du candidat, soit celui inféré à partir des résultats obtenus de TAARAC, ou celui fait par le candidat.

### 6.3.4.1. Feedback inféré

Avant de passer, au feedback réel du candidat, nous allons nous mettre à la place du candidat et voir si TAARAC satisfait ses attentes ou non. Autrement dit, faire une comparaison entre le niveau auto évalué et le niveau d'habileté estimé à la fin du test. Cette comparaison est montrée d'une façon plus détaillée dans le tableau 6.5 et la figure 6.21.

**Tableau 6.5 :** Les candidats répartis selon la différence entre le niveau auto évalué et celui estimé à la fin du test.

La différence entre le niveau auto évalué et le niveau estimé à la fin du test	Nombre de candidats	Leur pourcentage par rapport au nombre total de candidats
Inférieur à 5%	21	22%
Entre 5% et 10%	24	25%
Entre 10% et 15%	16	17%
Entre 15% et 20%	16	17%
Entre 20% et 25%	5	5%
Entre 25% et 30%	6	6%
Supérieur à 30%	8	8%
Total	96	



**Figure 6.21 :** Diagramme représentant le nombre des candidats selon les tranches de différence entre le niveau auto évalué et le niveau estimé à la fin du test.

Le tableau comme le diagramme montrent que presque la moitié des candidats verront leur niveau d'habileté estimé par TAARAC proche du niveau qu'ils pensaient avoir en anglais, avec une différence de 10% près. Ainsi, il y a une grande chance que le candidat soit satisfait du résultat obtenu par TAARAC. C'est ce que nous allons voir dans ce qui suit, en analysant le feedback réel des candidats.

#### 6.3.4.2. Feedback réel

Une fois que le candidat termine le test et visualise ses résultats, il est invité à travers un lien dans la page résultat à donner son avis sur le système. Ce lien le dirige directement vers la page de la figure 6.22.

TAARAC: Test Adaptatif d'Anglais par un Raisonnement à Base de Cas - Microsoft Internet Explorer

http://www.etud.ro.umontreal.ca/~lkhil/taaac/man.php?action=avis

**TAARAC**  
Test d'Anglais Adaptatif par Raisonnement A base de Cas

Veuillez si vous plaît donner votre avis sur le système.

**Légende :** 1 : En désaccord. 5 : Totalelement en accord.

L'interface est claire	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
La conception est bonne	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Le système est convivial	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Le but du système est compréhensif	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Vous êtes en accord avec le niveau estimé	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Commentaire :

Retour au test

Figure 6.22 : Le formulaire du feedback du candidat.

À partir de ce formulaire le candidat peut donner son appréciation sur, la clarté de l'interface, la conception, la convivialité et la compréhension du but du système. Comme il pourra nous communiquer son avis sur le résultat obtenu par TAARAC. Cette évaluation se fait dans une échelle de 1 à 5, 1 signifiait que son avis est complètement défavorable, alors que 5 signifie l'avis le plus favorable. La moyenne des appréciations de tous les candidats est rassemblée dans le tableau 6.6.



**Tableau 6.6** : Les moyennes des appréciations de 42 candidats.

	Moyenne des notes
L'interface est claire	4.40
La conception est bonne	4.19
Le système est convivial	4.21
Le but du système est compréhensif	4.41
Vous êtes en accord avec le niveau estimé	3.38

Nous remarquons ainsi que la plupart des appréciations sont favorables pour TAARAC. En plus les commentaires reçus sont encourageants et satisfaisants.

### **6.3.5. Conclusion**

Comme nous savons tous, contrairement au test papier crayon, le but essentiel d'un test adaptatif est d'estimer le niveau d'habileté du candidat de manière précise avec un test de longueur minimal. Dans cette évaluation, malgré la modestie de la base d'items, nous avons montré que le raisonnement à base de cas a amélioré le test adaptatif, en le rendant plus court, plus précis et plus efficace, répondant ainsi aux objectifs visés par cette étude.

## Chapitre 7 : Conclusion

L'objectif principal visé par ce travail, est de concevoir un test adaptatif capable d'estimer un niveau d'habileté de manière *efficace* et avec un nombre *minimal d'items*. Comme nous avons expliqué au chapitre 3, un test adaptatif est un test adapté selon le niveau d'habileté du candidat. Les questions sont choisies de telle sorte que leurs difficultés soient proches du niveau d'habileté du candidat. Ainsi ce dernier ne se sentira pas frustré devant un test qui est soit très dur ou trop facile pour lui.

