Sous la direction de

Jean-Guy Blais

Évaluation des apprentissages et technologies de l'information et de la communication

ENJEUX, APPLICATIONS ET MODÈLES DE MESURE

pul

Évaluation des apprentissages et technologies de l'information et de la communication

ENJEUX, APPLICATIONS ET MODÈLES DE MESURE

Page laissée blanche intentionnellement

Évaluation des apprentissages et technologies de l'information et de la communication

ENJEUX, APPLICATIONS ET MODÈLES DE MESURE

Sous la direction de
Jean-Guy Blais

Les Presses de l'Université Laval

Les Presses de l'Université Laval reçoivent chaque année du Conseil des Arts du Canada et de la Société d'aide au développement des entreprises culturelles du Québec une aide financière pour l'ensemble de leur programme de publication.

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

La publication de cet ouvrage a été rendue possible grâce à une subvention du programme équipe du Fonds Québécois de la recherche sur la société et la culture (FQRSC).

Mise en pages : Diane Trottier

Maquette de couverture : Hélène Saillant

© Les Presses de l'Université Laval 2009
Tous droits réservés. Imprimé au Canada
Dépôt légal 1er trimestre 2009

ISBN 978-2-7637-8711-4

Les Presses de l'Université Laval
Pavillon Pollack, bureau 3103
2305, rue de l'Université
Université Laval, Québec
Canada, G1V 0A6
www.pulaval.com

Table des matières

CHAPITRE 1
Évaluation des apprentissages et technologies de l'information et de la communication : un long chemin de parcouru et encore beaucoup à accomplir 1
Jean-Guy Blais

SECTION I
DES SYSTÈMES À DÉVELOPPER ET À TESTER

CHAPITRE 2
Vers une architecture de testing assisté par ordinateur pour l'évaluation des acquis scolaires dans les systèmes éducatifs orientés sur les résultats 13
Romain Martin, Gilbert Busana et Thibaud Latour

Introduction 13
Relations fortes entre *e*-learning et *e*-testing 15
Un bref historique du TAO. 16
L'évaluation comme outil de gestion des systèmes éducatifs et des compétences 19
Une approche intégrée pour l'évaluation des apprentissages 19
Nouveaux besoins pour des outils de TAO avancés 23
Défis fonctionnels pour une plateforme d'évaluation intégrée 27
Architecture logicielle 27
Diversité et hétérogénéité 28
Distribution, partage des ressources et connectivité. 29
Modularité et flexibilité. 30
Sécurité et protection de la vie privée 30
Capacité d'analyse et retour pédagogique 31
Conclusion 31

CHAPITRE 3
Le développement d'une application Web pour la passation d'épreuves adaptatives 35
Nancy Boiteau, Jean-Guy Blais et Martin Paré

Introduction 35
Caractéristiques théoriques de l'application 39
Modalité adaptative à deux niveaux 39
Modalité adaptative classique 40
Théorie de la réponse à l'item : un aperçu 41
Caractéristiques méthodologiques de l'application 42
Création de la banque d'items 42
Méthode d'estimation de l'habileté 43
La sélection des items 43
Temps de passation des tests 45
Caractéristiques fonctionnelles de l'application 45
Une utilisation avec une modalité traditionnelle 47
Développement de la deuxième génération de l'application 49

SECTION II
DES DÉVELOPPEMENTS ET
DES APPLICATIONS MULTIPLES

CHAPITRE 4
L'évaluation des compétences langagières 57
Réginald Burton, Monique Reichert et Romain Martin

Introduction 57
Le potentiel de l'ÉAO pour la mesure des compétences langagières 58
Possibilités du multimédia 58
Formats d'items 60
Le testing adaptatif 62
Interactions enrichies et systèmes d'auteur 63
Correction automatique 65
La neutralisation de la composante linguistique dans les systèmes multilingues 67

Validité des mesures. 69
Les études de comparaison 70
Les études de groupes contrastés 70
Les études de processus 71
Cadres de références et validité critérielle 71
Conclusions et perspectives 74

CHAPITRE 5
L'apport des technologies dans l'évaluation de la compétence à écrire 77
Michel D. Laurier et Luc Diarra

Introduction 77
L'évaluation de la compétence à écrire en mode TIC 79
Les TIC et le traitement de la réponse : la correction automatisée des productions écrites 86
Fiabilité et validité des systèmes de correction automatisée 93
Les avantages et les limites de la correction automatisée des productions écrites 97
Les avantages et les limites de la correction automatisée 97
Conclusion 101

CHAPITRE 6
L'évaluation des apprentissages en mathématiques – Révéler l'invisible. 105
France Caron et Nathalie Loye

Introduction 105
Des pratiques évaluatives en évolution. 105
Les pratiques évaluatives selon le renouveau pédagogique 108
Évaluer les compétences en mathématiques 109
La technologie et les pratiques évaluatives 112
Mise à contribution des outils informatiques pour l'évaluation de la compétence à résoudre une situation-problème complexe 113
Une approche diagnostique de l'utilisation des technologies par les élèves 115
Contraintes et défis 118
Conclusion 119

CHAPITRE 7
L'évaluation en ligne du raisonnement clinique médical par le test de concordance de script 123
Bernard Charlin et Driss Kazi-Tani

Introduction 123
Contexte théorique 124
Format du test 125
Méthode d'établissement des scores 126
Qualités psychométriques du test 127
Le TCS et la technologie informatique 129
Le contexte 130
La structure générale 130
Perspectives de développement 133
Conclusion 134

CHAPITRE 8
L'évaluation du profil psychosocial de demandeurs d'emploi, une population peu familiarisée à la technologie informatique 137
Claude Houssemand, Raymond Meyers et Romain Martin

Introduction 137
Comparaison entre scores obtenus en mode papier-crayon et en mode informatisé 139
Tests d'attitudes, de personnalité et d'opinions 139
Tests de performance ou d'aptitudes 140
Bilan de l'équivalence des versions 141
L'anxiété en face des tests, et le TAO auprès des populations défavorisées 142
Anxiété informatique et évaluations par ordinateur 144
L'anxiété informatique affecte-t-elle les résultats au TAO ? 144
Anxiété informatique et publics défavorisés 145
Un outil informatisé utilisé avec des chômeurs 145
Mise au point de l'outil 145
Mode de passation 147
Un outil utilisable avec des publics défavorisés 148
Validité des scores obtenus en mode informatisé auprès d'une population défavorisée 148
Équivalence du GHQ12 pour les non-qualifiés 150
Conclusions et recommandations 151

SECTION III
LES DÉFIS D'UNE APPROCHEADAPTIVE ET DIAGNOSTIQUE

CHAPITRE 9
Les modèles cognitifs 159
Nathalie Loye

Introduction 159
La matrice Q des attributs cognitifs 160
Les conditions expérimentales 163
Le modèle de fusion reparamétrisé 165
L'élaboration des matrices Q individuelles 168
L'élaboration des matrices Q synthèses 169
Les résultats liés à l'élaboration des matrices 171
La perception des experts 171
Les résultats liés à la modélisation RUM 174
Les matrices Q synthèses simplifiées 174
Les comparaisons des paramètres du RUM 176
La comparaison des trois conditions expérimentales de l'étude 181
Conclusion 182

CHAPITRE 10
Les modèles graphiques 187
Michel C. Desmarais et Xiaoming Pu

Introduction 187
Les modèles graphiques et l'évaluation de la connaissance 189
Concepts, items et nœuds cachés 189
L'approche de la TRI dans la perspective graphique 192
Le formalisme des réseaux bayésiens 193
L'approche POKS et la théorie des espaces de connaissances 194
La théorie des espaces de connaissances 195
L'inférence de la connaissance 199
D'un ordre partiel à un ensemble de réseaux à un niveau 201
Relations symétriques 202

Comparaison des prévisions des différents modèles 203
Comparaison de POKS avec la TRI 204
Comparaison de POKS avec un réseau bayésien 205
Combiner les approches 206
Conclusion 207

CHAPITRE 11
La médiatisation de récits d'expérience professionnelle 211
Kathye Bélanger et Michel Laurier

Introduction 211
Réfléchir sur l'action pour développer le jugement 212
Le REPM, un véhicule pour l'expérience professionnelle 218
RegulUM, la solution prototypée 222
La collecte des données 226
Les premiers résultats 228
Un cheminement individualisé 230
Le compagnonnage cognitif 231
Diagnostiquer pour mieux réviser 233
Conclusion 236

SECTION IV
REGARDS SUR DES DIMENSIONS
TECHNIQUES AVANCÉES

CHAPITRE 12
Des normes techniques pour les systèmes TIC associés à l'évaluation des apprentissages 243
Gilles Raîche, Martin Riopel, Martin Lesage et Komi Sodoké

Introduction 243
Les conventions du noyau de Dublin *(Dublin Core)* 244
La norme CMI *(Courseware Management Interface)* 247
La norme LOM (Learning Object Metadata) 248
La norme IMS (Instructional Management System) 249
La norme SCORM *(Sharable Content Object Reference Model)* 250

La norme *MINE SCORM* 251
Problèmes avec la normalisation des objets d'évaluation 251
Conclusion 253

CHAPITRE 13
La production automatisée de tâches d'évaluation. 257
Martin Riopel, Gilles Raîche, Mélissa Pilote et Patrice Potvin

Introduction 257
Premiers travaux 258
Travaux plus récents 260
Propriétés essentielles et propriétés accessoires. 261
Modélisation cognitive du raisonnement abstrait 263
Modélisation du raisonnement quantitatif 266
Modélisation de la compréhension de texte. 267
Modélisation dans un contexte informatisé. 269
Conclusion 269

CHAPITRE 14
La détection des patrons de réponses problématiques dans le contexte des tests informatisés. 275
Jean-Guy Blais, Gilles Raîche et David Magis

Introduction 275
Le contexte 276
Les indices de détection des patrons de réponses problématiques. 277
L'intégration de paramètres associés aux personnes 279
Conclusion et implications pour les tests adaptatifs. 288

CHAPITRE 15
La méthode alpha stratifiée pour le contrôle de la fréquence d'exposition des items dans les tests adaptatifs 293
Éric Dionne et Jean-Guy Blais

Introduction 293
Problématique. 295
Le contrôle de l'exposition des items : définitions et contexte 295
Les méthodes de contrôle de l'exposition des items. 296
La méthode alpha-stratifiée (ASTR) 297

Les statistiques pour la comparaison des méthodes 298
Le biais et le carré moyen de l'erreur (CME) 299
La statistique du khi-carré (χ^2) 299
Le nombre d'items sur-utilisés ou sous-utilisés (TEI) 299
Le taux de chevauchement des items (TCI) 300
La comparaison des résultats de la recherche 300
Les études retenues 300
Résultats et discussion 301
Conclusion 307

CHAPITRE 1

Évaluation des apprentissages et technologies de l'information et de la communication : un long chemin de parcouru et encore beaucoup à accomplir

Jean-Guy Blais
Université de Montréal

D'entrée de jeu, précisons que ce livre n'est pas un ouvrage motivé par une forme de « technotopie » ou de « technovangélisme »[1]. Il y a déjà suffisamment d'ouvrages de ce type disponibles actuellement en librairie, et il n'est pas nécessaire d'en ajouter un autre sur les rayons. Cet ouvrage ne vise pas donc à vendre à tout prix les « nouvelles » technologies comme une panacée aux différents problèmes que peut soulever l'évaluation des apprentissages en éducation/formation. Son objectif est d'un autre ordre. Il présente une vision actualisée et ciblée de différentes composantes propres au développement et à l'application de systèmes technologiques intégrés pour l'évaluation des apprentissages. Il se situe dans la lignée de publications semblables en langue anglaise éditées depuis 1980 et qui ont la particularité d'être des productions collectives posant un regard large sur les différents enjeux de l'utilisation de la technologie pour médiatiser les opérations de présentation des tâches, de récolte des données, de stockage, de correction et de notation en vue de l'évaluation des apprentissages (Weiss, 1983 ; Wainer *et al.*, 1990 ; Sands, Waters et McBride, 1997 ; van der Linden et Clas, 2000 ; Wainer *et al.*, 2000 ; Anderson *et al.*, 2002 ; Mills *et al.*, 2002 ; Shermis et

1. Expressions inspirées par Oppenheimer (2004) et résultant des contractions de *technologie* et *utopie* d'une part, et de *technologie* et *évangélisme,* d'autre part

Burstein, 2003 ; Hricko et Howell, 2006 ; Bartram et Hambleton, 2006 ; Williams, Howell et Hricko, 2006). Ce livre se retrouve donc en bonne compagnie, tout en exposant pour une première fois à un lectorat francophone certains des derniers développements en matière de *testing* informatisé. Il est en bonne partie le fruit des travaux de recherche des membres du GRIÉMÉtic, de l'Université de Montréal et de ceux de l'unité de recherche EMACS, de l'Université du Luxembourg[2].

Le livre est divisé en quatre sections thématiques. La première section présente deux systèmes/plateformes pour la gestion d'épreuves informatisées sur table ou en ligne, et elle propose une synthèse des caractéristiques souhaitables des systèmes technologiques dédiés à l'évaluation. La deuxième section présente un éventail de domaines d'applications (langue, écriture, mathématiques, santé, populations particulières) et renferme plusieurs informations intéressantes et utiles pour les profanes comme pour les spécialistes. La troisième section s'attarde aux défis et aux enjeux des systèmes adaptatifs et des systèmes diagnostiques, deux applications où, en principe, la « nouvelle » technologie pourrait présenter plusieurs de ses avantages. Finalement, la quatrième section présente des dimensions techniques plus avancées et pourrait être abordée plus tard par les personnes qui se sentent moins à l'aise avec les approches plus formelles (par exemple, les mathématiques). Le contenu des différents chapitres de l'ouvrage est résumé à la fin de cette introduction et permettra au lecteur de se l'approprier en fonction de ses intérêts et de ses priorités.

Le recours à la technologie comme solution aux problèmes de l'éducation, de l'enseignement, de l'apprentissage et de l'évaluation n'est pas apparu subitement, en même temps que les ordinateurs personnels devenaient plus accessibles et que le réseau Internet étalait sa puissance. Durant tout le vingtième siècle, l'éducation a été un terreau plus que fertile pour l'utilisation des technologies de support ou de substitution à l'enseignement. On y a vu défiler, les unes à la suite des autres, de « nouvelles » technologies censées révolutionner l'éducation et remplacer d'une part l'enseignant, et d'autre part, la technologie de la craie et du tableau. Ainsi, à tour de rôle, le cinéma, la radio, la télévision se sont vus portés aux nues et glorifiés quant aux miracles qu'ils pourraient faire en éducation. Cependant, lorsqu'est proposé un bilan des retombées de toutes ces technologies utilisées pour l'apprentissage en salle de classe, force est de constater qu'il est difficile à faire (Cuban, 2001 ; Oppenheimer, 2004). Le dernier prétendant en liste dans cette catégorie est

2. GRIÉMÉtic, pour Groupe de recherche interuniversitaire sur l'évaluation et la mesure en éducation à l'aide des technologies de l'information et de la communication ; EMACS, pour *Educational Measurement and Applied Cognitive Science*.

le réseau des réseaux, c'est-à-dire Internet, et seul le temps permettra de juger si ses résultats seront à la hauteur de ses ambitions. Naturellement, l'évaluation des apprentissages n'a pas été épargnée par la dernière vague technologique, et on assiste à une tendance lourde actuellement dans les opérations de *testing* à grande échelle en continu, où le crayon et le papier comme modalités de passation sont en voie d'être remplacés par le clavier et l'écran (Bennett, 2001). Ce changement de garde technologique accompagne également l'essor, considérable au dernier siècle, de la technologie des tests, épreuves et examens écrits pour apprécier les connaissances, les compétences et la performance des individus (Blais, 2002). La convergence et l'évolution de ces deux dimensions techniques depuis les vingt dernières années rendraient ainsi inexorable, pour certains, la substitution de l'écran et du clavier au papier et au crayon pour l'évaluation des apprentissages (Bennett, 2002).

Dans plusieurs domaines, les outils technologiques privilégiés pour compiler des données sur les apprentissages sont relativement simples et assez standards. En effet, les tâches contenues dans les épreuves et les tests sont traditionnellement présentées sur du papier, et on y répond de différentes façons avec un crayon sur des feuilles de papier (de là l'expression « épreuve papier-crayon »). Un autre médium, plus flexible et plus ouvert comme « l'ordinateur », supportant les tâches et servant à les présenter aux candidats, peut, selon les domaines, complètement transformer les pratiques en la matière à plus ou moins brève échéance. Grâce à ce médium de support des tâches, il est actuellement possible de concevoir des tâches plus complexes, utilisant les capacités du multimédia et intégrant des images et du son (en attendant la réalité virtuelle). L'information peut être renvoyée très rapidement aux personnes concernées (enseignants, formateurs, élèves, parents, administrateurs, employeurs). Ce support permettrait également de mieux répondre à certains besoins d'individualisation de l'évaluation en proposant des processus adaptatifs.

Les modalités d'enseignement/formation pourraient aussi changer passablement dans les années à venir. L'entrée en jeu de nouveaux instruments pour observer et recueillir les données influencera les connaissances et compétences valorisées, enseignées et, donc, évaluées. Certaines seront les mêmes qu'actuellement, mais il y en aura d'autres, nouvelles. Le jugement de l'humain ne sera pas éliminé du processus d'évaluation, mais son apport sera réservé à des tâches de jugement complexes plutôt que normatives, et à des interventions formatives qui devraient être facilitées par la présence d'outils pour le diagnostic.

Cet ouvrage propose donc un aperçu à la fois général et particulier de la problématique de l'évaluation des apprentissages en éducation/formation

avec les technologies de l'information et de la communication. Certains chapitres sont plus techniques et pourraient sembler de prime abord plus rébarbatifs, mais ils sont essentiels, car ils sont partie prenante des enjeux actuels et ne devraient pas être ignorés. D'autres chapitres font le tour complet de la question et mettent bien en lumière les instruments TIC développés pour évaluer les apprentissages dans certains domaines.

La seule combinaison clavier/écran/ordinateur ne suffit évidemment pas pour récolter des données ; c'est une question de quincaillerie, essentielle certes, mais pas suffisante. Il faut également disposer d'un système ou d'une plateforme qui permettent de gérer et de contrôler le processus, et c'est ce que nous proposent Martin, Latour et Busana au chapitre deux, et Boiteau, Blais et Paré au chapitre trois. Les premiers présentent succinctement le *testing* assisté par ordinateur et dressent le portrait des besoins liés à l'évaluation des compétences assistée par ordinateur dans le monde de l'éducation. Parmi ces besoins, les aspects de distributivité et de travail collaboratif pour le développement et la délivrance des tests se dégagent comme les plus fondamentaux. Pour leur part, Boiteau, Blais et Paré présentent une application Web de la gestion de tests, application développée dans le cadre des travaux de Boiteau (2004) qui visaient à comparer l'efficacité de plusieurs modalités de *testing* adaptatif. Le chapitre présente les caractéristiques théoriques, méthodologiques et fonctionnelles de l'application Web.

Une belle plateforme sans domaine et sans contenu n'est guère utile. La deuxième section, qui contient les chapitres quatre à huit, pose ainsi un regard sur diverses possibilités d'applications et sur divers développements dans plusieurs domaines. D'abord, pour l'évaluation des compétences linguistiques, Burton, Reichert et Martin illustrent bien au chapitre quatre que les potentialités offertes par l'ordinateur tiennent avant tout aux possibilités multimédias, aux nouveaux formats d'items, aux modalités d'interactions enrichies, aux gains de temps offerts par le *testing* adaptatif et aux possibilités de correction automatique. Ces caractéristiques ont conduit à l'émergence de nouveaux dispositifs d'évaluation prédictive, formative, certificative ou dynamique qui n'étaient pas envisageables dans le format papier-crayon. L'évaluation de la compétence à écrire est l'objet du chapitre cinq, par Laurier et Diarra, qui se demandent si elle doit, d'une part, prendre en compte la modalité TIC, autrement dit accepter la possibilité que l'élève rédige son texte à l'ordinateur, et, d'autre part, recourir à la correction automatisée de ce texte. Étant donné qu'écrire avec l'ordinateur est courant, et même incontournable dans beaucoup de situations de la vie courante, il serait plutôt paradoxal, voire anachronique, que les TIC soient mises de côté lorsqu'il s'agit de l'évaluation de la compétence à écrire.

Le chapitre six, par Caron et Loye, propose une réflexion sur l'évolution récente des pratiques évaluatives en mathématiques. Il met en évidence les enjeux et les difficultés liées à l'intégration de tâches complexes dans l'évaluation en mathématiques, et à l'observation des traces laissées par les élèves lors de leur utilisation de divers outils technologiques dans la résolution de telles tâches. La réflexion aboutit à une discussion sur la possibilité d'utiliser comme outils d'évaluation complémentaires en cours d'apprentissage des épreuves courtes qui intègreraient une dimension technologique et dont l'interprétation des résultats s'appuierait sur la modélisation des processus cognitifs impliqués. L'objectif serait de fournir un diagnostic pour chaque élève afin de réguler ses apprentissages en mathématiques et d'apporter de l'aide aux enseignants dans l'évaluation des tâches plus complexes. Au chapitre sept, Charlin et Kazi-Tani présentent le test de concordance de script (TCS), qui est basé sur une théorie cognitive mettant l'accent sur les réseaux de connaissances qu'utilisent les individus en situation de résolution de problème. Le TCS repose sur le principe que ces multiples jugements peuvent être sondés et que leur concordance avec ceux d'un panel de référence peut être mesurée, ce qui constitue un test du raisonnement clinique. La technologie décuple ses possibilités, aussi bien en matière de création de tests, d'inclusion d'images, ou d'évaluation d'habiletés particulières (la perception, par exemple) que de passation en ligne, de correction automatisée et de transmission de *feedback* aux candidats. Finalement, au chapitre huit, Houssemand, Mayers et Martin rappellent que les différentes populations avec lesquelles certains outils technologiques seront utilisés à des fins d'évaluation doivent aussi faire l'objet d'une attention particulière. Ces auteurs se sont intéressés à une population peu familiarisée à la technologie informatique dans le cadre d'une évaluation du profil psychosocial de demandeurs d'emploi. Leur recherche explore les hypothèses de la familiarisation avec l'outil informatique et de l'anxiété envers l'ordinateur, ces deux dimensions ayant déjà été avancées comme de potentielles sources de variation dans les performances enregistrées.

La troisième section explore à travers trois chapitres les méandres et les avancées du diagnostic automatisé et des interfaces adaptatives, le tout dans une perspective d'évaluation formative. Ainsi, Loye nous rappelle au chapitre neuf que, parallèlement au développement des tests adaptatifs, les vingt dernières années ont vu se multiplier des recherches qui joignent une approche « cognitive » à des modèles psychométriques, et qu'il en résulte l'existence de nombreux modèles visant à fournir un diagnostic en lien avec des variables latentes discrètes représentant des processus cognitifs. Les nouveaux modèles cognitifs qui se développent ont pour l'instant des ambitions modestes. Ils cherchent à modéliser la probabilité de répondre correctement à un item en fonction de la maîtrise d'un ensemble d'attributs

cognitifs propres à cet item. Ce chapitre présente les résultats d'une recherche exploratoire portant sur l'élaboration de la liste des attributs cognitifs en lien avec une épreuve à choix multiple en mathématiques à l'entrée de l'École Polytechnique de Montréal.

De leur côté, Desmarais et Pu, au chapitre dix, soulignent que les tests adaptatifs peuvent être considérés comme une des premières applications de ce qui est aujourd'hui un domaine en pleine effervescence, celui des interfaces adaptatives et personnalisées. Les tests adaptatifs sont peut-être la toute première application à grande échelle du principe de base de ces interfaces : construire un modèle de l'utilisateur et adapter le comportement de l'application en fonction de ce modèle. Mais, alors que la Théorie de la réponse à l'item (TRI) était en plein développement dans le domaine des tests adaptatifs, le domaine des tutoriels intelligents développait aussi ses propres approches en parallèle pour représenter et diagnostiquer les compétences dans des environnements d'apprentissage adaptatifs. Le lien entre les deux domaines a émergé avec les travaux sur les modèles graphiques de la connaissance. Le chapitre fournit une comparaison de différentes approches pour le cheminement adaptatif propres à chaque domaine.

Le chapitre onze, par Bélanger et Laurier, présente les résultats d'une recherche sur l'exploitation du Récit d'Expérience Professionnelle Médiatisé (REPM) comme stratégie interactive pour développer le jugement professionnel chez les futurs enseignants en formation initiale. Le REPM répond à deux grandes exigences : mettre en scène des situations d'expérience professionnelle authentiques, et placer l'apprenant en situation de résolution de problème pour faire appel à ses connaissances stratégiques. Un prototype d'un dispositif multimédia interactif a été développé, et une mise à l'essai a été réalisée auprès d'un groupe de futurs enseignants en formation. Le prototype vise la maîtrise de la pratique de l'évaluation en classe au primaire et au secondaire. Les premiers résultats font ressortir des pistes sur la prise en compte de facteurs pouvant stimuler la réflexion sur l'action dans ce type d'environnement d'apprentissage, plus particulièrement les interrogations sur le cheminement individualisé, le compagnonnage cognitif et la démarche de résolution de problème.

La quatrième et dernière section est la plus technique, et les thèmes qui y sont abordés sont au cœur des réflexions et des développements autour des prochaines générations de systèmes et de plateformes pour l'évaluation assistée par ordinateur. Le chapitre douze, de Raîche, Riopel, Lesage et Sodoke, a pour but de décrire les cinq normes techniques d'apprentissage par ordinateur les plus utilisées (les normes régissant les langages de support, entre autres), et de porter un jugement sur la capacité de ces normes à prendre en compte l'évaluation des apprentissages par ordinateur. Il est à souligner

qu'il s'agit uniquement des normes qui semblent les plus utilisées et qui paraissent connaître la croissance la plus importante. Au chapitre treize, Riopel, Raîche, Pilote et Potvin présentent la problématique de la production automatisée de tâches d'évaluation, qui consiste à utiliser un modèle formel pour prédire les caractéristiques de tâches nouvellement créées. Le chapitre présente les travaux qui ont jeté les bases théoriques de la production automatisée de tâches d'évaluation ainsi que les principales théories qui permettent de modéliser les paramètres d'items dans ce contexte. Au chapitre quatorze, Blais, Raîche et Magis soulignent que le résultat obtenu lors de l'évaluation des apprentissages d'un élève n'est pas toujours représentatif de son niveau d'habileté. Par exemple, l'élève peut avoir copié ses réponses sur celles d'un autre, son résultat étant alors surestimé. Dans d'autres situations, des élèves échouent intentionnellement à des tests de classement en langue seconde. Dans le cas des tests papiers-crayons conventionnels, des indices de détection ont été proposés pour identifier les patrons de réponses inappropriées. Toutefois, dans le cas de la passation de tests adaptatifs par ordinateur, la stratégie consiste à faire en sorte que le niveau de difficulté des questions se rapproche constamment du niveau d'habileté de l'étudiant. Le niveau de difficulté des items, contrairement à ce qui se passe avec les tests papiers-crayons conventionnels, varie peu. Ce chapitre propose donc des indices de détection de réponses inappropriées dans le contexte des tests adaptatifs. Finalement, au chapitre quinze, Dionne et Blais présentent une stratégie développée pour assurer un meilleur contrôle de l'exposition des items dans un test adaptatif : la méthode alpha-stratifiée. Plusieurs variantes associées à cette méthode sont abordées pour montrer les forces et les limites de chacune d'elles.

Depuis le début des années 1990, j'ai contribué de près à trois expériences mettant en scène l'évaluation des apprentissages à l'aide de la technologie informatique. Une première fois, alors que des tests informatisés ont été développés pour un suivi formatif des apprentissages dans le cadre d'un cours à l'Université de Montréal (Blais et Dassa, 1992). Puis, lors d'une deuxième occasion, à nouveau dans le cadre d'un cours à l'Université de Montréal, toujours avec comme objectifs le suivi formatif et le diagnostic, les questions à réponse choisie ont été mises de côté pour explorer le terrain des *Récits d'expériences professionnelles médiatisés* (Bélanger *et al.*, 2004 ; Bélanger et Laurier, chapitre 11). Finalement, lors d'une troisième occasion et dans une situation à plus grande échelle, une application Web et une plateforme de support de tests ont été développées pour une épreuve annuelle faisant figure de prétest en mathématiques pour les étudiants admis en première année à l'École Polytechnique de Montréal (Blais *et al.*, 2007 ; Boiteau *et al.*, chapitre 2). L'expérience du début des années 1990 visait, d'une part, à élaborer et à valider une stratégie d'évaluation formative dans

un contexte d'enseignement collectif et, d'autre part, à sensibiliser les étudiants aux problèmes d'ordre théorique et pratique rencontrés lors de l'implantation d'un modèle pédagogique centré sur l'optimisation de la réussite scolaire. À cette époque, tout de même pas si lointaine, les aspects novateurs de l'approche développée résidaient déjà dans la promesse de « l'intégration des dimensions pédagogiques, évaluatives, administratives et technologiques en une démarche unique ». Des tests informatisés hebdomadaires, fruit d'une collaboration entre plusieurs intervenants, ont été développés à partir d'une banque d'items où chaque item relevait d'un des objectifs spécifiques du cours.

Lorsque je regarde la distance entre le contexte et les résultats de la première expérience et ce que les deux expériences plus récentes ont permis de développer et d'expérimenter, il me semble qu'un bon bout de chemin a été parcouru, et que l'ère du bricolage est derrière nous. Une masse critique de recherches sur différentes composantes du processus d'évaluation des apprentissages à l'aide des technologies de l'information permet maintenant de mieux en saisir les enjeux et les limites, les applications pertinentes et les modèles à perfectionner. Cependant, malgré les avancées réalisées il reste encore beaucoup à accomplir à plusieurs égards et de nombreux défis à relever, comme s'emploient à l'illustrer les chapitres de cet ouvrage.

Références

Anderson, R.S., Bauer, J.F. et Speck, B.W. (dir.) (2002), *Assessment strategies for the on-line class : from theory to practice*, New direction for teaching and learning, **91**.

Bartram, D. et Hambleton, R.K. (dir.) (2006), *Computer-based testing. Issues and advances*, Hoboken, NJ, John Wiley and Sons, Ltd.

Bélanger, K., Blais, J-G., Giardina, M. et Laurier, M.D. (2004), « Le récit d'expérience professionnelle médiatisé, une stratégie interactive pour le développement des compétences », *Res Academica*, 22 (2), p. 197-214.

Bennett, R.E. (2001), « How the Internet Will Help Large-Scale Assessment Reinvent Itself », *Educational Policy Analysis Archives*, 9 (5).

Bennet, R.E. (2002), « Inexorable and Inevitable : The Continuing Story of Technology and Assessment », *The Journal of Technology, Learning and Assessment*, 1(1).

Blais, J.-G. et Dassa, C. (1992), « Enseignement universitaire et évaluation informatisée », *Mesure et évaluation en éducation*, 15, p. 33-52.

Blais, J.-G., Boiteau, N., Burney-Vincent, C., Caron, F., Desmarais, M., Gagnon, M., Grondin, J. et Paré, M. (2007), *La mise en place d'une plateforme de test informatisée pour l'administration et la gestion du prétest de mathématiques de l'École Polytechnique de Montréal*, Conférence présentée à la Maison des technologies pour la formation et l'apprentissage, École Polytechnique de Montréal, novembre.

Blais, J.-G. (2004), « L'obligation de résultats à la lumière de l'interaction entre le quantitatif et le social », dans C. Lessard et P. Meirieu (dir.), *L'obligation de résultats en éducation*, Québec, Les Presses de l'Université Laval, p.123-144.

Cuban, L. (2001), *Oversold and underused : Computers in classrooms* Cambridge, MA, Harvard University Press.

Hricko, M. et Howell, S.L. (dir.) (2006), *Online assessment and measurement : Foundations and challenge*, Londres, Information Science Publishing.

Mils, C.N., Potenza, M.T., Fremer, J.J. et Ward, W.C. (dir.) (2002), *Computer-based tsting. Building the foundations for future assessments*. Mahwah NJ, Lawrence Erlbaum associates.

Oppenheimer (2004), *The flickering mind. Saving education from the false promise of technology*, New York, Random House.

Sands, W.A., Waters, B.K. et McBride, J.R. (dir.) (1997), *Computer Adaptive Testing : From inquiry to operation*, Washington, DC, American Psychological Association.

Shermis, M.D. et Burstein, J.C. (dir.) (2003), *Automated Essay Scoring : A Cross-disciplinary Perspective*, Mahwah NJ, Lawrence Erlbaum associates.

Van der Linden, W.J. et Glas, C.A.W. (dir.) (2000), *Computerized Adaptive Testing : Theory and Practice*, Dordecht, Netherlands, Kluwer Academic Publishers.

Wainer, H, Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L. et Thissen, D. (dir.) (1990), *Computer adaptive testing : A primer*. Mahwah NJ, Lawrence Erlbaum associates.

Wainer, H, Dorans, N.J., Eignor,D., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L. et Thissen, D. (dir.) (2000), *Computer adaptive testing : A primer, 2e edition*. Mahwah NJ, Lawrence Erlbaum associates.

Weiss, D. (dir) (1983), *New Horizons in Testing. Latent trait test theory and computerized adaptive testing*, New York, Academic Press.

Williams, D.D., Howell, S.L. et Hricko, M. (dir) (2006), *Online assessment measurement and evaluation : emerging practices*, Londres, Information Science Publishing.

Page laissée blanche intentionnellement

SECTION I

Des systèmes à développer et à tester

Page laissée blanche intentionnellement

CHAPITRE 2

Vers une architecture de testing assisté par ordinateur pour l'évaluation des acquis scolaires dans les systèmes éducatifs orientés sur les résultats

Romain Martin, Gilbert Busana
Université du Luxembourg, Groupe de recherche EMACS

Thibaud Latour
Centre de Recherche Public Henri Tudor, Luxembourg

INTRODUCTION

Le Testing Assisté par Ordinateur (TAO) est une pratique qui connaît un intérêt croissant, particulièrement depuis l'avènement d'Internet comme infrastructure technologique privilégiée. Cette croissance est évidemment portée par la diffusion massive et la disponibilité de plus en plus importantes d'un matériel performant pour les TIC[1]. Cependant, force est de constater que cet intérêt n'est pas toujours suivi d'un effet concret. Pourtant, de manière fondamentale, le testing s'inscrit naturellement au cœur des processus d'apprentissage dans lesquels il constitue souvent l'unique moyen de déterminer la détention effective d'une connaissance ou d'une compétence. À l'instar de l'*e*-learning, l'utilisation de systèmes informatiques, leur mise en réseau et les progrès technologiques effectués dans les différentes disciplines associées aux TIC ces dernières années ouvrent un large éventail de nouvelles perspectives au testing, tant en termes de types de compétences testées qu'en termes de modes d'investigation de celles-ci.

1. TIC : Technologies de l'Information et de la Communication

Comme on le verra de manière plus détaillée dans la suite du texte, les contraintes liées au temps et à l'espace que l'on subit lors d'une approche plus traditionnelle papier-crayon peuvent être levées dès que les tests sont délivrés par ordinateur. En effet, dans certains contextes, il devient possible pour le candidat de gérer lui-même le moment où il se soumettra à l'évaluation. De même, la logistique associée à la passation répétée de tests s'en trouve simplifiée. De plus, l'informatisation des données relatives au test engendre un meilleur suivi des résultats et l'optimisation des processus de traitement. Cette traçabilité constitue un avantage à la fois pour les gestionnaires du système, et pour les sujets qui peuvent suivre leur propre progression. Enfin, la passation informatisée des tests permet éventuellement au sujet d'être informé de son résultat immédiatement après la fin du test, le score ayant été calculé en temps réel.

Le domaine du TAO est actuellement en pleine mutation. Alors que nous observons une tendance nette vers le choix de solutions informatisées pour le testing dans un grand nombre de contextes, il se dessine en même temps une relative réticence à investir dans ce type de dispositif. Les raisons en sont certainement multiples, et la façon dont nous envisageons notre relation avec les « machines » n'y est certainement pas étrangère. Cependant, au-delà de cet aspect plus fondamental et pour lequel nous ne chercherons pas de réponse ici, les motifs qui sont le plus souvent invoqués pour justifier ces hésitations sont sans conteste les coûts de développement et de mise en œuvre. Pour autant, si l'on pousse l'analyse un peu plus en profondeur, le problème plus fondamental de l'adéquation des outils existants avec les besoins se fait jour avec force. Notre propos n'est pas ici de montrer les raisons de cette inadéquation (elle résulte en grande partie de la mise en réseau des ordinateurs et de la révolution Internet qui n'a pas été intégrée dans toute sa mesure dans les évolutions successives des applications de *e*-testing), mais de montrer les nouveaux besoins et les réponses adéquates qui permettent de lever les freins à la diffusion de ces techniques. À défaut d'une analyse économique, que nous n'aborderons aucunement, nous proposons au terme de ce chapitre un ensemble de défis fonctionnels et techniques que devront relever les plateformes de la prochaine génération pour répondre aux besoins actuels dans le domaine de l'évaluation des acquis scolaires.

Mais avant d'aborder la problématique centrale de ce chapitre, nous poursuivrons cette introduction en replaçant le testing dans le contexte de l'apprentissage et en donnant un bref aperçu des évolutions du testing assisté par ordinateur au fil du temps.

Relations fortes entre *e*-learning et *e*-testing

Quel que soit leur objectif, que ce soit la validation de l'acquisition d'une compétence ou l'obtention d'une mesure indicative du bon déroulement du processus d'apprentissage, les possibilités offertes par la passation de tests par ordinateur, par rapport aux tests papier-crayon, ouvrent de nouvelles perspectives sur la mise au point d'outils pour récolter des données pertinentes à l'évaluation des apprentissages. Ces perspectives s'appliquent donc aux deux fonctions fondamentales qu'on distingue en général pour le processus d'évaluation, qu'il soit informatisé ou non : la fonction sommative et la fonction formative. Dans le premier cas, l'évaluation s'attachera à mesurer l'état des connaissances ou des compétences acquises par un sujet à un moment donné. Quel que soit le moment dans le processus d'apprentissage, qu'il soit terminal ou intermédiaire, cette évaluation aura un caractère certifiant. Elle s'attachera à déterminer la part du contenu enseigné qui aura été effectivement intégrée au fonctionnement cognitif du sujet. Le résultat de ce type d'évaluation permettra, selon les cas, d'autoriser la poursuite du processus par le passage à une étape ultérieure, de réorienter le processus ou de certifier sa bonne fin. Dans le deuxième cas, l'évaluation, bien que basée également sur le contenu de l'apprentissage, ne vise pas à certifier l'acquisition de ce contenu, mais s'intéresse plutôt au phénomène d'apprentissage en lui-même. Dans ce type d'évaluation, le résultat n'est pas la vérification de l'atteinte d'un niveau donné, mais la détermination d'un positionnement du candidat face au contenu d'apprentissage, à son processus de construction et aux principes pédagogiques appliqués pour l'enseigner ou pour l'initier. L'objectif de ce type d'évaluation est essentiellement dynamique et prospectif. Sa finalité est d'optimiser ou d'améliorer la suite du processus en permettant la détection de lacunes ou au contraire de points d'appui sur la base desquels le processus, la méthode pédagogique ou les objectifs à atteindre seront revus. Ce type d'évaluation permet donc une remédiation. Il a ainsi été démontré que, dans certains cas, le gain en efficacité de l'apprentissage est dépendant du nombre de tests formatifs qui ont été donnés durant le processus (Bangert-Drowns, Kulik et Kulik, 1991).

De manière différente, l'objectif fondamental d'une évaluation sommative est statique et surtout rétrospectif. Dès l'apparition des premiers systèmes d'apprentissage informatisés, cette relation a été maintenue. Ainsi, l'idée d'intégrer des outils de testing assisté par ordinateur dans les méthodes d'apprentissage scolaire n'est pas neuve et a été proposée dès 1989 sous la forme de programmes d'évaluation continue (Bunderson, Inouye et Olsen, 1989). Dans la pratique, ces programmes ont été peu appliqués et ont connu – au-delà de quelques mises à l'essai expérimentales – une retombée pratique assez limitée, faute, à cette époque, d'outils performants et abordables

permettant notamment la mise en réseau des acteurs-clefs (enseignants et apprenants), ainsi que du matériel pédagogique nécessaire (outils d'apprentissage et outils d'évaluation).

Un bref historique du TAO

Les premiers systèmes d'évaluation assistée par ordinateur ont simplement consisté en la transposition de tests papier-crayon existants et ayant déjà fait leur preuve. Il s'est d'ailleurs avéré dans ce contexte que la mise au point de versions parallèles papier-crayon/ordinateur était dans la plupart des cas assez facilement réalisable (Mead et Drasgow, 1993). La mise en ligne de ces tests a naturellement suivi l'évolution des technologies réseau et des architectures de type client-serveur qui ont mené à l'avènement d'Internet sans toutefois se libérer complètement de l'héritage du passé. L'avantage principal qui avait été perçu alors consistait surtout en l'automatisation de l'analyse des résultats issus de ces opérations de collecte de données. En effet, l'informatisation des tests permettait d'une part d'éviter le codage manuel des réponses, et d'autre part d'effectuer rapidement et automatiquement tous les traitements souhaités dans le calcul du score. L'informatisation avait également pour fonction d'assurer la persistance des données et leur gestion dans des bases de données. Pour cette première génération de tests informatisés dont on trouve encore bon nombre de représentants aujourd'hui, l'avantage se situait principalement pour l'administrateur des tests alors que, pour le sujet qui les passe, la différence entre le papier et l'ordinateur était somme toute minime. Cet avantage s'est mieux réparti dès lors que les tests ont été rendus accessibles à distance.

Cependant, durant de nombreuses années, l'évaluation des apprentissages a continué à être majoritairement pratiquée dans sa dimension sommative, et la simple transposition de tests papier-crayon existants restait la règle (Bergstrom et Lunz, 1999). Les fonctions formatives du testing restaient confinées au domaine de la recherche visant à évaluer les méthodes et dispositifs d'apprentissage. Mais dès que l'efficacité d'un dispositif d'apprentissage particulier était démontrée, l'évaluation était ensuite souvent laissée à l'enseignant sans qu'il ait à sa disposition de méthode d'évaluation validée et pouvant être mise en œuvre de manière continue afin d'accompagner et de piloter les processus d'apprentissage. Le manque d'intérêt dans le testing assisté par ordinateur du type formatif est également imputable au fait que la plupart des plateformes d'évaluation de l'époque ont été développées pour des opérations de certification de compétences à grande échelle, souvent avec un enjeu critique pour les personnes passant les tests (*high-stakes testing*). Le but ultime était de garantir qu'un niveau de compétence standardisé ait été acquis par l'apprenant, sans que soit examinée la manière avec

laquelle cet apprentissage avait été effectué (Martin, Latour, Burton, Busana et Vandenabeele, 2005). Ce manque d'intégration de l'évaluation dans les processus d'apprentissage a longtemps constitué le plus important désavantage de ces premiers systèmes à grande échelle (Plichart, Jadoul, Vandenabeele et Latour, 2004).

Dans le contexte des évaluations formatives, il est assez rapidement apparu que la seule prise en compte d'un nombre de réponses correctes ne permet pas d'atteindre les objectifs de guidage pédagogique prévus. La nécessité de mesurer, en complément à ce score, un ensemble de paramètres comportementaux liés aux questions posées et relatifs à la manière dont les réponses sont données, s'est très vite fait sentir. C'est ainsi que des mesures annexes, principalement chronométriques, ont été introduites dans les tests, même si l'exploitation de ces données chronométriques et leur signification en termes de mesure sont loin d'être univoques et nécessitent encore des efforts de recherche supplémentaires (Martin et Houssemand, 2002).

Avec l'évolution des technologies et surtout de la puissance des ordinateurs, de nouvelles possibilités sont apparues pour la réalisation de nouveaux types d'item d'une nature très différente de celle présentée jusqu'alors sur papier ou sur écran. L'inclusion de notions d'interactivité, de comportements dynamiques et de non-déterminisme dans la définition des items a ainsi ouvert la porte à un nombre quasi illimité de nouveaux champs de connaissances qu'il est désormais possible de traiter. Parmi les types d'items avancés, on peut citer ceux requérant la collaboration de plusieurs sujets (dans le but de tester les facultés de coordination des activités par exemple), ceux simulant un phénomène ou le fonctionnement d'un artefact quelconque, ceux qui font appel à d'autres modalités que la lecture d'un texte et la saisie d'une réponse en impliquant d'autres sens (par exemple à travers l'introduction du son ou de la praxie), ou encore les items qui incluent la réalisation d'une procédure, sans que cette liste de nouveaux formats d'items puisse être considérée comme exhaustive. La mise au point de systèmes d'évaluation formative de haut niveau passe par cette évolution. Ces nouveaux champs ont également fortement bénéficié de la recherche en sciences cognitives et en psychologie expérimentale (Martin, 2003).

L'avènement de nouvelles théories de mesure liées au testing, et dont l'application est facilitée dès lors que les tests sont informatisés, a également offert de nouvelles potentialités, principalement en termes de suivi de l'évolution des compétences d'un candidat et de l'optimisation des tests. Dans les tests papier-crayon à visée appliquée, la théorie classique du « score vrai » est souvent considérée comme le modèle de mesure canonique pour assurer l'objectivité et la fidélité des tests. Cette théorie implique la présentation d'un ensemble comparable de questions (souvent identiques) à un ensemble

d'individus. Le taux de réussite aux différents items va alors constituer un élément important dans le processus de mesure qui est avant tout basé sur une comparaison des individus avec leur groupe de référence. Cette vision est cependant prise en défaut sur plusieurs points. Dans ce système, l'équivalence de deux items ou la détermination de leur écart en termes de difficulté est dépendante du groupe de référence qui est à la base de l'étalonnage, et la mesure du niveau de compétence va directement dépendre du jeu d'items qui constituent le test. La nécessité d'une méthode et d'une métrique plus rigoureuses devient cependant cruciale dès lors que l'on désire comparer les résultats de tests construits sur des jeux d'items différents. Dans ce cas, la théorie classique du « score vrai » ne permet pas de proposer une solution méthodologiquement satisfaisante pour arriver à une mesure univoque. Un autre problème lié à la mesure classique basée sur les scores bruts et le taux de réussite se présente lorsqu'il s'agit de prendre en compte différents niveaux de compétence dans l'évaluation et d'adapter les tests en soi à ces niveaux, ce qui est par exemple également requis pour suivre l'évolution de niveaux de compétence successifs d'un individu au cours du temps. Ce problème est en général contourné par la mise au point de séries de tests différenciés par niveaux de compétence. Mais ici encore, le classement de ces tests par difficultés relatives reste l'affaire d'un étalonnage sur des groupes de référence distincts qui n'arrive pas à représenter la continuité de la dimension de compétence d'une manière méthodologiquement satisfaisante. De plus, cette méthode multiplie le coût de production des tests.

Le recours à la « Théorie de la réponse à l'item » (TRI) permet de concevoir des tests auxquels sont associées des échelles de compétences continues sur toute l'étendue des niveaux requis (voir Bertrand et Blais, 2004, pour une présentation plus détaillée). Elle permet en outre de concevoir des tests adaptatifs. La théorie de réponse à l'item est basée, entre autres éléments, sur la possibilité de juxtaposer l'échelle mesurant le niveau de compétence d'un individu et celle déterminant le niveau de compétence validé par un item. Elle repose également sur le postulat que, dans une population donnée, la probabilité de répondre correctement à un item dépend de la position de l'individu sur le trait latent de compétence qui est mesuré par cet item. Si on affecte à chaque item constituant le test une courbe de réponse (une fonction reliant la probabilité de réussir l'item au niveau de compétence sur le trait latent mesuré), les problèmes rencontrés dans l'évaluation classique et mentionnés ci-dessus peuvent être résolus. En effet, la capacité de mesure sur une seule échelle apportée par la TRI élimine la nécessité de concevoir plusieurs tests associés aux différents niveaux que l'on désire mesurer pour une même compétence. L'utilisation de tests adaptatifs permet en outre au sujet de passer plusieurs fois le même test à des moments différents sans que

le résultat soit significativement biaisé par le phénomène de mémorisation, puisqu'on peut se baser sur des jeux d'items différents tout en continuant à mesurer le même trait latent (voir le chapitre 3). De plus, tous ces tests restent comparables et permettent le suivi de l'évolution des compétences en référence à une échelle continue unique. Ce faisant, il est aisé de constater que les tests adaptatifs constituent à l'usage une solution économiquement intéressante puisqu'ils permettent de réduire le nombre de tests nécessaires, ainsi que la longueur de chaque test (Martin, 2003).

L'ÉVALUATION COMME OUTIL DE GESTION DES SYSTÈMES ÉDUCATIFS ET DES COMPÉTENCES

Une approche intégrée pour l'évaluation des apprentissages

Dans le but d'augmenter l'efficacité de leurs systèmes scolaires respectifs, de plus en plus de pays développés adoptent ou sont en voie d'adopter une approche d'évaluation centrée sur les résultats, et non sur les moyens mis en œuvre comme ce fut longtemps le cas (Klieme, 2003). Dans cette approche, un programme de base est souvent défini, et l'opérationnalisation pédagogique est laissée sous la responsabilité des écoles qui sont cependant tenues d'en atteindre les objectifs principaux. Ces derniers sont dans la plupart des cas définis selon un ensemble de niveaux de compétence qui doivent être atteints dans les domaines couverts par le programme de base. Au-delà de ces objectifs reposant sur ce programme de base, les écoles bénéficient souvent d'une certaine autonomie pour la définition d'autres programmes plus spécifiques. Cette autonomie couvre aussi le choix des méthodes pédagogiques qui seront mises en œuvre, et donne donc lieu à une grande diversité de projets pédagogiques prenant en compte les spécificités régionales ou locales. La diversification de l'offre scolaire qui en découle assure une meilleure adaptabilité par rapport à une population scolaire de plus en plus hétérogène dans beaucoup de pays.

L'ajustement et l'efficacité des systèmes scolaires orientés sur les résultats et bénéficiant d'une autonomie accrue sont largement dépendants de l'existence de programmes nationaux de monitorage scolaire qui soient adéquats. Ces programmes nationaux d'évaluation ont pour but de mesurer l'atteinte effective des objectifs de qualification sur la base des niveaux de compétences définis pour chaque domaine du curriculum de base. Ce monitorage doit, pour être pertinent à des fins de pilotage des politiques éducationnelles au niveau national ou régional, être effectué pour le système scolaire entier. Cependant, il reste nécessaire d'effectuer un retour sur les résultats à l'échelle des établissements eux-mêmes afin d'identifier, d'une part, les écoles qui peinent à atteindre leurs objectifs et, d'autre part, celles qui dépassent leurs

objectifs. Cette détermination d'écoles plus ou moins performantes doit tenir compte de plusieurs variables différentes, dont l'arrière-fond socio-culturel et linguistique des populations d'élèves des établissements. Dans les deux cas, il devrait être possible d'analyser les raisons du dépassement ou de la non-atteinte des objectifs. Cette analyse a en général pour but de proposer un ensemble de mesures correctrices aux écoles en difficulté. Ces méthodes de soutien peuvent s'inspirer des caractéristiques structurales des écoles performantes, pour autant qu'on ait pu identifier les facteurs déterminants de leur succès.

L'autonomie dans un tel système scolaire orienté sur les résultats signifie que chacune des écoles est individuellement responsable de l'atteinte des objectifs nationaux fixés par référence au curriculum de base, ainsi que de l'atteinte de ses propres objectifs, fixés par elle-même en référence à son projet pédagogique. Dans le but de suivre l'atteinte de ces objectifs et d'ajuster continuellement les caractéristiques locales des écoles vers un fonctionnement plus efficace, celles-ci devraient grandement bénéficier de programmes d'évaluation permettant de suivre l'évolution des niveaux de compétence au sein de leur population scolaire. Un tel programme local d'évaluation pourrait – dans une optique d'évaluation formative – ensuite être étendu au niveau individuel. Ce recentrage sur l'individu donnerait alors à l'enseignant un d'outil d'évaluation qui lui permettrait de prendre en considération les différences de niveaux de compétence existant au sein de son groupe d'élèves pour planifier ses interventions pédagogiques, qu'il pourrait ainsi ajuster sur des objectifs réalistes.

À l'échelle de l'école, les outils d'évaluation utilisés pour le suivi des résultats de l'apprentissage doivent non seulement être utiles pour les enseignants et utilisables pour eux (ce qui implique qu'ils soient capables d'utiliser l'outil et d'interpréter les résultats sans avoir recours à des modèles statistiques et psychométriques complexes), mais aussi être cohérents par rapport au système mis en place aux niveaux régional et national pour l'évaluation des objectifs fixés dans le programme de base.

Le système d'évaluation qui vient d'être décrit montre que le monitorage d'un système scolaire au moyen d'instruments d'évaluation implique la prise en compte de différents niveaux d'évaluation. De plus, ce système impose qu'une cohérence forte soit maintenue entre tous ces niveaux. Il est par ailleurs notoire que plusieurs systèmes de monitorage nationaux, ainsi que certains systèmes plus locaux, sont eux-mêmes intégrés dans des systèmes d'études comparatives internationales, telles que le PISA[2] (OCDE, 2004). Ces programmes internationaux d'évaluation proposent un ensemble de

2. *Program for International Student Assessment.*

méthodes et de modèles de compétence pour les dimensions centrales de l'enseignement scolaire. Ces cadres de référence doivent être non seulement considérés, mais également intégrés dans la définition des programmes nationaux.

La validation de l'expérience professionnelle est un autre aspect très important qui doit être considéré lors de la mise au point d'un programme national d'évaluation. Cette prise en compte est de nature à encourager les personnes à réintégrer les programmes qualifiants dans une perspective d'apprentissage tout au long de la vie. Pour ces dispositifs de validation, il est aussi nécessaire de disposer d'outils d'évaluation qui permettent une mise en correspondance des résultats qu'ils produisent avec ceux obtenus à d'autres niveaux (les différents niveaux d'enseignement, par exemple).

En complément de ces besoins de compatibilité entre niveaux, il est nécessaire, au vu des mutations démographiques profondes que vivent nos sociétés, de tenir compte des différences socioculturelles existant entre les individus évalués. Ce besoin est d'autant plus criant lorsque différentes langues cohabitent. En raison de cette problématique, les outils d'évaluation doivent autoriser une grande flexibilité linguistique et permettre de réduire au maximum les biais induits par les divers niveaux de maîtrise linguistique dans la mesure d'autres types de compétence. Ce critère de neutralité linguistique, sociale et culturelle des outils d'évaluation est d'ailleurs inscrit en bonne place dans le cahier des charges de l'OCDE pour les enquêtes internationales, telles que le PISA par exemple. L'enjeu essentiel dans les années à venir concernera le développement d'outils d'évaluation qui couvrent de manière cohérente les différentes dimensions décrites ci-après et illustrées dans la Figure 1.

En ce qui concerne la dimension chronologique, les instruments d'évaluation doivent être basés sur des dispositifs permettant d'assurer la continuité entre les différentes phases de l'apprentissage. Cette continuité est essentielle pour un suivi optimal du parcours d'apprentissage et de la progression des individus et des groupes d'individus depuis l'acquisition des compétences les plus élémentaires jusqu'à celle des compétences les plus complexes dans un domaine donné.

Le pilotage de l'apprentissage dans le contexte éducationnel, lui, s'inscrit dans une structure hiérarchique allant du niveau individuel jusqu'au niveau international. Les dispositifs d'évaluation des connaissances et des compétences appliquées à ces différentes strates doivent, dans la mesure du possible, assurer une certaine cohérence entre tous les niveaux.

Dans le contexte d'évolution rapide et d'introduction à grand rythme de nouvelles technologies, la gamme de compétences et de contenus susceptibles d'être évalués dans le contexte éducationnel est également en croissance

constante. Si, actuellement, seul un petit nombre de compétences de base (aptitude de lecture et connaissance des mathématiques, des sciences et des technologies de l'information, et de la communication) sont considérées dans le débat sur l'évaluation, un certain nombre d'autres domaines de compétences non cognitives, telles les habiletés sociales, et d'aptitudes professionnelles sont maintenant en passe d'être introduites dans les programmes d'évaluation internationale. L'évaluation de ces derniers types de compétence a, dans de nombreux cas, bénéficié du support d'outils de TAO adaptés (Bastians et Runde, 2002). La diffusion de plus en plus large de l'ordinateur tant dans les environnements d'apprentissage que dans les activités professionnelles, génère d'ailleurs le besoin de mesurer de nouveaux types de compétence qui sont directement liés à la maîtrise de l'outil informatique, tels que la littératie TIC, c'est-à-dire l'utilisation compétente et efficace des outils TIC (Lennon, Kirsch, Von Davier, Wagner et Yamamoto, 2003) ou la compétence en lecture de textes électroniques (Leu, 2007). Ces nouvelles compétences liées aux TIC sont actuellement en train d'être introduites dans les enquêtes internationales, et on peut supposer qu'elles en constitueront une partie intégrante bien établie dans un certain nombre d'années.

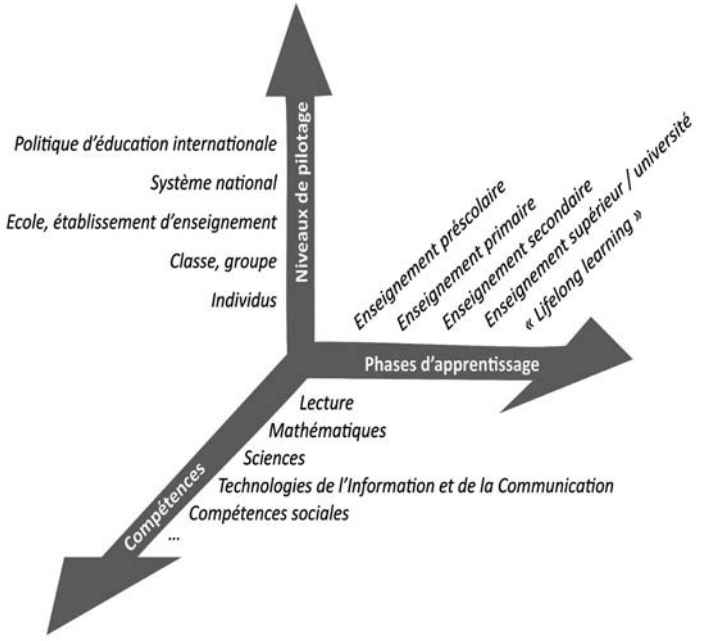


Figure 1 – L'espace des besoins en évaluation structuré selon trois axes majeurs : (1) la dimension chronologique, (2) la dimension du pilotage, et (3) la dimension des compétences

NOUVEAUX BESOINS POUR DES OUTILS DE TAO AVANCÉS

Le développement d'un dispositif d'évaluation intégré qui pourrait couvrir tout l'espace des besoins en évaluation tout en restant efficient et en évitant de faire exploser les coûts générés, ce développement est-il encore imaginable sans recours à l'outil informatique ? La nécessité de passer du format papier-crayon au format informatisé ne se justifiera donc à l'avenir plus en premier lieu par les bénéfices directs liés à la passation informatique – tels que la correction automatique –, mais bien plus par le besoin de maîtriser des processus-métiers qui font partie intégrante des activités de formation et qui sont en train de se complexifier. Des dispositifs d'évaluation, tels qu'ils sont décrits ci-dessus, nécessitent la mise en réseau d'un grand nombre d'acteurs ; ils impliquent la gestion, le croisement et le traitement d'un grand nombre de données, tout en générant le besoin – induit par le type de compétence évaluée – d'inclure des modalités d'interaction multimédia incompatibles avec le support papier. Ces dispositifs intégrés devront bien entendu permettre la gestion d'un grand nombre de formats d'item différents, présentant des contenus très divers. Il est évident que, dans de tels dispositifs, la composition des tests devra pouvoir se faire par l'agrégation d'items sélectionnés en fonction d'une mise en correspondance des objectifs d'évaluation et des caractéristiques des items. Cette tâche seule rend nécessaires la création et la maintenance d'une base d'items spécifiquement conçue à cette fin. L'alimentation de ces bases d'items devra mobiliser la collaboration de nombreux développeurs (Way, Steffen et Anderson, 2002). Cette équipe de développeurs devra être multidisciplinaire. En effet, la création d'un item requiert l'intervention conjointe de pédagogues et de didacticiens pour la définition du contenu, de psychologues (notamment des experts en psychologie cognitive) pour les aspects liés aux modalités d'interaction entre le dispositif de test (et donc l'item) et le sujet évalué, de psychométriciens pour la détermination des caractéristiques métriques de l'item, probablement de spécialistes en technologies multimédia ainsi que des graphistes dans le cas où l'item fait intervenir du contenu plus riche que le texte seul, et finalement de spécialistes des TIC pour l'implémentation de ces items. Ces intervenants dans la chaîne de production des items se trouvent souvent dans différentes institutions et organisations, elles-mêmes appartenant à des niveaux différents sur l'axe des niveaux de pilotage. Dans le cas d'enquêtes à grande échelle, la dispersion géographique vient s'ajouter à la diversité des métiers. Dès lors, la création collaborative et distribuée de contenus, dans une perspective de partage des ressources, et plus particulièrement des items, devient un élément crucial dans ce type de programme d'évaluation. En

effet, le nombre et la nature des candidats impliqués dans l'opération nécessitent la production et le renouvellement d'un grand nombre d'items.

Le cloisonnement des ressources peut également avoir pour origine des considérations de sécurité et de protection de la vie privée. En fonction du niveau de pilotage et de la phase d'apprentissage considérée, ainsi que des objectifs de l'évaluation en termes psychométriques, un certain nombre d'informations concernant les participants à la campagne de test doivent être collectées, gérées et stockées. Il est bien évident que ces informations doivent faire l'objet du maximum de garantie en ce qui concerne leur sécurisation et le respect des dispositions légales en matière de protection de la vie privée. Ces renseignements se trouvent souvent dans des bases de données distribuées, et le fait de les maintenir au plus près de leur source facilite grandement cette gestion. En cette matière également, les législations différentes entre les pays impliqués et les points de vue différents sur le type d'information concernant les candidats, qui peuvent apparaître entre les divers intervenants (administration, TIC, pédagogues, chercheurs) rendent la gestion plus difficile. La prise en compte de cette diversité est aussi facilitée si on adopte une gestion distribuée de ces informations.

Si la création et la gestion des tests et des items, ainsi que la gestion des informations relatives aux sujets et aux groupes cibles de l'évaluation, sont grandement facilitées par l'utilisation d'une plateforme informatique ayant les caractéristiques présentées, il en va de même avec la délivrance des tests et la gestion de leurs résultats. En effet, si l'on prend la peine de faire une analyse plus approfondie des tests, on s'aperçoit qu'ils sont, en tant que tels, constitués d'éléments (les items) les plus souvent indépendants entre eux et même indépendants de certaines caractéristiques du test (par exemple le mode de gestion de la séquence des items adopté pour un test donné). On s'aperçoit également que la passation d'un test nécessite la conjonction d'informations relatives aux sujets et aux groupes de sujets, et que ces informations sont aussi indépendantes les unes des autres. Il apparaît en conséquence qu'un test n'est pas une structure statique formant un ensemble fixe et fermé (vision la plus répandue jusqu'ici), mais au contraire un objet dynamique qui peut être assemblé à la demande et en fonction des besoins. La constitution de tests à la demande implique la collecte de tous ses éléments constitutifs, là où ils se trouvent, et leur assemblage en un tout cohérent qui peut être présenté au candidat. Ce principe a de multiples avantages et corollaires. Premièrement, les possibilités de réutilisation sont maximales, par le partage contrôlé des éléments constitutifs des tests au sein d'un réseau d'acteurs différenciés gérant des ressources de natures très diverses. Deuxièmement, la construction d'un test par un des acteurs du réseau ne nécessite plus la mise en œuvre et la maîtrise de l'ensemble des facettes du testing et la mobilisation de tous les types d'acteurs dans un processus

synchrone. Cette désynchronisation de la production d'un test est rendue possible par le fait que chaque élément est produit indépendamment des autres de manière désynchronisée, et que la création d'un test se résume finalement à l'assemblage des éléments produits par d'autres, à d'autres moments.

Pour preuve de la pertinence de cette analyse, sans qu'il y ait eu de remise en question de l'intérêt majeur porté au TAO, l'approche du test isolé sous forme d'une application indépendante, qu'elle soit délivrée en ligne ou non, a maintes fois montré ses limites en termes de flexibilité et de coût. Ainsi, dans l'étude préliminaire menée par l'OCDE sur la faisabilité de tels tests, deux des conclusions majeures ont été la lourdeur des développements et la difficulté de capturer les comportements des utilisateurs, malgré qu'un besoin fort ait été formulé à ce sujet (Lennon *et al.*, 2003). Ces tests fermés ont bien souvent des coûts très élevés, même au regard des budgets alloués à ces grandes campagnes de test. Ce constat avait déjà été effectué dans le cadre de l'Enquête Internationale sur l'Alphabétisation et les Compétences des Adultes (EIACA, voir Murray, Clermont et Binkley, 2005) et il est notoire que la campagne PISA 2006 n'incluait pas de test assisté par ordinateur dans sa partie commune (la campagne comportait toutefois un test optionnel sur la culture scientifique qui devait servir à évaluer la faisabilité technique et psychométrique de ce genre d'outil). Le coût supérieur d'un test assisté par ordinateur par rapport au test papier-crayon a été également ressenti comme un frein à l'usage du TAO dans les écoles aux États Unis (Bennett, 2001). En effet, à la différence de tests d'évaluation à haute valeur commerciale ou à enjeux critiques (telle l'obtention d'un diplôme ou d'un certificat), dans les enquêtes à grande échelle, les tests produits ne sont en général utilisés qu'une seule fois (au contraire des items). Et de ce fait, le manque de réutilisabilité et de désynchronisation de la production constitue un handicap majeur.

La mise en réseau des acteurs intervenant dans la création et la délivrance des tests, ainsi que la possibilité d'effectuer un partage contrôlé d'une partie des ressources générées par ces acteurs, couplée à la composition dynamique des tests finalement soumis aux sujets laisse présager d'une substantielle réduction des coûts de production et d'exploitation.

Eu égard à la très forte pénétration d'Internet dans les foyers, dans les entreprises et dans les écoles des pays industrialisés, la création, la gestion et la délivrance des tests par ce médium permettent une accessibilité maximale à ces outils d'évaluation. Ce faisant, tout en réduisant les coûts et en augmentant l'accessibilité, les bénéfices apportés par l'usage d'une telle plateforme subsistent : calcul automatique des résultats, environnement de test standardisé, possibilité de collecter des informations temporelles et comportementales

sur le sujet durant le test, et nouvelles modalités d'interaction ouvrant la voie à l'évaluation d'une très large gamme de compétences. Rappelons également que l'utilisation de modèles avancés (tels ceux de la TRI) et d'algorithmes avancés de calcul des résultats, et la sélection des items (tels les tests adaptatifs), dépendent largement de la mise à disposition d'une plateforme informatisée.

La mise en œuvre des grandes campagnes d'évaluation internationale a également mis en lumière la forte composante logistique qui y est rattachée. Si, sur ce point, l'intérêt de disposer de tests informatisés et accessibles par le réseau paraît évident en ce sens que la dématérialisation des tests élimine de facto un grand nombre de problèmes logistiques (tels que le transport, la sécurisation, la distribution, la récupération, la centralisation et le traitement des tests papier), il subsiste néanmoins un problème essentiel lié à l'analyse des données. L'analyse fine des résultats, dans une perspective d'évaluation formative, requiert plus que la manipulation des scores individuels et collectifs obtenus par les sujets dans différents tests (voir, par exemple, les chapitres 10 et 11). Ces analyses exigent que des mises en relation puissent être faites entre les caractéristiques des connaissances et compétences évaluées au moyen des tests, et les caractéristiques des groupes et des individus qui sont évalués. De plus, ces analyses doivent être affinées par la prise en compte des comportements du sujet durant la passation des tests. Pour que ces mises en relation puissent être réalisées, et surtout pour que l'outil puisse réellement et efficacement soutenir la recherche en sciences cognitives et en sciences de l'éducation, et ainsi devenir une réelle aide aux pédagogues, il est essentiel que des modèles englobant ces différentes sources de données puissent être construits. La possibilité de modéliser les ressources du testing au moyen d'ontologies (la couche de métadonnées) répond à ce besoin.

Même s'il existe un grand nombre d'outils d'évaluation assistée par ordinateur, tels *ItemBanker*, *TestWare-AnyWere*, *CQuest*, *FastTestPro*, *InQsit*, *CASTLE-Computer ASsisted Teaching and LEarning*, *Interactive Question Server*, *WebQuiz*, *Moodle*, *OpenSurveyPilot*, *Questiontools*, *Hot Potatoes*, *Adaptex Assessment System*, aucun d'entre eux ne semble actuellement pouvoir être utilisé comme outil générique au sens où nous l'avons défini plus haut. Effectivement, aucun de ces outils n'autorise une grande flexibilité en termes de gestion des métadonnées qui caractérisent les différentes ressources du test. La distributivité n'est pas assurée dans la majeure partie des cas, et ces outils n'offrent pas la possibilité d'étendre les algorithmes de testing et les modèles d'items. De plus, aucun d'eux ne semble capable de supporter la modélisation des domaines de compétence qui peuvent être, le cas échéant, corrélés avec les comportements de réponse (Plichart *et al.*, 2004).

DÉFIS FONCTIONNELS POUR UNE PLATEFORME D'ÉVALUATION INTÉGRÉE

Après avoir décrit les contextes d'application de l'évaluation des compétences en éducation, ainsi que les nouveaux besoins générés par une vision intégrée du système d'évaluation tout au long des différentes phases d'apprentissage, à tous les niveaux de pilotage et pour un large ensemble de compétences, nous allons maintenant parcourir brièvement les grandes orientations fonctionnelles que devraient adopter selon nous les nouveaux outils d'évaluation assistée par ordinateur.

Architecture logicielle

Une des caractéristiques majeures du processus global d'évaluation est, d'une part, sa composition faite d'un ensemble de sous-processus (création de contenu, gestion de ressources informationnelles, génération du test par assemblage de contenus, délivrance et exécution du test, analyse et exploitation des résultats) et, d'autre part, l'intervention d'un nombre d'acteurs divers pouvant être localisés à des endroits différents. Une analyse plus poussée montre qu'il existe cinq, et potentiellement six, grands domaines dans le TAO. À chacun de ces domaines correspondent un type d'acteur et une catégorie de ressource. Ainsi le système complet comprend un espace où sont gérés les sujets devant être évalués, le plus souvent par des acteurs administratifs ; un espace où sont gérés les groupes de sujets soumis à l'évaluation, ces groupes étant souvent formés par les enseignants ; un espace où sont gérés et construits les items, ce travail étant effectué conjointement par des pédagogues, des spécialistes de la compétence évaluée, assistés éventuellement par des spécialistes du multimédia ; un espace où sont gérés et définis les tests que les groupes peuvent passer, ces tests étant composés d'items orchestrés au moyen d'un ensemble d'algorithmes ; et finalement un espace où les résultats sont collectés, gérés et analysés. Avec le développement de nouveaux modèles de mesure, la gamme de modèles et d'algorithmes envisageables pour la mise en œuvre des items s'étend régulièrement. S'ajoute à cela une nouvelle génération d'algorithmes pour l'évaluation automatique de tâches complexes. L'utilisation de ces tâches complexes exploitant les possibilités d'affichage et d'interaction offertes par l'ordinateur comme outils d'évaluation devient justement possible grâce à l'existence d'algorithmes performants qui permettent de transformer les données comportementales enregistrées en indicateurs de performance (Williamson, Mislevy et Bejar, 2006). Dès lors, il peut devenir intéressant de prévoir un espace spécifique où ces modèles et algorithmes seront gérés.

Comme on a pu le voir précédemment, tous ces espaces sont indépendants, et certains sont reliés par des relations de composition. Il semble donc logique de proposer une architecture modulaire dans laquelle chacun de ces espaces de ressources est pris en charge par un module spécialisé. Chaque module doit proposer les fonctionnalités de base pour la modélisation du domaine considéré, la création de ressources et les fonctions de recherche. Les relations existant entre les différents espaces de ressources sont réalisées par des fonctions de communication équipant ces modules et permettant d'associer des ressources provenant de deux domaines différents. En plus de ces fonctionnalités de base, chaque module doit proposer un ensemble de fonctionnalités propres à l'univers de ressources qu'il gère et au processus métier qu'il supporte.

Techniquement, une telle architecture distribuée peut être réalisée sous la forme d'un réseau *Peer-to-Peer* (P2P) de modules interconnectés partageant les ressources des différents domaines qui constituent le TAO et le processus d'évaluation global.

Diversité et hétérogénéité

La nature même du TAO et l'architecture idéale des plateformes qui en découle impliquent la collaboration d'acteurs très différents gérant des ressources spécifiques dans des contextes variés. Cela entraîne inévitablement une grande hétérogénéité de modèles. Nous avons vu dans les sections précédentes qu'un système d'évaluation efficace doit intégrer l'ensemble de l'espace des besoins en évaluation (cf. Figure 1). Afin d'assurer la cohérence sur l'ensemble des dimensions, et étant donné la diversité des composantes du domaine, il est crucial que le système puisse prendre en charge l'hétérogénéité des modèles. Pour ce faire, il faut d'une part considérer le système comme un système ouvert, et d'autre part éviter au maximum d'imposer des modèles a priori. Cette exigence va de pair avec la mise à disposition d'outils de modélisation expressifs et simples à utiliser.

La possibilité d'utiliser la même plateforme pour un ensemble varié de types de compétence et un public diversifié lors des différentes phases d'apprentissage impose une grande flexibilité quant à la définition des types d'item. En effet, selon le type de compétence (cognitive, procédurale, psychomotrice ou autre), ce sont la modalité d'interaction du dispositif de test avec le candidat, le type d'activité demandée au sujet, et le type de réponse enregistrée qui seront tous potentiellement différents. À cette hétérogénéité s'ajoute la prise en compte des éventuels biais linguistiques et culturels dans la manière dont l'information nécessaire à l'évaluation est présentée au sujet. Les outils traditionnels ne permettent pas de gérer cette

diversité autrement qu'en redéveloppant complètement une grande partie du système. Par ailleurs, ces outils n'autorisent généralement pas la combinaison de plusieurs types d'items différents dans un seul et même test. En ce qui concerne plus particulièrement la problématique linguistique, un des défis les plus importants vu dans les enquêtes PISA est la gestion de l'aspect multilingue, tant en termes de modélisation de la mesure qu'en termes de logistique. La plateforme intégrée qui permettra de gérer cette problématique devra inclure des fonctions avancées de gestion des ressources multilingues, la possibilité de livrer des tests multilingues, des facilités d'incorporation de nouvelles langues dans les ressources et les tests mêmes, ainsi que la capacité de prendre en compte différents alphabets. Le travail collaboratif nécessaire à la réalisation de traductions de haute qualité (impliquant un processus de contrôle-qualité explicite) pourrait également être organisé et accompagné à travers la même plateforme intégrée.

Le développement de nouvelles théories et de modèles avancés pour les items et les tests a également rendu nécessaire leur gestion. Les outils traditionnels cependant implantent directement ces modèles et laissent peu de place aux évolutions possibles. Si, dans un contexte très opérationnel et spécifique, cette implantation directe ne pose pas de problème, dans un contexte moins ciblé, il est parfois nécessaire de pouvoir investiguer sur plusieurs modèles et algorithmes. Les besoins des chercheurs dans le domaine de l'éducation et de l'évaluation sont certainement les plus pointus en cette matière. En effet, ce sont eux qui mettent au point et qui valident ces modèles. Pour cela, il leur faut une plateforme qui leur permette une très grande flexibilité sur ce thème ; ils doivent donc pouvoir gérer ces algorithmes et ces modèles comme des ressources du test, au même titre que les items. Aucun outil actuel ne propose ce type de fonctionnalité.

Il apparaît donc clairement que la plateforme d'évaluation de nouvelle génération doit être suffisamment ouverte (et donc conçue et développée à plusieurs niveaux d'abstraction) pour permettre les multiples spécialisations et paramétrisations nécessaires à la prise en compte de la diversité et de l'hétérogénéité.

Distribution, partage des ressources et connectivité

En lien avec l'architecture proposée, les relations entre les différents modules pourraient se faire sur base d'abonnements mutuels. Ces abonnements devraient permettre la gestion fine des droits d'accès aux données qui sont gérées dans le module fournisseur. Un des grands avantages d'une telle architecture est qu'elle permet de partager et d'assembler des ressources de même nature. Ainsi, non seulement la plateforme d'évaluation intégrée mais

aussi la mise en commun de ressources de même nature doivent permettre une grande modularité en fonction des domaines de ressources et des acteurs. Ce faisant, il devient possible de générer et de livrer rapidement des tests incluant des items conçus et gérés par différents partenaires qui possèdent chacun leur propre module de gestion d'item.

De manière tout à fait similaire, et cet aspect revêt une importance capitale lorsqu'on désire intégrer l'évaluation sur plusieurs niveaux de pilotage du système, il est également possible de constituer des groupes de sujets dont les membres sont eux-mêmes gérés dans différents modules administrés par des organisations différentes. Ce mode de distribution des ressources complètement décentralisée dans son concept de base, est le garant d'une très grande flexibilité. Ce type de réseau n'est cependant pas incompatible avec une éventuelle volonté de centralisation, puisque chaque nœud peut être considéré comme localement central dans le réseau.

Modularité et flexibilité

Étant donné la grande diversité des acteurs et l'étendue des besoins potentiels que ceux-ci pourraient exprimer, pour répondre valablement à ces besoins et assurer la continuité du monitorage sur l'ensemble des dimensions illustrées à la Figure 1, la plateforme d'évaluation devra inclure de bonnes capacités d'extension et de personnalisation. Typiquement, l'outil devra faciliter la définition de procédures automatisées de traitement des données facilement transférables entre modules, l'accès à distance à un ensemble de services, la possibilité d'adapter les interfaces graphiques, de créer de nouveaux environnements de langue pour les interfaces, et enfin d'autoriser la spécialisation de modules pour la gestion de ressources appartenant à d'autres domaines que ceux actuellement définis.

Sécurité et protection de la vie privée

Dans le contexte de l'évaluation, les aspects de sécurité sont principalement de deux ordres : les aspects de protection de la vie privée, d'une part, et la protection des items, d'autre part. Ces besoins en sécurité se déclinent en termes de droits d'accès et de cryptage des données.

Le corollaire au travail collaboratif est une gestion des droits d'accès efficace. Sur ce point, la plateforme doit permettre la gestion simultanée de plusieurs utilisateurs pour lesquels des privilèges (lecture, écriture, etc.) ont été attribués à chaque ressource. En conjonction avec les autres mesures classiques de sécurisation des systèmes informatiques, ces privilèges d'accès doivent permettre de gérer les aspects liés à la protection de la vie privée dans

le module de sujets. En effet, seul ce module contient des informations à caractère privé, ce qui facilite le contrôle sur ces données. L'utilisation de ces données est faite en fonction de droits d'accès attribués aux utilisateurs du module lui-même, mais aussi aux autres modules qui peuvent, en tant que modules, accéder à ces ressources. Afin d'assurer ce contrôle, il doit être possible de masquer certaines propriétés des ressources (des candidats, dans notre exemple).

Étant donné l'architecture proposée, un grand nombre d'échanges d'information auront lieu entre les modules. Le contenu de ces communications pouvant contenir soit des données à caractère privé, soit des données relatives aux items, il conviendra de les sécuriser en cryptant le contenu des messages. De manière tout à fait équivalente, les items, parce qu'ils contiennent dans leur définition la réponse correcte, et parce qu'ils constituent le composant essentiel des tests et qu'ils représentent donc l'essentiel de sa valeur, doivent être protégés. La protection des items contre une utilisation abusive doit être incluse dans la plateforme.

Capacité d'analyse et retour pédagogique

Dans un contexte d'évaluation formative et dans l'optique de suivi des processus d'apprentissage et de l'efficacité des organisations qui les définissent et les dispensent, il est nécessaire de pouvoir relier le score obtenu lors de la passation d'un test avec un ensemble d'informations de guidance pédagogique. Cette fonctionnalité requiert la prise en compte d'une série de modèles incluant l'apprenant, la compétence enseignée, les méthodes d'enseignement et d'apprentissage, les modalités d'évaluation, ainsi que tout autre élément qui fait sens dans un contexte d'évaluation donné. Deux conditions fonctionnelles sont nécessaires pour réaliser cela : disposer de capacités de formalisation de ces modèles, quels qu'ils soient, et pouvoir les mettre en relation avec les mesures effectuées durant le test (score, mesures chronométriques et mesures comportementales). Dans ce sens, les fonctionnalités d'une telle plateforme rejoindraient donc celles qui sont décrites pour les tuteurs intelligents (Orey et Nelson, 1994).