Afin d'atteindre cet objectif, nous avons effectué une combinaison de deux techniques totalement différentes : *la théorie des réponses aux items* et *le raisonnement à base de cas*. La première technique nous a permis de construire notre banque d'items, faire la sélection de l'item et estimer le niveau d'habileté au cours du test et à la fin du test. La deuxième technique est utilisée pour estimer le niveau d'habileté initial du candidat.

La banque d'items est conçue à partir d'un test d'anglais de la faculté de l'éducation permanente de l'Université de Montréal (section 5.3.1). Au cours du test, l'item est sélectionné en maximisant sa fonction d'information au voisinage du niveau d'habileté estimé, et en minimisant le taux d'exposition (section 5.4.2). Après chaque réponse du candidat, son niveau d'habileté ainsi que l'erreur type sont estimés à l'aide de la méthode de *l'espérance à postériori* (l'exemple 3 de la section 5.4.2). Le test est arrêté lorsque le nombre des items présentés au candidat est égal à 20 ou lorsque l'erreur type est de l'ordre de 0.2.

Le niveau d'habileté initial est inféré à partir des niveaux d'habileté des candidats qui ont déjà passé le test. Ainsi une comparaison est faite entre le profil du nouveau candidat et ceux stockés dans la base des profils. La moyenne des niveaux d'habileté des candidats dont le profil se rapproche le plus de celui du nouveau candidat sera son niveau d'habileté initial (voir la section 5.4.1). C'est la technique du raisonnement à base de cas.

L'intégration du raisonnement à base de cas a rendu le test adaptatif *plus court et plus efficace et précis*. Ceci est confirmé dans la partie validation avec les résultats obtenus après avoir testé le système sur une centaine de candidats. TAARAC est bien un test adaptatif puisque tout le long du test les questions sélectionnées sont de difficulté proche du niveau d'habileté estimé du candidat (la section 6.3.2). De plus, au fur et à mesure que la longueur du test diminue, le pourcentage de l'utilisation du raisonnement à base de cas augmente. Ainsi pour les candidats dont la longueur du test était de l'ordre de 16, 67% avaient pour niveau d'habileté *initial* celui déduit à partir du raisonnement à base de cas, et par conséquent la précision désirée est atteinte rapidement (la section 6.3.3.1). De même avec le raisonnement à base de cas nous avons une grande probabilité que le niveau d'habileté *initial* soit proche du niveau d'habileté *estimé* à la fin du test, ainsi toutes les questions choisies au cours du test seront de difficulté très proche du niveau *réel* du candidat. Le tableau 6.4 montre que le pourcentage des candidats dont le niveau d'habileté *initial* est proche de celui *estimé* à la fin du test est égal à 38% pour ceux dont le niveau d'habileté *initial* est déduit à partir du raisonnement à base de cas, et de l'ordre de 26% pour les autres candidats (section 6.3.3.2).

Dans ce travail, afin de montrer l'efficacité du raisonnement à base de cas dans un test adaptatif, nous nous sommes contentés de faire une comparaison entre les candidats dont le niveau d'habileté initial est celui déduit par le raisonnement à base de cas avec ceux dont le niveau d'habileté initial est celui auto évalué. Mais nous comptons utiliser une autre évaluation, notamment un niveau d'habileté initial déduit à partir de l'historique du candidat.

En comparant TAARAC avec trois des tests les plus connus au niveau international, notamment le TOEFL (the Test Of English as a Foreign Language), Le GMAT (Graduate Management Admission Test) et le MCSE (Microsoft Certified Systems Engineer), nous avons constaté que seul TAARAC fait une estimation du niveau d'habileté initial du candidat. Toutefois, les autres tests choisissent comme première question, celle de difficulté moyenne ou faible. Un choix qui peut rendre le test plus long, puisque le candidat est obligé de répondre au début du test à quelques questions de difficulté éloignée du niveau du candidat avant de recevoir celles de difficulté proche de son niveau d'habileté.