CONCLUSION

L'accroissement de l'efficacité des systèmes de formation est considéré aujourd'hui comme indissociable de la mise en œuvre d'outils d'évaluation fiables et valides qui permettent de vérifier si des objectifs de formation rendus opérationnels en termes de niveaux de qualification/compétences ont effectivement pu être atteints. À partir de cette orientation sur les résultats

se dégage la nécessité de mettre en place des systèmes d'évaluation intégrés susceptibles de couvrir d'une manière cohérente les besoins en évaluation qui seront ainsi générés à des niveaux institutionnels différents dans les systèmes éducatifs. Les systèmes d'évaluation intégrés qui en découlent seront d'une complexité qui est telle qu'ils ne sauront à terme être gérés que par des systèmes efficaces de testing assisté par ordinateur. Or, l'analyse des nouveaux besoins pour des outils de TAO avancés que nous avons essayé de réaliser dans le cadre du présent chapitre fait clairement ressortir que ce sera seulement par une architecture ouverte et collaborative que l'ensemble des besoins énoncés pourront trouver une réponse satisfaisante. Cette analyse fait également ressortir qu'il est indispensable de réaliser une séparation des différents « métiers » du processus d'évaluation sous forme de modules spécialisés communiquant entre eux. Une telle architecture dépasse donc très largement, aussi bien techniquement que conceptuellement, les fonctionnalités qu'on trouve actuellement dans des outils informatisés destinés à la passation de questionnaires ou de tests, que ce soit en ligne ou localement. Sa réalisation impliquera un travail de développement important, mais ce travail se rentabilisera rapidement si on prend en considération qu'une telle architecture permettra une subdivision efficace du travail de mise au point de nouveaux tests. Une première tentative d'implantation dans une optique de développement *open-source* est actuellement en cours (Martin *et al.*, 2005). Un autre avantage très important dans la rentabilisation de la plateforme sera le fait que, pour le développement d'un outil répondant à un nouveau besoin d'évaluation, il suffira de développer les éléments vraiment nouveaux dans cet outil, car toute l'architecture de base nécessaire à la réalisation d'une évaluation va être fournie par la plateforme générique.

Références

Bangert-Drowns, Robert L., James A. Kulik et Chen-Lin C. Kulik (1991), « Effects of Frequent Classroom Testing », *Journal of Educational Research,* vol. 85, nº 2, p. 89-99.

Bastians, Frauke et Bernd Runde (2002), « Instrumente zur Messung sozialer Kompetenzen », *Zeitschrift für Psychologie,* vol. 210, nº 4, p. 186-196.

Bennett, Randy E. (2001), « How the Internet Will Help Large-Scale Assessment Reinvent Itself », *Education Policy Analysis Archives,* vol. 9, nº 5.

Bergstrom, Betty A. et Mary E. Lunz (1999), « CAT for certification and licensure », dans Fritz Drasgow et Julie B. Olson-Buchanan (dir.), *Innovations in computerized assessment*, Mahwah, NJ, Lawrence Erlbaum Associates.

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

Bunderson, Victor C., Dillon K. Inouye et James B. Olsen (1989), « The four generations of computerized educational measurement », dans Robert L. Linn (Ed.), *Educational measurement : Third edition*, New York, Macmillan.

Klieme, Eckhard. (2003), *Zur Entwicklung nationaler Bildungsstandards : eine Expertise*, Berlin, Bundesministerium für Bildung und Forschung.

Lennon, Marylou, Irwin Kirsch, Matthias Von Davier, Michael Wagner et Kentaro Yamamoto (2003), *Feasibility Study for the PISA ICT Literacy Assessment, Report to Network A*, Paris, OECD.

Leu, D. (2007), *Expanding the Reading Literacy Framework of PISA 2009 to include Online Reading Comprehension*, Manuscrit inédit.

Martin, Romain. (2003), « Le testing adaptatif par ordinateur dans la mesure en éducation : potentialités et limites », *Psychologie et Psychométrie,* vol. 24, n^{os} 2-3, p. 89-116.

Martin, Romain et Claude Houssemand (2002), « Intérêts et limites de la chronométrie mentale dans la mesure psychologique », *Bulletin de Psychologie,* vol. 55, n^{o} 6, p. 605-614.

Martin, Romain, Thibaud Latour, Reginald Burton, Gilbert Busana et Luc Vandenabeele. (2005), « Covering Different Levels of Evaluation Needs by an Internet-Based Computer-Assisted Testing Framework for Collaborative Distributed Test Development and Delivery », dans Piet Kommers et Griff Richards (dir.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2005*, Chesapeake, VA, AACE.

Mead, Alan D. et Fritz Drasgow (1993), « Equivalence of computerized and paper-and-pencil cognitive ability tests : A meta-analysis », *Psychological Bulletin,* vol. 114, n^{o} 3, p. 449-458.

Murray, T. Scott, Yvan Clermont et Marilyn R. Binkley (2005), *Mesurer la littératie et les compétences des adultes : des nouveaux cadres d'évaluation*, Ottawa, Statistique Canada.

OECD. (2004), *Learning for Tomorrow's World : First Results from PISA 2003*, Paris, OECD.

Orey, Michael. A. et Wayne. A. Nelson (1994), « Development principles for intelligent tutoring systems : Integrating cognitive theory into the development of computer-based instruction », *ETR&D – Educational Technology Research & Development,* vol. 41, p. 59-72.

Plichart, Patrick, Raynald Jadoul, Luc Vandenabeele et Thibaud Latour (2004). *TAO, a collaborative distributed computer-based assessment framework built on Semantic Web standards.* Paper presented at the International Conference on Advances in Intelligent Systems – Theory and Applications, Luxembourg.

Way, Walter D., Manfred Steffen et G.S. Anderson (2002), « Developing, Maintaining, and Renewing the Item Inventory to Support CBT », dans Craig N. Mills, Maria T. Potenza, John J. Fremer et William C. Ward (dir.), *Computer-based testing : building the foundation for future assessments*, Mahwah, N.J., L. Erlbaum Associates.

Williamson, David M., Robert J. Mislevy et Isaac I. Bejar (2006), *Automated scoring of complex tasks in computer-based testing*, Mahwah, N.J., Lawrence Erlbaum Associates.

CHAPITRE 3

Le développement d'une application Web pour la passation d'épreuves adaptatives

Nancy Boiteau
Université du Québec à Rimouski

Jean-Guy Blais
Université de Montréal

Martin Paré
meZur.com

INTRODUCTION

En éducation, la plupart des outils contemporains pour l'évaluation des apprentissages ont été élaborés de façon à ce que tous les individus à qui on fait passer une épreuve répondent en même temps au même endroit à un même groupe d'items. Durant le 20e siècle, cette forme d'épreuves (des épreuves papier-crayon à longueur et à contenu fixes) a été la technologie privilégiée pour récolter des données sur les apprentissages réalisés par des élèves. Olsen, Maynes, Slawson et Ho (1989) affirmaient ainsi que la passation d'épreuves papier-crayon était devenue, au fil des ans, la méthode usuelle pour obtenir des données concernant les apprentissages réalisés par des élèves. Bien que cette approche possède certains avantages non négligeables, par exemple celui de standardiser la passation des épreuves (Wainer, 2000), elle présente certains problèmes, dont celui de la validité des résultats et ceux liés à la précision de l'estimation du niveau d'habileté obtenu (Hambleton, Swaminathan et Rogers, 1991 ; Laurier, 1993 ; Wainer, 2000). En effet, lors

d'une épreuve traditionnelle papier-crayon, les individus d'habileté élevée risquent d'être moins motivés pour des items trop faciles, et les individus d'habileté faible risquent également d'être démotivés devant des items trop difficiles (Hambleton et Swaminathan, 1985 ; Laurier, 1993 ; Lord, 1980). Deux attitudes qui peuvent remettre en question la validité et la précision des résultats obtenus.

Dans ses écrits, Lord (1980) mentionne que l'obtention d'une mesure qui reflète avec une grande précision l'habileté d'un individu nécessiterait que la difficulté des items présentés corresponde au niveau d'habileté de l'individu (Lord avait déjà proposé l'expression « testing sur mesure », *tailored testing*, pour ce type d'épreuve). Dans les épreuves traditionnelles papier-crayon, ce but ne peut être atteint pour l'ensemble des individus que s'ils sont relativement homogènes en termes d'habileté. Les recherches de Lord ont ainsi montré que les épreuves traditionnelles papier-crayon se soldent nécessairement en une mesure non optimale pour la majorité des individus.

Il faut remonter au début du siècle précédent pour voir apparaître ce qu'on détermine comme les premières épreuves dites « adaptatives » papier-crayon (des épreuves où les individus se voient présenter des items dont la difficulté correspond à leur habileté). Par exemple, dans sa forme initiale, le test de Binet (Binet et Simon 1911) possédait les caractéristiques d'une épreuve adaptative, car le niveau de difficulté des items destinés à un enfant devait correspondre à son niveau d'âge mental, celui-ci étant déterminé pendant le processus de gestion des items. Ainsi, le test pouvait être d'une longueur différente d'un enfant à un autre et le groupe d'items donnés pouvait également différer, puisque les items étaient sélectionnés sur la base des réponses de l'enfant aux items précédents. Bien que ce type de passation du test de Binet ait été en place pendant plusieurs années, ce n'est pas avant la période des années 1950 à 1970 que des tentatives ont été faites pour adapter les épreuves aux compétences des individus lors d'opérations de testing à plus grande échelle (Reckase, 1988 ; Weiss, 1985). La littérature montre clairement que ce type d'épreuve peut fournir une meilleure précision de l'estimation du niveau d'habileté des individus que ne le font les épreuves traditionnelles, et ce, à tous les niveaux d'habileté. Les épreuves adaptatives peuvent ainsi apporter une solution au problème de précision associé aux épreuves traditionnelles en permettant l'exploitation des réponses antérieures des individus pour sélectionner les items subséquents de sorte que ces derniers soient de difficulté appropriée à l'habileté de chaque individu (Hambleton et Swaminathan, 1985 ; Hulin, Drasgow et Parsons, 1983 ; Kim et Plake, 1993 ; Lord, 1980 ; McBride, 1997 ; Segall et Moreno, 1999 ; Wainer, 2000 ; Weiss, 1985 ; Weiss et Betz, 1973). Un autre des avantages des épreuves adaptatives concerne leur efficacité. En effet, ces épreuves

permettent d'obtenir des estimations concernant la performance ou l'habileté des individus en utilisant moins d'items et en moins de temps que ne le requièrent les épreuves traditionnelles papier-crayon (Weiss, 1982 ; Bock et Zimowski, 1998 ; Sands et Waters, 1997 ; Wainer, 2000). Finalement, le testing adaptatif en temps réel permet, dans certaines situations, de communiquer rapidement un diagnostic et, le cas échéant, de procéder à une intervention de remédiation mieux ciblée.

Au fil des ans, l'évolution de la technologie associée aux propositions de nouvelles théories de la mesure ont permis aux chercheurs de développer des épreuves adaptatives reposant sur un support technologique avancé, c'est-à-dire un support informatique (Hambleton, Jones et Rogers, 1990 ; Kim et Plake, 1993 ; Lord, 1980 ; McBride, 1997 ; Wainer, 2000 ; Weiss, 1985 ; Weiss et Betz, 1973). Reckase (1996) affirme ainsi qu'au cours de la dernière décennie, une des plus importantes innovations dans le processus de développement des épreuves a été l'utilisation de l'ordinateur pour la sélection des items. Embretson (1999) rappelle que plusieurs épreuves largement utilisées s'effectuent maintenant sous une modalité adaptative informatisée : qu'il suffise de mentionner le CAT-ASVAB (Segall et Moreno, 1999), le Graduate Record Examinations (GRE CAT) General Test (Mills, 1999) et la dernière révision du Stanford-Binet (Thorndike, Hagen et Slatter, 1986). Depuis le début des années 1980, les épreuves adaptatives informatisées font l'objet de nombreuses recherches aux États-Unis. En 1984, les écoles publiques de Portland ont implanté un système informatisé de passation d'épreuves adaptatives (CAT system) dans cinq matières différentes. D'autres organismes ont également emboîté le pas : la Psychological Corporation, le College Board, l'American College Testing, la Société américaine des pathologistes et l'American Board of Internal Medicine (Raîche, 2000). Au Québec, Laurier (1993) faisait figure de précurseur en développant le programme CAPT (Computerized adaptive placement test) visant le classement en français langue seconde au niveau postsecondaire. Les entreprises impliquées en *e*-formation montrent aussi leur intérêt pour l'utilisation d'épreuves adaptatives informatisées. C'est le cas notamment de Microsoft qui a implanté des versions adaptatives de ses épreuves de certification. Il est évident que le recours aux épreuves adaptatives informatisées prend de l'ampleur et qu'il pourrait devenir incontournable dans un proche avenir (Embretson, 1999).

Lors du développement d'épreuves adaptatives informatisées, il est nécessaire de disposer d'items dont les caractéristiques métriques sont connues. À cet égard, Hambleton, Jones et Rogers (1990) affirment que les écrits supportent grandement l'utilisation des modèles issus de la théorie de la réponse à l'item (TRI) pour l'estimation des caractéristiques métriques des items. Cette théorie, qui est en fait une proposition de modélisation de

la mesure, a une caractéristique qui lui confère un avantage sur d'autres modèles de mesure. Lorsque les données s'ajustent adéquatement aux modèles, les valeurs estimées des paramètres d'items et d'habileté des individus sont invariantes. Cela signifie, notamment, que les estimations d'habileté émanant de la passation d'épreuves adaptatives informatisées peuvent être comparées sur la même échelle, et ce, même si les individus n'ont pas répondu aux mêmes items.

Actuellement, les logiciels disponibles sur le marché pour des travaux de recherche en lien avec les épreuves adaptatives informatisées sont plutôt limités, et ils sont conçus pour répondre à des besoins spécifiques. Dans le cadre d'une étude destinée à comparer l'efficacité d'épreuves passées sous des modalités adaptatives informatisées différentes, Boiteau (2004) a développé, en collaboration avec une équipe d'experts en informatique, une application Web permettant la création et la passation en ligne d'épreuves adaptatives. Cette étude s'intéressait particulièrement à une forme d'épreuve adaptative connue sous le nom de modalité adaptative à deux niveaux et qui a été utilisée sous la forme papier-crayon dans les enquêtes du Programme d'indicateurs de rendement scolaire (PIRS) au Canada. En combinant les modèles de la TRI à l'usage des TIC, cette étude visait à créer une modalité adaptative « optimale » à deux niveaux qui soit d'efficacité comparable à celle offerte par la modalité adaptative classique. Soulignons que la modalité adaptative à deux niveaux présente des caractéristiques que l'on ne trouve pas habituellement dans la modalité adaptative classique. Par exemple, on peut penser à la possibilité, pour un individu, de réviser ou de modifier ses réponses aux items qui constituent l'épreuve.

Pour répondre à l'objectif de cette étude, une plate-forme de création et de passation d'épreuves adaptatives supportée par l'infrastructure réseau d'Internet a été développée en collaboration avec des spécialistes dans le développement d'applications Web. Le travail des architectes, analystes et programmeurs consistait à modéliser une structure de données souple capable d'accueillir des banques d'items calibrés par l'entremise d'un modèle de la TRI. La logique programmée, quant à elle, consistait à coder une procédure de sélection des items ainsi que l'algorithme permettant d'estimer l'habileté des individus en fonction d'un modèle de réponses aux items de la TRI. Une fois la modélisation de la banque de données et le codage de la logique effectués, les spécialistes en informatique ont développé une interface utilisateur conviviale pour rendre le produit fini facile à utiliser et attrayant pour les individus qui devaient effectuer l'une ou l'autre des épreuves adaptatives. Une période de mise à l'essai a permis de valider la qualité du développement réalisé. Par la suite, l'application a été transférée sur un serveur Web. À partir d'une simple adresse Internet, autant les individus qui devaient effectuer une épreuve adaptative que les administrateurs du système avaient accès de

façon sécuritaire au système via un nom d'utilisateur et/ou un mot de passe. Ainsi, il a été possible d'administrer les épreuves adaptatives à plus de 1000 individus, et ce, en trois semaines.

Les trois prochaines sections de ce chapitre présentent les caractéristiques théoriques, méthodologiques et fonctionnelles de l'application Web développée dans le cadre de l'étude effectuée par Boiteau (2004).

CARACTÉRISTIQUES THÉORIQUES DE L'APPLICATION

Toutes les formes d'épreuves adaptatives fonctionnent à partir d'une banque d'items dont les caractéristiques métriques, par exemple la difficulté, ont été calibrées (Weiss, 1974). Elles diffèrent cependant par leurs règles de départ, de poursuite et de fin de test. Certaines utilisent un nombre fixe d'items, et d'autres un nombre variable. Plusieurs auteurs ont présenté ces formes d'épreuves adaptatives et ont comparé leurs caractéristiques (Dechef et Laveault, 1999 ; Hambleton et Swaminathan, 1985 ; Hulin, Drasgow et Parsons, 1983 ; Lord, 1980 ; Raîche 2000 ; Weiss, 1974). Dans le cadre de l'application Web, nous avons modélisé et programmé les fonctionnalités pour deux formes d'épreuves informatisées : la modalité adaptative à deux niveaux et la modalité adaptative classique.

Modalité adaptative à deux niveaux

La modalité à deux niveaux est la plus simple des modalités adaptatives. Elle consiste à faire passer une première épreuve (un test de classement[1]) commune à tous les individus et, par la suite, à assigner une seconde épreuve (test d'évaluation[2]), dont le niveau de difficulté moyen dépend de l'estimation de l'habileté obtenue à partir de l'évaluation des réponses au test de classement. Le test de classement peut comporter des items couvrant tous les niveaux de difficulté sur l'échelle d'habileté (distribution rectangulaire), c'est-à-dire un test constitué d'items faciles, moyens et difficiles. Il peut aussi comporter des items dont la difficulté correspond au niveau moyen du groupe d'individus à qui on le donne (distribution unimodale). Le test de classement est donc un test court qui contribue à fournir une estimation initiale du niveau d'habileté d'un individu. Le test d'évaluation, lui, comporte des items dont le niveau de difficulté se situe dans un intervalle de valeurs correspondant à différents niveaux d'habileté des candidats. Par exemple, une modalité adaptative à deux niveaux pourrait comporter trois tests d'évalua-

1. Résulte de la traduction du terme « routing test ». On entend aussi l'expression « test directeur ».
2. Résulte de la traduction du terme « evaluation test ».

tion dont le premier est de faible difficulté, le deuxième de difficulté moyenne, et le troisième de difficulté élevée. Si le niveau d'habileté de l'individu, obtenu à partir des réponses aux items du test de classement, se situe dans l'intervalle pour lequel il a été décidé que le niveau d'habileté serait considéré comme étant faible, l'individu en question sera assigné au test d'évaluation composé d'items de faible difficulté. Le même principe s'applique pour un individu dont le niveau d'habileté estimé se situe dans l'intervalle d'habileté moyenne ou pour celui dont le niveau d'habileté estimé est élevé. Bien que cette modalité puisse être vue comme étant formée de deux épreuves traditionnelles, il y a une estimation provisoire de l'habileté de l'individu sur laquelle on se base pour adapter la suite de l'épreuve à son niveau d'habileté : elle est donc bel et bien une modalité adaptative.

Modalité adaptative classique

Cette modalité fonctionne également à partir d'une banque d'items calibrés, c'est-à-dire dont les paramètres d'items ont été estimés à partir d'un modèle de mesure adéquat (dans le cas présent, un modèle de la TRI). Cette banque d'items est structurée en fonction de caractéristiques préalablement établies. Le niveau d'efficacité de la modalité adaptative classique dépend du nombre, de la variété et de la qualité des items dans la banque pour chaque niveau d'habileté. Sous cette modalité, l'habileté de chaque individu est estimée de façon itérative après chaque réponse par l'intermédiaire d'un algorithme mathématique pendant le processus de passation de l'épreuve.

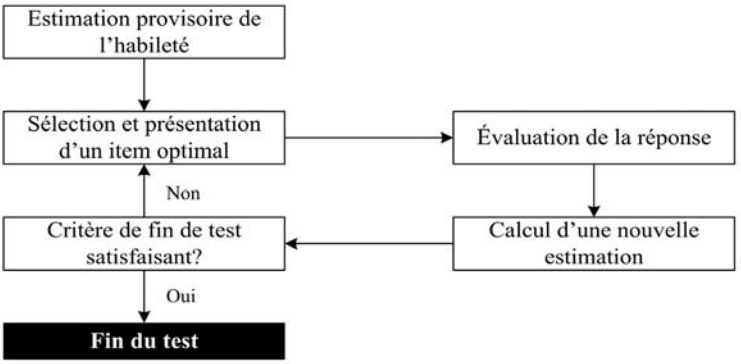


Figure 1 – La logique de la modalité adaptative classique selon Thissen et Mislevy (2000)

La Figure 1 présente la logique de la modalité adaptative classique selon Thissen et Mislevy (2000). Le processus débute avec une estimation d'habileté provisoire de l'individu qui peut être établi sur la base d'une information disponible en ce qui le concerne, sur l'habileté moyenne du groupe, ou encore en supposant que l'individu est d'habileté moyenne. Dans son fonctionnement le plus simple, cette modalité assigne habituellement à l'individu un premier item de difficulté moyenne (Wainer et Kiely, 1987). Si la réponse à cet item est correcte, le prochain item qui lui sera proposé sera plus difficile. De la même façon, si la réponse au premier item est incorrecte, l'item suivant sera plus facile. Basé sur la réponse au premier item, l'algorithme lui assignera l'item de la banque qui sera susceptible de fournir le plus d'information (item optimal) concernant l'habileté de l'individu. Une fois la réponse à ce deuxième item obtenue, une nouvelle estimation de son habileté sera déterminée par la méthode d'estimation utilisée. Cette nouvelle estimation d'habileté contribuera à sélectionner le prochain item de la banque dont le potentiel d'information concernant l'habileté de l'individu est le plus élevé. Et ainsi de suite jusqu'à la fin de l'épreuve. Une erreur de mesure est estimée à chaque niveau du processus de testing. L'épreuve se termine lorsque l'erreur de mesure atteint un niveau de précision prédéterminé ou qu'un certain nombre d'items a été passé.

La modalité adaptative classique peut aussi s'appuyer sur les développements des modèles de mesure et, comme il a déjà été fait mention, ceux de la théorie des réponses aux items ont été privilégié jusqu'à maintenant (McBride, 1997 ; Wainer, 2000 ; Weiss, 1983). Pour optimiser l'efficacité de cette modalité, on doit disposer de modèles de mesure éprouvés pour le calibrage d'une banque de plusieurs items, et c'est ce que proposent les modèles de la TRI

Théorie de la réponse à l'item : un aperçu

Tel que cela a été mentionné précédemment, les épreuves adaptatives nous permettent d'obtenir une estimation d'habileté précise avec un minimum d'items. Pour obtenir une précision élevée, une banque d'items très étoffée est nécessaire. Pour une banque d'items de cette ampleur, apte à fournir des estimations d'habileté précises, l'utilisation d'estimations de paramètres d'items basés sur la TRI s'avère un moyen pour atteindre la précision souhaitée (Bertrand et Blais, 2004 ; Laurier, 1990 ; McBride, 1997 ; Wainer, 2000 ; Weiss, 1983). La TRI offre des modèles probabilistes permettant d'obtenir une estimation de l'habileté des individus à partir de leurs réponses aux différents items. Elle fait partie de la famille de modèles qui pose l'hypothèse de l'existence d'un trait latent pour lequel le résultat à une épreuve n'est qu'une manifestation observable de ce trait (dans le cas qui

nous intéresse, ce trait est ce qu'on appelle « l'habileté »). Les modèles plus connus de l'analyse factorielle font aussi partie de cette catégorie. Plusieurs concepts à la base de la TRI ont permis le développement de grands courants méthodologiques reliés au développement d'épreuves adaptatives (Wainer et Mislevy, 2000). C'est en présentant les caractéristiques méthodologiques de l'application Web (prochaine section) que nous ferons état des concepts de la TRI intégrés dans les trois grandes étapes de développement des deux modalités étudiées : la création de la banque d'items, la méthode d'estimation de l'habileté et le processus de sélection des items.

CARACTÉRISTIQUES MÉTHODOLOGIQUES DE L'APPLICATION

Dans le processus de développement d'une épreuve adaptative, les éléments à considérer touchent principalement la création d'une banque d'items, la méthode d'estimation de l'habileté et la sélection des items pendant le processus d'administration. Que ce soit pour la modalité adaptative à deux niveaux ou la modalité adaptative classique, nous avons attribué les mêmes caractéristiques en ce qui a trait aux trois étapes de développement. Toutefois, la stratégie adaptative utilisée sous chacune des modalités étant différente, nous avons respecté les principes qui régissent leurs fonctionnements respectifs.

Création de la banque d'items

Une banque d'items consiste en un regroupement d'items organisés en fonction de caractéristiques bien définies. Cette structure permet de repérer et de sélectionner rapidement les items qui sont présentés à chaque sujet. La validité et la précision d'une modalité adaptative informatisée basée sur les modèles TRI dépendent de la quantité et de la qualité des items qui composent la banque (Weiss, 1985). Cette banque doit regrouper un certain nombre d'items, préférablement un grand nombre, dont les estimations de paramètres sont connues (van der Linden et Eggen, 1986). C'est ce que signifie l'expression « banque d'items calibrés ». L'utilisation d'une banque d'items calibrés permet d'obtenir, pour tous les individus à qui on fait passer une épreuve adaptative, des estimations d'habileté comparables entre elles, et ce, peu importe les items de la banque qui seront proposés.

Afin de construire une banque d'items calibrés, étape essentielle pour créer les deux modalités étudiées, nous avions à notre disposition des données provenant de plusieurs milliers d'élèves canadiens ayant répondu à des centaines d'items dans le cadre d'une évaluation à grande échelle effectuée

en mathématique en 1997 et en 2001. Il s'agit des données provenant du Programme d'indicateurs de rendement scolaire (PIRS) ; cette banque de données a été mise à notre disposition par le Conseil des ministres de l'Éducation du Canada. Selon Weiss (1982), un minimum d'une centaine d'items est recommandé pour élaborer une banque d'items en vue de faire passer une épreuve sous la modalité adaptative classique. Ce minimum vise à assurer un nombre suffisant d'items à différents points de l'échelle d'habileté. En tout, 116 items ont été retenus en vue de constituer notre banque et ils ont été calibrés par l'entremise du logiciel BILOG-MG (Zimowski, Muraki, Mislevy et Bock, 1996) en utilisant la méthode d'estimation du maximum de vraisemblance marginal (Embretson et Reise, 2000 ; Wainer, 2000) appliqué au modèle logistique à trois paramètres (paramètres de discrimination « a », de difficulté « b » et de pseudochance « c »). La calibration s'est effectuée à partir d'un échantillon de 44 522 sujets.

Méthode d'estimation de l'habileté

Les modèles de la TRI ont été aussi utilisés pour estimer le paramètre d'habileté des individus pendant la passation des épreuves sous chacune des deux modalités. L'estimation du paramètre d'habileté est conditionnelle à la connaissance préalable des valeurs des paramètres d'items. Autant pour la modalité adaptative à deux niveaux que pour la modalité adaptative classique, la méthode d'estimation provisoire du niveau d'habileté est la même que la méthode d'estimation finale. Dans l'application Web, c'est la méthode de l'espérance a posteriori (EAP) qui est utilisée dans l'estimation du niveau d'habileté des individus. Cette méthode utilise la moyenne de la distribution de probabilité a posteriori comme estimateur du niveau d'habileté. En testing adaptatif, elle constitue une des méthodes de choix, étant donné la simplicité de ses calculs et la précision obtenue dans l'estimation d'habileté (Raîche, 2000). Cette méthode semble être celle qui fournit la plus petite part de biais dans l'estimation d'habileté (Embretson et Reise, 2000 ; Thissen et Mislevy, 2000 ; Wang et Vispoel, 1998), et Thissen et Orlando (2000) suggèrent fortement d'utiliser cette méthode dans le cas du modèle à trois paramètres.

La sélection des items

Un algorithme de sélection des items est un ensemble de règles visant à déterminer les items qui seront proposés à un individu pendant le processus de testing. Autant pour la modalité adaptative à deux niveaux que pour la modalité adaptative classique, la sélection des items à inclure repose sur la stratégie du maximum d'information (Chang et Ying, 1996 ; Hau et Chang,

1998 ; Kingsbury et Zara, 1989 ; Thissen et Mislevy, 2000). La stratégie de sélection des items basée sur le maximum d'information consiste à choisir le ou les items pour lesquels l'information, au sens de Fisher (1922), évaluée au niveau d'habileté provisoire de l'individu est maximale (Lord, 1980, p.199). Les paramètres d'items nous renseignent sur la quantité d'information que nous procure un item. Lorsque la difficulté d'un item correspond au niveau d'habileté de l'individu, l'information donnée par cet item est maximale autour de ce niveau d'habileté. De plus, l'information est d'autant plus grande que la discrimination de l'item est élevée : un item peu discriminant nous donne peu d'information. Finalement, plus le risque de réussir un item au hasard est faible, plus l'item est informatif. Les paramètres TRI de notre banque d'items étant connus, un algorithme de calcul des valeurs d'information pour chaque item a été programmé. Ainsi, les valeurs d'information pour chaque item peuvent être calculées à chacun des niveaux d'habileté sur l'échelle. Ces valeurs indiquent également le degré de précision de la mesure à chaque point de l'échelle d'habileté.

Pour la modalité adaptative à deux niveaux, le choix des items à inclure dans le test de classement s'effectue également sur la base de l'information fournie par chacun à des points précis sur l'échelle d'habileté. Par exemple, il est possible de sélectionner un nombre prédéterminé d'items, les plus informatifs, à différentes places sur le continuum de l'échelle d'habileté (niveaux faible, moyen et fort). En ce qui a trait aux tests d'évaluation, ils se composent des items les plus informatifs, à des points précis de l'échelle d'habileté, parmi ceux qui n'ont pas été sélectionnés au test de classement. L'échelle d'habileté étant divisée en intervalles égaux en fonction du nombre de tests d'évaluation souhaité, l'algorithme de sélection des items permet de sélectionner un certain nombre d'items, les plus informatifs, au point milieu de chaque intervalle. Les intervalles servant à constituer les tests d'évaluation servent également à déterminer les points de coupure permettant d'assigner le test d'évaluation dont le niveau de difficulté correspond le mieux à l'estimation d'habileté de l'individu obtenu à partir de ses réponses au test de classement.

Pour la modalité adaptative classique, l'algorithme de sélection des items détermine le premier item à donner, sélectionne les items subséquents et spécifie quand le test est terminé. En ce qui a trait à la règle de départ, la moyenne de la population ciblée par le test fournit une estimation initiale raisonnable du niveau d'habileté lorsqu'il n'y a pas d'autres informations disponibles concernant l'habileté des individus et permet d'établir le niveau de difficulté du premier item présenté aux individus (Thissen et Mislevy 2000). Afin de contrôler la fréquence d'exposition du premier item, l'algorithme permet d'assigner aléatoirement, en guise de premier item, ceux dont le niveau de difficulté se situe autour du niveau d'habileté moyen de la

population ciblée par le test. La sélection des items subséquents repose sur la stratégie du maximum d'information qui consiste, rappelons-le, à choisir l'item pour lequel l'information est maximale à l'estimation du niveau d'habileté de l'individu. La règle d'arrêt se base, entre autres, sur la suggestion de Thissen et Mislevy (2000) qui consiste à fixer un nombre maximal d'items à donner au cas où l'estimation d'habileté serait trop longue à calculer eu égard à la précision à atteindre. De plus, pour assurer une certaine validité de contenu lorsqu'il est possible d'atteindre le niveau de précision souhaité avec un nombre d'items trop petit, l'application permet de préciser un nombre minimal d'items pour tous les individus.

Temps de passation des tests

Un avantage des épreuves adaptatives réside dans l'assouplissement du temps de passation. Chaque individu peut prendre, en principe, le temps nécessaire pour répondre aux items de l'épreuve. Cependant, le cas d'un individu prenant une journée pour répondre à un item, par exemple, est une situation qui, à l'heure actuelle, est plutôt irréaliste. Comme, de toute façon, lorsque la modalité papier-crayon est utilisée. Il y a aussi des contraintes de temps qui peuvent régir la passation d'épreuves adaptatives informatisées. Dans la modalité adaptative à deux niveaux, l'individu connaît la plupart du temps dès le début de l'épreuve le nombre d'items auquel il sera soumis ainsi que le temps qu'on lui alloue pour y répondre. Dans la modalité adaptative classique, un problème se pose : l'individu connaît le temps total qu'il a pour effectuer l'épreuve, mais ne connaît pas le nombre d'items auquel il aura à répondre. Pour administrer les épreuves sous la modalité adaptative classique, l'algorithme de sélection des items doit être en mesure de chronométrer la durée de présentation de chaque item. Ainsi, une programmation intégrée dans l'algorithme de sélection des items a été effectuée pour assurer le contrôle du temps de présentation de chaque item. Un petit sablier clignote au bas de l'écran à partir du moment où il reste à l'individu une minute pour répondre à l'item.

CARACTÉRISTIQUES FONCTIONNELLES DE L'APPLICATION

L'application Web dispose de trois volets fonctionnels : le volet *Administration*, le volet *Gestion* et le volet *Passation de tests*. Ces trois volets sont accessibles à partir de la page d'accueil du site Web. De façon générale, le volet *Administration* permet de définir les utilisateurs qui auront accès au volet *Gestion*. Le troisième volet supporte la passation des épreuves adaptatives. Trois profils d'utilisateurs sont ainsi définis dans le système :

l'administrateur (le chercheur, le responsable, etc.), l'intervenant (l'utilisateur) et le répondant (le candidat, l'élève, etc.).

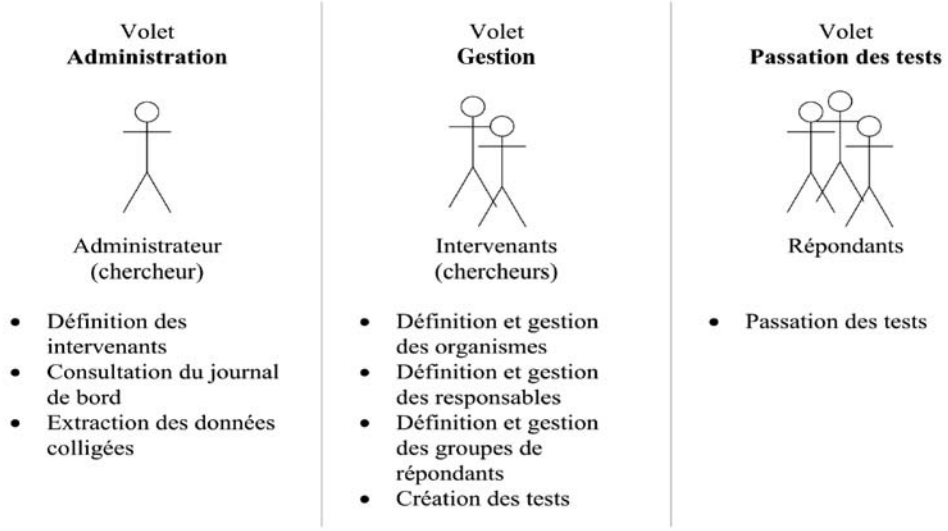


Figure 2 – Volets fonctionnels de l'application Web

La Figure 2 décrit sommairement les profils d'utilisateurs aptes à exploiter le système ainsi que les fonctionnalités propres à chacun des trois volets de l'application Web. En plus de créer la liste des intervenants, le volet *Administration* permet à l'utilisateur de consulter un journal de bord. Ce journal affiche les sessions de travail de chaque utilisateur qui accède à l'application à partir du volet *Gestion* ainsi que toutes les tentatives d'intrusion infructueuses au système. Ce volet offre également la capacité d'extraire l'ensemble des données colligées des répondants qui ont réalisé une épreuve adaptative. De plus, quelques rapports statistiques permettent de constater l'évolution de la passation des épreuves en ce qui a trait, par exemple, à la précision moyenne des épreuves adaptatives de chaque modalité, à la note moyenne et au temps moyen requis par celles-ci.

Le volet *Gestion* permet aux utilisateurs de créer les épreuves adaptatives sous chacune des deux modalités intégrées dans le système. En fait, ce volet permet de définir les critères pour la construction des tests sous chaque modalité. Pour la création des tests sous la modalité adaptative à deux niveaux, il faut, dans un premier temps, définir le nombre de tests d'évaluation ainsi que le nombre d'items requis pour chacun. Par la suite, l'utilisateur construit le contenu du test de classement en choisissant des items, les plus informatifs,

à un ou plusieurs niveaux de l'échelle d'habileté. Ensuite, le système peut générer le contenu des tests d'évaluation en choisissant les items les plus informatifs, non inclus dans le test de classement, au point milieu de chaque intervalle. Rappelons que l'échelle d'habileté est divisée en intervalles égaux en fonction du nombre de tests d'évaluation souhaité. Pour la modalité adaptative classique, il faut définir les conditions d'arrêt de l'épreuve telles que le nombre maximum d'items à proposer au répondant, le niveau de précision à atteindre ainsi que le nombre minimum d'items à donner. Le volet *Gestion* offre aussi aux utilisateurs la possibilité d'intégrer les coordonnées des organismes où seront administrées les épreuves. Une fois l'organisme ajouté au système, l'utilisateur peut définir les responsables (enseignantes ou enseignants) des groupes de répondants qui auront à réaliser une épreuve adaptative. Pour un organisme, un ou plusieurs responsables peuvent être ajoutés au système.

Le volet *Gestion* permet aussi de définir les groupes de répondants (associés à chacun des responsables). L'utilisateur doit préciser certaines informations relatives à chacun des groupes de répondants, comme le nom du groupe, le nombre d'individus, le code d'accès ainsi que la période de validité à l'intérieure de laquelle l'accès au système par le groupe est assuré par le code d'accès qui lui est désigné. Une fonctionnalité de ce volet permet de préciser si les épreuves adaptatives définies dans le système seront aléatoirement attribuées à chacun des répondants ou encore, de préciser si un groupe de répondants aura à effectuer une même épreuve adaptative.

Le troisième volet *Passation de tests* est facilement accessible par tous les répondants via la page d'accueil de l'application. Par l'entremise du code d'accès qui leur est désigné, les répondants peuvent effectuer une épreuve adaptative. Précédés d'une notice explicative sur le fonctionnement de l'épreuve qu'ils auront à passer, les items sont présentés aux répondants en fonction des principes de fonctionnement propres à chacune des deux modalités. À la fin de l'épreuve, le niveau de performance du répondant est affiché sur une page accompagné d'un court texte qui le remercie d'avoir participé à l'étude.

UNE UTILISATION AVEC UNE MODALITÉ TRADITIONNELLE

L'application Web supporte actuellement la modalité adaptative à deux niveaux, la modalité adaptative classique ainsi que la modalité de passation traditionnelle (nombre d'items et contenu fixes). Cette dernière modalité a été utilisée dans le cadre d'un projet d'informatisation du prétest en mathématiques de l'École Polytechnique de Montréal (Blais, Boiteau,

Burney-Vincent, Caron, Desmarais, Gagnon, Grondin et Paré, 2007). Depuis 1989, l'École Polytechnique de Montréal fait passer une épreuve diagnostique (prétest) de mathématiques aux étudiants nouvellement admis dans un de leurs programmes universitaires. Chaque année, l'épreuve diagnostique est donnée en format papier-crayon et comprend 60 items puisés dans une banque composée de 985 items à choix multiples portant sur six domaines des mathématiques : fonctions élémentaires, trigonométrie, géométrie, algèbre, calcul différentiel et calcul intégral. À la suite de la correction de l'épreuve, un profil de performance pour les six domaines évalués est envoyé par la poste à chaque étudiant. En fonction de la performance obtenue, l'étudiant peut se voir proposer un cours de mise à niveau intensif en mathématiques avant d'entreprendre sa session d'études.

Avec comme objectifs d'une part de développer une plateforme générique pour recueillir des données issues de la passation en ligne d'épreuves et, d'autre part, de faciliter la gestion et la passation de l'épreuve de mathématiques, une collaboration entre l'École Polytechnique de Montréal, le Groupe de recherche interuniversitaire pour l'évaluation et la mesure en éducation à l'aide des technologies de l'information et des communications (Griémétic) et l'entreprise Mezur.com a permis l'informatisation du prétest en mathématiques de l'École Polytechnique de Montréal au printemps et à l'été 2007 et 2008. Pour la réalisation de ce projet, les fonctions génériques de gestion des répondants, des items, des tests et des rapports d'une nouvelle version de la plateforme (V2, voir la section suivante) ont été mises à contribution de manière à développer une solution spécifique à l'École Polytechnique de Montréal. Également, l'image de marque du site Internet de l'organisation a été reproduite fidèlement de manière à faciliter la navigation lors de la passation de l'épreuve.

La passation du prétest informatisé s'est déroulée en milieu contrôlé dans les locaux informatisés de l'institution. L'accueil était assuré par la présence de surveillants dans chacun des locaux. L'étudiant se présentant au prétest se voyait octroyer un code d'accès par lequel il pouvait accéder à l'épreuve. Après la saisie du code d'accès, l'étudiant devait s'authentifier et lire des consignes générales sur le déroulement du prétest. L'étudiant disposait alors de deux heures pour répondre aux soixante questions de l'épreuve. À la fin du prétest, l'étudiant obtenait automatiquement son diagnostic à l'écran (profil de performance et suggestions) qui lui était également transmis par courriel. L'utilisation des fonctions génériques de la nouvelle version (V2) de la plateforme a permis à l'École Polytechnique de Montréal de tirer profit des avantages relatifs à un système technologique contemporain, comme ceux de la correction instantanée, de la communication rapide et efficace des résultats, de la gestion simplifiée des aspects administratifs et de la production de rapports automatisée.

Présentement, grâce à la collaboration de différents groupes de recherche, l'application Web est optimisée afin de faciliter la création et la passation d'épreuves traditionnelles et adaptatives résultant de différentes modalités existantes, et aussi de permettre l'analyse de données en ligne. Le caractère généralisable de la future application offrira aux utilisateurs la possibilité d'utiliser des procédures de sélection et de notation des items diversifiées résultant de l'application des modèles de la TRI. À moyen terme, ce système informatisé sera disponible pour les chercheurs désirant effectuer ou poursuivre leurs travaux de recherche dans le domaine du testing adaptatif. La section suivante présente quelques-unes des principales orientations fonctionnelles adoptées pour la conception et le développement de la deuxième génération de l'application.

DÉVELOPPEMENT DE LA DEUXIÈME GÉNÉRATION DE L'APPLICATION

La deuxième génération de l'application (V2) permettra l'accès à plusieurs formes d'épreuves. Dès la première phase de développement, les formes développées et offertes par défaut seront : les épreuves adaptatives, les enquêtes (sondages) et les formes traditionnelles (questionnaires et tests). Selon les besoins de chaque utilisateur (individu ou organisation), cette architecture ouverte offrira l'opportunité de développer avec l'équipe informatique une forme d'épreuve personnalisée.

Chaque épreuve, peu importe la modalité de passation, doit reposer sur une banque d'items d'où proviennent les items qui la composent. Il est aussi possible de disposer de plus d'une banque d'items. Cependant, selon les modalités d'épreuves à exploiter, les items de ces banques devront ou non disposer de certaines caractéristiques. Par exemple, actuellement, les modalités adaptatives informatisées requièrent que les items utilisés soient calibrés, mais toutes les organisations n'ont pas nécessairement les ressources pour produire ce type d'informations techniques, faute de pouvoir expérimenter les items sur le terrain avec un nombre raisonnable de répondants. De plus en plus d'organismes et de groupes de recherche se spécialisant dans le développement de banques d'items de qualité, la nouvelle application permettra de partager les banques d'items spécifiques et développées par différentes entités.

Les organisations qui exploitent des épreuves informatisées doivent disposer d'un dépôt de données qui recensent la population (l'échantillon) d'individus qui subissent les épreuves. Or, chaque organisation est assujettie à une structure organisationnelle qui caractérise chacun de ses répondants. Afin de répondre aux besoins spécifiques de chaque organisation, l'application

intègrera une fonctionnalité qui définira la structure de l'organisation dans laquelle les données des répondants (nom, prénom, genre, etc.) pourront être distribuées. Le caractère générique de cette approche permet de répondre aux besoins spécifiques de chaque organisation.

La nouvelle application prévoit aussi la gestion multilingue de chacune des fonctionnalités prévues dans le système (de la gestion du système à la passation des items). Des procédures sont développées afin de faciliter le changement instantané d'une langue à une autre, et ce, autant pour l'interface d'utilisation que pour l'affichage des items. Dans un contexte de dualité ou de pluralité linguistique, pour les enquêtes à grande échelle par exemple, cet avantage peut grandement réduire l'effet des biais associé à la traduction des items dans un nombre limité de langues.

L'image de marque d'une organisation qui fait passer des épreuves informatisées via Internet peut être intégrée à l'application. Il existe aujourd'hui sur le marché de nombreux systèmes de passation d'épreuves, mais rares sont ceux qui offrent la possibilité d'intégrer une image de marque identique (ou très similaire) à l'image du site corporatif d'une organisation. De cette façon, les utilisateurs de la plateforme identifiée à une organisation (intervenants et répondants) peuvent mieux s'y retrouver étant donné qu'elle est à l'image du site corporatif de l'organisation et qu'elle respecte les habitudes de navigation dans celui-ci.

Ces principes d'évolutivité, de partage des banques d'items, de personnalisation des épreuves et de l'image de marque ainsi que de gestion multilingue se raffineront avec le temps. Ces principes constituent le point de mire des appréciations qui ont été communiquées tout au long de la validation des travaux de l'architecture fonctionnelle de la deuxième génération de la plate-forme. En fait, ils transcendent son fort caractère novateur.

Références

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

Blais, J.-G., Boiteau, N., Burney-Vincent, C., Caron, F., Desmarais, M., Gagnon, M., Grondin, J., Paré, M. (2007), *L'informatisation du prétest de mathématiques de l'École polytechnique de Montréal*, Conférence présentée à la Maison des technologies de formation et d'apprentissage, École Polytechnique de Montréal, novembre.

Bock, R. D. et Zimowski, M. F. (1998), *Feasibility studies of two-stage testing in large-scale educational assessment : Implications for NAEP*, Chicago, National Opinion Research Center (NORC).

Boiteau, N. (2004), *Comparaison d'une modalité optimale de testing à deux niveaux et de la modalité de testing adaptatif par ordinateur*, Thèse de doctorat inédite, Université Laval, Canada.

Chang, H. H. et Ying, Z. (1996), « A global information approach to computerized adaptive testing », *Applied Psychological Measurement*, vol. 20, p. 213-229.

Dechef, H. et Laveault, D. (1999), « Le testing adaptatif par ordinateur », *Psychologie et Psychométrie*, vol. 20, p. 151-180.

Drasgow, F., Olson-Buchanan, J. B. et Moberg, P.J. (1999), « Development of an interactive video assessment : Trials and tribulations », dans F. Drasgow et J. B. Olson-Buchanan (dir.), *Innovations in computerized assessment*, Mahwah, NJ, Lawrence Erlbaum Associates.

Embretson, S. E. (1999), « Issues in the measurement of cognitive abilities », dans S.E. Embretson et S. L. Hershberger (dir.), *The new rules of measurement : What every psychologist and educator should know*, Mahwah, NJ, Lawrence Erlbaum Associates.

Embretson, S. E. et Reise, S. P.(2000), *Item response theory for psychologist,* Mahwah, NJ, Lawrence Erlbaum Associates.

Fisher, R.A (1922). « On the mathematical foundations of theoretical statistics », *Philosophical transactions of the Royal Society of London (A), 222*, p. 309-368.

Hambleton, R. K., Jones, R. W. et Rogers H. J. (1990), « Influence of item parameter errors in test development », Texte présenté lors de la rencontre annuelle du APA, Boston.

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

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

Hau, K. T. et Chang, H. H. (1998), « Item selection in computerized adaptive testing : Should more discriminating items be used first ? », Texte présenté lors de la rencontre annuelle du National Council on Measurement in Education, San Diego, États-Unis.

Hulin, C. L., Drasgow, F. et Parsons, C. K. (1983), *Item response theory : Application to psychological measurement*, Homewood, IL, Dow Jones-Irwin.

Kim, H. et Plake, B. S. (1993), « Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing », Texte présenté lors de la rencontre annuelle du National Council on Measurement in Education, Atlanta, États-Unis.

Kingsbury, G. G. et Zara, A. R.(1989), « Procedures for selecting items for computerized adaptive tests », *Applied Measurement in Education*, vol. 2, p. 359-375.

Laurier, M. (1993), *L'informatisation d'un test de classement en langue seconde*, Québec, Université Laval, Faculté des lettres.

Lord, F. M. (1980), *Applications of item response theory to practical testing problems*, Hillsdale, NJ, Lawrence Erlbaum Associates.

McBride, J. R. (1997), « Technical perspective », dans W. A. Sands, B. K. Waters et J. R. McBride (dir.), *Computerized adaptive testing : From inquiry to operation,* Washington, DC, American Psychological Association.

Mills, C. N. (1999), « Development and introduction of a computer adaptive graduate record examinations general test », dans F. Drasgow et J. B. Olson-Buchanan (dir.), *Innovations in computerized assessment*, Mahwah, NJ, Lawrence Erlbaum Associates.

Olsen, J. B., Maynes, D. D., Slawson, D. et Ho, K. (1989), « Comparisons of paper-administered, computer-administered and computerized adaptive achievement tests », *Journal of Educational Computing Research*, vol. 5, p. 311-326.

Raîche, G. (2000), *La distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt : selon l'erreur-type et selon le nombre d'items administrés*, Thèse de doctorat inédite, Université de Montréal, Québec.

Reckase, M. D. (1988), « Computerized adaptive testing : A good idea waiting for the right technology », Texte présenté lors de la rencontre annuelle du American Educational Research Association, Nouvelle-Orléans, États-Unis.

Reckase, M. D. (1996), « Test construction in the 1990s : Recent approaches every psychologist should know », *Psychological Assessment*, vol. 8, p. 354-359.

Sands, W. A. et Waters, B. K. (1997), « Introduction to ASVAB and CAT », dans W. A. Sands, B. K. Waters et J. R. McBride (dir.), *Computerized adaptive testing : From inquiry to operation*, Washington, DC, American Psychological Association.

Segall, D. O. et Moreno, K. E. (1999), « Development of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery », dans F. Drasgow et J. B. Olson-Buchanan (dir.), *Innovations in computerized assessment*, Mahwah, NJ, Lawrence Erlbaum Associates.

Thissen, D. et Mislevy, R. J. (2000), « Testing algorithms », dans H. Wainer (dir.), *Computerized adaptive testing : A primer*, Hillsdale, NJ, Lawrence Erlbaum Associates.

Thissen, D. et Orlando, M. (2000), « Likeliood-based item fit indices for dichotomous item response theory models », *Applied Psychological Measurement*, *24*, p. 50-64.

Thorndike, R. L., Hagen, E. P. et Slatter, J.M. (1986), *Technical manual : Stanford-Binet intelligence scale,* Chicago, Riverside.

van der Linden, W. J. et Eggen, T. J. H. M. (1986), « An empirical bayesian approach to item banking », *Applied Psychological Measurement*, vol. 10, p. 345-354.

Wainer, H. (2000), *Computerized adaptive testing : A primer*, Hillsdale, NJ, Lawrence Erlbaum Associates.

Wainer, H. et Kiely, G. L. (1987), « Item clusters and computerized adaptive testing : A case for testlets », *Journal of Educational Measurement*, vol. 24, p. 185-201.

Wainer, H. et Mislevy, R. J. (2000), « Item response theory, item calibration and proficiency estimation », dans H. Wainer (dir.), *Computerized adaptive testing : A primer*, Hillsdale, NJ, Lawrence Erlbaum Associates.

Wang, T. et Vispoel, W. P. (1998), « Properties of ability estimation methods in computerized adaptive testing », *Journal of Educational Measurement*, vol. 35, p. 109-135.

Weiss, D. J. (1974), *Strategies of adaptive ability measurement*, (Research Report n° 74-5). Minneapolis, University of Minnesota.

Weiss, D. J. (1982), « Improving measurement quality and efficiency with adaptive testing », *Applied Psychological Measurement*, vol. 6, p. 473-492.

Weiss, D. J. (1983), *New horizons in testing : Latent trait test theory and computerized adaptive testing*, New York, Academic Press inc.

Weiss, D. J. (1985), « Adaptive testing by computer », *Journal of Consulting and Clinical Psychology*, vol. 6, p. 774-789.

Weiss, D. J. et Betz, N. E. (1973), *Ability measurement : Conventional or adaptative*?, (Research Report n° 73-1), Minneapolis, University of Minnesota.

Zimowski, M. F., Muraki, E., Mislevy, R. J. et Bock, R. D. (1996), *BILOG-MG. Multiple-group IRT analysis and test maintenance for binary items,* Mooresville, IN, Scientific Software.

Page laissée blanche intentionnellement

SECTION II

Des développements et des applications multiples

Page laissée blanche intentionnellement

CHAPITRE 4

L'évaluation des compétences langagières

Réginald Burton, Monique Reichert, Romain Martin
Université du Luxembourg, unité de recherche EMACS

INTRODUCTION

L'évaluation des compétences langagières assistée par ordinateur[1] s'est considérablement développée depuis l'introduction des premiers micro-ordinateurs dans le courant des années 70 grâce aux progrès de la technologie et de la psychométrie. Selon Bunderson, Inouye, et Olsen (1989), quatre générations d'Évaluation Assistée par Ordinateur (ÉAO) se sont ainsi succédé. Pendant la première génération, nous avons assisté à la transposition pure et simple de tests papier-crayon sur un support informatisé. La deuxième génération de tests informatisés a vu le jour selon les modalités du testing adaptatif par ordinateur. Le caractère adaptatif des tests constitue une innovation et un atout majeur puisqu'il permet de raccourcir considérablement la longueur des tests sans perdre en précision de mesure. La troisième génération des tests informatisés décrite par Bunderson *et al.* (1989) est appelée mesure continue[2]. Son fonctionnement technique est calqué sur celui du *testing* adaptatif par ordinateur, mais l'évaluation est censée être livrée en continu, de manière à être intégrée dans le processus même d'apprentissage en vue de guider les apprenants. Le résultat de la mesure devient ainsi dynamique dans l'optique d'une évaluation formative. La quatrième génération des tests informatisés, qualifiée de mesure intelligente[3], va encore aller plus

1. En anglais : Computer-Assisted Language Testing (CALT).
2. En anglais : Continuous Measurement (CM).
3. En anglais : Intelligent Measurement (IM).

loin dans cette direction en essayant d'automatiser les mesures de régulation à adopter à la suite des évaluations continues qui sont réalisées.

À l'heure actuelle, nous assistons à l'émergence d'une nouvelle génération d'ÉAO qui intègre de nouvelles options ou, du moins, dans laquelle se développent des potentialités jusque là sous-exploitées. Les progrès technologiques réalisés ces vingt dernières années ont non seulement favorisé l'intégration de contenus multimédias et le développement de nouveaux formats d'items, mais ont aussi permis d'envisager de nouvelles possibilités d'interaction et de correction automatique. Nous verrons comment ces potentialités ont été appliquées dans le cadre spécifique de l'évaluation des compétences langagières à travers la présentation de divers dispositifs d'ÉAO.

En dehors de l'évaluation de la maîtrise des compétences en langue à proprement parler, nous verrons, en outre, que l'ordinateur permet de neutraliser les composantes linguistiques qui entrent en jeu dans les contextes d'apprentissage multilingues en proposant aux sujets de choisir, à tout moment de la passation des tests, la langue en fonction de la tâche à effectuer.

La fin du chapitre sera consacrée à deux questions fondamentales soulevées par l'exploitation des nouvelles possibilités d'ÉAO en langues. La première concerne la validité des mesures réalisées avec les systèmes informatiques. Les systèmes d'ÉAO, plus particulièrement en langues, mesurent-ils les mêmes construits que les tests papier-crayon ? La seconde a trait à l'intégration des cadres de référence pour l'évaluation des compétences langagières dans les dispositifs d'ÉAO. Est-il possible de faire coïncider les cadres de référence pour l'évaluation des compétences langagières avec les modèles sous-jacents à l'ÉAO ? Enfin, dans les conclusions, nous évoquerons brièvement une dernière perspective concernant l'exploitation des traces.

LE POTENTIEL DE L'ÉAO POUR LA MESURE DES COMPÉTENCES LANGAGIÈRES

Possibilités du multimédia

De prime abord, le potentiel du *testing* assisté par ordinateur (TAO) repose avant tout sur la possibilité d'insérer des contenus multimédias sous des formes multimodales. Il est ainsi possible de présenter aux sujets des questions comprenant des vidéos, des textes, des sons, des graphiques, des simulations, des jeux, etc. Les applications sont évidentes dans le cadre de l'évaluation des compétences langagières notamment pour l'évaluation de la compréhension de l'oral et de l'écrit. On peut ainsi imaginer présenter des extraits d'émissions radiophoniques ou télévisées pour l'évaluation de la

compréhension de l'orale ou encore des environnements virtuels simulés pour l'évaluation de la compréhension de l'écrit. A l'heure actuelle, la plupart des dispositifs d'ÉAO (voir par exemple, DIALANG[4] et TAO[5]) proposent des items comportant des bandes sonores ou vidéo pour l'évaluation de la compréhension de l'oral.

Dans l'exemple ci-dessous (figure 1) extrait du dispositif DIALANG, l'utilisateur peut visionner à loisir un extrait vidéo avant de répondre à une question à choix multiple. DIALANG est le premier système d'ÉAO majeur mis gratuitement à disposition sur Internet. Il a été conçu à partir du Cadre européen commun de référence pour les langues. Il a été développé par plus de vingt universités européennes sous la tutelle de la Commission européenne. Il permet de réaliser des évaluations formatives équivalentes dans quatorze langues de l'Union européenne pour l'écoute, la lecture, la rédaction, la grammaire et le vocabulaire. La caractéristique principale de DIALANG, qui le distingue des autres dispositifs d'ÉAO, est de combiner des phases de *testing* et d'autoévaluation. Il propose ainsi, avant la phase de test proprement dite, un test de niveau sur l'étendue du vocabulaire et un questionnaire d'autoévaluation. Cette phase initiale permet de prépositionner les sujets dans un des trois niveaux de difficulté du test final (Demeuse et al., 2007).

Figure 1 – Item de compréhension de l'oral (DIALANG).

Par ailleurs, dans sa prochaine édition en 2009, le programme international pour le suivi des acquis des élèves (PISA) prévoit un module facultatif ayant pour objectif d'évaluer la compréhension de la lecture des textes électroniques. Dans cette perspective, les concepteurs des épreuves envisagent des items qui simuleront un environnement Internet[6]. Avec au départ un

4. http://www.dialang.org
5. http://www.tao.lu
6. cf. Brochure de présentation du Programme International pour le Suivi des Acquis des élèves (2007) : http://www.pisa.oecd.org

nombre restreint de pages web construites pour les besoins de l'évaluation (moteurs de recherche, forum, blogues …) et interconnectées par des hyperliens, les candidats sont amenés à répondre à des questions de compréhension de niveaux taxonomiques variés.

Selon Blais, Dionne et Raîche (2005), l'intégration de sons, d'images, de scénarios ou de simulations rendent les épreuves plus réalistes et plus authentiques. Autrement dit, la possibilité d'insérer des contenus multimédias permet d'augmenter la validité écologique des situations d'évaluation. Ces potentialités ne se limitent pas aux seules modalités de question, mais s'étendent tout naturellement aux modalités de réponse. Il est donc également possible d'enregistrer des sons, des images, des textes, etc.

Formats d'items

Le dernier exemple cité au sujet de l'évaluation de la compréhension de la lecture des textes électroniques dans le programme international pour le suivi des acquis des élèves (PISA 2009) laisse déjà transparaître que l'atout majeur de l'ÉAO n'est pas tant la possibilité d'insérer des contenus multimédias mais bien, grâce à cette dernière, la possibilité de formats d'items nouveaux. Si, comme nous l'avons rappelé dans l'introduction, les premières tentatives de TAO se sont limitées à transposer des items classiquement utilisés dans les tests papier-crayon (questions ouvertes, QCM…) sur un support informatisé, nous assistons, à l'heure actuelle, au développement de nouveaux formats d'items. Ainsi, plusieurs dispositifs d'ÉAO expérimentent des formats d'items qu'il n'était pas possible de concevoir dans le cadre de tests papier-crayon ou du moins qu'il était difficile de mettre en œuvre. À titre d'exemple, DIALANG propose dix-huit formats pour l'évaluation de la compréhension de l'oral/de l'écrit, de l'expression orale/écrite et du vocabulaire (voir encart 1 ci-dessous).

Au-delà des possibilités d'insérer des contenus multimédias, l'ÉAO permet, à travers ces nouveaux formats d'items, d'appréhender des compétences complexes difficilement mesurables par des tests papier-crayon, telles que des compétences sociales ou professionnelles. Cela est notamment rendu possible grâce à une interaction accrue entre le test et les sujets. Chapelle et Douglas (2006) notent à ce propos qu' « en dehors du fait que l'évaluation des compétences langagières assistée par ordinateur n'offre pas les mêmes types d'interactivité que d'autres techniques d'évaluation telles que les interviews, elle offre beaucoup d'avantages par rapport aux méthodes statiques de mesure des compétences linguistiques en termes d'adaptabilité (adaptation de la difficulté de la tâche, choix du mode de questionnement et réponse par le sujet, autoévaluation…) et de réactivité (correction automatique et

ENCART 1
Formats d'items expérimentaux proposés par DIALANG.

Type 1 – choix multiple d'images avec son
Avec au départ une description orale, le sujet est amené à choisir une image correspondante selon un choix multiple.

Type 2 – image interactive avec son
Avec au départ une description orale, le sujet est amené à choisir l'élément d'une image interactive.

Type 3 – vidéo-clips pour l'écoute
Avec au départ une vidéo, le sujet est amené à répondre à des questions de compréhension à choix multiple.

Type 4 – activité de *drag & drop*
Avec au départ des directives orales, le sujet est amené à déplacer des objets à des endroits déterminés.

Type 5 – réorganisation
Le sujet est amené à remettre dans l'ordre une série de phrases constitutives d'un texte.

Type 6 – mise en évidence/soulignement
Le sujet est amené à mettre en évidence une phrase d'un texte en fonction de critères déterminés. En cliquant sur certains mots, la définition de certains mots de vocabulaire apparaît.

Type 7 – insertion
Le sujet doit dans un premier temps repérer l'endroit d'une phrase où il manque un mot et, dans un second temps, écrire une proposition pour le mot manquant.

Type 8 – effacement
Le sujet est amené à repérer un mot superflu dans une phrase.

Type 9 – groupement thématique
Parmi une liste de mots, le sujet est amené à réaliser des groupements thématiques.

Type 10 – réponses multiples
Le sujet est amené à répondre à une question à choix multiple qui attend plusieurs réponses.

Type 11 – expression indirecte avec bandes sonores
Le sujet est amené à répondre à une question à choix multiple dont les possibilités sont des bandes sonores.

Type 12 – repérage en écriture directe
Dans un premier temps, le sujet est amené à écrire un texte dans une fenêtre prévue à cet effet selon des consignes données. Il a ensuite la possibilité d'autoévaluer le niveau de son texte par comparaison avec des exemples triviaux se référant aux niveaux de performance du CECR.

Type 13 – autoévaluation de l'expression orale
À la suite d'une consigne de départ (par exemple, indiquer le chemin à une personne), des bandes sonores de niveaux de complexité croissants sont proposées aux sujets de manières successives. Ceux-ci doivent alors spécifier le niveau de réponse qui leurs correspond en identifiant la bande sonore qu'ils jugent être de leur niveau.

Type 14 – correspondance *drag & drop*
Le sujet est amené à faire correspondre deux listes de mots par *drag & drop*.

Type 15 – transformation
Le sujet est amené à transformer par écrit une phrase initiale selon des critères donnés.

Type 16 – tâches combinées
Au moyen d'un texte, le sujet peut choisir indifféremment de répondre à des questions de vocabulaire ou de compréhension.

Type 17 – cartes conceptuelles et diagrammes fléchés
Le sujet est amené à construire des cartes conceptuelles ou des diagrammes fléchés synthétisant un texte de départ.

Type 18 – degrés de certitude
En plus de sa réponse, le sujet est amené à évaluer son degré de certitude sur une échelle.

rapide des réponses, aide à la résolution de la tâche…) », comme le montrent les formats d'items proposés par DIALANG.

En développant encore les capacités d'interactivité des systèmes d'ÉAO, Jadoul *et al.* (2006) prévoient la conception d'un nouveau format d'item pour procéder à des évaluations dynamiques des compétences. Initialement conçu pour l'évaluation des compétences scientifiques, le système peut être facilement transférable à l'évaluation des compétences langagières, ce qui ouvre la porte à la troisième génération de TAO décrite par Bunderson *et al.* (1989) réalisée dans une perspective d'évaluation formative. Le système comporte trois phases :

- une phase de « prétest » au cours de laquelle les connaissances initiales du sujet dans un domaine précis sont évaluées ;
- une phase de « présentation des ressources d'information » au cours de laquelle le sujet a accès à un matériel multimédia concernant la tâche sur laquelle les questions de la première phase ont porté ;
- une phase d'« autoformation » pendant laquelle le sujet va pouvoir utiliser à sa guise le matériel qui est mis à sa disposition en vue de retravailler et de corriger les réponses données aux questions de la phase initiale.

Le système informatique comprend quant à lui quatre fenêtres interactives : une fenêtre principale d'affichage, une fenêtre contenant une arborescence permettant à tout moment de consulter les questions posées et d'y répondre, une fenêtre d'accès à des ressources d'informations (vidéos, textes, graphiques…) contenant des éléments de réponse aux questions posées et, enfin, une fenêtre où s'inscrit en temps réel l'ensemble des opérations réalisées par le sujet qui peut ainsi définir avec précision les moments-clés de son exploration. Lors de l'évaluation, les comportements d'exploration des sujets vis-à-vis des ressources d'information sont enregistrés par l'ordinateur (type de matériel consulté, temps passé sur chaque type de matériel, stratégie d'exploration). Ces données comportementales pourront être utilisées par la suite pour la réalisation d'analyses typologiques permettant de déterminer des stratégies d'utilisation et d'exploration des ressources d'informations (Burton et Martin, 2006).

Le testing adaptatif

Un autre intérêt majeur de l'ÉAO réside dans sa modalité adaptative : le *testing* adaptatif par ordinateur[7] (voir N. Boiteau et M. Paré, Développement

7. En anglais : Computer Adaptive Testing (CAT).

d'une application Web pour l'administration d'épreuves adaptatives, chapitre 3). Le *testing* adaptatif par ordinateur consiste à présenter aux sujets des items dont la difficulté est adaptée à leur niveau de compétence. Dans une procédure de *testing* adaptatif par ordinateur, les items présentés aux sujets vont dépendre des réponses des sujets aux items précédemment passés, et des sujets différents vont donc se voir donner des items différents. Le premier avantage du *testing* adaptatif par ordinateur est de réduire considérablement le temps de passation des tests puisqu'il est possible d'estimer, avec une même précision, la compétence des sujets au départ d'un nombre restreint d'items. Le second avantage concerne la qualité des mesures obtenues en limitant la démotivation des sujets lors de la passation des tests. La perte de motivation des sujets peut être ainsi évitée puisque seul un nombre restreint d'items dont la difficulté se situe dans leur zone proximale de développement (Vygotsky, 1984) sont passés. La procédure est rendue possible grâce à l'application de modèles issus de la Théorie de la Réponse à l'Item (TRI), car cette dernière propose une modélisation de la mesure permettant d'estimer la compétence des sujets indépendamment des items qui sont passés. Wainer (2000) indiquait un accroissement exponentiel dans la passation de tests assistés par l'ordinateur sous la modalité adaptative depuis le début des années 90.

En langues, on trouve plusieurs applications adaptative du TAO, dont le « Test of English as a Foreign Language[8] » (TOEFL) est sans doute le plus connu. Le TOEFL est un test standardisé qui vise à évaluer l'aptitude à comprendre et à utiliser l'anglais pour des sujets dont ce n'est pas la langue maternelle. La première version papier du TOEFL (PBT ou *Paper-Based* TOEFL) a été élaborée en 1964 par la firme *Educational Testing Service* (ETS) et a été transposée par la suite sur ordinateur (CBT ou *Computer-Based* TOEFL). La version initiale PBT est composée de trois sections : compréhension de l'oral, structure et expression écrite, compréhension de l'écrit. En 2005, le TOEFL a été décliné dans une version Internet (iBT ou *Internet-Based* TOEFL) qui comporte quatre parties : lecture, écoute, oral, rédaction. La firme ETS affirme sur son site internet que le test a été utilisé par plus de 5000 institutions réparties dans 110 pays.

Interactions enrichies et systèmes d'auteur[9]

Les systèmes d'ÉAO présentés jusqu'à présent ont la particularité d'être fermés aux concepteurs d'évaluations dans la mesure où ils n'offrent aucune

8. http://www.fr.toefl.eu
9. En anglais : *authoring system*.

possibilité d'intervention ou de création, et se limitent à délivrer des tests préconstruits. À côté de ces systèmes fermés, on trouve des systèmes d'auteurs ouverts aux concepteurs désireux de construire eux-mêmes leur propre dispositif d'évaluation. De tels systèmes mettent à la disposition d'auteurs une palette d'outils informatiques différents pour la construction, la mise en œuvre et le traitement de tests d'évaluation. On peut dès lors imaginer, par exemple, des enseignants mettant au point au sein d'une école un dispositif d'évaluation formative assistée par ordinateur. Les systèmes d'auteurs peuvent être caractérisés selon deux critères (Chanier, 2007) : la complexité de développement et la liberté de création. L'auteur peut décider de faire de la programmation et de créer un produit qui répondra spécifiquement à ses besoins. Le processus de développement sera alors complexe et technique, mais la liberté de l'auteur importante. Il peut aussi décider d'utiliser un langage d'auteur, qui offre des fonctions à visées pédagogiques toutes prêtes tout en fermant certaines possibilités de programmation. À l'autre extrémité, l'auteur pourra développer son logiciel à partir d'un système d'auteur de façon beaucoup plus simple et rapide, mais cet environnement pourra lui imposer des schémas et des séquencements d'activités qui peuvent s'avérer incompatibles avec son scénario pédagogique initial. Ainsi, on peut devenir « auteur » à partir de plusieurs bases, lesquelles détermineront le temps à consacrer aux travaux et l'étendue des applications.

À titre d'exemple, citons la plateforme TAO[10] mise au point conjointement par l'unité de recherche EMACS de l'Université du Luxembourg et le département CITI (Centre d'Innovation par les Technologies de l'Information) du Centre de Recherche Public Henri Tudor (Martin *et al.*, 2005). La plateforme TAO est une plateforme ouverte (*open-source*) qui fournit aux auteurs une architecture ouverte pour le développement et la passation de tests assistés par ordinateur. Sa nature générique et polyvalente permet à tout utilisateur de créer des items et des tests, de délivrer des tests, de gérer des programmes d'évaluation, de gérer des populations de sujets testés et d'analyser les résultats quelle que soit la discipline envisagée. Elle se compose de six modules (figure 2) : un module de développement des items, un module de développement de tests, un module de gestion des sujets, un module de gestion de groupes de sujets (classes, écoles, entreprises…), un module de passation des tests, et un module de gestion des résultats. Sa structure modulaire et distribuée favorise la collaboration entre les différents utilisateurs situés en des lieux différents à des moments différents.

10. http://www.tao.lu

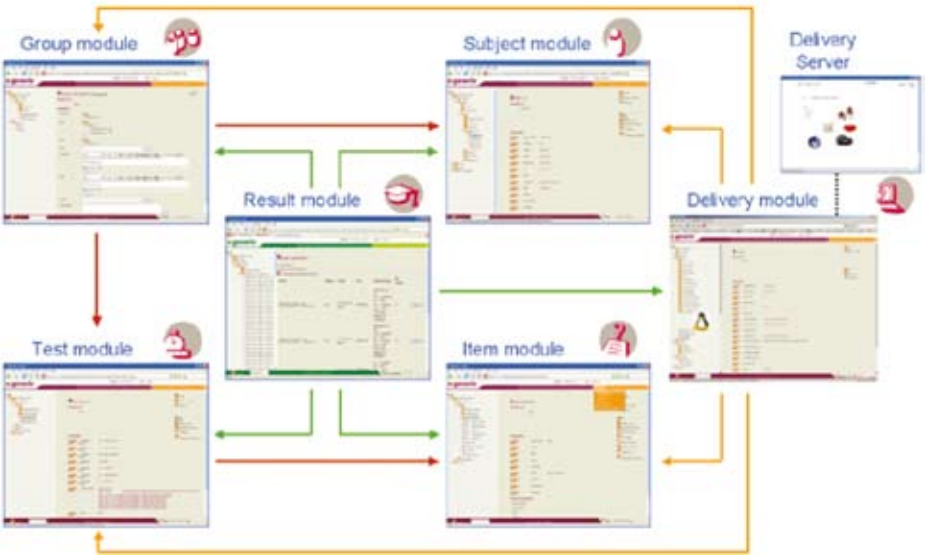


Figure 2 – Architecture de la plateforme TAO.

En ce qui concerne l'évaluation des compétences langagières, un système de *testing* adaptatif par ordinateur a notamment été mis au point pour l'évaluation de la compréhension de l'oral et de l'écrit. Ce système a été appliqué au Centre de Langues Luxembourg pour les besoins de placement des apprenants dans des groupes de niveau définis par le CECR (Burton, 2004).

Correction automatique

De nombreux formats d'items utilisés dans les évaluations en langues attendent comme réponse des productions écrites plus ou moins élaborées. À l'heure actuelle, la correction de tels items est généralement effectuée manuellement par des correcteurs humains. L'ÉAO fournit cependant des solutions alternatives. Ces dernières années ont été marquées par le développement de méthodes de correction automatique basées sur des analyses linguistiques. Un chapitre complet de cet ouvrage (le chapitre 5) se penchant sur la question de la correction automatisée, cette section ne s'attarde pas à un traitement en profondeur. Cependant, il apparaît important d'en résumer quelques enjeux.

Les dernières années ont été marquées par le développement de méthodes de correction automatique basées sur des analyses linguistiques qui permettent de déterminer les aspects significatifs des productions écrites. Deane

(2006) distingue trois types d'analyses linguistiques qui peuvent être employées seules ou simultanément dans les systèmes d'ÉAO :

Les méthodes basées sur les caractéristiques linguistiques

Dans ces approches, des caractéristiques linguistiques spécifiques sont extraites de textes et des analyses de régression sont ensuite effectuées pour déterminer leur relation avec des indicateurs de performance. Les caractéristiques sont répertoriées dans des « catégories de niveau conceptuel », c'est-à-dire des catégories de structures linguistiques telles que des constructions grammaticales particulières, des catégories de mots particuliers. Ces méthodes incluent les simples statistiques lexicales telles que la fréquence de mots ou le nombre de mots par phrase, mais aussi des caractéristiques basées sur une liste de mots appartenant à une classe sémantique particulière, ou encore toute une série de caractéristiques grammaticales extraites par des algorithmes sophistiqués. Elles sont employées, en combinaison avec d'autres méthodes, dans des systèmes de correction automatique des dissertations tels que le « Project Essay Grade » (PEG, cf. Page, 1994 ; Page, 2003 cités par Deane, 2006) qui utilise principalement les méthodes basées sur les caractéristiques linguistiques associées aux analyses de structure linguistique, ou le « E-rater » (cf. Burstein, 2003 ; Burstein, Kukich, Wolff, Lu, Chodorow *et al.*, 1998 cités par Deane, 2006) qui utilise une combinaison des trois méthodes. Elles sont également employées pour des analyses de genre (cf. Biber, 1998 ; Biber, Reppen, Clark et Walter, 2001 cités par Deane, 2006).

Les méthodes des espaces vectoriels

Ces approches consistent à représenter un texte par un vecteur de caractéristiques linguistiques – qui peuvent par exemple être simplement des occurrences de mots utilisés – dans un espace vectoriel approprié. Un modèle prédictif empirique est ensuite élaboré par le calcul de similarité entre les vecteurs. Dans ce type de méthode, l'attention est portée davantage sur la synthèse des caractéristiques linguistiques plutôt que sur leur extraction. Ces méthodes doivent être distinguées des premières dans la mesure où elles n'emploient pas directement des représentations des catégories de niveau conceptuel. La « Latent Semantic Analysis » (LSA) est sans doute la plus connue des méthodes des espaces vectoriels (cf. Landauer et Dumais, 1997 cités par Deane, 2006). Elles sont utilisées dans diverses tâches de correction automatique des dissertations comme l'« Intelligent Essay Assessor (IEA, cf. Landauer, Laham et Foltz, 2003 ; Foltz, Laham et Landauer, 1999 cités par Deane, 2006), mais aussi pour mesurer la cohérence de textes (Foltz, Kintsch

et Landauer, 1998 cités par Deane, 2006), pour évaluer des résumés des étudiants (Foltz, Gilliam et Kendall, 2000 ; Wade-Stein et Kintsch, 2003 cités par Deane, 2006) et pour guider des systèmes de tutorat automatisés (Lemaire, 1999 ; Wiemer-Hastings, P., Wiemer-Hastings, K. et Graesser, A.C., 1999 a et b cités par Deane, 2006).

Les analyses de structure linguistique

Dans ces approches, les techniques du processus du langage naturel[11] sont appliquées pour construire un modèle partiel des représentations mentales associées aux concepts pertinents de la tâche. Ce qui distingue les analyses linguistiques de cette sorte des simples méthodes d'extraction est qu'elles extraient une structure, un ensemble d'éléments et leurs relations, plutôt que de simples propriétés quantifiables telles que l'occurrence de mots ou la fréquence de catégories concepts. Les analyses de structure linguistique servent notamment à l'élaboration de systèmes de correction automatique des réponses courtes, tels que le « R-rater scoring system » développé par le « Educational Testing Service » (cf. Leacock et Chodorow, 2004 cités par Deane, 2006) et le « Automark system » développé par « intelligent assessment Technologies » (cf. Mitchell, Russell, Broomhead et Aldridge, 2002 cités par Deane, 2006).

LA NEUTRALISATION DE LA COMPOSANTE LINGUISTIQUE DANS LES SYSTÈMES MULTILINGUES

Les modèles de mesure utilisés dans le cadre de l'évaluation des compétences – qu'elle soit assistée par ordinateur ou non – supposent généralement l'utilisation d'instruments de mesure produisant des ensembles de données unidimensionnels qui permettent de situer d'une manière non ambigüe les sujets le long d'un continuum. C'est le cas pour la théorie classique des tests, mais aussi pour une grande majorité de modèles de la TRI, du moins pour les plus exploités aujourd'hui. L'unidimensionnalité d'un test signifie qu'il doit mesurer un et un seul construit à la fois. « Dans la pratique, ce critère n'est jamais parfaitement satisfait du fait des inévitables erreurs de mesure et de la complexité des traits mesurés » (Laveault et Grégoire, 1997). En effet, il arrive souvent que d'autres facteurs cognitifs ou affectifs viennent influencer les performances des élèves au test, tels que la compréhension en lecture dans le cas d'un test de mathématiques, la motivation ou l'anxiété. En fait, « il est raisonnable de considérer qu'un ensemble de *n* tests ou de *n* items

11. En anglais : Natural Language Processing (NLP).

dichotomiques est unidimensionnel si et seulement s'il s'ajuste à un modèle factoriel non linéaire avec un seul facteur commun » (McDonald, 1981 cité par Laveault et Grégoire, 1997). Dans les contextes multilingues, la langue dans laquelle sont donnés les tests peut avoir un impact sur l'unidimensionnalité du trait latent mesuré.

Citons à titre d'exemple les épreuves standardisées réalisées dans le cadre de l'orientation des élèves de 6e primaire au Luxembourg. Les épreuves standardisées réalisées pour les besoins de la procédure d'orientation portent sur trois disciplines : le français, les mathématiques et l'allemand. Dans chaque discipline, des compétences différentes sont évaluées afin de dresser un large profil de compétences pour chaque enfant. Ainsi, les épreuves de langues (français et allemand) tentent généralement d'appréhender la capacité des élèves dans les domaines de la compréhension de l'écrit et de l'oral, de la grammaire, de l'orthographe et de la conjugaison. L'épreuve de mathématiques tente, quant à elle, d'appréhender les aptitudes des élèves à reproduire des procédures qui ont fait l'objet d'un apprentissage et celles à les transférer dans d'autres contextes. Dans la pratique, l'allemand constitue la langue véhiculaire de l'apprentissage des mathématiques, et les tests de mathématiques sont principalement délivrés dans la même langue (moins de 10 % des tests de mathématiques sont délivrés en français). Dans ces conditions, une analyse en composante principale réalisée au départ des scores des élèves dans chaque discipline, montre l'existence d'une composante linguistique allemande dans l'évaluation des compétences en mathématiques (Burton, 2003). La prépondérance du facteur relatif aux compétences langagières en allemand se traduit dans la pratique par une discrimination envers les élèves romanophones (portugais, français, belges…) tant au niveau de leurs performances aux épreuves standardisées qu'au niveau des décisions qui en découlent, c'est-à-dire l'avis d'orientation. En fait, les élèves ont la possibilité de choisir une version française où allemande du test de mathématiques, mais pour les élèves romanophones, le choix d'une version française pose également le problème de la présence d'un facteur allemand dans la mesure où les mots de vocabulaire en mathématiques sont connus dans la langue véhiculaire.