Comme nous l'avons déjà dit précédemment, la banque d'items de TAARAC est conçue à partir d'un test papier d'anglais de la faculté de l'éducation permanente de l'université de Montréal qui ne comprend que 50 questions de grammaire. Pour estimer les paramètres de ces questions, il nous fallait un échantillon d'au moins mille candidats qui avaient passés le test papier. Nous nous sommes quand même contenté de ce test-ci, puisque les responsables du test étaient prêts à nous fournir tous les données dont nous avons besoin. Ces responsables ne gardent généralement pas de grandes quantités de données puisqu'ils s'en débarrassent rapidement. Or, pour nous, ces données sont utiles. Après l'estimation des paramètres des items, et leurs analyses, nous nous sommes retrouvés avec 46 items de difficulté moyenne ou faible. Par conséquent, seuls les candidats qui sont de niveau moyen ou faible en anglais seront bien classés par TAARAC.

La méthodologie de TAARAC peut être généralisée à n'importe quel domaine d'application : il suffit de remplir la base d'items par des données propres au domaine, et de remplacer les attributs qui constituent le profil du candidat par ceux qui sont significatifs pour ce domaine. De plus, le langage de programmation utilisé est *multiplateforme*, donc nous pouvons intégrer TAARAC dans différents systèmes d'apprentissage. Enfin il existe plusieurs applications où TAARAC peut être exploité, notamment les systèmes d'apprentissage à distance ou les tests de classement en ligne.

Dans un environnement d'apprentissage à distance où nous intégrerons TAARAC, l'apprenant peut être évalué à tout moment par TAARAC et ceci tout le long de sa formation indépendamment du niveau d'apprentissage atteint par celui-ci. Ainsi nous évitons la conception d'un nouveau test pour chaque niveau atteint par l'apprenant.

La plupart des formations nécessitent un ou plusieurs tests préalables, pour évaluer le niveau des apprenants. D'autres formations comme celles des langues secondes exigent le passage d'un test de classement afin d'affecter le candidat dans le niveau adéquat. Avec TAARAC, le test est plus efficace et plus court et le candidat ne sera pas ennuyé. Puisque la plupart des personnes qui ont évalué TAARAC le trouvent agréable, convivial et clair comme nous l'avons vu dans la partie validation. Nous avons

omis de prendre l'avis des enseignants sur TAARAC à cause de la contrainte du temps, mais c'est un cas que nous envisageons faire prochainement.

Actuellement TAARAC se focalise sur la partie grammaire en anglais. Nous comptons intégrer dans le futur, d'autres aspects d'anglais à TAARAC, notamment le vocabulaire, la compréhension orale et écrite. Nous comptons également étendre les fonctionnalités de TAARAC à d'autres domaines autres que l'anglais.

## Bibliographie

[Aamodt & Plaza 1994] Aamodt, A., Plaza, E. *Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches*. AI Communications. IOS Press, Vol. 7, No. 1, 1994, pp. 39-59.

[Acros & Lòpez de Màntaras 2001] Acros, J. L., Lòpez de Màntaras, R. *SaxEx: An interactive case-based reasoning system for generating expressive music*. Applied Intelligence, Vol. 14, No. 1, 2001, pp. 115-129.

[Aha *et al.* 2005] Aha, D. W., Molineaux, M., Ponsen, M. *Learning to win: case-based plan selection in a real-time strategy game*. Dans Munoz-Avila, H. et Ricci, F. Eds. *Proceedings of the sixth International Conference on case-Based Reasoning*. Berlin: Springer, 2005, pp. 5-20.

[Andrich 1978] Andrich, D. *A rating formulation for ordered response categories*. Psychometrika, Vol. 43, No. 4, 1978, pp. 561-573.

[Ashley & Rissland 1988] Ashley, K. D., Rissland, E. L. *A case-based approach to modeling legal expertise*. IEEE Expert, Vol. 3, No. 3, 1988, pp. 70-77.

[Azuaje *et al.* 2000] Azuaje, F., Dubitzky, W., Black, N., Adamson, K. *Retrieval strategies for case-based reasoning: a categorised bibliography*. The Knowledge Engineering Review, Vol. 15, No. 4, 2000, pp. 371-379.

[Bergmann *et al.* 2006] Bergmann, R., Kolodner, J., Plaza, E. *Representation in case-based reasoning*. The Knowledge Engineering Review, Vol. 20, No. 3, pp. 209-213.

[Bertrand & Blais 2004] Bertrand, R., Blais, J. *Modèles de mesure : L'apport de la théorie des réponses aux items*, Eds. Presses de l'université de Québec, Québec, 2004.

[Bock *et al.* 1982] Bock, R. D., Mislevy, R. J., & Woodsen, C. E. M. *The next stage in educational assesement*. Educational Researcher, Vol. 11, 1982, pp. 4-11.

[Boyd 2003] Boyd A. M. *Strategies for Controlling Testlet Exposure Rates in Computerized Adaptive Testing Systems*. Thèse de doctorat. Austin : Université de Texas, 2003, 161 pages.

[Chang & Ying 1999] Chang, H., Ying, Z. *a-Stratified Multistage Computerized Adaptive Testing*. *Applied Psychological Measurement*, Vol. 23, No. 3, 1999, pp. 211–222.

[Chang *et al.* 2001] Chang, H., Qian, J., Ying, Z. *a-Stratified Multistage Computerized Adaptive Testing With b Blocking*. *Applied Psychological Measurement*, Vol. 25, No. 4, 2001, pp. 333–341.

[Chang & Ying, 2003] Chang, Y., Ying, Z. *Sequential estimation in variable length computerized adaptive testing*. *Journal of Statistical Planning and Inference*, Vol. 121, 2004, pp. 249-264.

[Cheetham 2005] Cheetham, W. *Tenth Anniversary of Plastics Color Matching*. *Artificial Intelligence Magazine*, Vol. 26, No. 3, 2005, pp. 51-61.

[Choppin 76] Choppin B. *Recent developments in item banking*. *Advances in psychological and educational measurement*, Eds. D. N. de Gruijter & L. J. van der Kamp, London, 1976, pp. 233-245.

[Collins *et al.* 1996] Collins, J. A., Greer, J. E., Huang, S. X. *Adaptive assessment using granularity hierarchies and bayesian net*. In *Intelligent Tutoring Systems. Lecture Notes in Computer Science*, Vol. 1086, (Proceedings of Third International Conference on Intelligent Tutoring Systems, ITS 96, Montreal, June, 1996) Berlin: Springer Verlag, pp. 568-577.

[Conejo *et al.* 2004] Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-De-La-Cruz, J. L., Ríos, A. *SIETTE: A Web-Based Tool for Adaptive Testing*. *International Journal of Artificial Intelligence in Education*, Vo. 14, 2004, pp. 1-33.

- [De Ayala *et al.* 1995] De Ayala, R.J., Schafer, W.D., Sava-Bolesta, M. *An investigation of the standard errors of expected a posteriori ability estimates*. British journal of mathematical and statistical psychology, Vol. 48, No. 2, 1995, pp. 385-405.
- [Desmarais & Maluf 1996] Desmarais, M. C., Maluf, A. *User-Expertise Modeling with Empirically Derived Probabilistic Implication Networks*. User Modeling and User-Adapted Interaction, Vol. 5, No 3-4, 1996, p 283-315.
- [Desmarais & Pu 2005] Desmarais, M. C., Pu, X. *A bayesian inference adaptive testing framework and its comparison with item response theory*. International Journal of Artificial Intelligence in Education, Vol. 15 , pp. 291-323.
- [Desmarais *et al.* 2006] Desmarais, M. C., Meshkinfam, P., Gagnon, M. *Learned Student Models with Item to Item Knowledge Structures*. User Modeling and User-Adapted Interaction, Vol. 16, No 5, 2006, pp. 403-434.
- [Finnie & Sun 2003] Finnie, G., Sun, Z., *R<sup>5</sup> model for case-based reasoning*. Knowledge-Based Systems, Vol. 16, 2003, pp. 59-65.
- [Frederiksen *et al.* 93] Frederiksen, N., Mislevy, R. J., Bejar I. I. *Test Theory for a new generation of Tests*, Eds. Lawrence Erlbaum Associates, Hillsdale, 1995.
- [Gómez-Martín *et al.* 2005] Gómez-Martín, P. P., Gómez-Martín, M. A., Díaz-Agudo, B., Gongález-Calero, P. A. *Opportunities for CBR in learning by Doing*, dans Munòz-Avila, H. et Ricci, F. Eds. Cas-based Reasoning Research and development, Proceeding of 6<sup>th</sup> International Conference on Case-based Reasoning (lecture Notes in Artificial Intelligence, 3620). Berlin: Springer, 2005, pp. 267-281.
- [Hambleton & Swaminathan 85] Hambleton, R. K., Swaminathan, H. *Item Response Theory: Principles and application*, Eds. Kuwer.Nijhoff, Boston, 1985.
- [Hammond 1989] Hammond, K. J. *Case-based planning: Viewing planning as a memory task*, Eds. Academic Press, Boston, 1989.