Il existe bien entendu, dans le cadre des propositions de la TRI, des modèles multidimensionnels (voir Mc Donald, 1997 ; Reckase, 1997 ; Kelderman, 1997) qui permettraient de tenir compte des composantes linguistiques qui entrent en jeu dans la compréhension des questions d'évaluation. Mais, souvent très complexes sur le plan psychométrique, ils sont difficiles à mettre en œuvre notamment aussi à cause de leur exigence en terme d'effectifs de sujets nécessaires pour l'estimation des paramètres (Laveault et Grégoire, 1997 ; Bertrand et Blais, 2004). Les systèmes de TAO permettent quant à eux d'offrir la possibilité à tous les élèves de choisir, à

tout moment de la passation des tests, la langue qu'il préfère en fonction de la tâche à effectuer.

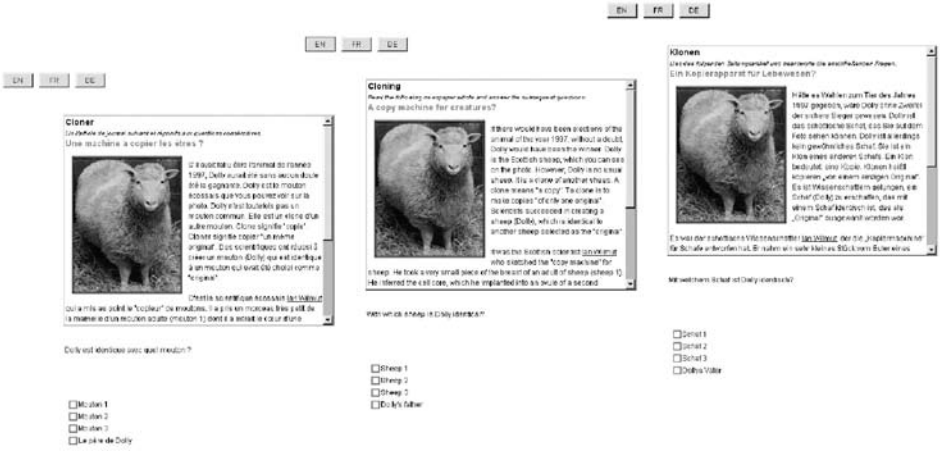


Figure 3 – Possibilité de changer de langue sur la plateforme TAO

Le système TAO, par exemple, permet aux sujets de changer la langue à n'importe quel moment de l'évaluation grâce à des boutons. Le sujet peut ainsi passer d'une langue à l'autre en fonction de ses caractéristiques linguistiques (figure 3). Les systèmes de *testing* assisté par ordinateur permettent également, comme dans le cas de DIALANG, de fournir des aides automatiques lors de la résolution d'une tâche (des dictionnaires par exemple).

VALIDITÉ DES MESURES

Une des critiques souvent avancée à propos de l' ÉAO concerne la validité des mesures obtenues. Selon Chapelle et Douglas (2006), le TAO en langues offre à l'heure actuelle un vaste éventail de possibilités. En même temps, ces nouvelles options conduisent à s'interroger sur la nature des mesures qui sont réalisées et sur ce que mesurent réellement ces nouveaux tests. Wainer *et al.* (2000), cités par Chapelle et Douglas (2006), résument cette position de la manière suivante : « Le fait que les items soient présentés sur un écran d'ordinateur, plutôt que sur un morceau de papier, modifie les processus mentaux requis pour répondre correctement à l'item et la validité des inférences réalisées au départ de ces scores peut s'en retrouver altérée ». Autrement dit, ces auteurs émettent l'hypothèse que la familiarité des sujets avec les technologies de l'information et/ou certains formats d'items particuliers pourrait influencer leur performance aux tests d'habiletés langagières.

Ces deux dernières décennies, plusieurs recherches se sont attachées à étudier l'impact de l'ÉAO sur les performances des sujets et sur les processus cognitifs invoqués dans les tâches d'évaluation des compétences langagières. Dans une revue de la littérature menée par Chapelle et Douglas (2006), ces recherches sont classées en trois types.

Les études de comparaison

Les études de comparaison consistent à comparer les performances de sujets qui se voient passer un même test sous deux formats différents : le plus souvent sous format papier-crayon d'une part, et sous format de test assisté part ordinateur d'autre part. Une première étude de ce type menée par Schaeffer *et al*. (1993), cités par Chapelle et Douglas (2006), aux Etats-Unis montre que les résultats obtenus par des sujets au *Graduate Record Examination* (GRE) ne sont pas affectés de manière claire et déterminante par le mode de passation des tests. Par ailleurs, une autre étude menée par Taylor *et al.* (1999), cités par Chapelle et Douglas (2006), indique un coefficient de corrélation de 0,84 entre les scores obtenus par les sujets aux versions PPT et CBT du TOEFL. Ils en concluent qu'il n'existe pas d'effet significatif du mode de passation du test sur les performances des sujets. D'une manière générale, on peut affirmer que les études de comparaison semblent concordantes à ce sujet (voir, par exemple, Coniam, 2001 ; Choi, Kim et Boo, 2003).

Les études de groupes contrastés

Les études de groupes contrastés proposent, quant à elles, de voir jusqu'à quel point des sujets peu familiers avec les technologies de l'information sont désavantagés lorsqu'ils se voient passer un test assisté par ordinateur. Dans une étude intitulée *TOEFL computer familiarity study*, Taylor *et al.* (1999), cités par Chapelle et Douglas (2006), ont réparti les sujets en deux groupes (les sujets familiers et les sujets non familiers avec l'informatique) selon leurs réponses à un questionnaire de familiarité envers les applications informatiques telles qu'Internet ou les traitements de textes. Après avoir pris en compte la compétence des sujets en langues mesurée par la version papier-crayon du TOEFL, aucune différence de performance au test assisté par ordinateur n'a été constatée entre les deux groupes. Cependant, Chapelle et Douglas (2006) signalent que les effets relatifs à la familiarité envers l'informatique ont pu être minimisés dans la mesure où les sujets ont profité, avant la passation du test assisté par ordinateur, d'une séance de tutoriel.

Les études de processus

Les études de processus ont pour objectif d'étudier l'effet du mode de passation des tests sur les processus cognitifs utilisés par les sujets. Les méthodes employées relèvent généralement d'analyses du mouvement des yeux, de protocoles d'analyses verbales, d'interviews ou de questionnaires. Mais Sawaki (2001), cité par Chapelle et Douglas (2006), note que de telles recherches n'apparaissent pas encore à l'heure actuelle dans la littérature de recherche relative à l'évaluation des compétences langagières.

CADRES DE RÉFÉRENCES ET VALIDITÉ CRITÉRIELLE

Depuis l'année 2000, l'Union européenne s'est dotée d'un cadre de référence commun pour l'apprentissage, l'enseignement et l'évaluation des langues vivantes. Le Cadre européen commun de référence (CECR) offre une base commune pour l'élaboration de programmes en langues vivantes, le développement des activités d'apprentissage et l'organisation de l'évaluation. Il décrit ce que les apprenants d'une langue doivent apprendre afin de l'utiliser dans le but de communiquer. Il énumère également les connaissances et les habiletés pour que les apprenants acquièrent un comportement langagier efficace. La description englobe aussi le contexte culturel qui soutient la langue. Le CECR envisage la compétence à communiquer à travers la réalisation d'activités langagières variées pouvant relever de la réception, de la production, de l'interaction, de la médiation (notamment les activités de traduction et d'interprétation). Chacun de ces modes d'activités est susceptible de s'accomplir soit à l'oral, soit à l'écrit, soit à l'oral et à l'écrit.

À cette première dimension « horizontale », le CECR fournit une dimension « verticale » en définissant six niveaux de référence hiérarchisés (voir l'encart 2) pour décrire la compétence de l'apprenant et pour mesurer leurs progrès à chaque étape de l'apprentissage. À côté du CECR, subsistent d'autres cadres de référence comme le cadre du PISA, élaboré pour l'évaluation de la compréhension de l'écrit qui définit également des niveaux de compétence.

ENCART 2
Niveaux de compétence en compréhension de l'oral selon le CECR.

		Descripteurs
Utilisateur Expérimenté	C2	Peut comprendre sans effort pratiquement tout ce qu'il entend ou lit.
	C1	Peut comprendre une grande gamme de textes longs et exigeants, ainsi que saisir des significations implicites.
Utilisateur indépendant	B2	Peut comprendre le contenu essentiel de sujets concrets ou abstraits dans un texte complexe, y compris une discussion technique dans sa spécialité.
	B1	Peut comprendre les points essentiels quand un langage clair et standard est utilisé et s'il s'agit de choses familières dans le travail, à l'école, dans les loisirs…
Utilisateur élémentaire	A2	Peut comprendre des phrases isolées et des expressions fréquemment utilisées en lien avec des domaines immédiats de priorité (informations personnelles et familiales simples, achats, environnement proche, travail…).
	A1	Peut comprendre et utiliser des expressions familières et quotidiennes ainsi que des énoncés très simples qui visent à satisfaire des besoins concrets.

En dehors de la revue de la littérature entreprise par Chapelle et Douglas (2006) qui concerne davantage la validité de construit, l'intégration d'un cadre de référence dans les dispositifs d'ÉAO pose de nouveaux problèmes méthodologiques relatifs à la validité critérielle des outils d'évaluation. Ces difficultés sont non seulement liées aux modèles psychométriques utilisés lors des évaluations, mais aussi au caractère multidimensionnel de la notion de compétence langagière. Le premier problème méthodologique concerne le caractère multidimensionnel de la compétence langagière notamment illustré par la définition qu'en donne le CECR. Or, les MRI actuellement utilisés en ÉAO supposent l'unidimensionnalité du trait latent mesuré. Comment dès lors prendre une décision sans appel de placement d'un sujet dans un groupe de niveau unique malgré le caractère multidimensionnel de la compétence langagière ? Autrement dit, comment déterminer le niveau global d'un apprenant, si celui-ci possède des niveaux différents selon les diverses facettes de la compétence évaluée ? Le second problème méthodologique concerne la fixation des seuils qui permettront de définir les niveaux de compétence sur le continuum du trait latent finalement mesuré et la correspondance de ces niveaux entre les divers cadres de référence.

À ces questions méthodologiques fondamentales, les cadres de référence existants répondent généralement par un manque de clarté évident, et les

concepteurs de dispositifs d'ÉAO sont généralement réduits à utiliser des règles empiriques ou des procédures locales. À titre d'exemple, Burton (2004) a étudié la validité d'un classement effectué au départ de tests adaptatifs assistés par ordinateur dans le cadre d'une procédure de placement des apprenants dans des groupes de compétences définis selon les quatre premiers niveaux du CECR. Dans cette perspective, le classement des apprenants réalisé au départ de tests adaptatifs en compréhension de l'oral et de l'écrit (voir la figure 4) pour une seconde langue a été comparé au classement des mêmes individus effectué initialement par des experts aguerris au CECR via des méthodes classiques d'évaluation (entretiens et tests papier-crayon). Les résultats montrent que 82 % des classements des apprenants en français, 83 % en allemand et 85 % en luxembourgeois correspondent à l'avis émis initialement par les enseignants. Dans cet exemple, les chercheurs ont utilisé un critère externe, l'avis des experts, comme point de référence. Les seuils entre chaque niveau ont été déterminés par analyse discriminante pour prédire au mieux le classement initial des experts.

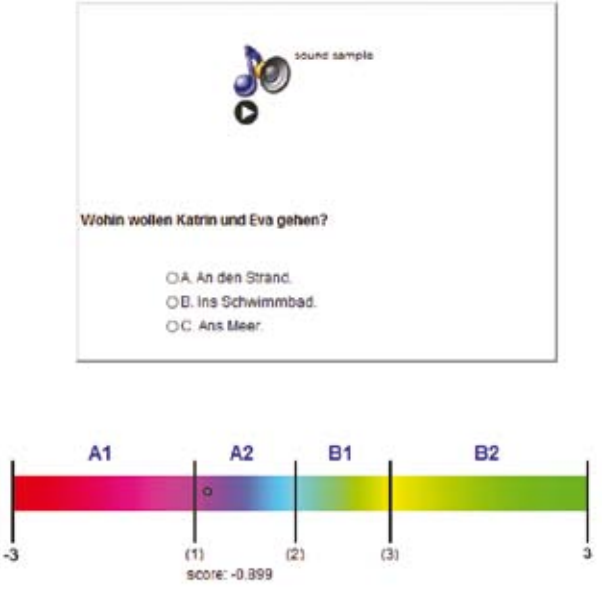


Figure 4 – Item de placement des apprenants selon les niveaux de compétence du CECR sur la plateforme TAO

Les résultats internes obtenus dans cette étude ne sauraient cependant masquer un positionnement arbitraire de l'instrument par rapport au CECR. Ces considérations prônent en faveur de l'élaboration d'items de références

afin de clarifier les procédures d'étalonnage. Cela permettra non seulement d'étalonner les dispositifs d'ÉAO mais aussi d'assurer la transition entre les différents cadres.

CONCLUSIONS ET PERSPECTIVES

L'évaluation des compétences langagières à l'aide des technologies de l'information et de la communication a pris un essor considérable ces dernières années. Depuis les premiers essais d'ÉAO en langues qui consistaient en de simples transpositions des tests papier-crayon sur support informatique, nous avons assisté à une évolution rapide des dispositifs selon un principe général d'interactivité accrue. Le premier stade du processus d'évolution concerne les formats d'items : d'abord statiques, ils deviennent dynamiques par l'intégration de multimédias. Les sujets peuvent alors interagir avec les tâches dont la validité écologique augmente. Il est possible dès lors de procéder à des évaluations de processus cognitifs complexes. La quête d'une interactivité accrue concerne ensuite le mode de passation des tests. Les processus de passation linéaire de tests laissent la place à des dispositifs dans lesquels la difficulté de la tâche s'adapte automatiquement à la compétence du sujet grâce aux développements récents des modèles de réponse aux items (MRI). Le procédé permet de réduire considérablement la longueur des tests tout en maintenant le degré de précision de l'estimation de la compétence du sujet. Le processus d'évolution concerne ensuite les modes de correction des items qui deviennent de plus en plus interactifs. Aujourd'hui, la correction automatique ne concerne plus uniquement les questions à choix multiple, mais aussi les questions ouvertes à réponse courte. Il est également possible de procéder à des analyses linguistiques de productions écrites complexes. Enfin, le principe d'interactivité touche également les architectures des plateformes informatiques elles-mêmes qui servent de support à l'évaluation. Même si ces plateformes étaient initialement « fermées », nous assistons à l'heure actuelle à l'émergence de systèmes d'auteur aux architectures distribuées qui favorisent l'interaction entre les concepteurs des évaluations. Il devient possible pour n'importe qui de construire un dispositif d'évaluation en collaboration avec d'autres personnes.

Au-delà d'une interactivité accrue, l'ÉAO permet également un traçage comportemental très fin à travers la récolte de données chronométriques et l'enregistrement des interactions comportementales. Il s'agit de recueillir des données sur l'activité des sujets soit pour en déduire des informations sur les apprentissages réalisés et établir des diagnostics, soit pour étudier les processus cognitifs invoqués et établir des profils comportementaux. La traçabilité ouvre ainsi la voie à des possibilités de nouvelles analyses typologiques (par exemple, profils comportementaux performants) ou stratégiques

(par exemple, analyses des stratégies de lecture). Cependant, force est de constater qu'à l'heure actuelle, peu de travaux ont été réalisés sur le sujet. La difficulté majeure à laquelle se heurtent les chercheurs n'est pas tant la récolte des traces, mais bien leur exploitation. Il existe pourtant des modèles théoriques qui mériteraient d'être exploités pour procéder au traitement psychométrique de ce type de données. Nous pensons notamment aux MRI avec temps de réponse (Verhelst, Verstralen, Jansen, 1997 ; Roskam, 1997) pour le traitement des données chronométriques et les réseaux neuronaux pour le traitement des interactions comportementales.

Références

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

Blais, J.-G., Dionne, É., et Raîche, G. (2005), « Défis et enjeux de l'évaluation des apprentissages à l'aide des technologies de l'information et de la communication », dans *Les Actes du 18e Colloque International de l'ADMÉÉ-Europe*, Reims, France, octobre.

Bunderson, V. C., Inouye, D. K. et Olsen, J. B. (1989), « The four generations of computerized educational measurement », in R. L. Linn (Ed.), *Educational measurement : Third edition*, New York, Macmillan.

Burton, R. et Martin, R. (2006), « L'évaluation assistée par ordinateur », dans G. Figari et L. Mottiez Lopez (Ed.), *Recherche sur l'évaluation en éducation*, Paris, L'Harmattan.

Burton, R. (2004), « L'évaluation des compétences langagières dans la formation professionnelle selon le Cadre Européen Commun de Référence », dans *Les Actes du 17e Colloque International de l'ADMÉÉ-Europe*, Lisbonne, septembre.

Burton, R. (2003), « Le système scolaire luxembourgeois est-il équitable ? », dans *Les Actes du 16e Colloque International de l'ADMÉÉ-Europe*, Liège, septembre.

Chanier, T. et Lamy, M.-N. (2007), « Glossaire français-anglais sur l'apprentissage des langues et les systèmes d'information et de communication », http ://alsic.u-strasbg.fr/Info/gloss.htm, Université de Franche-Comté.

Chapelle, C. A., et Douglas, D. (2006), *Assessing Language through Computer Technology,* Cambridge, Cambridge University Press.

Choi, I.-C., Kim, K. S. et Boo, J. (2003), « Comparability of a paper-based language test and a computer-based language test ». *Language Testing*, 20(3), p. 295-320.

Coniam, D. (2001), « The use of audio or video comprehension as an assessment instrument in the certification of English language teachers : A case study », *System*, 29, p. 1-14.

Demeuse, M., Artus, F., Casanova, D., Crendal, A., Desroches, F., et Renaud, F. (2007), « Évaluation et autoévaluation en français langue étrangère : approche empirique du Cadre Européen Commun de Référence pour les langues (CECR) du Conseil de l'Europe à travers le TEF et l'échelle d'autoévaluation DIALANG », dans *Les Actes du 19e Colloque International de l'ADMÉÉ-Europe*, Luxembourg, septembre.

Jadoul, R., Merche, J.F., Martin, R., et Latour, T. (2006), « CASCADE – a prospective formative testing tool », dans *Acts of the IV International Conference On Multimedia And Information And Communication Technologies Education (m-ICTE)*, Seville Spain, November.

Kelderman, H. (1997), « Loglinear Multidimensional Item Response Model for Polytomously Scored Items », dans W. J van der Linden et R. K. Hambleton (Ed.), *Handbook of Modern Item Response Theory*, New-York, Springer-Verlag.

Laveault, G., et Grégoire, J. (1997), *Introduction aux Théories des Tests en Sciences Humaines*, Bruxelles, De Boeck.

Martin, R., Latour, T., Burton, R., Busana, G., et Vandenabeele, L. (2005), « Une plate-forme collaborative et ouverte de testing assisté par ordinateur permettant l'élaboration et la passation d'instruments d'évaluation par voie d'Internet », dans *Les Actes du 18e Colloque International de l'ADMÉÉ-Europe*, Reims, France, octobre.

McDonald, R. P. (1997), « Normal-Ogive Multidimensional Model », dans W. J. van der Linden & R. K. Hambleton (Ed.), *Handbook of Modern Item Response Theory*, New-York, Springer-Verlag.

Reckase, M. D. (1997), « A Linear Logistic Multidimensional Model for Dichotomous Item Response Data », dans W. J. van der Linden et R. K. Hambleton (Ed.), *Handbook of Modern Item Response Theory*, New-York, Springer-Verlag.

Roskam, E. E. (1997), « Models for Speed and Time-Limit Tests », dans W. J. van der Linden et R. K. Hambleton (Ed.), *Handbook of Modern Item Response Theory*, New-York, Springer-Verlag.

Verhelst, N. D., Verstralen, H. H. F. M., et Jansen, M. G. H. (1997), « A Logistic Model for Time-Limit Tests », dans W. J. van der Linden et R. K. Hambleton (Ed.), *Handbook of Modern Item Response Theory*, New-York, Springer-Verlag.

Vygotsky, L. S. (1984), *Pensée et langage*, Paris, Editions Sociales.

Wainer, H. (2000), « Cats : Whither and whence », *Psicologica*, 21(1-2), p. 121-133.

CHAPITRE 5

L'apport des technologies dans l'évaluation de la compétence à écrire

Michel D. Laurier et Luc Diarra
Université de Montréal

INTRODUCTION

La correction en général est un travail ardu, et cette tâche est encore plus difficile lorsqu'il s'agit de performances complexes. Dans cette catégorie, les productions écrites en vue de faire une inférence sur la compétence à écrire trouvent une place de choix. En effet, l'élaboration d'une grille de correction pertinente ne va pas de soi, et son application n'est pas non plus facile vu la multitude des aspects à prendre en considération. Ainsi, on doit juger du contenu, du discours et de la langue, chacun de ces aspects pouvant se décomposer en critères particuliers. En outre, vu le temps que prend la correction d'une production écrite, l'enseignant est plus exposé à des facteurs externes : effets de fatigue/saturation, de contraste, d'ordre, de contamination, etc.

Dans ce contexte, l'apparition de la correction automatisée par ordinateur des productions écrites se présente comme une solution attrayante. Même si, du côté francophone, la recherche a peu investi dans ce domaine, du côté anglo-saxon, on trouve depuis quelques années plusieurs travaux en ce sens. Valenti *et al.*, (2003) ont dénombré une dizaine de logiciels de correction automatisée, produits d'expérimentation de la recherche dans le domaine ou purs produits commerciaux : *Project Essay Grade* (PEG), *Intelligent Essay Assessor* (IEA), *Educational Testing Service 1* (ETS I), *Electronic Essay Rater* (*E-Rater*), *C-Rater*, BETSY, *Intelligent Essay Marking System* (IEMS), SEAR, *Paperless School free text Marking Engine* (PS/ME) et *Automark*. Il y en a probablement plus à l'heure actuelle. Curieusement d'ailleurs, un des systèmes les plus connus dans la littérature, *IntelliMetric*,

créé en 1998, ne figure pas sur cette liste[1]. Le développement spectaculaire des programmes dans ce domaine traduit l'intérêt que revêt désormais la correction automatisée de production écrite.

Dans une vision naïve des choses, l'utilisation de la correction automatisée pourrait aller de soi : on utilise l'ordinateur pour accomplir d'innombrables tâches, pourquoi ne l'utiliserait-on pas alors pour corriger des copies ? Cependant, la question de l'utilisation de la correction automatisée est plus complexe. Comment fonctionne la correction automatisée ? Quel en est le degré de fiabilité et de validité ? Quels en sont les avantages et les limites ? Telles sont les questions qui préoccupent les spécialistes de l'évaluation, les enseignants et d'autres acteurs de l'éducation, et auxquelles nous voulons apporter une réponse dans le présent chapitre. Nous nous proposons de décrire le processus de la correction automatisée et d'examiner ses résultats à partir des expériences menées par certains chercheurs.

Une production écrite est une réponse de l'élève à une question qui exige de lui l'élaboration d'une réponse personnelle sous la forme d'un texte écrit en faisant appel à sa propre capacité à organiser et à défendre ses idées, à argumenter, à avoir un esprit critique, à faire preuve de créativité, etc. L'évaluation de ce type de production est, lorsqu'on préserve le caractère authentique de la situation d'écriture, un cas typique d'évaluation basée sur la performance au sens de Wiggins (1998). Elle est considérée comme la plus appropriée pour évaluer la compétence en écriture par opposition à l'évaluation par le biais d'items à réponses choisies (*indirect writing assessment*), où l'apport des technologies est plus évident – comme dans le cas des tests adaptatifs (Laurier, 2004).

L'usage de la correction automatisée dans l'évaluation des productions écrites suppose que celles-ci soient sous forme électronique, c'est-à-dire saisies : 1) en les retranscrivant ou en utilisant un logiciel spécial de reconnaissance d'écriture pour le cas où elles auraient été rédigées manuellement ; ou 2) en utilisant l'ordinateur pour les rédiger. En général, lorsqu'on parle de correction automatisée comme c'est le cas ici, on imagine aisément que c'est le second cas de figure qui est envisagé, c'est-à-dire un texte produit avec l'ordinateur. La question de la correction automatisée des productions écrites va donc de pair avec celle de l'utilisation du traitement de texte pour écrire (modalité informatisée) dans l'évaluation de la compétence à écrire. C'est pourquoi nous jugeons utile de faire une esquisse de la problématique de l'intégration de cette modalité dans l'évaluation de la compétence à écrire avant de nous pencher sur la correction automatisée proprement dite.

1. Cette absence peut probablement se justifier par la disponibilité tardive de publications sur ce système de correction automatisée.

L'ÉVALUATION DE LA COMPÉTENCE À ÉCRIRE EN MODE TIC

L'absence paradoxale du traitement de texte dans l'évaluation de la compétence à écrire

Dans l'évaluation de la compétence à écrire, doit-on prendre en compte la modalité TIC, autrement dit accepter la possibilité que l'élève rédige son texte à l'ordinateur ? Il y a de plus en plus d'arguments pour répondre affirmativement à cette question. En effet, l'ordinateur est devenu relativement accessible. Le multimédia s'est dorénavant imposé dans l'enseignement/apprentissage de la langue (présence accrue de logiciels éducatifs, de tests adaptatifs, de portfolios, etc.). L'enseignement assisté par ordinateur connaît un développement sans précédent. Qui plus est, dans un monde en mutation où la technologie occupe de plus en plus une place centrale, le constat est que l'acte d'écrire est désormais associé à l'ordinateur. La compétence en traitement de texte apparaît aujourd'hui comme une compétence valable, souhaitable et même le plus souvent indispensable à l'individu dans certaines sociétés pour la rédaction de n'importe quel texte, en lieu et place de l'usage du papier-crayon traditionnel. En d'autres termes, écrire avec l'ordinateur est maintenant une activité usuelle, et même incontournable, dans beaucoup de situations de la vie courante. Parallèlement, les nouveaux programmes de formation construits en fonction d'une approche par compétences font la promotion d'une évaluation revêtant un caractère authentique, c'est-à-dire ancrée dans des situations proches de la vie réelle. Il est donc paradoxal, et même anachronique, que les TIC soient mises de côté lorsqu'il s'agit de l'évaluation de la compétence à écrire. C'est pourquoi de plus en plus de travaux (Russell et Haney, 1997 ; Russell, 1999 ; Russell et Plati, 2000 ; Russell 2006) suggèrent que la modalité informatisée soit acceptée en plus de la modalité qui utilise le papier-crayon (modalité manuscrite).

Des raisons d'intégrer la modalité TIC dans l'évaluation de la compétence à écrire

Un des arguments en faveur de l'intégration des TIC dans l'évaluation de la compétence à écrire est que l'élève peut avoir développé d'autres habiletés pour écrire avec l'ordinateur dont on ne tient pas compte dans l'évaluation de sa compétence à écrire. Du même coup, l'inférence que l'on veut faire à partir de sa performance serait biaisée. Dans ce sens, les travaux de Russell et Haney (1997), de Russell (1999), ainsi que de Russell et Plati (2000) ont révélé que les élèves habitués à l'utilisation de l'ordinateur pour écrire, lorsqu'ils étaient soumis à des épreuves utilisant le traitement de texte, avaient des notes

meilleures que dans une évaluation de la modalité papier crayon. Les élèves concernés se trouveraient donc pénalisés dans le second cas.

La première étude sur cette question, celle de Russell et Haney (1997), a abouti à des résultats qui montraient que, lorsque les étudiants habitués au traitement de texte étaient contraints à la modalité manuscrite, seulement 30 % d'entre eux passaient au niveau supérieur. Par contre, 67 % des mêmes élèves passaient au niveau supérieur lorsque le traitement de texte était utilisé comme mode d'évaluation. On remarquera alors le paradoxe : dans le premier cas, une bonne partie des élèves se retrouvent pénalisés parce qu'ils ont probablement développé des stratégies d'écriture adaptées au mode TIC, mais malheureusement ces habiletés ne sont pas prises en compte. Pourtant, non seulement, l'objectif de l'évaluation est de mesurer la compétence à écrire mais, de surcroît, les habiletés en mode traitement de texte sont celles-là mêmes qui sont (ou qui seront) les plus sollicitées au quotidien dans les situations authentiques. C'est pourquoi, face à ce constat, Russell et Haney (2000) ont suggéré que la possibilité soit donnée au candidat de choisir la modalité d'évaluation qui lui convient le mieux – modalité informatisée et modalité manuscrite. La province de l'Alberta a déjà pris cette voie dans son système de sanction des études (*graduation testing program*) et aurait vu le taux des candidats optant pour la modalité informatisée croître de 6,7 % en 1996 à 24,5 % en 2000 (Russell et Tao, 2004).

Dans une deuxième étude, en comparant les performances entre les modalités d'évaluation manuscrite et informatisée, Russell (1999) a abouti à un résultat qui montre que les étudiants qui saisissaient environ 20 mots par minute étaient les plus nombreux parmi ceux qui réussissaient à passer de la 7e année à la 8e année.

Dans une troisième étude, Russell et Plati (2000) se sont intéressés aux résultats du *Massachusetts Comprehensive Assessment Systems Language Arts Tests* de 1999 pour montrer que le changement du mode de passation des items, selon que les élèves utilisaient la modalité informatisée ou la modalité manuscrite pour rédiger leurs productions écrites, pouvait avoir un impact important. Les résultats de l'étude ont permis d'estimer qu'en utilisant l'ordinateur, 19 % des étudiants de 4e année qui étaient classés au niveau « a besoin d'amélioration » passeraient au niveau de performance « compétent » et que 5 % des étudiants qui étaient classés au niveau « compétent » seraient admis au niveau « avancé ».

La problématique de la comparabilité entre modalités

À l'étape actuelle, l'intégration réussie des TIC dans l'évaluation de la compétence à écrire suppose que l'on connaisse suffisamment les effets de

cette modalité par rapport à la modalité traditionnelle manuscrite qui est la plus courante et qui prévaut toujours. On s'interroge, d'une part, sur la comparabilité des aspects liés au processus d'écriture mis en œuvre par l'élève et, d'autre part, sur des aspects relatifs à la comparabilité des résultats (notes attribuées aux réponses de l'élève).

La comparabilité au niveau des stratégies d'écriture

La problématique de la comparabilité entre les deux modalités nous amène à nous demander si les scripteurs mettent en œuvre les mêmes stratégies d'écriture dans les deux cas. En d'autres termes, le processus d'écriture est-il modifié ou non ? Malgré une abondante littérature sur le processus d'écriture mis en jeu par l'apprenant dans le cas de l'écriture manuelle, on ne connaît pas grand-chose de ce processus d'écriture lorsque l'apprenant utilise le traitement de texte. Certaines études ont éclairé l'un ou l'autre aspect du processus rédactionnel utilisé lors du traitement de texte pour montrer certaines de ses spécificités par rapport au processus d'écriture mis en œuvre en contexte de production manuelle.

Dans le processus d'écriture en contexte de traitement de texte, la planification[2], selon plusieurs auteurs, ne serait pas systématiquement facilitée (Piolat *et al.* ; 1993 ; Bangert-Drownz, 1993 ; Barker, 1987 ; Daiute, 1985). Cette thèse est étayée par l'impossibilité d'utiliser des flèches ou autres signes graphiques pour relier les idées, de faire des métacommentaires, etc. Cependant, il faut relever que ces études datent de plusieurs années, parfois même d'avant la généralisation des environnements graphiques. La technologie progresse rapidement, et la plupart des logiciels ont remédié à ces insuffisances. De même, les scripteurs se sont mieux familiarisés avec l'utilisation de logiciels avancés de traitement de texte.

À propos de la « textualisation », selon Coen (2000), le traitement de texte par le biais des fonctions de suppression, ajout et déplacement d'éléments facilite la réorganisation et la transformation du texte. Paradoxalement cependant, l'auteur affirme que la manière d'écrire à l'ordinateur est le plus souvent linéaire et, sous cet angle, tout à fait analogue à un environnement papier-crayon.

C'est au sujet de la révision qu'on trouve le plus de recherches. Les corrections apportées par les sujets à l'aide du traitement de texte seraient le plus souvent périphériques (Bridwell, Johnson et Brehe, 1987 ; Case, 1985 ; Harris, 1985). De plus, Piolat *et al.* (1987) cités par Coen (2000),

2. Nous pouvons distinguer la planification, la textualisation et la révision en tant qu'étapes du processus d'écriture (Flower, 1981).

en décrivant les stratégies de correction des scripteurs experts et novices, ont montré que les opérations de suppression, d'ajout et de permutation correspondaient à des niveaux de difficulté différents. Certains auteurs affirment que l'affichage à l'écran est moins efficace pour la lecture et relecture. Pour Coen (2000), cette thèse est appuyée par le fait que bon nombre de rédacteurs disent imprimer leur texte pour pouvoir en faire une lecture efficace. En outre, d'autres auteurs (Daiute, 1986 ; Kurth, 1987) mettent en doute l'efficacité de la révision sur la qualité du produit final lors du traitement de texte. Il convient toutefois de rester prudent au sujet de certaines de ces affirmations, car le degré d'habileté du rédacteur en matière de révision est un facteur déterminant qu'il faudrait prendre en compte. D'autres auteurs ne vont d'ailleurs pas dans ce sens : les corrections seraient, par exemple, plus judicieuses, notamment sur les plans syntaxique et stylistique (Nicolet *et al.,* 1987, cités par Coen, 2000).

L'examen des résultats de ces travaux laisse croire que le processus d'écriture utilisé lors du traitement de texte diffère de celui utilisé dans les conditions de la production écrite manuelle. La recherche nous apporte des renseignements utiles sur le processus rédactionnel mis en œuvre lors du traitement en se limitant à un ou à quelques aspects de ce processus. Mais l'ensemble du processus n'a pas été véritablement au centre des travaux.

La comparabilité des résultats

Une des questions essentielles à laquelle l'on doit répondre afin d'éviter les biais possibles est la suivante : les notes attribuées par les correcteurs humains sont-elles équivalentes selon qu'il s'agit d'une production manuscrite ou d'une production informatisée ?

Quelques études ont déjà été menées pour comparer les résultats obtenus selon les deux types de production. Même si leurs résultats divergent, certaines d'entre elles ont montré que les notes attribuées aux copies des étudiants varient en fonction de la modalité d'évaluation utilisée. Autrement dit, le comportement du correcteur n'est pas le même suivant qu'il a devant lui une copie rédigée à la main ou une copie rédigée à l'ordinateur par le biais du traitement de texte.

Manalo et Wolfe (2000) ont fait une analyse statistique des résultats au TOEFL[3], dont une nouvelle version donne au candidat le choix entre la modalité informatisée et la modalité manuscrite. Ils ont fait l'hypothèse que l'introduction des deux modalités peut entraîner un biais dans l'inférence

3. Le TOEFL est un test largement utilisé dans les universités pour évaluer la maîtrise de l'anglais comme langue seconde. Il est produit par le Educational Testing Service.

faite sur la compétence en écriture des candidats. L'analyse de variance effectuée a donné des résultats qui montrent qu'il y a des différences systématiques entre, d'une part, ceux qui optent pour l'ordinateur et, d'autre part, ceux qui optent pour la modalité manuscrite : les copies manuscrites reçoivent des notes plus élevées que les copies informatisées. Manalo et Wolfe (2000) ont estimé cet écart en faveur des copies manuscrites à 0,30 point sur une échelle de 6 points. Ils expliquent cette situation en recourant alors à une des hypothèses de Powers *et al.* (1994) selon laquelle les examinateurs ont des attentes plus élevées à l'égard des copies informatisées.

Bunch *et al.* (2006) se sont intéressés aux productions écrites de 205 élèves de 6e année, volontaires, mais cette fois-ci répartis en quatre groupes, chacun des quatre devant traiter deux sujets selon une combinaison libre avec les modalités : manuscrit/manuscrit, manuscrit/informatisé, informatisé/manuscrit et enfin informatisé/informatisé. Une fois rédigées, les copies ont été toutes saisies pour éliminer l'effet éventuel du mode de présentation. Elles ont été ensuite corrigées et notées. L'analyse de variance n'a pas décelé d'écarts significatifs suivant les sujets traités, ce qui est rassurant pour l'équivalence des sujets. Par contre, l'analyse montre que la modalité du test, selon qu'elle est manuscrite ou informatisée, a un effet relativement petit mais significatif sur les notes obtenues : les moyennes aux tests informatisés sont légèrement plus élevées que celles obtenues aux tests manuscrits. Puisque l'effet du mode de présentation a été éliminé, on peut déduire que les productions informatisées sont de meilleure qualité.

Le choix des modalités ayant été fait librement, les performances plus élevées pour les copies informatisées peuvent toutefois s'expliquer par le fait que ceux qui ont choisi d'utiliser l'ordinateur ont tout simplement de meilleures performances habituellement – par exemple, en bénéficiant de conditions socio-économiques propices à l'apprentissage, indépendamment de la modalité de passation de l'épreuve. Les auteurs, avant de clore leur travail, n'excluent d'ailleurs pas le recours à des facteurs externes pour expliquer les différences de performance observées entre ceux qui ont opté pour la rédaction manuelle et ceux qui ont choisi l'ordinateur.

La comparabilité : le cas spécifique du format de présentation

Dans l'étude de la comparabilité entre modalités informatisée et manuelle, certains auteurs se sont penchés sur le cas spécifique du format de présentation (Wolfe *et al.,* 1993 ; Powers *et al.,* 1994 ; Russell et Tao, 2004) pour voir si le correcteur humain attribue des notes équivalentes quand il a devant lui une copie imprimée à l'ordinateur.

Wolfe *et al.* (1993) ont mené deux expériences auprès de 157 élèves de 10e année pour comparer les évaluations de productions écrites en modalité manuscrite et informatisée. La deuxième expérience avait spécifiquement pour but de voir si les correcteurs notent de façon différente une copie manuscrite et une copie informatisée. Chacune des 157 copies, dont la moitié avait été rédigée en modalité manuscrite et l'autre en modalité informatisée, a été soumise à deux correcteurs. Avant la correction, on a saisi et imprimé les copies manuscrites en variant la police et la qualité de l'impression de l'ensemble des copies pour éviter de les différencier. De même, les copies rédigées à l'ordinateur ont été transcrites manuellement avec des variations d'écriture pour éviter qu'elles soient reconnues. Après la correction et la notation, l'examen des résultats montre que les copies originales, qu'elles soient rédigées à l'ordinateur ou qu'elles soient manuscrites, reçoivent des notes plus élevées que les copies retranscrites. En outre, les résultats montrent que les copies rédigées avec l'ordinateur reçoivent des notes plus élevées que les copies manuscrites.

On peut se demander si l'expérience, telle qu'elle a été conduite par Wolfe *et al.* (1993), n'a pas été entachée de quelques biais susceptibles d'influencer le jugement du correcteur au détriment des copies non originales reproduites. L'hypothèse de biais est possible dans la mesure où les auteurs affirment que, dans les copies reproduites par verbatim, l'on pouvait relever en moyenne deux mots mal orthographiés par rapport à l'original. En outre, dans le cas de la reproduction en manuscrit, l'écriture mal soignée peut avoir également introduit un biais difficile à contrôler. De ce fait, il faudrait prendre avec réserve les résultats de l'expérience de Wolfe *et al.* (1993) qui ont conclu que leurs résultats sont en contradiction avec ceux de Powers *et al.* (1994)[4], lesquels ont trouvé dans leur étude que les copies manuscrites recevaient des notes plus élevées, qu'elles soient originales ou retranscrites. Pour mesurer l'effet du mode de présentation, ils ont tout simplement mené des expériences sur des copies de productions écrites de 500 étudiants de 1re, 2e ou 3e année d'université qui ont rédigé chacun une copie manuscrite et une copie informatisée. À partir des résultats, la paire de copies de chacun des seize étudiants représentatifs du groupe dans l'épreuve manuscrite et celle de chacun de seize autres représentatifs dans l'épreuve informatisée ont été choisies pour la suite de l'étude. Les 64 copies ainsi concernées ont été doublées dans le mode de présentation inverse pour être corrigées par la suite selon les deux formats. Les correcteurs ne pouvaient donc pas différencier une copie originalement manuscrite ou originalement saisie. L'examen des résultats montre que, pour

4. Il convient de noter que Wolfe *et al.* (1993) disposaient déjà de la première version du travail de Powers *et al.* (1994) – puisqu'ils les citent.

les mêmes réponses, les copies manuscrites ont reçu de la part des correcteurs des notes plus élevées que celles produites avec le traitement de texte. Pour expliquer cette variation, les chercheurs ont émis plusieurs hypothèses, dont la suivante : les réponses produites manuellement apparaissent plus longues et, de ce fait, se présentent comme le résultat d'un plus grand effort[5].

Pour vérifier cette dernière hypothèse, Powers *et al.* (1994) ont mené une expérience complémentaire dans laquelle les textes produits à l'ordinateur étaient doublement espacés pour apparaître plus longs. En outre, de nouveaux correcteurs ont été retenus et ils ont été formés pour appliquer les mêmes critères aux copies manuscrites et aux copies produites avec le traitement de texte. Les résultats ont montré que l'effet du mode de présentation a diminué sans pour autant disparaître.