[Hinkle & Toomey, 1994] Hinkle, D., Toomey, C. N. *CLAVIER: Applying case-based reasoning on to composite part fabrication*. Proceeding of the Sixth Innovative Application of AI Conference, Seattle, WA, AAAIPress, 1994, pp. 55-62.

[Holt & Benwell 1999] Holt, A., Benwell, GI. *Applying case-based reasoning techniques in GIS*. International Journal of Geographical Information Science, Vol. 13, No. 1, 1999, pp. 9-25.

[Jurisica & Glasgow, 2004] Jurisica, I., Glasgow, J. L. *Applications of case-based reasoning in molecular biology*. AI Magazine, Vol. 25, No. 1, 2004, pp. 85-96.

[Kolodner 1993] Kolodner, J. *Case-based reasoning*, Eds. Morgan Kaufmann, San Mateo, 1993.

[Kolodner et al. 2006] Kolodner, J. L., Cox., M. T., Gonzalez-Calero, P. A. *Case-based reasoning-inspired approaches to education*. The Knowledge Engineering Review, Vol. 20, No.3, 2006, pp. 299-303.

[Koton 1988] Koton, P. *Reasoning about evidence in causal explanations*. In Proceedins of AAAI-88. Cambridge, MA: AAAI Press/MIT Press, 1988, pp. 256-261.

[Laurier 1999] Laurier, M. *The development of an adaptive test for placement in French*. Studies in Language Testing, Vol. 10, pp.122-135.

[Leake 1996] Leake, D. *Case-Based Reasoning Experiences, Lessons, & Future Directions*, Eds. AAAI Press/ The MIT Press, Menlo Park, California/ Cambridge, Massachusetts/ London, England, 1996.

[Legree et al. 1998] Legree, P. J., Fischl, M. A., Gade, P. A., Wilson, M. *Testing Word Knowledge by Telephone to Estimate General Cognitive Aptitude Using an Adaptive Test*. Intelligence, Vol. 26, No. 2, 1998, pp. 91-98.

[Li & Schafer, 2005] Li, Y. H., Schafer, W. D. *Increasing the Homogeneity of CAT's Item-Exposure Rates by Minimizing or Maximizing Varied Target Functions While*

*Assembling Shadow Tests*. ProQuest Psychology Journals, Vol. 42, No. 3, 2005, pp. 245-269.

[Liao *et al.* 1998] Liao, T. W., Zhang, Z., Mount, C. R. *Similarity measures for retrieval in case-based reasoning*. Applied Artificial Intelligence, Vol. 12, 1998, pp. 267-288.

[Lopez de Mantaras *et al.* 2006] Lopez de Mantaras, R., Mcsherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M. L., Cox, M. T., Forbus, K., Keane, M., Aamodt, A., Watson, I. *Retrieval, reuse, revision and retention in case-based reasoning*. The Knowledge Engineering Review, Vol. 20, No. 3, 2006, pp. 215-240.

[Lord 1974] Lord, F. M. *Estimation of latent ability and item parameters when there are omitted responses*. Psychometrika, Vol.39, 1974, pp. 247-264.

[Lord 1980] Lord, F. M. *Application of item response theory to practical testing problems*, Eds. Lawrence Erlbaum Associates, Hillsdale, 1980.

[Lord & Novick 1968] Lord, F. M., Novick, M. R. *Statistical theories of mental test scores*, Eds. Addison-Wesley, Reading, 1968.

[Marling & Cheetham 2006] Marling, C., Cheetham, B. *Case-Based Reasoning: Theory and Application*. AAAI-06, 2006, [http://oucsace.cs.ohiou.edu/~marling/aaai06\\_tutorial.pdf](http://oucsace.cs.ohiou.edu/~marling/aaai06_tutorial.pdf)

[Marvelde *et al.* 2006] Marvelde, J. M., Van Landeghem, G., Van Damme, J. *Application of Multidimensional Item Response Theory Models to Longitudinal Data*. Educational and Psychological Measurement, Vol. 66, No 1, 2006, pp. 5-34.

[Meijer & Nering 1999] Meijer, R. R., Nering, M. L. *Computerized Adaptive Testing: Overview and Introduction*. Applied Psychological Measurement, Vol. 23, No. 3, 1999, pp.187–194.