L'étude menée par Russell et Tao (2004) intervient comme une suite et une réponse à Powers *et al.* (1994) au sujet de l'effet du mode de présentation sur la note attribuée. En effet, dans le cadre d'une plus large étude qui examine l'effet de la modalité de l'épreuve d'écriture du *Massachusetts Comprehensive Assessment System* sur les résultats, Russell et Tao (2004), ont prélevé du corpus de 240 productions écrites de Russell et Plati (2000) un échantillon aléatoire de 60 copies de la huitième année et de la dixième année respectivement, et l'ensemble des 52 copies de la quatrième année. Dans la première expérience – celle qui nous intéresse ici –, ils ont saisi toutes les productions écrites manuscrites retenues (172 copies). Ils ont pris le soin de relire à plusieurs reprises la version imprimée afin que les mots et les phrases soient orthographiés exactement comme dans la copie originale manuscrite. Au bout du compte, chaque copie était disponible en trois formats : manuscrit, imprimé en espace simple, imprimé en espace double. Même si le groupe cible était différent, ils sont parvenus à la même conclusion que Powers *et al.* (1994), à savoir que les notes reçues par les copies manuscrites sont meilleures que celles des copies informatisées. Cependant, à l'inverse de l'hypothèse émise par Powers *et al.* (1994), Russell et Tao (2004) n'ont pas trouvé que l'espacement pour allonger le texte réduisait l'effet de la présentation.

Comme on le constate, les résultats divergent quelques fois, et la problématique de la comparabilité entre les deux modalités d'évaluation reste posée. En outre, les études recensées, dans leur quasi-totalité, ont porté sur les textes

5. Les autres hypothèses sont : 1) Les correcteurs s'attendent à un produit final entièrement édité et raffiné avec le traitement de texte et, par conséquent, ils ont des attentes plus élevées pour les textes réalisés par ce biais ; 2) Le texte écrit à la main fait que le correcteur se sent plus proche de celui qui écrit, ce qui permet d'identifier étroitement la voix individuelle de l'auteur comme un aspect important du texte écrit ; 3) Les correcteurs donnent aux réponses produites manuellement le bénéfice du doute quand ils voient des copies mal soignées ou difficiles à lire.

écrits en anglais. Des études de comparabilité sur l'évaluation de productions écrites pourraient permettre de confirmer ou infirmer l'une ou l'autre hypothèse et de tirer des conclusions d'une plus grande validité.

LES TIC ET LE TRAITEMENT DE LA RÉPONSE : LA CORRECTION AUTOMATISÉE DES PRODUCTIONS ÉCRITES

Il convient de rappeler que l'origine de la correction automatisée des productions écrites remonte aux années 60 (Page, 1966). L'examen de la littérature nous révèle que les recherches et les applications dans ce domaine connaissent actuellement un engouement spectaculaire qui a débuté à partir des années 90 du côté anglo-saxon. Jusque-là, aucune utilisation officielle et exclusive de cet outil lors d'un test à enjeux critiques n'a été rapportée. Cependant, l'épreuve d'écriture du GMAT (*Graduate Management Admission Test)*, un examen de haute importance pour l'admission aux grandes écoles de commerce internationales aux États-Unis, a franchi un pas important en proposant la correction automatisée avec le logiciel *E-Rate*r en association avec la correction humaine. Dans les lignes qui suivent, nous allons décrire la correction automatisée et montrer comment elle fonctionne avant d'en examiner la fiabilité et la validité dans l'évaluation de la compétence à écrire.

Le fonctionnement de la correction automatisée des productions écrites

La correction automatisée repose sur une modélisation du processus de correction sur la base de celle réalisée par un humain. Cette représentation formelle du processus de correction associe à des aspects d'une production écrite (unité lexicale, ponctuation, paragraphe, etc.) des valeurs variables dont la prise en compte, grâce à des procédures statistiques et au TALN (traitement automatique de langues naturelles), permet de générer un score correspondant à la note accordée par un humain à la copie concernée.

Préliminaires à partir de deux types d'évaluation

La description de l'activité de correction, telle qu'elle est explicitée par Dessus et Lemaire (2004) qui s'inspirent de certains travaux antérieurs, peut servir de point de départ pour comprendre le mécanisme de la correction automatisée. Cette description part du constat que l'activité de correction a beaucoup de similitudes avec l'activité de révision telle qu'elle est décrite

par Hayes (1998) : dans les deux cas, « le lecteur d'un texte à des fins de révision ou d'évaluation le lit pour le comprendre, pour détecter d'éventuels problèmes, pour proposer d'éventuelles alternatives » (Dessus et Lemaire, 2004 : 4). Aussi distingue-t-on la correction centrée sur le texte de celle centrée sur le contenu[6]. La correction centrée sur le texte est orientée vers des caractéristiques formelles du texte (calligraphie, orthographe, ponctuation, grammaire, longueur des paragraphes, etc.). La méthode qui consiste à évaluer les productions écrites en mesurant leur lisibilité relève de ce type de correction[7]. Le programme de correction automatisée PEG s'inscrit aussi dans cette option en appliquant des calculs statistiques sur des aspects de surface. Quant à la correction centrée sur le jugement d'expert, autrement dit le contenu, elle porte beaucoup plus sur des aspects liés au contenu traité. L'expert a recours soit au score holistique[8] soit au score du trait pertinent ou des traits multiples[9]. La logique de ce deuxième type d'évaluation va elle aussi donner naissance à un type de correction automatisée. La correction humaine et la correction automatisée sont donc similaires. Dessus et Lemaire (2004) soulignent d'ailleurs que la seconde est une activité de simulation de la première. Cependant, cela ne doit pas occulter la pluralité des approches et des techniques qui entrent en jeu dans la correction automatisée.

Aperçu typologique des programmes de correction automatisée

La correction automatisée doit son essor notamment aux progrès enregistrés en linguistique informatique, en traitement automatique des langues naturelles (*Natural language processing*)[10] et au développement des techniques de recherche d'information (*Information retrieval* et *Data mining*, Kukich 2000). Plusieurs systèmes existent actuellement et peuvent être classés selon

---

6. Outre l'évaluation centrée sur le texte et celle centrée sur le jugement d'expert, Dessus et Lemaire (2004) en distinguent une troisième, l'évaluation centrée sur le lecteur, qui fait référence au public prévu pour la production écrite. Mais ils la jugent peu pertinente dans le contexte scolaire où l'auditoire reste l'enseignant lui-même.
7. Ainsi, Dale et Chall (1948), en partant de deux critères, phrase longue (LP) et difficulté des mots (MD), en étaient arrivés à la formule : Score de lisibilité = (0,1579 x MD) + (0,0496 x LP) + 3,6365.
8. La distinction entre « score holistique » et « classement par impression générale » ne nous paraît pas pertinente.
9. Il s'agit dans ce cas d'une évaluation critériée où le correcteur applique un critère principal ou une liste de critères.
10. En partant de l'exemple de la liste des systèmes de correction automatisée établi par Valenti *et al.* (2003), nous pouvons relever que la plupart d'entre eux ont eu recours aux techniques du traitement des langues naturelles : *Intelligent Essay Assessor* (IEA), *Educational Testing Service I*, *Electronic Essay Rater* (*E-Rater*), *C-Rater*, SEAR, *Paperless School Free Text Marking Engine* et *Automark*.

trois critères, même s'il ne s'agit pas toujours de catégories étanches : 1°) la distinction entre les systèmes de correction automatisée centrée sur le texte et ceux où la correction est centrée sur le jugement d'expert ; 2°) le classement des logiciels de correction automatisée selon l'approche utilisée pour mesurer les caractéristiques du style ou celles du contenu ; et 3°) la distinction des programmes de correction automatisée selon les techniques utilisées.

À défaut de pouvoir décrire tous les programmes de correction automatisée qui existent, nous allons décrire quatre d'entre eux (PEG, IAE, *IntelliMetric*, et E-Rater), choisis pour leur représentativité relative quant aux approches et techniques utilisées (voir le tableau 1). Ces quatre logiciels sont les plus fréquents dans la littérature sur la correction automatisée. Ils sont d'ailleurs cités par Attali et Burstein (2006) comme les systèmes les plus importants.

Tableau 1 – Classification des quatre logiciels étudiés

	Aspects évalués	Procédés de mesure	Outils de mesure
PEG	Texte	Indicateurs indirects	Statistiques
IAE	Contenu	Indicateurs indirects	Algèbre / TALN
IntelliMetric	Contenu	Indicateurs réels	Statistiques (approche bayésienne) / TALN
E-Rater	Contenu	Indicateurs réels	Statistiques / TALN

Description d'exemples de programmes de correction automatisée

Un modèle pionnier : le Project Essay Grader *(PEG) de Page (1966)*

Avec une longueur d'avance, Page a inventé un système de correction automatisée, le *Project Essay Grader* (PEG). À partir d'un ensemble de copies corrigées par des enseignants, il a recensé des traits à partir desquels il pourrait prédire la note attribuée par l'enseignant. Ces traits se limitant à des aspects de surface d'un texte (lisibilité, syntaxe, etc.) sont mesurés par des indicateurs indirects qui sont, entre autres : la longueur moyenne des mots, la longueur du texte rédigé par l'élève à travers le nombre de mots, le nombre de points (ponctuation), le nombre de prépositions et le nombre de mots non usuels (n'appartenant pas au lexique courant).

Après avoir trouvé que chacun de ces traits de surface corrélait avec l'appréciation de la qualité d'un texte par un humain, Page a ajusté un modèle de régression multiple, pour prédire, à partir de ces traits présents dans un écrit, la note attribuée par l'enseignant à cette copie. Il est alors parvenu à déterminer une combinaison linéaire de ces traits qui pouvait prédire une note se rapprochant le plus de la note attribuée par l'enseignant. La formule ainsi trouvée pour prédire la note attribuée pouvait servir de base à la conception d'un programme informatique, le PEG.

La corrélation entre les notes attribuées automatiquement par le système de PEG et celles attribuées par l'humain était estimée à l'époque à 0,78, ce qui représentait un résultat encourageant. Cependant, la communauté des chercheurs et des praticiens de l'éducation a été très réticente à l'utilisation de cette découverte de Page. Une des raisons de l'enthousiasme modéré réservé au PEG à l'époque est sans doute le contexte de la découverte : dans les années 1960, l'ordinateur était certainement moins accessible, et on peut imaginer que son utilisation pour écrire n'était pas une pratique courante chez les étudiants. Les ordinateurs avaient d'ailleurs des capacités et des fonctionnalités relativement limitées. Tout cela fait dire à Page (2003) que des contraintes d'ordre pratique n'ont pas permis au PEG d'être favorablement accueilli. Cette explication plausible ne saurait toutefois occulter les critiques formulées contre ce système de correction automatisée : on a reproché à ce système, d'une part, d'utiliser des mesures indirectes de la compétence à écrire que les élèves mal intentionnés pourraient utiliser pour hausser leurs notes et, d'autre part, de ne pas prendre en compte des aspects importants de la qualité de l'écrit tels que le contenu, l'organisation et le style, ce qui empêchait de donner une rétroaction formative aux élèves pour améliorer leur écrit (Kukich, 2000).

Utiliser des mesures plus directes de la qualité de l'écrit, en d'autres termes le contenu sémantique du texte, restait donc un défi. Il y a eu un regain d'intérêt pour la correction automatisée à partir des années 1990, et les différents programmes informatiques de correction automatisée ont essayé d'aller plus loin en exploitant les outils du traitement automatique des langues naturelles.

*L'*Intelligent Essay Assessor *(IEA)*

L'*Intelligent Essay Assessor* (IEA) est un produit commercial issu de l'approche de l'analyse de la sémantique latente (LSA pour *Latent semantic analysis*), développée par Landauer et son équipe (Landauer *et al.*, 1998). Un principe de base régit le fonctionnement de la LSA : définir le sens d'un mot à partir de son contexte d'apparition. C'est d'ailleurs de cette façon que

le locuteur apprend à comprendre le sens de la plupart des mots d'une langue.

La LSA représente chaque mot par un vecteur dans un espace vectoriel qui a plusieurs dimensions, plusieurs centaines selon Dessus et Lemaire (2004). Cela revient donc à attribuer plusieurs centaines de valeurs numériques à chaque mot. Cette représentation permet à son tour de représenter un texte entier en additionnant tout simplement les vecteurs de l'ensemble des mots qui le composent. Comparer plusieurs textes revient alors à comparer les vecteurs qui les représentent respectivement. Un tel modèle de représentation sémantique peut être exploité dans la correction automatisée. C'est ce que fait justement IEA en comparant la copie de l'élève à des copies-types sélectionnées auparavant par un enseignant.

Au départ conçu pour évaluer un résumé, IEA s'est par la suite étendu à l'évaluation des productions écrites avec la même technique, c'est-à-dire grâce à la comparaison des textes par le biais des vecteurs pour leur attribuer un score en utilisant deux techniques (Foltz *et al.*, 1999, cités par Dessus et Lemaire (2004) :

- la méthode holistique, qui consiste à comparer la copie à une série d'autres notées au préalable par un jury. La note de la copie sera celle avec laquelle elle entretient la plus grande proximité ;
- la méthode « étalon-or » (*gold standard*), qui consiste à comparer la copie de l'élève avec une copie modèle qui peut être rédigée par l'enseignant.

Il faut noter qu'en dehors de l'aide qu'il apporte à l'enseignant dans sa correction des copies, le produit IEA peut être utile à l'étudiant grâce aux rétroactions qu'il donne instantanément.

IntelliMetric

Le programme de correction automatisée *IntelliMetric* (Elliot, 2003 ; Rudner *et al.*, 2006) est un produit de *Vantage Learning*. Ayant été retenu par le *Graduate Management Admission Council* (GMAC) pour être, à partir de janvier 2006, l'institution responsable du développement et de la correction du test *Graduate Management Admission Test* (GMAT), *American College Test* (ACT) a proposé l'utilisation de *IntelliMetric* en remplacement de *E-Rater*, utilisé depuis 1999 et produit par la firme *Educational Testing Service* (ETS).

IntelliMetric essaie d'imiter le processus de correction tel qu'il se déroule chez les humains et trouve ses bases théoriques dans le traitement cognitif (*cognitive processing*), dans la linguistique informatique (*computational linguistics*) et dans la classification (Elliot, 2003). Le système a recours à la

fois à des outils du TALN et de la statistique (approche bayésienne). *IntelliMetric* nécessite, pour un sujet donné, une préparation à partir de copies déjà corrigées et notées par des correcteurs humains. À l'issue de cette préparation, un modèle utilisant un nombre réduit de traits est retenu pour la correction.

La démarche comporte trois étapes :

1) Dans un premier temps, un ensemble de copies déjà évaluées et notées par des correcteurs humains est soumis au système. Celui-ci en dérive alors les caractéristiques correspondant à chaque point de l'échelle de notation. En d'autres termes, s'il s'agit d'une échelle de 6 points, *IntelliMetric* déterminera quelles sont les caractéristiques associées à chacun des scores possibles sur cette échelle. Le système analyse plus de 300 traits, sémantiques et syntaxiques, ainsi que des aspects du discours. Elliot (2003) classe ces traits en cinq catégories : a) Centration et cohésion ; b) Développement et élaboration ; c) Organisation et structure du discours ; d) Structure syntaxique des phrases ; e) Code linguistique (qualité de la langue)
2) Dans un deuxième temps, le modèle ainsi construit à partir de l'échantillon de copies déjà évaluées est testé sur un petit échantillon de copies dont les scores sont connus. Ce test permet de valider le modèle.
3) S'il est confirmé qu'il peut être généralisé, le modèle est alors dans un troisième temps appliqué pour corriger les copies à évaluer dont les scores ne sont pas encore connus.

IntelliMetric a été utilisé pour évaluer des textes écrits en anglais, en espagnol, en hébreu et en bahasa, et des fonctionnalités seraient disponibles pour la correction de textes en français, en allemand, en néerlandais, en portugais, en italien, en arabe et en japonais (Elliot, 2003, p. 72). Selon Eliot (2003), plus de 140 études ont été menées pour vérifier la validité du modèle.

Electronic Essay Rater *(E-Rater et E-Rater2)*

E-Rater

Le programme *E-Rater* (Burstein *et al.*, 1998, Powers *et al.* 2002) qui a donné naissance à *E-Rater V.2* (Attali et Burstein, 2006) est un système de correction automatisée, produit par l*e Educational Testing Service*, qui utilise des outils statistiques et des outils du TALN. Grâce à ces outils, *E-Rater* analyse automatiquement toutes les phrases d'une production écrite et en extrait des traits linguistiques pertinents pour prédire le score le plus proche

possible de celui d'un correcteur humain. Au total, plus d'une cinquantaine de traits ont été identifiés pour le fonctionnement de *E-Rater*. Selon le thème, le système déterminera un sous-ensemble de ces traits comme pertinent pour prédire un score correspondant à celui de la performance d'un correcteur humain.

E-Rater nécessite d'abord, sur un thème donné, un important échantillon de centaines, voire de milliers de copies déjà évaluées et notées par des correcteurs humains. Cet échantillon comporte des productions écrites variées correspondant à chaque score possible. L'analyse par les techniques du TALN et par la régression permet de déterminer les traits pertinents pour prédire le score humain considéré comme l'idéal à atteindre. Les plus pertinents dans la prédiction – leur nombre varie entre 8 et 12 selon Powers *et al.*, (2002a) – sont ensuite retenus pour former le modèle de base de la correction. Après cette phase de préparation du système, les copies peuvent être soumises pour correction. Pour évaluer une copie donnée, le système y cherche les similarités avec les copies déjà corrigées. Il est nécessaire dans *E-Rater* de préparer de nouveau le système pour la correction d'un autre thème. En outre, un des avantages de *E-Rater* est qu'il donne également une rétroaction.

De E-Rater *à* E-Rater2

E-Rater 2 n'est pas une simple version améliorée de *E-Rater.* Le processus de correction automatisée y est lui-même modifié pour l'utilisation directe des traits pertinents qui régissent un jugement d'expert. Dans la version antérieure et dans les autres systèmes de correction automatisée, une multitude de traits, variables selon le thème, est utilisée. Le système *E-Rater* initial avait déterminé une soixantaine de traits pertinents. De même, *IntelliMetric* utilise des centaines de traits, tout comme IEA utilise une multitude de vecteurs. Ce qui importe dans ces programmes est la prédiction d'un score correspondant à celui attribué par un correcteur humain.

Dans *E-Rater2*, le nombre de traits est fortement réduit. Attali et Burstein (2006) font état d'un noyau de dix traits :

1) Grammaire : accord sujet-verbe, etc. ;
2) Usage : erreur d'article, mot impropre, etc. ;
3) Orthographe d'usage : fautes d'orthographe ;
4) Style : répétition de mots, nombre de phrases courtes et de phrases longues, etc. ;
5) Organisation : plan, agencement des idées ;
6) Développement : arguments ; thèse, illustration, etc. ;

7) Vocabulaire : niveau de langue du vocabulaire utilisé ;
8) Longueur moyenne des mots ;
9) Vocabulaire courant ou usuel : similitude du vocabulaire de la copie avec le vocabulaire courant (appelé « cosinus maximal ») ;
10) Vocabulaire spécifique lié au thème : similitude du vocabulaire de la copie avec celui de la meilleure copie (appelé « corrélation cosinus »).

L'utilisation de plus de mesures directes par opposition à l'utilisation des indicateurs indirects était un des défis de la correction automatisée (Kukich, 2000). On peut dire que ce défi est relevé plus qu'auparavant dans *E-Rater2*. Les dix traits de *E-Rater 2* seraient pratiquement les mêmes que ceux que l'humain utilise pour apprécier la qualité d'une copie et lui attribuer une note (Attali et Burstein, 2006). Une des conséquences est que les deux types de correction se rapprocheraient et qu'on pourrait donc se baser sur l'accord entre le système et un seul correcteur humain (Yang *et al.*, 2002) comme critère de validité de la correction automatisée, comme le font Attali et Burstein (2006). Jusque-là, on ne pouvait que comparer l'accord entre deux correcteurs humains, d'une part, et celui entre un correcteur humain et un système de correction automatisée, d'autre part. Une deuxième conséquence est que ces traits utilisés par *E-Rater*2 ne changeront pas d'un thème à un autre, tout comme les critères d'appréciation d'une bonne copie par un correcteur humain restent les mêmes d'un thème à un autre.

Fiabilité et validité des systèmes de correction automatisée

La fiabilité et la validité restent une préoccupation primordiale dans l'utilisation de tout système de mesure et d'évaluation. La correction automatisée n'est pas en reste. Qu'est-ce qui prouve que les résultats de la correction automatisée sont fiables et peuvent servir à faire une inférence sur la compétence à écrire des élèves ?

Fiabilité

Dans la vie de tous les jours, une personne confrontée à une même situation à différentes reprises ne prend pas toujours la même décision. Ainsi, certaines fois, elle peut faire preuve d'inconstance dans ses jugements. Pour apprécier le degré d'inconstance, on étudiera ce qu'on appelle la fidélité intrajuge. De même, deux personnes qui portent un jugement sur une situation ne sont pas toujours d'accord. Cette situation ouvre donc la porte à l'étude de la fidélité interjuges

Dans la correction automatisée, la fidélité intrajuge ne pose pas de problème. L'ordinateur donne une appréciation constante. Seule la fidélité interjuges est pertinente comme preuve de fiabilité de la correction automatisée. À ce niveau, chaque système de correction automatisée a fait son apparition comme produit de recherche ou produit commercial accompagné d'un minimum de preuves pour soutenir sa fiabilité. Puisque les modèles de la correction automatisée sont conçus sur le modèle de la correction humaine, on peut justifier leur fiabilité à partir des corrélations qui existent entre les scores qu'ils attribuent et ceux qu'attribue un correcteur humain entraîné et pourvu d'instruments, en d'autres termes à partir de la relation des scores « correction automatisée-correcteur humain ». Ainsi, le PEG tire sa fiabilité de la forte corrélation qui existe entre les scores de son système et ceux de la correction humaine. Les dernières expériences de Page ont donné une corrélation de l'ordre de 0,87 entre les scores attribués par PEG et ceux attribués par l'humain (Valenti *et al.*, 2003). L'expérience menée par Page et Petersen (1995) sur 1 314 productions écrites du *Praxis* (un examen classique de la firme ETS) a même révélé que le système de correction automatisée de PEG pouvait servir à prédire les scores attribués par les humains, mieux que certains correcteurs humains entre eux. En ce qui concerne le système IEA, son expérimentation sur des copies du GMAT a donné un taux d'accord (*adjacent agreement*) estimé entre 85 % et 91 %. En l'expérimentant sur quinze questions de vérification, l'accord entre le correcteur humain et le *E-Rater* a donné des taux d'accord allant de 87 % à 94 %. Sur 750 000 textes écrits corrigés du GMAT, cet accord est passé à 97 % !

En ce qui concerne toujours *E-Rater*, l'expérience menée par Bridgeman (2004) et relatée par Monaghan et Bridgeman (2005) a même donné des résultats où l'accord entre des correcteurs humains et le système *E-Rater* était plus élevé que l'accord interjuges pour certains correcteurs humains entre eux (94,2 % contre 94,0 %).

La plupart des systèmes de correction automatisée ont donc des performances fiables lorsqu'on les compare à ceux des correcteurs humains dans le contexte d'évaluation à grande échelle où la correction se fait selon des normes ou des critères bien définis. Mais le plus difficile n'est-il pas de prouver leur validité ?

Validité

Plusieurs travaux ont été totalement ou partiellement consacrés à l'étude de la validité de la correction automatisée (Bennett et Bejar, 1998 ; Power *et al.*, 2002, Elliot, 2003 ; Keith, 2003 ; Yang *et al.*, 2002 ; Attali et Burstein, 2006). L'examen de la littérature fait par Yang *et al.* (2002) a permis à ces

auteurs, suivis par Attali et Burstein (2006), de distinguer trois approches dans les travaux sur la validité de la correction automatisée :

- Tout d'abord, en lien avec la démonstration de fiabilité qui vient d'être évoquée, la correction automatisée tire principalement sa validité de la forte corrélation entre les scores attribués par ce système et ceux attribués par la correction humaine. Cette première approche de validation consiste donc à démontrer une validité concurrente. Certains auteurs ont mis en évidence cette corrélation entre correction automatisée et correction humaine pour prouver la validité des systèmes de correction automatisée. Bennett et Bejar (1998) critiquent cette approche de validation, qui consiste à faire une comparaison avec la correction humaine qui, elle-même, n'est pas infaillible. De même, Attali et Burstein (2006) estiment que ce type d'argument est insuffisant pour mettre en évidence la validité de la correction automatisée.
- La deuxième approche pour valider la correction automatisée consiste à examiner les relations entre, d'une part, les résultats de la correction automatisée d'un test et, d'autre part, des mesures distinctes du même construit ou d'un construit similaire. Cela revient à comparer les résultats d'un test utilisant la correction automatisée et ceux d'autres mesures de la compétence à écrire des mêmes élèves. Ainsi, pour aller plus en profondeur dans la quête des fondements de la validité de la correction automatisée avec *E-Rater*, Powers *et al.* (2002) ont mené une expérience pour mesurer à la fois la relation qui existe entre chacune des deux corrections (humaine et *E-Rater*) et des variables indépendantes (*external variables*) de la compétence à écrire qui ne sont pas liées au test. Les résultats de l'expérience ont mis en évidence qu'il y a une corrélation entre ces variables externes et les résultats de la correction automatisée. Les auteurs ont cependant trouvé que les corrélations étaient plus faibles avec la correction automatisée du système *E-Rater* qu'avec la correction humaine. De plus, une autre étude utilisant la même approche a voulu comparer les résultats de la correction automatisée du système *IntelliMetric* avec ceux d'un autre test d'écriture utilisant des questions à choix multiples, ainsi qu'avec les jugements des enseignants sur les différents niveaux de compétence des élèves en écriture (Elliot, 2003). Cette étude a montré que les résultats issus de la correction utilisant *IntelliMetric* avaient une corrélation aussi bonne avec les résultats du test à choix multiple qu'avec les mesures humaines.
- Indépendamment des tests statistiques pour faire des comparaisons avec la correction humaine prise comme modèle valide, la correction

automatisée devrait pouvoir, de façon autonome, apporter la preuve de sa validité. Une telle perspective constitue la troisième approche qui voudrait que l'on se base sur la compréhension du processus de correction automatisée en soi. On comprend aisément pourquoi ce type d'approche avait été jusque-là mal exploré : ce qui importait dans les systèmes de correction automatisée était d'utiliser les traits pour prédire le score le plus proche possible du score attribué par l'humain. Les traits utilisés pouvaient varier d'un sujet à l'autre. Avec *E-Rater* 2 (*cf.* ci-dessus au2.1.3.4.2), la troisième approche est désormais possible, puisque ce système utilise un ensemble standard de traits. Qui plus est, ce modèle comporte des rubriques quasiment identiques à celles qui guident l'attribution des scores chez l'humain. Attali et Burstein (2005) ont appliqué ce modèle standard de dix traits du système *E-Rater 2* à 64 thèmes de différents programmes d'examen de la 6e à la 12e année (section argumentation du GMAT, section discussion du GMAT, TOEFL), soit à plus de 25 000 productions en tout. Les résultats montrent que chaque trait (grammaire, usage, style, organisation, etc.) a une corrélation pratiquement stable avec l'appréciation du correcteur humain, quel que soit l'examen. On en déduit que les mêmes critères servent de la même manière à apprécier une copie. En outre, les résultats de la régression pour prédire les scores du premier correcteur humain (H1) montrent que chaque critère de la grille standard a sensiblement le même poids dans la prédiction du score, quel que soit le programme. Les rubriques relatives au contenu du domaine spécifique traité ne pèsent pas beaucoup dans la balance (15 %), ce qui veut dire que le système *E-Rater 2* peut facilement apprécier la qualité d'une bonne copie, quel que soit le thème traité. On connaît désormais ce qui est évalué à l'image de l'utilisation d'une grille de correction humaine. On en déduit que le modèle de *E-Rater 2* est plus compréhensible et plus explicite quant au processus d'évaluation utilisé et quant à ce qui est évalué, ce qui est en soi une preuve de validité.

Même si la question de la validité reste longtemps encore un sujet de débat, force est de reconnaître que les démonstrations de cette validité ont beaucoup progressé. La correction automatisée apparaît donc de plus en plus comme une alternative viable avec ses avantages et ses limites.

LES AVANTAGES ET LES LIMITES DE LA CORRECTION AUTOMATISÉE DES PRODUCTIONS ÉCRITES

Les avantages et les limites de la correction automatisée

Avant de les expliciter, les éléments essentiels retenus comme avantages et limites de la correction automatisée sont présentés dans le tableau suivant :

Tableau 2 – Synthèse des avantages et des limites de la correction automatisée

Avantages	Gain en fiabilité (fidélité intrajuge)
	Gain en temps et en argent (une fois le système mis en place)
	Efficacité possible dans la détection de plagiat
Limites	Efficacité insuffisante pour la correction de copies d'élèves expérimentés et mal intentionnés (objection défensive)
	Rétroaction limitée pour l'amélioration de la compétence évaluée (objection de construit)
	Impossibilité de comprendre et d'apprécier la production comme un humain (objection humaniste)

Alors que la correction humaine peut être influencée par certains facteurs (fatigue/ saturation, humeurs du correcteur, effet du halo, moment de la correction, ordre des copies, etc.), la correction automatisée permet un gain en fiabilité en générant des scores sans reproche sur la fidélité intrajuge. Elle se révèlerait également efficace dans la détection des cas de plagiat (voir le cas avec le système IEA dans Williams, 2001). Cependant, l'avantage le plus considérable est certainement autant l'aspect financier que celui du temps consacré à cette activité. Reconnues comme étant plus appropriées que les réponses à des questions à choix multiples pour évaluer la compétence à écrire, les productions écrites exigent malheureusement pour leur correction beaucoup de temps et d'efforts (Burstein et Ward cités par Attali et Burstein, 2006). Initialement professeur d'anglais au début des années 60, Page (2003) rappelle que l'expérience du pensum de la correction vécue avait été à l'origine de son engagement dans des recherches qui ont abouti plus tard au PEG. L'historique sur le texte écrit de l'examen du GRE (*Graduate Record Examination*) aux États-Unis, relaté par Monaghan et Bridgeman (2005), en est également révélateur. Ces auteurs arrivent à la conclusion que la réduction du coût substantiel de cet examen a été une des raisons qui ont

motivé ETS à investir dans la recherche sur la correction automatisée. En somme, l'utilisation de la correction automatisée peut rendre une évaluation efficiente en en réduisant le coût – ce qui dégage des ressources à utiliser pour couvrir d'autres besoins de l'éducation.

Néanmoins, la correction automatisée a des limites. Page et Petersen (1995), suivi par Page (2003), ont distingué trois objections à l'égard de la correction automatisée. La première objection est liée au comportement du sujet. Elle avance que les systèmes de correction automatisés ne sont efficaces que pour les copies rédigées par des élèves de bonne foi. L'étude menée par Powers *et al.* (2002b) a effectivement révélé des insuffisances de la correction automatisée dans ce sens. Leur expérience a montré notamment que les scripteurs expérimentés peuvent dans une certaine mesure « tromper » le système pour obtenir des notes plus élevées que leurs copies ne méritent. Cette étude de Powers *et al.* (2002b) interpellait les promoteurs de la correction automatisée afin qu'ils améliorent leurs systèmes respectifs. La deuxième objection est liée au construit. Elle se fonde sur le principe que, même si elles ont une forte corrélation avec celles qui guident l'attribution de la note chez l'humain, les variables utilisées par la correction automatisée pour évaluer une copie ne sont pas ce qui est réellement le plus important dans l'évaluation des productions écrites. Une conséquence en est qu'une rétroaction, notamment un diagnostic formatif adressé à l'élève pour lui permettre d'améliorer sa compétence à écrire, ne peut être attendue de la correction automatisée. Malgré les progrès enregistrés, la correction automatisée a certainement encore des défis à relever dans le domaine de la rétroaction à l'élève ou à l'enseignant. La troisième objection est liée à une perspective humaniste. Elle consiste à refuser la correction automatisée parce qu'un tel système ne pourra jamais comprendre et apprécier un texte écrit comme le ferait un humain. Cette objection semble la plus tenace à l'encontre de la correction automatisée. Elle semble cependant tenir davantage de la validité apparente. Elle est toutefois importante, car l'acceptabilité sociale d'une épreuve en dépend.

Plus fondamentalement, on peut douter de la capacité de la correction automatisée à :

- détecter les erreurs de fond que font les scripteurs très compétents.
- porter un jugement aussi nuancé et averti sur le contenu que celui d'un expert.
- porter un jugement sur le style et l'organisation du contenu dans des situations non prévisibles (l'appréciation d'un texte pourrait être biaisée parce qu'il sort de l'ordinaire par rapport à l'échantillon de copies qui a servi à la préparation du système

- examiner la créativité qui est pourtant une des qualités importantes dans l'évaluation de la performance d'un scripteur.

On manque de connaissances sur l'applicabilité des systèmes de correction automatisée à des productions écrites rédigées dans des langues autres que l'anglais. On peut cependant émettre comme dernière limite une objection d'ordre pratique et économique passée sous silence par les différents auteurs, mais indéniable. En effet, la mise en place de ces systèmes est coûteuse et nécessite d'importantes ressources au départ. En outre, l'utilisation de la correction automatisée exige un nombre important de candidats, et elle ne sera économique que si ce nombre est beaucoup plus élevé que le minimum requis. Or, malheureusement, les effectifs et les ressources ne suffisent pas toujours. Ces considérations d'ordre pratique et économique peuvent en partie expliquer pourquoi la correction automatisée n'est pas utilisée jusqu'à maintenant en dehors des États-Unis, où le testing à grande échelle fait partie de la tradition éducative et où l'on peut mobiliser des ressources importantes pour le réaliser.

La complémentarité entre corrections automatisée et humaine comme solution

Les opposants à la correction automatisée continueront certainement à reprocher au système de ne pas pouvoir lire et comprendre une copie comme le fait un correcteur humain. Cet argument d'ordre humaniste selon lequel l'ordinateur n'a pas la capacité intrinsèque de l'humain pour distinguer une bonne copie d'une mauvaise copie peut persister pendant longtemps encore. Au vu des réticences liées à cet état d'esprit, et compte tenu de certaines inquiétudes bien fondées, il ne peut être envisagé dans un proche avenir de substituer exclusivement la correction automatisée à la correction humaine. Par contre, dans une vision de complémentarité, l'assistance que peuvent apporter les systèmes de correction automatisée au correcteur humain est considérable. Toutes les études scientifiques mettent en évidence une fiabilité accrue sur certains aspects. Les gains en argent et en temps sont certains. L'utilisation de la correction automatisée en complément à la correction humaine se présente donc comme une possibilité viable et très prometteuse, tout en étant un compromis intéressant pour lever les inquiétudes qui relèvent de l'objection humaniste et pour tenir compte des contraintes et des ressources. Toutefois, dans cette vision de complémentarité, deux approches paraissent possibles : soit utiliser la correction automatisée comme l'un des deux correcteurs (cas de l'application de la correction automatisée au GMAT), soit utiliser la correction automatisée comme instrument d'appui à la correction humaine.

Dans la seconde option, l'utilisation de la correction automatisée vient en appui pour vérifier la correction humaine. Il s'agit d'une possibilité originale suggérée par Monaghan et Bridgeman (2005) dans le but d'accroître la fiabilité de l'évaluation tout en réduisant le coût. Dans le cadre du GRE, ces auteurs ont mené des expériences sur les productions écrites de 5950 candidats qui ont traité les deux thèmes du test d'écriture de la section *Analytical Writing* du GRE. Leur étude a montré que la combinaison correcteur humain et correction automatisée avec *E-Rater* donnait une fidélité élevée. En effet, la fidélité, lorsqu'une correction humaine est vérifiée par *E-Rater* est comparable à l'accord entre deux correcteurs humains. L'originalité de l'étude de Monaghan et Bridgeman (2005) réside surtout dans l'utilisation du système *E-Rater* comme vérificateur à côté d'un correcteur humain. Cette option permet d'aboutir à des résultats aussi fiables que dans le cas de deux correcteurs humains, tout en permettant un gain d'argent et de temps. De plus, elle paraît acceptable par ceux qui s'opposent à l'utilisation de la machine comme correcteur.

Avant que l'utilisation de la correction automatisée ne soit suggérée, une autre perspective de complémentarité, qui consiste à remplacer tout simplement un des correcteurs humains par la correction automatisée, était déjà illustrée. Il s'agit de l'exemple du GMAT déjà mentionné. En effet pour l'évaluation des productions écrites de cet examen, la tâche d'un des deux correcteurs humains était assurée depuis 1999 par la correction automatisée utilisant *E-Rater*. Correction humaine et correction automatisée étaient ainsi utilisées en complémentarité. Les deux notes étaient additionnées et divisées comme on a coutume de le faire lorsque les résultats ont un effet sur l'avenir des candidats. En cas de désaccord important entre les deux correcteurs, c'est-à-dire lorsque les deux notes distinctement attribuées étaient très différentes, on avait recours à un autre correcteur humain comme cela se passe habituellement dans les situations de haute importance pour les candidats. *E-Rater* a été remplacé en janvier 2006 par *IntelliMetric.*

Une autre forme de complémentarité est aussi possible. Il s'agit de celle où l'ordinateur serait utilisé pour accomplir certaines tâches qu'il peut mieux faire que le correcteur humain dans l'évaluation des productions écrites tandis que ce dernier s'occuperait des aspects du contenu sollicitant la capacité intrinsèque d'un humain pour porter un jugement d'expert. Par exemple, jusqu'à preuve du contraire, l'humain est celui qui peut mieux évaluer les aspects de fond tels que l'originalité et la créativité d'un écrit. Par contre, la correction des aspects du texte liés à l'usage de la langue (orthographe, syntaxe, etc), tâche routinière, ennuyeuse et mécanique est parfois plus difficile pour l'humain qui peut se tromper par moments sous l'effet de l'inattention, de la fatigue, etc. L'ordinateur pourrait avec une extrême rapidité assurer plus efficacement de telles tâches. Curieusement cependant,

cette forme de complémentarité n'est envisagée par aucun auteur, ni dans aucun des programmes de correction automatisée disponibles sur le marché.

CONCLUSION

Le présent chapitre aura permis de montrer les potentialités de la mise à contribution des technologies dans l'évaluation de la compétence à écrire. Tout en embrassant la problématique liée à la prise en compte de la modalité informatisée dans la production de la réponse de l'élève, cette esquisse aura surtout mis en exergue les possibilités que l'ordinateur offre à la phase du traitement de cette réponse. En effet, nonobstant le défi préalable qu'il reste à relever, celui d'accepter et de réussir la prise en compte de l'acte d'écrire avec l'ordinateur dans l'évaluation de la compétence à écrire, les outils de correction automatisée sont déjà disponibles. L'examen des résultats des différentes expériences menées notamment par des chercheurs américains montre l'efficacité de ces outils quant à l'assistance qu'ils peuvent apporter dans la correction des productions écrites.

Cependant, la présente esquisse aura également montré que cette technologie a des limites, et que des inquiétudes subsistent. Par conséquent, d'autres études doivent être menées non seulement en vue de fournir des renseignements utiles pour accroître la validité et la fiabilité de la correction automatisée, mais aussi pour affiner les connaissances sur la faisabilité de l'évaluation de la compétence à écrire à l'aide d'outils technologiques. Dans cette perspective, pour la communauté scientifique et pour les différents acteurs de l'éducation, et particulièrement pour la Francophonie, ce serait une œuvre très utile que les investigations sur l'exploitation de ces technologies puissent être poursuivies et élargies à des applications sur l'évaluation de la compétence à écrire dans d'autres langues, telles que le français.

Références

Attali, Y. et Burstein, J. (2006), « Automated Essay Scoring With e-rater® V.2 », *Journal of Technology, Learning, and Assessment,* 4(3), disponible à l'adresse URL [http://www.jtla.org].

Bangert-Drowns, R. L. (1993), « The word processor as an instructional tool : a meta-analysis of word processing in writing instruction », *Review of Educational Research*, 63(l), p. 69-93.

Barker, T. T. (1987), « Studies in word processing and writing », *Computer in the school,* 4, p. 109-121.

Bennett, R. E. et Bejar, I. I. (1998), « Validity and Automated Scoring : It's Not Only the Scoring », *Educational Measurement : Issues and Practice*, 17(4), p. 9–17.

Bridwell, L. S., Johnson, P., et Brehe, S. (1987), « Computers and composing : case studies of experienced writers », dans Matsuhashi, A. (Éd.), *Writing in real time modeling production processes,* New York, Logman

Bunch, M. B., Ridge, K., et Burke, J. (2006), « Does mode of composition (paper versus online) have an effect on sixth-grade writing performance ? », Communication présentée lors du congrès annuel de *l'American Educational Research Association*, San Francisco, avril.

Case, D. (1985), « Processing professiorial words : personnal computers and the writing habits of university professors », *College Composition and Communication*, 36, p. 317-322.

Coen, P. F. (2000), A quoi pensent les enfants quand ils écrivent *? Analyse des Processus cognitifs et métacognitifs en jeu lors d'une tâche d'écriture*, Thèse de doctorat inédite, Université de Fribourg.

Daiute, C. (1985), *Writing and computers*, Reading, MA, Addison-Wesley Publishing Co.

Dessus, P. et Lemaire, B. (2004), « Assistance informatique à la correction de copies », dans Gentaz E. et Dessus P. (Eds.), *Comprendre les apprentissages, Sciences cognitives et éducation*, Paris, Dunod.

Elliot, S. (2003), « IntelliMetric : From here to validity », dans Shermis, M. D. et Burstein, J., *Automated Essay Scoring : A Cross-Disciplinary Perspective*, Hillsdale, N.J., Lawrence Erlbaum Associates.

Harris, J. (1985), « Student writers and word processing : a preliminary evaluation », *College Composition and Communication, 36*, p. 323-330.

Hayes, J.R., et Flower, L. (1980), « Identifying the organization of writing processes », dans Gregg, L.W et E.R. Steinberg (Ed.), *Cognitive processes in writing*, p. 3-30.

Hayes, J. R. (1998), « Un nouveau cadre pour intégrer cognition et affect dans la rédaction », dans Piolat, A. et Pélissier, A. (Ed.), *La rédaction de textes, approche cognitive,* Lausanne, Delachaux et Niestlé.

Keith, T. Z. (2003), « Validity and automated Essay scoring Systems », dans Shermis, M. D. et Burstein, J., *Automated Essay Scoring : A Cross-Disciplinary Perspective*, Hillsdale, N.J., Lawrence Erlbaum Associates.

Kukich, K. (2000), « Beyond Automated Essay Scoring », *IEEE Intelligent Systems, 15*(5), p. 22-27.

Kurth, R. J. (1987), « Using word processing to enhance revision strategies during student writing activities », *Educational Technology, 27*, p. 13-19.

Landauer, T. K., Foltz, P. W., et Laham, D. (1998), « Introduction to latent semantic analysis », *Discourse Processes, 25, p.* 259–284.

Landauer, T., Laham, D. et Foltz, P. (2003), « Automated scoring and annotation of essays with the Intelligent Essay Assessor », dans Shermis, M. D. et Burstein, J., *Automated Essay Scoring : A Cross-Disciplinary Perspective*, Hillsdale, N.J., Lawrence Erlbaum Associates.

Laurier, M.D. (2004), « Évaluation et multimédia dans l'apprentissage d'une L2 », *Recall*, 16(2), p. 475-487.

Manalo, J. R. et Wolfe, E. W. (2000), « A comparison of word-processed and handwritten essays written for the Test of English as a Foreign Language », Communication présentée lors du congrès annuel de l'American Educational Research Association, Nouvelle-Orléans, États-Unis.

Monaghan, W. et Bridgeman, B. (2005), « E-rater as a quality control on human scorers » dans R et D Connections, Princeton, NJ : Educational Testing Services, disponible à l'adresse URL [http ://www.ets.org/Media/Research/pdf/RD_Connections2.pdf].

Page, E. B. (1994), « Computer grading of student prose, using modern concepts and software », *Journal of Experimental Education, 62,* p. 127–142.

Page, E. B. (1996), « The imminence of grading essays by computer », *Phi Delta Kappan, 48,* p. 238–243.

Page, E. B. et Petersen, N. S. (1995) « The Computer Moves into Essay Grading : Updating the Ancient Test », *Phi Delta Kappan*, 76, p. 561–565.

Page, E, B. (2003), « Project Essay grade : PEG », dans Shermis, M. D. et Burstein, J.. *Automated Essay Scoring : A Cross-Disciplinary Perspective*, Hillsdale, N.J., Lawrence Erlbaum Associates.

Piolat, A., Isnard, N., et Della Valle, V. (1993), « Traitement de texte et stratégies rédactionnelles », *Le Travail Humain, 56*, p. 79-99.

Powers, D., Fowles, M, Farnum, M. et Ramsey, P. (1994), « Will they think less of my handwritten essay if others word process theirs ? Effects on essay scores of intermingling handwritten and word-processed essays », *Journal of Educational Measurement*, 31 (3), p. 220-223.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E. et Kukich, K. (2002a), « Stumping e-rater : challenging the validity of automated essay scoring », *Computers in Human Behavior*, 18, p. 103-134.

Powers, D.,E., Burstein, J. C., Chodorow, M., Fowles, M.E, et Kukich, K. (2002b), « Comparing the validity of automated and human scoring of essays », *Journal of educational computing research*, 26(4), p. 407-425.

Rudner, L. M., Garcia, V. et Welch, C. (2006), « An Evaluation of the IntelliMetric Essay Scoring System », *Journal of Technology, Learning, and Assessment, 4*(4), disponible à l'adresse URL http ://www.jtla.org].

Russell, M. (1999), « Testing writing on computers : A follow-up study comparing performance on computer and on paper », *Educational policy analysis Archives*, 7 (20), disponible à l'adresse URL [http :epaa.asu.edu/epaa/v7n20].

Russell, M. (2006), *Technology and assessment. The tale of two interpretations*, Geenwich, Information Age publishing.

Russell, M. et Haney, W. (1997), « Testing writing on computers : an experiment comparing student performance on tests conducted via computer and via paper-and-pencil », *Educational policy analysis Archives*, 5 (3), disponible à l'adresse URL [http ://olam.ed.asu.edu/epaa/v %n/.html].

Russell, M. et Haney, W., (2000), « Bridging the Gap Between Testing and Technology in Schools », *Education Policy Analysis Archives*, 8(19), disponible à l'adresse URL [http ://epaa.asu.edu/epaa/v8n19.html].

Russell M. et Plati T.(2000), « Mode of administration effets on MCAS Composition performance for Grades Four, Eight and Ten », Rapport soumis au Massachusetts Departement of education par le National Board on Educational Testing and public policy, disponible à l'adresse URL [http//nbetpp.bc.edu/reports.html].

Russell M. et TAO, W. (2004), « Effects of handwriting and computer-print on composition scores : A follow-up to Powers, Fowles, Farnum & Ramsey », *Pratical Assessment, Research and Evaluation*, 9 (1), [htt ://PAREonline.net/getvn.asp ?v=9&n=1].

Shermis, M. D. et Burstein, J. (2003), *Automated Essay Scoring : A Cross-Disciplinary Perspective*, Hillsdale, N.J., Lawrence Erlbaum Associates.

Valenti, S., Neri F., et Cucchiarelli, A. (2003), « An Overview of Current Research on Automated Essay Grading » *Journal of Information Technology Education* Volume 2, disponible à l'adresse URL [http ://jite.org/documents/Vol2/v2p319-330-30.pdf].

Wiggins, G. (1998), *Educative Assessment : Designing Assessments to Inform and Improve Student Performance* », San Francisco, CA, Jossey-Bass.

Williams, R. (2001), « Automated essay grading : An evaluation of four conceptual models », dans Herrmann, A. and Kulski, M. M. (Ed.), *Expanding Horizons in Teaching and Learning,* Proceedings of the 10th Annual Teaching Learning Forum, 7-9 February, Perth : Curtin University of Technology. Disponible à l'adresse URL [http ://lsn.curtin.edu.au/tlf/tlf2001/williams.html]

Wolfe, E. W. Bolton, S., Feltovich, B., et Welch, C. (1993), « A comparison of word-processed and handwritten form a standardized writing assessment ». ACT Research 93-8.

Yang, Y., Buckendahl, C. W., Jusziewicz, P. J. et Bhola, D. S. (2002), « A Review of Strategies for Validating Computer-Automated Scoring », *Applied measurement in education*, 15(4), p. 391-412.

CHAPITRE 6

L'évaluation des apprentissages en mathématiques – Révéler l'invisible

France Caron et Nathalie Loye
Université de Montréal

INTRODUCTION

Ce chapitre procède d'une réflexion sur l'évolution récente des pratiques d'évaluation des apprentissages en mathématiques. Il met en évidence les enjeux et les difficultés liées à l'intégration de tâches complexes et à l'observation des traces laissées par les élèves lors de leur utilisation de divers outils technologiques dans la résolution de telles tâches. Cette réflexion aboutit à une discussion sur la possibilité d'utiliser, comme outils d'évaluation complémentaires en cours d'apprentissage, des épreuves courtes qui intègreraient l'outil technologique et dont l'interprétation des résultats s'appuierait sur une modélisation des processus cognitifs impliqués. L'objectif serait de fournir un diagnostic pour chaque élève afin de réguler ses apprentissages en mathématiques et d'apporter de l'aide aux enseignants dans l'évaluation des tâches plus complexes. Les apports, limites et difficultés d'une telle approche sont également présentés.

DES PRATIQUES ÉVALUATIVES EN ÉVOLUTION

Depuis 2005, le Québec vit une réforme du programme au secondaire. En s'inscrivant dans une approche par compétences, ce « renouveau pédagogique » redéfinit les enjeux de l'apprentissage, ainsi que les connaissances et les pratiques que l'élève aura à développer. Le paradigme d'apprentissage remplace le paradigme d'enseignement et amène les enseignants à revoir non seulement leurs méthodes pédagogiques, mais aussi leurs méthodes d'évaluation.

Cette réforme n'implique pas uniquement d'utiliser de nouveaux outils d'évaluation, mais également de reconsidérer les enjeux de l'évaluation. Même si cette dernière garde sa fonction certificative, elle doit de plus en plus être utilisée en salle de classe comme levier pour aider les élèves à apprendre. Pour assurer que les buts généraux poursuivis et formulés en termes de compétences soient bien au cœur des apprentissages des élèves, on préconise d'utiliser régulièrement des situations complexes qui permettront aux compétences visées d'être sollicitées. Cet aspect du renouveau pédagogique implique des changements importants dans les pratiques évaluatives, même si le retour au bulletin chiffré qui se met en place actuellement peut sembler remettre partiellement en question certains de ces changements.

Du côté des mathématiques à l'école secondaire, il est frappant de constater que les compétences visées par le nouveau programme (Gouvernement du Québec, 2003 et 2007), c'est-à-dire « Résoudre une situation problème », « Déployer un raisonnement mathématique » et « Communiquer à l'aide du langage mathématique », se trouvaient déjà dans les quatre « objectifs globaux » visés par les programmes des années 90 : « Établir des liens », « Gérer une situation problème », « Communiquer » et « Raisonner ». Toutefois, l'évaluation des années 90 avait bien souvent tendance à laisser de coté ces objectifs globaux au profit d'objectifs terminaux et intermédiaires, qui relevaient plus clairement d'un sous-ensemble bien circonscrit de concepts et processus, et dont la maîtrise est ainsi plus facile à déterminer. Cette réalité constitue un legs des programmes par objectifs qui ont connu leur apogée au Québec dans les années 80. Le but alors visé consistait à faire disparaître le flou et la subjectivité du processus d'évaluation en décomposant et en précisant au maximum les diverses manifestations des connaissances :

Dans le contexte scolaire, évaluer l'apprentissage, c'est essentiellement porter un jugement en comparant ce qui est arrivé à ce qui avait été prévu. Plus ce qui est prévu est exprimé clairement et plus ce qui est arrivé est analysé formellement, alors mieux on sera éclairé sur la réalité et sur la qualité de l'apprentissage.[1]

Dans le cadre de ce programme de mathématiques, « ce qui est prévu » est exprimé à l'aide des objectifs généraux, terminaux et intermédiaires. Les objectifs terminaux et intermédiaires étant les plus spécifiques, ce sont eux qui feront l'objet de mesure de la part de l'enseignant afin de lui permettre d'effectuer une saine évaluation pédagogique. Il apparaît alors important d'intégrer l'évaluation pédagogique aux apprentissages de l'élève afin de

1. M.E.Q., *Politique générale d'évaluation pédagogique*, septembre 1981, nº 16-7500, p. 6.

favoriser un enseignement correctif adapté à celui-ci (Gouvernement du Québec, 1984, p.33).

L'approche par objectifs des années 80 définissait donc les apprentissages en fonction de ce qui serait facilement observable et évaluable de façon objective, comme en témoignent les verbes choisis pour définir les objectifs intermédiaires en mathématiques : construire, énumérer, identifier, calculer, définir, trouver, appliquer, énoncer, etc. Si plusieurs de ces objectifs se prêtaient sans difficulté à des épreuves simples comme des questionnaires à choix multiples (QCM), ils ne rendaient pas compte de l'inévitable complexité des disciplines enseignées et des situations dans lesquelles les savoirs disciplinaires peuvent être mis à contribution.

En mathématiques, la définition du programme en termes de comportements observables associés à des éléments spécifiques du contenu pouvait conduire à une compartimentation des connaissances. Cette dernière ne laissait de place ni à la structure du savoir mathématique ni à l'importance pour l'élève de développer une capacité à mettre à contribution différents concepts et processus pour résoudre un problème un peu plus complexe. C'est ce qui explique que les programmes de mathématique aient été les premiers à être revus au début des années 90. Tout en augmentant le contenu couvert, ces nouveaux programmes réduisaient de moitié le nombre d'objectifs intermédiaires. Ils les rendaient plus englobants, avec un niveau de complexité plus élevé, visible dans le choix des verbes utilisés : déduire, appuyer, justifier, distinguer, décrire, déterminer et même démontrer. En principe, l'évaluation devait s'ajuster en conséquence. Tout en reconnaissant, à l'instar du Conseil supérieur de l'éducation (1992, p.1), que « la réflexion et la pratique en matière d'évaluation » avaient connu « un essor considérable dans le système scolaire québécois au cours de la dernière décennie », le ministère de l'Éducation visait avec ces nouveaux programmes à « atteindre une plus grande cohérence entre l'esprit des programmes d'études et les pratiques d'évaluation » (Gouvernement du Québec, 1994, p.16-17) :

L'apprentissage dans le présent programme est plus que l'acquisition de connaissances. C'est plutôt l'investigation, la communication, la représentation, le raisonnement et l'utilisation d'une variété d'approches pour résoudre un problème. C'est également l'acquisition d'autres habiletés et attitudes.

Ce que l'on voudrait évaluer, c'est le savoir, le savoir-faire et le savoir-être de l'élève, objets plus ou moins en mouvement. Il faut donc créer des situations permettant de recueillir des éléments d'information qui, après interprétation critériée ou normative, puissent révéler un portrait fiable à propos du savoir et du savoir-faire personnels ou collectifs des élèves.

Le programme des années 90 proposait donc déjà pour évaluer les élèves d'utiliser des journaux de bord, des présentations orales, des jeux-questionnaires, des entrevues, des travaux d'équipe, des épreuves-synthèses, des grilles d'observation, des tests ou des observations durant l'enseignement assisté par ordinateur, etc. Cette variété des moyens devait refléter la variété des activités d'apprentissage : manipulation, découverte, communication (orale ou écrite, individuelle ou en groupe), utilisation d'une calculatrice ou d'un ordinateur, etc.

Dans les faits, à part l'ajout, aux QCM des épreuves ministérielles, de quelques problèmes à résoudre pour lesquels on évaluait l'ensemble de la démarche, force est de constater que les pratiques évaluatives développées avec l'approche par objectifs ont continué à dominer dans bien des établissements. La difficulté d'évaluer de façon objective la performance à des épreuves plus complexes, associée à une certaine perception du travail mathématique (la réponse est soit juste, soit fausse, et il n'y en a qu'une seule possible), a semblé faire obstacle au déploiement de telles épreuves. Il apparaît ainsi que, dans l'évaluation des apprentissages en mathématiques, les notions de complexité et d'objectivité ont quelques difficultés à cohabiter.

Du côté de la technologie, si la calculatrice (scientifique ou graphique) est admise depuis plusieurs années à bien des épreuves d'évaluation, celles-ci n'ont pas été conçues pour refléter la complexité des situations auxquelles de tels outils peuvent donner accès et qui s'approchent de la pratique mathématique actuelle. En effet, sans doute par souci d'équité envers ceux qui ne disposeraient pas de l'outil, les tâches pour lesquelles son usage est permis peuvent en général être réalisées sans y recourir. Dans de telles conditions, il devient légitime de s'interroger sur la pertinence de l'intégration des outils technologiques dans les épreuves évaluatives.

LES PRATIQUES ÉVALUATIVES SELON LE RENOUVEAU PÉDAGOGIQUE

Avec le renouveau pédagogique, l'approche par compétences valorise la complexité des apprentissages et préconise l'utilisation de situations complexes et authentiques autant pour l'apprentissage que pour l'évaluation des élèves. On vise ainsi la mobilisation par l'élève des ressources utiles telles que des savoirs, des savoir-faire, des stratégies, etc., pour aborder et résoudre ces problèmes complexes aux solutions possiblement multiples.

Indépendamment de l'approche utilisée, on peut voir la démarche évaluative comme un processus en cinq étapes : (a) la planification, (b) la prise d'informations, (c) l'interprétation, (d) le jugement et (d) la communication. Cette démarche a été adaptée aux nouvelles contraintes de

l'approche par compétences. Les informations quantitatives des programmes par objectifs (nombre de bonnes réponses, nombre de points attribués, etc.) se voient transformées en informations qualitatives constituées par les traces laissées par les élèves et observées par l'enseignant (processus, démarche, production de l'élève à divers stades de réalisation, stratégie, autoévaluations, etc.). L'interprétation normative et quantitative fait place à une interprétation critériée en fonction des critères d'évaluation fournis par le programme de formation. Ces critères, qui sont reformulés par l'enseignant en des manifestations observables génériques ou appropriées à chaque situation, doivent permettre la comparaison entre ce que réalise l'élève et ce qui est attendu.

L'utilisation de grilles d'évaluation permet un jugement sous la forme d'une appréciation qualitative de la performance de l'élève à la tâche[2]. Si l'objectivité du jugement paraissait primordiale dans les années 80, le renouveau pédagogique valorise aujourd'hui les nuances et la richesse du jugement qui découlent d'une approche utilisant une grille d'évaluation analytique descriptive, même si cette approche conduit presque inévitablement à une part de subjectivité. On fait l'hypothèse que cette dernière peut être minimisée par une bonne planification de la part de l'enseignant et par un soin particulier apporté à l'élaboration de ses instruments de collecte des données, mais cela représente évidemment un coût en temps, surtout si chacune des situations d'évaluation, aux différents moments de l'apprentissage, fait appel à de telles grilles appliquées à une ou plusieurs situations complexes.

ÉVALUER LES COMPÉTENCES EN MATHÉMATIQUES

En mathématiques, comme dans les autres disciplines, les observations qui sont à la base du processus d'évaluation doivent porter autant sur le processus ou les démarches adoptées par l'élève que sur les résultats auxquels il parvient (Gouvernement du Québec, 2006). Les critères d'évaluation des trois compétences mathématiques incluent notamment la capacité à comprendre et à modéliser une situation-problème, à valider une solution, à formuler des conjectures, à utiliser des concepts ou des processus, à structurer et à réaliser des preuves ou des démonstrations, à interpréter des résultats ou des graphiques, et à produire des messages à caractère mathématique. Autant de facettes de l'activité mathématique qui avaient peut-être été perdues de vue ou qui ont pris une plus grande ampleur avec la complexité

2. Le retour au bulletin chiffré annoncé pour l'année scolaire 2007-2008 implique une conversion en pourcentage du résultat obtenu ainsi qu'une pondération des compétences pour aboutir à un résultat disciplinaire, mais il ne semble pas remettre en question cette pratique évaluative.

des problèmes qu'une pratique mathématique instrumentée par la technologie permet aujourd'hui d'aborder.

La mobilisation de connaissances mathématiques pour résoudre des problèmes issus de situations réelles est emblématique de l'orientation des nouveaux programmes de formation de l'école québécoise. Cela constitue l'enjeu de la Compétence 1 : « Résoudre une situation-problème » et reflète une tendance importante de l'évolution récente de la pratique mathématique. En effet, avec les technologies qui permettent d'explorer et de résoudre des problèmes de plus en plus complexes, et donc d'aborder des pans de la réalité de plus en plus vastes, la modélisation mathématique de situations réelles a considérablement gagné en importance dans la pratique mathématique (Caron, 2007). Elle est aujourd'hui considérée par certains mathématiciens comme une des composantes essentielles des mathématiques et comme un des enjeux principaux de leur enseignement (Bouleau, 2002).

La modélisation mathématique peut être envisagée comme une forme particulièrement complexe et exigeante de la résolution de problèmes. Selon le paradigme adopté, empirique ou théorique (Maul et Berry, 2001), elle vise essentiellement à prédire ou à expliquer un aspect de la réalité. Elle se traduit par un processus cyclique (Blum *et al.*, 2002) qui inclut :

- La structuration et la simplification d'une situation issue du monde réel pour en faire un problème spécifique avec un but, des objets et des relations qui les lient ; ces dernières pouvant être inférées à partir de données ou dérivées de principes sous-jacents connus ;
- La traduction du problème en un modèle mathématique ;
- Le travail mathématique sur le modèle pour aboutir à une solution au problème, avec la possibilité (de plus en plus répandue) d'une mise à contribution des technologies en fonction de la complexité du problème ;
- L'interprétation et la validation des résultats à l'égard de la situation de départ.

La figure 1 présente une illustration détaillée du cycle de modélisation. Elle s'inspire de la description produite pour l'Étude 14 de la Commission internationale pour l'enseignement des mathématiques (Blum *et al.*, 2002). Le nombre de fois où le cycle est parcouru peut être considéré comme un indicateur de la complexité de la tâche et du degré d'authenticité de l'application (Caron et Muller, 2005).

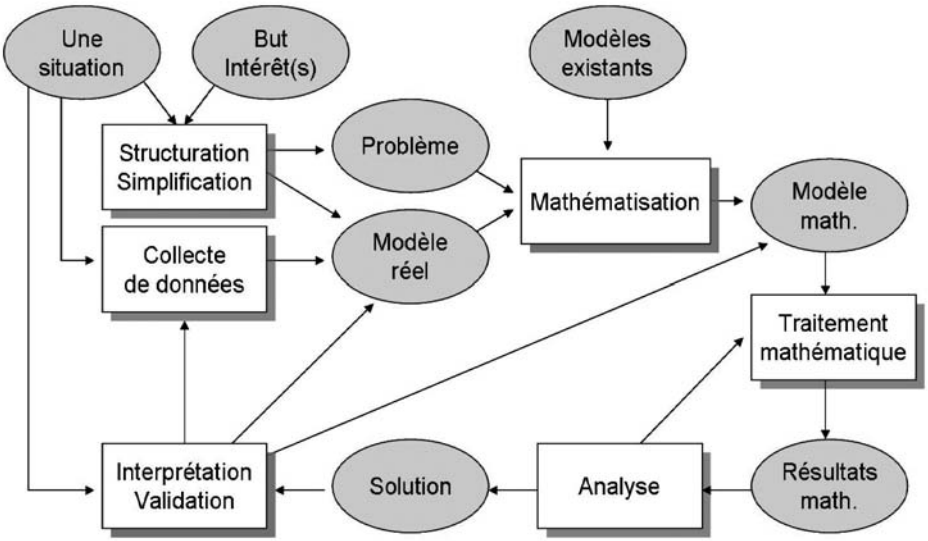


Figure 1 – Le cycle de modélisation

Une intégration féconde du cycle de modélisation dans l'apprentissage des mathématiques, dans lequel il revient à l'élève de déterminer les ressources qu'il met à contribution, demande un subtil dosage de la complexité. D'une part, les situations retenues doivent être suffisamment complexes pour être représentatives de la réalité, favoriser le recours par l'élève à un modèle mathématique du niveau des mathématiques enseignées, et possiblement justifier la mise à contribution d'outils technologiques dans le traitement du modèle. D'autre part, le contexte dans lequel elles s'inscrivent doit être suffisamment connu de l'élève pour favoriser l'appropriation du problème et éviter que les savoirs qui relèvent de l'extramathématique ne s'instaurent en obstacle à sa résolution. La recherche de situations qui répondent à ces critères constitue un défi en soi.

La modélisation par les élèves de telles situations demande du temps, de la recherche, un accès à des données et à des outils technologiques et, par conséquent, se prête mieux à une approche par projets qu'aux formes traditionnelles d'évaluation en mathématiques (Nyman, et Berry, 2002 ; Caron, et Bélair, 2007).

Les critères d'évaluation doivent être reformulés en manifestations observables par l'enseignant afin d'être rendus opérationnels. Or, les manifestations liées aux compétences mathématiques dans la résolution de problèmes complexes aux solutions possiblement multiples sont loin d'être

aussi facilement prévisibles que lors de l'application directe de processus simples et de concepts bien circonscrits. Qui plus est, la prise en charge par les outils technologiques de certains aspects du travail mathématique donne souvent lieu à une réduction de la documentation de la démarche par l'élève et donc à un ensemble restreint d'éléments observables. Il devient alors difficile pour l'enseignant de savoir avec précision ce que l'élève a fait et de déterminer ce qui a pu donner lieu à un résultat erroné. Pour l'enseignant de mathématiques, encore assez peu habitué à fonctionner avec une évaluation de nature critériée, la tâche d'évaluation peut paraître, et à raison, à la fois passablement plus lourde et plus floue qu'auparavant.

LA TECHNOLOGIE ET LES PRATIQUES ÉVALUATIVES

L'utilisation de la technologie en mathématique offre la possibilité d'aborder des problèmes beaucoup plus complexes que par le passé. En même temps, elle a tendance à rendre invisibles certaines parties du processus de résolution utilisé par l'élève. À l'avantage de mettre l'élève dans des situations réellement complexes s'oppose donc le défi de n'avoir ensuite que des traces parcellaires pour évaluer des processus associés au déploiement de stratégies de haut niveau. Il devient alors nécessaire de définir des moyens pour juger si les difficultés de l'élève se situent dans la compréhension du problème ou dans sa modélisation, dans l'élaboration de sa stratégie ou dans l'application des processus, et si ces difficultés sont liées à l'utilisation de l'outil technologique ou non.

Pour ce dernier aspect, il est assez facile d'imaginer la possibilité de sauvegarder toutes les actions menées par un élève sur une calculatrice ou sur un ordinateur. De nouveaux modèles de calculatrices graphiques et symboliques offrent d'ailleurs cette option. La véritable difficulté vient plutôt de trouver un moyen d'interpréter et de porter un jugement sur la quantité importante d'informations qui sont alors à la disposition de l'enseignant. Il convient de se demander si une grille d'évaluation adaptée constitue un outil adéquat dans une telle situation, et dans quelle mesure cette grille peut rendre compte de ce qui s'est réellement passé. De plus, la quantité de travail demandée à l'enseignant devient alors énorme, car l'utilisation de l'outil informatique n'est qu'une des nombreuses facettes de ce qu'il doit observer.

Ces considérations nous amènent à nous questionner sur la possibilité de disposer de moyens plus économiques pour détecter les erreurs typiques des élèves dans le processus de modélisation et dans leur utilisation des outils technologiques afin de réguler l'enseignement en cours d'apprentissage et de mieux contribuer au développement des compétences visées. Les

technologies, en partie responsables de cette nouvelle situation, pourraient-elles apporter une partie de la solution ? C'est à cette question que nous tenterons, dans ce qui suit, d'apporter quelques pistes de réponses.

MISE À CONTRIBUTION DES OUTILS INFORMATIQUES POUR L'ÉVALUATION DE LA COMPÉTENCE À RÉSOUDRE UNE SITUATION-PROBLÈME COMPLEXE

Des questions à réponse courte qui cherchent à mesurer des éléments de processus complexes en allant au-delà de la simple reproduction de comportements associés à un contenu très limité pourraient soutenir l'enseignant dans sa tâche d'évaluation en prenant en charge une partie de la dimension diagnostique ou formatrice. En particulier, il paraît réalisable d'élaborer des questionnaires à choix multiples (QCM) ou à réponse courte (QRC) afin de diagnostiquer les forces et les faiblesses des élèves dans les différentes phases de la modélisation de situations complexes, et l'utilisation des outils informatiques habituellement mis à leur disposition lors de la réalisation de projets (calculatrices numériques, graphiques, symboliques, etc.). Précisons d'emblée, pour éviter toute confusion, qu'une question à réponse courte peut présenter dans son énoncé une information complexe et exiger dans la réponse des habiletés de haut niveau.

Du côté de la modélisation, Haines et Crouch (2005) proposent en ce sens des exemples de questions à choix multiples qui permettent d'isoler, de travailler ou d'évaluer les habiletés liées aux différentes phases du processus de modélisation[3] : circonscrire le problème à partir d'une situation réelle, spécifier un modèle pour comparer différents scénarios (qu'on pourrait simuler par la suite), établir des relations entre des graphiques et des aspects de la réalité qu'ils représentent, interpréter des données, etc. De telles questions ne peuvent évidemment pas se substituer aux projets plus vastes pour le développement et l'évaluation des compétences de modélisation, mais elles peuvent se révéler fort utiles dans le cadre d'une épreuve diagnostique (ou formatrice) en début ou en cours d'apprentissage.

Du côté de l'utilisation des technologies dans le traitement mathématique, la recherche en didactique a mis en évidence dans les dernières années plusieurs phénomènes récurrents chez les élèves. Ces phénomènes témoignent

3. Les travaux de Haines et Crouch s'inscrivent dans ceux du *International Community of Teachers of Mathematical Modelling and Applications* où l'on note un intérêt croissant à mieux circonscrire les compétences mobilisées dans l'exercice de modélisation mathématique, de façon à en permettre le développement et l'évaluation. En témoignent pas moins de huit contributions sur ce sujet aux Actes de la rencontre de ICTMA 12 (Haines, Galbraith, Blum et Khan, 2007).

tantôt de nouvelles techniques spontanément mises à contribution par les élèves, lesquelles utilisent ou court-circuitent le savoir mathématique en jeu, tantôt de conceptions erronées à l'endroit de l'outil utilisé ou des objets mathématiques à travers les représentations qu'en donne l'outil (Artigue, 1997, 2002 ; Doerr et Zangor, 2000 ; Haspekian, 2005 ; Hölz, 1996 ; Lagrange, 2000 ; Ruthven, 1990 ; Trouche, 2000 ; etc.). Ces nombreuses études fournissent un répertoire riche concernant les processus cognitifs mis à contribution dans l'utilisation des outils technologiques en mathématiques. Elles pourraient aider à constituer des QCM ou QRC qui intègrent ces outils et qui permettent de faire ressortir de façon plus ciblée les forces et difficultés individuelles des élèves dans leur utilisation.

De telles tâches peuvent facilement être informatisées et les épreuves peuvent être ou non adaptatives. Notons que la réflexion qui suit ne prendra pas cette dernière option en considération en raison des difficultés supplémentaires qu'elle implique et qui, typiquement, font que la dimension adaptative ne s'intègre aux outils d'évaluation que dans une seconde phase de leur développement.

L'utilisation de QCM ou QRC informatisés présenterait l'avantage de fournir rapidement à l'enseignant des informations spécifiques sur ses élèves pour planifier, élaborer, gérer et évaluer leurs activités dans des situations plus complexes, qui demeurent nécessaires au développement des compétences et à l'exercice d'une évaluation globale des compétences. Une telle approche fournirait un moyen technologique économique (en temps et en énergie) pour détecter les erreurs typiques, réguler l'utilisation des outils par les élèves et, par une rétroaction rapide, contribuer au développement de leur jugement en situation de modélisation.

En outre, la mise en place d'une plateforme informatique pour passer les tests permettrait d'assurer l'accès aux outils technologiques dans la mesure où ces derniers seraient intégrés à même la plateforme. Il est important de rappeler que les outils informatiques font de plus en plus partie de la pratique mathématique et des apprentissages qui y préparent. Refuser leur présence en situation d'évaluation signifie qu'on ne reconnaît pas un statut institutionnel aux techniques potentiellement très puissantes qui ont pu être développées avec ces outils, et limite la complexité des tâches qu'il est possible de soumettre à l'étudiant. Par ailleurs, les admettre pour des tâches où ils s'avèrent peu pertinents engendre inévitablement des tensions et des contradictions.

Les informations propres à chaque élève issues des QCM/QRC pourraient par la suite aider à analyser ses productions dans la résolution de problèmes plus complexes qui intègrent l'outil technologique. Le fait d'évaluer dans un processus distinct des habiletés plus circonscrites liées à la modélisation ou à l'utilisation des outils permet de pallier l'éventuelle absence

de traces de la démarche utilisant l'outil ou la difficulté d'interpréter ces traces si elles sont disponibles. Le temps nécessaire à cette interprétation pourrait être diminué. Une telle approche favoriserait l'intégration des outils informatiques dans des tâches pour lesquelles leur utilisation s'avère pertinente, voire indispensable, et donc un authentique accroissement en complexité des problèmes auxquels on prépare l'élève dans ses apprentissages.

UNE APPROCHE DIAGNOSTIQUE DE L'UTILISATION DES TECHNOLOGIES PAR LES ÉLÈVES

La démarche d'évaluation implique toujours une façon d'interpréter les résultats obtenus. La mise en place de tests informatisés et la rapidité de la rétroaction qu'elle permet n'ont d'intérêt dans une perspective diagnostique ou formatrice que si les résultats de chaque élève aboutissent à des informations utiles à la régulation de ses apprentissages. Dans le cas où l'évaluation cherche à aider un élève à développer des compétences (par exemple liées à l'utilisation de divers outils technologiques), il ne suffit pas de lui dire s'il réussit ou s'il échoue (Mislevy, Wilson, Ercikan et Chudowsky, 2001). Les réponses fournies aux QCM/QRC doivent donc être modélisées de manière à établir un lien entre les performances et les habiletés utilisées dans le processus de réponse. Il faut tenir compte du fait qu'une tâche puisse être résolue de diverses manières et par le truchement d'habiletés liées entre elles dans le temps ou logiquement (Corter, 1995 ; Mislevy et al, 2003).

Depuis le milieu des années 1980, des recherches se sont penchées sur la possibilité de relier certains développements de la psychométrie à ceux des sciences cognitives. L'objectif consiste à fournir un diagnostic en lien avec des variables latentes discrètes représentant des processus cognitifs (Chipman, Nichols, et Brennan, 1995). De nouveaux modèles de mesure (Loye, 2005 et chapitre 9), appelés modèles cognitifs, se sont développés afin de modéliser la probabilité de répondre correctement à un item en fonction de la maîtrise d'un ensemble d'attributs cognitifs propres à cet item.

Ces attributs cognitifs peuvent être interprétés comme des savoirs, des savoir-faire, des habiletés qui peuvent être combinés afin d'obtenir la bonne réponse à une question. La perspective d'une interprétation diagnostique implique de retenir des attributs simples et pas trop nombreux (Chipman *et al.*, 1995 ; Corter, 1995 ; Gierl, 1997). Il faut évidemment aussi qu'il y ait un intérêt à baser le diagnostic des sujets sur la maîtrise ou la non-maîtrise de ces attributs. Gierl (1997) définit un attribut comme une procédure, une habileté, une stratégie ou une connaissance qu'un sujet doit posséder pour résoudre un test, mais qu'on ne peut pas observer directement. Le degré de maîtrise des attributs doit donc être inféré à partir des réponses des sujets.

Buck et ses collègues (1998) soulignent qu'un attribut fait référence à tout ce qui peut affecter la performance lors d'une tâche : cela peut être une caractéristique de la tâche ou des connaissances ou des habiletés nécessaires pour la mener à bien. Pour Tatsuoka (1995), l'attribut cognitif fait référence à une notion ou à un microniveau d'habileté en plus d'inclure des variables plus générales liées au contexte, à la forme de la question ou au processus de réponse.

Les modèles cognitifs s'appuient sur l'élaboration, par des experts, d'une liste d'attributs cognitifs nécessaires et suffisants pour répondre aux items du test. Les experts peuvent, par exemple, déterminer au préalable les processus cognitifs qu'ils veulent diagnostiquer, et fabriquer un test approprié. Ils peuvent également utiliser un test existant et définir *a posteriori* quels sont les attributs en lien avec chacun des items. Le modèle fournit ensuite des informations sur la maîtrise ou non de chacun des attributs ciblés pour chacun des sujets.

Le modèle RUM[4] (*Reparametrized Unified Model*) (Dibello, Stout et Rousseau, 1995 ; Hartz, 2002 ; Rousseau *et al.*, 2007) est un modèle cognitif qui se base sur une liste de k attributs reliés aux items d'un test, mais qui tient aussi compte du fait qu'un certain nombre d'attributs non spécifiés jouent un rôle dans la performance des sujets. Il estime des paramètres liés aux sujets, des paramètres liés aux items, des paramètres liés aux attributs ainsi que des paramètres liés aux relations spécifiques qui existent entre un attribut et un item puisqu'un attribut n'est associé qu'à quelques-uns des items (voir figure 2). Ce modèle est donc particulièrement intéressant si l'on cherche à cibler certaines habiletés sans forcément tenir compte de tous les aspects de la résolution du problème. Par exemple, il serait possible de n'intégrer au modèle que des attributs cognitifs liés à l'utilisation des outils technologiques, en sachant que d'autres ressources sont nécessaires à l'élève, sans pour autant les détailler.

La perspective d'utiliser des QCM ou QRC informatisés pour diagnostiquer les forces et faiblesses des élèves dans la modélisation mathématique et l'utilisation d'outils technologiques pourrait donc fort bien se prêter à une application du modèle RUM. L'enjeu consisterait, dans un premier temps, à mettre à profit les travaux en didactique des mathématiques pour cibler les différents processus cognitifs que l'élève doit mobiliser lorsqu'il modélise ou qu'il utilise chacun de ces outils (calculatrices, tableur, ou autre logiciel). Le défi consisterait ensuite à élaborer ou à déterminer des questions à réponses courtes reliées à un ou plusieurs des processus identifiés. L'avantage du modèle RUM vient du fait qu'il offre de n'établir le diagnostic que

4. Voir les équations du modèle RUM dans le chapitre 9 de cet ouvrage.

relativement à des processus ciblés, les autres processus pouvant être pris en compte par le modèle de manière séparée (par les paramètres c et θ, voir figure 2).

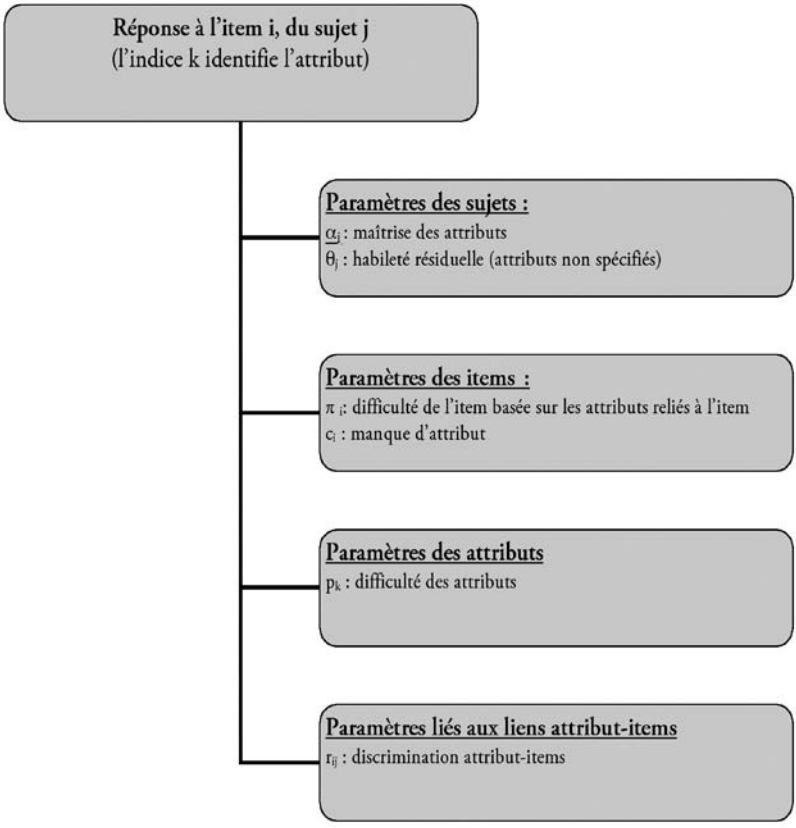


Figure 2 – Les paramètres estimés dans le modèle RUM

À partir des réponses des élèves aux items, le modèle propose un diagnostic sous la forme d'un vecteur (α) des probabilités de maîtrise de chacun des k attributs pour chaque élève. Ces informations peuvent servir à la régulation des apprentissages de l'élève qui obtient alors des informations sur ses forces et ses faiblesses. Ces informations pourraient donc aider l'enseignant à porter un jugement sur les traces récoltées lors des mises en situations plus complexes.