[Messick et al. 1983] Messick, S., Beaton, A. E., & Lord, F. M. *National Assessment of Educational Progress reconsidered: A new design for a new era* (NAEP Rep. 83-1), Eds. National Assessment for Educational Progress, Princeton, 1983.

[Millàn & Pérez-de-la-Cruz 2002] Millàn, E., Pérez-de-la-Cruz, J. L. *A Bayesian Diagnostic Algorithm for Student Modeling and its Evaluation*. User Modeling and User-Adapted Interaction, Vol. 12, 2002, pp. 281-330.

[Raïche 2000] Raïche, G. *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, 2000, 407 pages.

[Raïche & Blais 2002] Raïche, G. et Blais, J.-G. *Practical considerations about expected a posteriori estimation in adaptive testing: adaptive a priori, adaptive correction for bias, and adaptive integration interval*. Communication dans le 11e Biennial international objective measurement workshop. New Orleans, 2002, LO: IOMW. [ERIC DOCUMENT NO ED 464 110]

[Raymond et al. 1997] Raymond, J. A., Wilson, M. and Wang, W. *The multidimensional random coefficients multinomial logit model*. Applied Psychological Measurement, Vol. 21, No. 1, 1997, pp.1-23.

[Reinartz et al. 2001] Reinartz, T., Iglezakis, I., Roth-Berghofer, T. *Review and restore for case-based maintenance*. Computational Intelligence, Vol. 17, No. 2, 2001, pp. 214-234.

[Riesbeck & Schank 1989] Riesbeck, C. K., Schank, R. C. *Inside Case-Based Reasoning*, Eds. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1989.

[Salcedo et al. 2005] Salcedo, Pinninghoff, M. A. and Contreras, R. *Computerized Adaptive Tests and Item Response Theory on a Distance Education Platform*.

International Work-conference on the Interplay between Natural and Artificial Computation, IWINAC 2005, LNCS 3562, pp. 613–621.

[Van Der Linden 1998] Van Der Linden, W. J. *Bayesian item selection criteria for adaptive testin.*, Psychometrika, Vol. 63, NO. 2, 1998, pp. 201-216.

[Van der Linden & Reese 1998] Van Der Linden, W. J., Reese, L. M. *A model for optomal constrained Adaptive testing.* Applied psychological measurement, Vol. 22, No. 3, 1998, pp. 259-270.

[Van Rijn *et al.* 2002] Van Rijn, P. W., Eggen, T. J. H. M., Hemker, and Sanders, P. F. *Evaluation of Selection Procedures for Computerized Adaptive Testing With Polytomous Items.* Applied Psychological Measurement, Vol. 26, No. 4, 2002, pp. 393–411.

[Vomlel 2004] Vomlel, J. *Bayesian networks in educational testing.* Internatinal journal of uncertainty fuzziness and knowledge-based Systems, Vol. 12, 2004, pp. 83-100.

[Wainer 2000] Wainer, H. *Computerized Adaptive Testing: A Primer*, 2nd Edition, Eds. Lawrence Erlbaum, Hillsdale, New Jersey, 2000.

[Weiss 1984] Weiss, D. J. *Application of computerized adaptive testing to educational problems.* Journal of educational measurement, Vol. 21, 1984, pp. 361-376.

[URL1] <http://www.childrens-mercy.org/stats/definitions/biserial.htm>

[URL2] <http://www.ime.usp.br/~mbranco/paperAPMfinal.pdf>

[URL3] <http://www.ssicentral.com/>

[URL4] <http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/?vgnextoid=b195e3b5f64f4010VgnVCM10000022f95190RCRD>

[URL5] [http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/?vgnextoid=69c0197a484f4010VgnVCM10000022f95190RCRD&WT.ac=Redirect\\_ets.org\\_toefl](http://www.ets.org/portal/site/ets/menuitem.fab2360b1645a1de9b3a0779f1751509/?vgnextoid=69c0197a484f4010VgnVCM10000022f95190RCRD&WT.ac=Redirect_ets.org_toefl)

[URL6] <https://www.ncsbn.org/1200.htm>

[URL7] <http://edres.org/irt/param/>

[URL8] <http://www.mba.com/mba/takethegmat>

[URL9] <http://www.vue.com/>

[URL10] <http://www.microsoft.com/france/formation/cert/mcse.mspx>

[URL11] <http://www-etud.iro.umontreal.ca/~lakhilz/taarac/main.php>