CONTRAINTES ET DÉFIS

Les contraintes liées à l'opérationnalisation d'une telle approche en salle de classe sont multiples. Tout d'abord, elle nécessite que les élèves aient un accès à des ordinateurs en nombre suffisant et que les enseignants soient prêts à utiliser différentes technologies dans leur approche pédagogique. Les coûts liés au développement de plateformes de tests doivent également être pris en considération. Il est essentiel que les outils technologiques intégrés au test soient les mêmes que les outils habituellement utilisés dans la salle de classe. Leur disponibilité sous forme de logiciel ou d'émulateurs doit donc être vérifiée. En outre, ces outils évoluent rapidement, nécessitant une mise à jour fréquente des plateformes informatiques.

Étant donné que l'utilisation de ces outils est une des dimensions de l'activité mathématique sur lesquelles porte l'évaluation, la mise en page à l'écran doit être la plus semblable possible à la réalité. Par exemple, l'utilisation d'un émulateur de calculatrice intégré à la plateforme du test doit, pour l'élève, être identique (ou très semblable) à l'utilisation de sa propre calculatrice.

En ce qui concerne la sécurité, elle n'est pas essentielle ici, puisque les enjeux de l'évaluation ne sont pas très élevés. La mise en page à l'écran doit néanmoins chercher à minimiser le risque de plagiat. De nombreuses décisions doivent être prises au moment de l'informatisation du test : adaptatif ou non, possibilité de revoir les items déjà complétés ou non, choix de l'ordre des réponses dans le cas où le test n'est pas adaptatif, etc.

Toutefois, le défi le plus important tient à la modélisation des réponses des élèves à l'aide du modèle cognitif. Afin que des enseignants puissent utiliser une telle approche, il est important que le processus se fasse automatiquement, et que l'enseignant et l'élève n'aient accès qu'aux informations pertinentes. Par exemple, le diagnostic de l'élève pourrait lui être fourni après analyse des données sous la forme d'un graphique ou de la liste des attributs associés à leur probabilité de maîtrise. Étant donné que c'est la modélisation des données à l'aide du modèle cognitif qui aboutit à ce portrait de l'élève, il serait intéressant de l'intégrer à la plateforme pour éviter à l'enseignant de manipuler les données.

Enfin, rien ne peut être mis en place sans le travail d'experts à l'élaboration des listes des attributs cognitifs et des items qui leur sont reliés. Cette étape est la première d'un processus qui est long et complexe, mais qui pourrait être prometteur. Il s'agit là d'une occasion intéressante pour que se rencontrent didacticiens et docimologues dans l'élaboration d'épreuves visant à soutenir l'apprentissage.

CONCLUSION

Ce chapitre propose de faire jouer un nouveau rôle aux QCM et aux QRC, en les mettant à contribution dans le repérage des processus moins facilement observables qui sont typiquement mobilisés pour résoudre des situations-problèmes complexes en mathématiques, et ainsi soutenir l'enseignant dans sa tâche d'évaluation. Le croisement des dernières recherches sur les modèles cognitifs avec celles portant sur l'intégration de la modélisation et des technologies dans l'apprentissage des mathématiques pourrait donner lieu à l'outil recherché. Il serait donc intéressant de mener des études en ce sens afin de vérifier si l'approche proposée peut être rendue opérationnelle. Cette volonté de croiser deux perspectives de recherche en éducation (l'évaluation et la didactique) nous semble prometteuse, autant pour la recherche que pour les acteurs sur le terrain, élèves et enseignants.

Références

Artigue, M. (1997), « Le logiciel DERIVE comme révélateur de phénomènes didactiques liés à l'utilisation d'environnements informatiques pour l'apprentissage », *Educational Studies in Mathematics, 33*(2), p. 133-169.

Artigue, M. (2002), « L'intégration de calculatrices symboliques à l'enseignement secondaire : les leçons de quelques ingénieries didactiques », dans, D. Guin, et L. Trouche (Ed.), *Calculatrices symboliques, Transformer un outil en un instrument du travail mathématique : un problème* didactique, Grenoble, La Pensée Sauvage, chapitre 9, p. 277-349.

Bouleau, N. (2002), « Sur le rôle des mathématiques dans la société d'aujourd'hui », *Bulletin de l'APMEP* n° 440, p. 309-322.

Blum, W., *et al.* (2002), « ICMI Study 14 : Applications and modelling in mathematics education – Discussion Document », *Educational Studies in Mathematics*, *51*(1-2), p. 149-171.

Caron, F. (2007), « Mise à contribution de la notion de compétence pour guider le développement d'une pratique mathématique instrumentée », dans, R. Floris et F. Conne (Ed.), *Environnements informatiques, enjeux pour l'enseignement des mathématiques*, Bruxelles, Éditions De Boeck Université, p. 185-200.

Caron, F. et Bélair, J. (2007), « Exploring university students' competencies in modelling », dDans, C. Haines, P. Galbraith, W. Blum, S. Khan (Ed.), *Mathematical Modelling : Education, Engineering and Economics*, Chichester, Horwood Publishing, p. 120-129.

Caron, F. et Muller, E. (2005), « Integrating Applications and Modelling in Secondary and Post secondary Mathematics », dans, E. Simmt et B. Davis (Ed.), *Proceedings of the 28th Annual Meeting of the Canadian Mathematics Education Study Group*, p. 63-80.

Chipman, S.F., Nichols, P.D., et Brennan, R.L. (1995), « Introduction », dans, P.D. Nichols, S.F. Chipman et R.L. Brennan (Ed.), *Cognitively diagnostic assessment*, Hillsdale, NJ, Erlbaum, p. 1-18.

Corter, J.E. (1995), « Using clustering methods to explore the structure of diagnostic tests », dans, P.D. Nichols, S.F. Chipman et R.L. Brennan (Ed.), *Cognitively diagnostic assessment*, Hillsdale, NJ, Erlbaum, p. 305-326.

Conseil supérieur de l'éducation (1992), *Évaluer les apprentissages au primaire : un équilibre à trouver,* Québec, Direction des communications du CSE.

DiBello, L.V., Stout, W.F., et Roussos, L.A (1995), « Unified cognitive/psychometric diagnostic assessment likeklihood-based classification techniques », dans, P.D. Nichols, S.F. Chipman et R.L. Brennan (Ed.), *Cognitively diagnostic assessment*, Hillsdale, NJ, Erlbaum, p. 361-389.

Doerr, H.M. et Zangor, R. (2000), « Creating meaning for and with the graphing calculator », *Educational Studies in Mathematics, 41*, p. 143–163.

Gierl, M.J. (1997), *An investigation of the cognitive foundation underlying the rule-space model*, Thèse de doctorat non publiée, Champaign, IL, University of Illinois.

Gouvernement du Québec (1984), *Programme d'études – Secondaire – Mathématique, Second Cycle*, Québec, ministère de l'Éducation.

Gouvernement du Québec (1994), *Programme d'études – Enseignement secondaire – Mathématique 314*, Québec, ministère de l'Éducation.

Gouvernement du Québec (2003), *Programme de formation de l'école québécoise. Enseignement secondaire, premier cycle,* Québec, ministère de l'Éducation.

Gouvernement du Québec (2006), *Échelles des niveaux de compétence. Enseignement secondaire, premier cycle,* Québec, ministère de l'Éducation, du Loisir et du Sport.

Gouvernement du Québec (2007), *Programme de formation de l'école québécoise. Enseignement secondaire, deuxième cycle,* Québec, ministère de l'Éducation, du Loisir et du Sport.

Haines, C., Galbraith, P., Blum, W. et Khan, S. (Ed.), *Mathematical Modelling (ICTMA 12) : Education, Engineering and Economics*, Chichester, Horwood Publishing.

Haspekian, M. (2005), « An instrumental approach to study the integration of a computer tool into mathematics teaching : the case of spreadsheets », *International Journal of Computers for Mathematical Learning*, *10*, p. 109–141.

Hartz, S. (2002), *A Bayesian framework for the unified model for assessing cognitive abilities : Blending theory with practicality*, Thèse de doctorat non publiée, Champaign, IL, University of Illinois.

Hölz, R. (1996), « How does dragging affect the learning of geometry », *International Journal of Computers for Mathematical Learning*, *1*, p. 169-187.

Lagrange, J.B. (2000), « L'intégration d'instruments informatiques dans l'enseignement : Une approche par les techniques », *Educational Studies in Mathematics*, *43*(1), p. 1-30.

Legendre, R. (2005), *Dictionnaire actuel de l'éducation, 3ème édition*, Montréal, Guérin.

Loye, N. (2005), « Quelques modèles modernes de mesure », *Mesure et évaluation en éducation, 28*(3), p. 51-68.

Maull, W. et Berry, J. (2001), « An investigation of student working styles in a mathematical modelling activity », *Teaching Mathematics and its Applications*, *20*(2), p. 78-88.

Mislevy, R.J., Steinberg, L.S., et Almond, R.G. (2003), « On the structure of educational assessments », *Measurement : Interdisciplinary Research and Perspectives*, *1*(1), p. 3-62.

Nyman, M. et Berry, J. (2002), « Developing transferable skills in undergraduate mathematics students through mathematical modelling », *Teaching Mathematics and its Applications*, *21*(1), p. 29-45.

Rousseau, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., et Templin, J. L. (2007), « The fusion model skill diagnosis system », dans, J. P. Leighton and M. J. Gierl (Ed.), *Cognitive diagnostic assessment for education*, New York, Cambridge University Press, p. 241-316.

Ruthven, K. (1990), « The influence of graphic calculator use on translation from graphic to symbolic forms », *Educational Studies in Mathematics*, *21*, p. 431-450.

Tatsuoka, K.K. (1995), « Architecture of knowledge structures and cognitive diagnosis : A statistical pattern recognition and classification approach », dans, P.D. Nichols, S.F. Chipman et R.L. Brennan (Ed.), *Cognitively diagnostic assessment*, Hillsdale, NJ, Erlbaum, p. 305-326.

Trouche, L. (2000), « La parabole du gaucher et de la casserole à bec verseur : étude des processus d'apprentissage dans un environnement de calculatrices complexes », *Educational Studies in Mathematics*, *41*, p. 239-264.

Page laissée blanche intentionnellement

CHAPITRE 7

L'évaluation en ligne du raisonnement clinique médical par le test de concordance de script

Bernard Charlin et Driss Kazi-Tani
Université de Montréal

INTRODUCTION

Les problèmes posés par certaines situations professionnelles peuvent être résolus par l'application simple de règles et de connaissances mais, le plus souvent, les professionnels doivent affronter des situations complexes ou ambiguës qui demandent une réflexion pour interpréter les données et pour progresser vers la solution. La capacité à raisonner en contexte d'incertitude et à résoudre des problèmes mal définis caractérise l'exercice professionnel. Si les institutions d'enseignement forment assez bien à l'utilisation des savoirs techniques qui permettent la résolution des problèmes bien définis, elles forment et évaluent plus difficilement le raisonnement en contexte d'incertitude (Schön, 1983). Les connaissances nécessaires à la résolution des problèmes simples liés à des situations pour lesquelles il y a une seule bonne réponse peuvent être adéquatement et fidèlement évaluées à partir des données recueillies avec des outils standardisés tels que les questions à choix multiples basés sur des cas (Case et Swanson, 2001). Il existe cependant un besoin de méthodes innovatrices permettant d'effectuer des évaluations dans les contextes d'incertitude fréquemment vécus en pratique (Dinant, 2004 ; Gerrity, Earl, et deVellis, 1992 ; Hall, 2002). C'est ce qui est visé avec l'utilisation du test de concordance de script pour l'évaluation du raisonnement clinique médical.

CONTEXTE THÉORIQUE

En matière d'investigations cognitives sur l'expertise médicale, on a assisté à un déplacement de la recherche d'habiletés générales de résolution de problème vers la focalisation envers l'organisation en mémoire, l'utilisation des connaissances, la représentation de problèmes et leurs évolutions en fonction de l'expérience. Dans le domaine de l'évaluation des apprentissages, ce changement de focalisation n'a pas conduit à une modification des pratiques. Dans cette perspective, certains auteurs ont suggéré que l'évaluation se concentre sur le jugement des qualités des opérations cognitives ou des structures de connaissances en comparant la représentation de problème des étudiants, leurs jugements et leurs choix à ceux de groupes expérimentés (voir, par exemple, Elstein, Shulman et Sprafka, 1990).

Le test de concordance de script (TCS) s'inscrit dans cette approche. Il est basé sur une théorie cognitive, la théorie des scripts, qui met l'accent sur les réseaux de connaissances qu'utilisent les individus en situation de résolution de problème (Charlin, Boshuizen, Custers et Feltovich, 2007 ; Charlin, Tardif et Boshuizen, 2000). Bordage (1994) a montré que les cliniciens les plus efficaces dans une situation donnée sont ceux qui établissent de nombreux liens dans le réseau de leurs connaissances cliniques. Selon la théorie des scripts, les professionnels de la santé progressent vers la solution des problèmes cliniques en ayant présent à l'esprit des hypothèses (ou des options de prise en charge) avec les réseaux de connaissances (scripts) qui leurs sont liées. Ils utilisent activement ces connaissances pour poser en permanence des jugements sur l'effet de chaque nouvel élément d'information sur le statut de l'hypothèse ou de l'option. Le TCS repose sur le principe que ces multiples jugements peuvent être appréciés formellement et que leur concordance avec ceux d'un panel de référence peut être mesurée, ce qui constitue un test d'évaluation du raisonnement clinique (Fournier, Demeester et Charlin, 2007). Les bases théoriques concernant les scripts, et leurs caractéristiques en sciences de la santé, sont décrites dans d'autres publications qui peuvent être consultées par le lecteur voulant en apprendre plus sur l'application du concept de script en sciences de la santé (Charlin *et al.*, 2007 ; Charlin, Tardif *et al.*, 2000).

Le TCS est utilisable en format papier-crayon, mais les développements récents en informatique décuplent ses possibilités, aussi bien en matière de création de test, d'inclusion d'images, ou d'évaluation d'habiletés particulières (la perception par exemple) que de passation en ligne, de correction automatisée et de transmission des résultats aux candidats. Ce format de test a

surtout été utilisé en sciences de la santé (médecine, chirurgie, radiologie, sages-femmes, pharmacie, ergothérapie). Il peut toutefois être utilisé dans toutes les professions qui impliquent une certaine incertitude. Le but de ce texte est de présenter le format du test, la méthode de correction, les qualités psychométriques du test, ainsi que les possibilités qu'offre l'ordinateur pour la construction, la passation et la correction en ligne du test.

FORMAT DU TEST

La figure 1 illustre le format du test. Chaque item (c'est-à-dire un cas avec les questions qui lui sont liées) comporte une vignette décrivant une situation problématique. Les situations choisies impliquent délibérément de l'incertitude, même pour des experts du domaine, et plusieurs options sont pertinentes (Fournier *et al.*, 2007). Chaque question contient une option de résolution du cas, suivie par la présentation d'une nouvelle donnée. Le candidat doit porter un jugement sur l'effet qu'a cette nouvelle donnée sur le statut de l'option proposée. Les réponses sont captées sur une échelle s'apparentant à une échelle de Likert, et reflétant la façon dont l'information est traitée en situation de résolution de problème selon la théorie des scripts (Charlin *et al.*, 2007 ; Charlin, Tardif *et al.*, 2000). Les questions suivantes, au sein d'un même cas, explorent l'effet d'autres données sur d'autres options.

Dans des situations comportant de l'incertitude, les cliniciens ne suivent pas les mêmes cheminements dans leur raisonnement et n'interprètent pas les données de façon identique (Barrows et Feltovich 1987 ; Elstein AS, Shulman, L.S., Sprafka 1978, Grant et Marsden, 1988). Pour tenir compte de cette variabilité, le processus d'attribution du score consiste à faire passer le test à un panel de référence et à donner au candidat un résultat reflétant le nombre de membres du panel qui ont donné la même réponse que lui. Ces principes permettent d'introduire dans le processus d'évaluation des situations de vie réelle, avec toute leur complexité, situations pour lesquelles il n'y a pas clairement une seule bonne réponse. Cela permet de créer du matériel d'examen comportant des tâches qui, du fait de l'absence d'une indiscutable « bonne réponse », ne sont pas utilisées dans les tests usuels. Fournier et des collaborateurs ont récemment décrit les étapes de construction du test de concordance de script (Fournier *et al.*, 2007).

Item XX

Vous êtes à l'urgence. Une patiente de 65 ans y est envoyée par une clinique pour une suspicion d'embolie pulmonaire. Elle a des antécédents de diabète, d'insuffisance cardiaque et de bronchopneumopathie chronique obstructive.

Si vous pensiez faire (option d'examen complémentaire)	**Et qu'alors vous trouvez (nouvelle information obtenue par examen clinique ou examen complémentaire)**	**L'effet sur la nécessité de demander ce test est le suivant**
Q1- Un angioscanner thoracique	Un traitement par metformine (Glucophage retard®)	-2 – 1 0 + 1 + 2
Q2- Un dosage des D dimères	Un cancer de l'ovaire en cours de traitement	-2 – 1 0 + 1 + 2
Q3- Un écho-Doppler veineux	Un signe de Homans	-2 – 1 0 + 1 + 2

- 2 : absolument contre-indiqué
- 1 : peu utile ou plutôt néfaste
0 : non pertinent dans cette situation
+ 1 : utile et souhaitable
+ 2 : indispensable

Figure 1 – Exemple d'item (Fournier *et al.*, 2006)

MÉTHODE D'ÉTABLISSEMENT DES SCORES

Pour établir les scores, on demande d'abord à des professionnels expérimentés (des experts) du domaine évalué de remplir le test individuellement, ce qui constitue le panel de référence. Chacune des réponses des membres du panel contribue à la création de la grille de correction (Gagnon, 2005). Les candidats reçoivent ainsi un score qui reflète le nombre de membres du panel qui ont répondu comme eux. Un point est alors attribué lorsque le candidat choisit l'option qui correspond à la réponse modale des membres du panel à la question. Les autres réponses reçoivent un score proportionnel au nombre de membres du panel ayant privilégié l'option choisie par le candidat et les réponses non choisies par le panel se voient attribuées un score de 0. Le processus consiste, pour chaque question, à diviser le nombre de réponses à chaque ancrage de l'échelle des réponses par la valeur modale pour cette question. Dans l'exemple des figures 1 et 2, si, pour une question donnée, quatre membres du panel (parmi neuf) ont choisi l'ancrage « - 2 », cette réponse reçoit 1 point (4/4). Si trois membres du panel ont choisi

« - 1 », cette réponse reçoit 0,75 point (3/4) et, si 2 membres ont choisi l'ancrage « 0 », cette réponse reçoit 0,5 (2/4). Le score total pour le test est la somme de tous les crédits obtenus pour chaque question. L'échelle des scores est ensuite transformée afin d'obtenir une limite supérieure de 100.

Si vous pensiez faire	**Et qu'alors vous trouvez**	**L'effet sur la nécessité de demander ce test est le suivant**				
	Échelle de réponse	**- 2**	**- 1**	**0**	**+ 1**	**+ 2**
Un angioscanner thoracique (Q1)	Un traitement par metformine (Glucophage retard®)					
Distribution des réponses des membres du panel d'experts		4	3	2	0	0
Points attribués selon la réponse du candidat		1	0,75	0,5	0	0
Un dosage des D dimères (Q2)	Un cancer de l'ovaire en cours de traitement					
Distribution des réponses des membres du panel d'experts		0	0	4	4	1
Points attribués selon la réponse du candidat		0	0	1	1	0,25
Un écho-Doppler veineux (Q3)	Un signe de Homans					
Distribution des réponses des membres du panel d'experts		0	1	5	2	1
Points attribués selon la réponse du candidat		0	0,2	1	0,4	0,2

Figure 2 – Création des clés de correction pour chaque question (panel de neuf experts)

QUALITÉS PSYCHOMÉTRIQUES DU TEST

Les études publiées sur le test de concordance de script montrent que, dans les contextes où ce type de mesures est mis en œuvre, une soixantaine de questions (test d'une durée d'environ une heure) permettait d'atteindre une fidélité de la mesure, estimée par le coefficient de cohérence interne de Cronbach, atteignant ou dépassant 0,75 (Charlin, Desaulniers, Gagnon, Blouin et van der Vleuten, 2002 ; Charlin, Gagnon, Sauvé et Coletti, 2006 ; Charlin, Gagnon, Sibert et Van der Vleuten, 2002 ; Charlin, Roy, Brailovsky et Van der Vleuten, 2000 ; Charlin et van der Vleuten, 2004 ; Fournier *et al.*, 2006 ; Gagnon, 2005). Des études de « généralisabilité » ont révélé que le test peut être optimisé en termes de fidélité, pour cette durée de test, par la création de combinaisons de deux à quatre questions par cas nichées dans quinze à vingt cas (Gagnon, 2007).

Un test qui mesure réellement les capacités de raisonnement professionnel devrait donner des scores plus élevés aux personnes qui ont plus d'expérience professionnelle (validité de construit). Or, les tests usuels

donnent des résultats paradoxaux : ce ne sont pas les professionnels expérimentés qui obtiennent les meilleurs scores, mais les personnes de niveau intermédiaire, les résidents en fin de formation, proches de leurs examens de certification. Cet effet est nommé « effet intermédiaire » dans la littérature en éducation médicale (van der Vleuten et Newble, 1995). Pour le test de concordance de script, toutes les études (Charlin, Desaulniers *et al.*, 2002 ; Charlin *et al.*, 2006 ; Charlin, Gagnon *et al.*, 2002 ; Charlin, Roy *et al.*, 2000 ; Charlin et van der Vleuten, 2004 ; Fournier *et al.*, 2006 ; Gagnon, 2005) ont montré une progression des scores depuis les novices (les étudiants), vers les médecins expérimentés, en passant par les différents niveaux de résidence. Cette absence de l'effet intermédiaire a par exemple été démontrée à propos du raisonnement des radio-oncologues (Lambert, 2005) et des médecins d'urgence (Carrière, 2005).

La validité prédictive du test a été illustrée par une étude dans laquelle un test de concordance a été donné à tous les étudiants d'une faculté immédiatement avant leurs examens de fin de formation (Brailovsky, Charlin, Beausoleil, Coté et Van der Vleuten, 2001). Le test portait sur les connaissances cliniques en chirurgie. La cohorte d'étudiants qui s'est ensuite engagée en formation de médecine familiale (donc dans un domaine tout à fait différent) a été suivie jusqu'à l'examen d'obtention de la spécialité, deux ans plus tard. Compte tenu de l'objet de mesure du test, c'est-à-dire l'organisation des connaissances adaptées au raisonnement clinique, l'hypothèse était que le test prédirait bien les résultats des tests de raisonnement clinique (cette fois-ci dans le domaine de la médecine familiale) et moins bien ceux concernant les habiletés cliniques, mesurées par un examen clinique objectif et structuré (ECOS). Les résultats ont confirmé cette hypothèse, suggérant que les étudiants qui organisent bien leurs connaissances à un moment de leur formation continuent de le faire aux stades ultérieurs de formation, et ce, même dans des domaines de connaissance différents.

Le processus de mise en forme du score, composante clé des tests de concordance, incorpore la variabilité du jugement des experts lorsqu'ils sont confrontés aux problèmes complexes et mal définis qui caractérisent la vie professionnelle. Cette variabilité soulève d'ailleurs des questions pertinentes à propos du nombre de membres d'un panel requis pour obtenir des scores stables, de l'influence de la composition du panel sur les scores ou du degré de variabilité des réponses parmi les experts.

Il a été démontré que, pour obtenir la stabilité de scores indispensables pour des examens à enjeu critique (examens de certification, par exemple), le panel doit réunir une quinzaine de membres (Gagnon, 2005). Pour des examens à enjeu moindre, un minimum de huit à dix membres est nécessaire. L'expérience acquise dans d'autres types d'examen peut faire paraître difficile

l'atteinte de tels nombres d'experts, mais on demande en fait à des personnes expérimentées de répondre à des questions qui sont semblables à celles qu'ils se posent quotidiennement dans leur pratique. Ces personnes doivent compléter le test exactement comme des étudiants le feraient. Leurs réponses sont ensuite compilées pour bâtir la grille de correction. Ce processus peu exigeant en temps est, de façon générale, très bien accepté.

Une étude menée en médecine générale (Charlin *et al.*, 2006) a comparé les scores de résidents obtenus à partir de deux panels, l'un composé d'enseignants, l'autre de médecins de pratique privée sans fonction d'enseignement, mais suivant régulièrement des sessions de formation continue. Le rang des résidents était le même (corrélation à 0,98). Par contre, les scores étaient un peu plus élevés lorsque la correction utilisait le panel des enseignants. Il a par ailleurs été démontré que les questions qui discriminent bien en fonction du degré d'expérience sont celles qui induisent une variabilité entre les réponses des membres du panel. Il s'agit des questions liées à des tâches cognitives suffisamment complexes et ambigües pour qu'elles induisent une variabilité entre membres du panel. En effet, si tous les membres du panel donnent la même réponse, la question devient similaire à une question à choix multiple, or ce type de question induit l'effet intermédiaire décrit ci-dessus. Par contre, si les réponses divergent totalement, il s'agit vraisemblablement d'une question mal construite.

Précisons enfin qu'étudiants, résidents et professionnels trouvent le test intéressant à passer, car les tâches cognitives ressemblent à celles de la pratique réelle. Le test, qu'il soit présenté sur papier ou sur écran d'ordinateur, est relativement facile à gérer, requérant peu de ressources matérielles ou humaines. Le processus de correction est cependant complexe et requiert l'utilisation de programmes informatisés.

LE TCS ET LA TECHNOLOGIE INFORMATIQUE

Les développements des technologies de l'information et de la communication peuvent apporter beaucoup à l'utilisation d'une épreuve comme le TCS et permettre des applications difficilement concevables dans le cadre d'une approche conventionnelle papier-crayon. Nous présenterons ici le contexte qui a entraîné la mise en ligne d'un certain nombre de TCS, la structure générale du logiciel qui a été mis à contribution pour ce faire et les développements que permet l'utilisation du multimédia en situation d'examen.

Le contexte

Les premières publications concernant le TCS sont parues à la fin des années 90 et au début des années 2000. Le concept a depuis lors suscité un intérêt croissant à travers des disciplines et des pays divers. Cet intérêt, associé au renforcement de l'usage d'Internet à des fins éducatives[1], a poussé au développement de moyens informatiques permettant de faciliter la création et la passation de tests en ligne, tout en offrant ressources et soutien, tout au long de leur cheminement, aux créateurs de tests situés à distance. Par ailleurs, bien des concepteurs de tests n'étant pas particulièrement férus en informatique, les solutions proposées devaient être simples à utiliser, tout en prenant en considération que le TCS représente une modalité d'évaluation très différente des modalités usuelles et qu'il n'était donc pas possible d'utiliser des logiciels existants sur le marché. Le CPASS (Centre de pédagogie appliquée en sciences de la santé), ou plus exactement la structure qui l'a précédé à l'Université de Montréal (l'URDESS), a pris en 2003 la décision de développer totalement le logiciel nécessaire. La première version opérationnelle a vu le jour la même année.

La structure générale

Le système développé présente deux sections, une pour l'administrateur du test, l'autre pour les participants. Chacune possède son interface et chacune est constituée d'un assemblage de modules offrant un service particulier.

La section « Administrateur », destinée au responsable du test, comprend les modules suivants :

- Le module « gestion du contenu » du test permet a) d'intégrer les cas cliniques et leurs questions ; b) de personnaliser les pages d'accueil des participants, de créer des consignes à lire avant de commencer le test et des messages de fin de test ; c) de paramétrer le test en mode de simple mesure (le système se contente de donner un score) ou en mode informatif et qui affiche (en instantané ou en différé) les résultats, une comparaison avec le reste du groupe, des références vers la littérature pertinente. Un temps limite pour passer le test ou des dates de début et de fin de période de passage du test peuvent être également déterminés.

1. Statistiques Canada, 2005 : Utilisation d'Internet par les individus à des fins éducatives http://www40.statcan.ca/l02/cst01/comm22_f.htm

- Le module « gestion des participants » permet d'intégrer les candidats à évaluer et les membres du panel de référence. L'appartenance d'un individu à l'un ou l'autre groupe peut être modifiée, avec possibilité de retour dans le groupe d'origine. La compilation des résultats est ajustée instantanément (pour des motifs de recherche, il peut être intéressant d'étudier l'effet du passage d'un individu ou d'un sous-groupe dans le panel ou vice versa). Il est également possible d'envoyer des messages courriel pour contacter individuellement ou collectivement les participants afin d'annoncer les modalités de déroulement du test ou d'effectuer un rappel aux participants n'ayant pas terminé le test.
- Le module « gestion du déroulement de l'activité » verrouille la configuration choisie et assure l'enregistrement des réponses des participants. Il produit a) un rapport concernant les réponses enregistrées pour les participants (bilan des « Questions ayant reçu une réponse » ou « Questions restantes, « Candidats à évaluer » ou « Membres du panel de référence ») ; b) la clé de correction détaillée élaborée à partir des réponses des membres du panel de référence ; c) les données de mesure pour chaque participant ; d) les données concernant l'ensemble des participants ; e) l'historique de navigation des participants, incluant le temps passé sur chaque question ; f) la communication des résultats aux participants après qu'ils ont finalisé leur test.
- Le module « téléchargement/impression des données » permet de télécharger les données dans les logiciels MS Excel ou SPSS (pour effectuer des traitements de données non réalisables par le logiciel TCS) ou de réaliser une impression en format PDF.

La section « participant » comprend les modules suivants :

- Le module « passage du test » affiche pour les participants (panel ou candidats) les cas et les questions selon les paramètres définis par l'administrateur (en ordre préétabli ou en ordre aléatoire, avec ou sans restriction de temps ou de date de passage du test, etc.). Si les questions ont été entrées dans deux langues, français et anglais par exemple, il est possible de basculer de l'une à l'autre facilement.
- Le module de « compilation personnalisée des résultats » montre les résultats du test à chaque candidat, en affichant une comparaison avec la performance du groupe entier. Des références vers la littérature pertinente peuvent être affichées si le mode « informatif » est activé.

Cette activation peut être faite en différé de façon à ce que les candidats puissent y revenir à une date ultérieure pour en bénéficier.

Résultats et perspectives liées à l'informatisation du test

L'ordinateur facilite de beaucoup toutes les étapes de réalisation d'un test de concordance, soit la construction, la passation par les candidats et la compilation des résultats. Il procure une économie de temps, tant aux constructeurs du test qu'aux personnes évaluées, et permet de communiquer immédiatement les résultats aux participants. Ces résultats peuvent être enrichis de liens vers des moyens de formation variés en lien avec les questions dont la réponse était mauvaise. Le succès de la mise en ligne du test est certainement dû à sa facilité d'utilisation et à la rapidité avec laquelle, du fait de l'informatisation, il est possible pour un responsable de test de réaliser toutes les étapes de la production d'un test en ligne. La logistique de mise en œuvre est ainsi reléguée à l'arrière-plan, ce qui permet au responsable de se concentrer davantage sur les éléments cruciaux de création d'un TCS de qualité, tels que la composition de cas/questions de qualité ou la sélection des membres du panel de référence.

Le système est adapté aux possibles imprévus et erreurs ; des messages sont alors affichés à l'administrateur afin qu'il puisse intervenir, ou simplement prendre en considération des cas de figure lors des interprétations finales des résultats. En cas de modification du panel, par exemple, la compilation s'ajuste automatiquement aux nouvelles réponses enregistrées et affiche instantanément les résultats modifiés. La possibilité de faire passer le test en ligne aux membres du panel comme aux candidats évite d'avoir à les regrouper physiquement, ou de le faire par vagues successives, ce qui est particulièrement bien adapté au contexte nord-américain, où les médecins sont peu disponibles.

L'utilisation de l'ordinateur permet de faire passer le test en ligne à tous les candidats d'un pays. Sibert et ses collaborateurs(Sibert, Darmoni, Dahamna, Weber et Charlin, 2005) ont réalisé une expérience de passation d'un test de concordance en ligne pour tous les urologues français. Leur panel comportait 26 membres, et la population testée était composée de 100 urologues, 100 résidents et 50 étudiants. Tous les participants remplissaient le test à partir d'un site Internet. L'intention des auteurs était d'aider les urologues à maintenir leur qualité de raisonnement clinique en se testant à partir de leur lieu de pratique ou de leur domicile. Le test de concordance, conçu pour exposer les participants à des tâches de la vie professionnelle réelle dans toute leur complexité et vérifier dans quelle mesure leurs décisions sont conformes à celles prises par les membres d'un panel de référence,

permet d'évaluer des domaines qui étaient jusqu'à maintenant difficilement évaluables de façon standardisée et objective. Citons, par exemple, la simulation du raisonnement préopératoire des chirurgiens (Meterissian, Zabolotny, Gagnon et Charlin, 2006).

L'utilisation du multimédia

L'ordinateur permet également d'enrichir la situation d'évaluation en présentant aux candidats des tâches cognitives impliquant images, sons, ou enregistrements vidéo. Chalk et Lubarsky (2007) ont ainsi mis au point un test d'évaluation du raisonnement diagnostic en neurologie qui amène les candidats à raisonner en face de scanners cérébraux, d'images cliniques ou encore de films vidéo rapportant des symptômes complexes, tels que tremblement ou mouvements automatiques des yeux. L'utilisation du multimédia permet ainsi de rapprocher la situation d'évaluation de la complexité des situations cliniques réelles. Le site Internet http ://www.fpcmed2.umontreal.ca/sctdemo/ permet de visualiser la richesse des stimuli auxquels sont exposés les participants au test.

Par ailleurs, les diagnostics dans les spécialités médicales basées sur l'image (radiologie, anatomo-pathologie, dermatologie) impliquent des habiletés perceptuelles, largement non analytiques. À l'exception de cas simples, la perception des données varie selon les cliniciens. Cette variabilité rend difficile la mesure de cette capacité avec les outils d'évaluation habituels. L'utilisation de l'ordinateur et la possibilité d'enregistrer automatiquement les opérations avec le curseur des personnes qui passent le test, associés à des méthodes innovatrices pour mesurer et attribuer un score de perception ouvrent de nouveaux champs d'évaluation et notamment la possibilité de créer un test de perception. Le test comporte quinze clichés correspondant à des présentations communes et aux objectifs éducationnels au niveau de l'externat et de la résidence. Les coordonnées des « clics » enregistrés sont comparées à celles d'un panel de quinze radiologistes expérimentés pour déterminer si l'anomalie a bien été détectée et produire des scores de perception (exemple dans le site http ://www.fpcmed2.umontreal.ca :8080/sctimagesdemo3/).

Perspectives de développement

La mise en ligne des TCS a permis d'établir des partenariats de recherche ou de passation de test avec différentes universités et différents pays. Les contacts avec ces partenaires ont fait apparaître de nouveaux besoins, tels que la personnalisation des interfaces pour chaque lieu de test ou la possibilité

de sauvegarder un TCS finalisé dans un format permettant de le réintroduire ultérieurement d'un clic du curseur et de démarrer ainsi une nouvelle session avec les mêmes participants ou avec une cohorte différente.

L'intégration de la technologie dite « *Web Services* » permettra de construire et de gérer un test de façon automatisée et sécuritaire sur des serveurs Internet appartenant à d'autres universités. Ce sont ces autres serveurs qui décideront de la présentation du résultat d'un traitement quelconque aux participants. Il sera alors possible à nos partenaires à distance de personnaliser leurs projets (langage utilisé, logos, charte graphique de l'institution, masquage de certaines informations ou au contraire accentuation du traitement avant l'affichage final).

Enfin, nous travaillons à appliquer les technologies de mesure de la perception, mesure effectuée actuellement sur des images fixes et sur des films vidéo interactifs. Cette fonctionnalité nouvelle permettra de donner des scores de perception dans des domaines difficilement mesurables jusqu'à présent, tels les endoscopies, l'échographie, le CT-Scan dynamique ou des interventions chirurgicales filmées. Le participant pourra avancer ou reculer dans la vidéo présentée, puis sélectionner à un moment donné une zone d'anomalie perçue, ce qui permettra de générer des scores de perception dans ces domaines d'imagerie dynamique.

CONCLUSION

Le test de concordance de script permet de confronter les participants à des situations authentiques reflétant la complexité de leur pratique professionnelle, tandis que le processus de mise en forme du score conduit à des résultats à la fois valides et fiables. Le test s'inscrit tout à fait dans le contexte éducatif actuel des sciences de la santé, centré sur le développement des compétences. L'utilisation de l'ordinateur en facilite la construction, la passation et la correction. L'ordinateur permet également d'introduire des stimuli multimédias qui enrichissent la situation d'évaluation et d'aborder des habiletés (par exemple la perception sur images de données anormales) difficilement mesurables de façon pratique et standardisée jusqu'à maintenant. Enfin, la récolte des données en temps réel, plutôt qu'en temps différé comme avec la technologie papier-crayon, permet d'informer immédiatement les candidats sur la qualité de leur performance et de leur proposer des pistes de formation adaptées aux lacunes détectées. L'expérience acquise dans les sciences de la santé pousse à recommander son utilisation dans toute autre profession qui implique de raisonner en contexte d'incertitude.

Références

Barrows HS, et Feltovich PJ. (1987), « The clinical reasoning process ». Medical Education, 21(2), p. 86-91.

Brailovsky, C., Charlin, B., Beausoleil, S., Coté, S. et Van der Vleuten, C. (2001), « Measurement of Clinical Reflective Capacity Early in Training as a Predictor of Clinical Reasoning Performance at the End of Residency : An Exploratory Study on the Script Concordance Test », *Medical Education*, 35, p. 430-436.

Carrière, B. (2005), *Development and initial validation of a script concordance test for residents in a pediatric emergency medicine rotation*, Mémoire de maîtrise non publié, Université de Illinois à Chicago.

Case, S. M. et Swanson, D. B. (2001), *Constructing written test questions for basic and clinical sciences* (3e édition.), Philadelphie, The National Board of Medical Examiners.

Chalk, C. et Lubarsky, S. (2007), SCT on neurology. Montréal, Université McGill.

Charlin, B., Boshuizen, H. P. A., Custers, E. J. F. M. et Feltovich, P. (2007), « Scripts and clinical reasoning », *Medical Education, Submitted*.

Charlin, B., Desaulniers, M., Gagnon, R., Blouin, D. et van der Vleuten, C. (2002), « Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity », *Teaching and Learning in Medicine, 14*(3), p. 150-156.

Charlin, B., Gagnon, R., Sauvé, E. et Coletti, M. (2006), « Composition of the panel of reference for concordance tests : do teaching functions have an impact on examinees' ranks and absolute scores ? », *Medical Teacher*.

Charlin, B., Gagnon, R., Sibert, L. et Van der Vleuten, C. (2002), « Le test de concordance de script : un instrument d'évaluation du raisonnement clinique », *Pédagogie Médicale*, 3, p. 135-144.

Charlin, B., Roy, L., Brailovsky, C. et Van der Vleuten, C. (2000), « The Script Concordance Test, a Tool to Assess the Reflective Clinician », *Teaching and Learning in Medicine*, 12, p. 189-195.

Charlin, B., Tardif, J. et Boshuizen, H. P. A. (2000), « Scripts and Medical Diagnostic Knowledge : Theory and Applications for Clinical Reasoning Instruction and Research », *Academic Medicine*, 75, p. 182-190.

Charlin, B. et van der Vleuten, C. (2004), « Standardized assessment of reasoning in contexts of uncertainty : the script concordance approach », *Evaluation & The Health Profession*, 27(3), p. 304-319.

Dinant, G. J. (2004), « Diagnosis and decision making : Undifferentiated illness and uncertainty in diagnosis and management », dans, R. Jones, N. Britten, L. Culpepper et coll. (Ed.), *Oxford Textbook of Primarry Care* (Volume 2), Oxford : Oxford Universitity Press, p. 201-203.

Elstein AS, Shulman, L.S. et Sprafka, S.A. (1978), Medical Problem Solving : An Analysis of Clinical Reasoning, Cambridge, MA, Harvard University Press.

Elstein, A. S., Shulman, L. S. et Sprafka, S. A. (1990), « Medical Problem Solving, a Ten-Year Retrospective », *Evaluation & The Health Profession*, 13, p. 5-36.

Fournier, J. P., Demeester, A. et Charlin, B. (soumis), « Script Concordance Tests : Guidelines for construction », *Medical Teacher*.

Fournier, J. P., Thiercelin, D., Pulcini, C., Alunni-Perret, V., Gilbert, E., Minguet, J.-M. et al. (2006), « Évaluation du raisonnement clinique en médecine d'urgence : les tests de concordance des scripts décèlent mieux l'expérience clinique que les questions à choix multiples à contexte riche », *Pédagogie Médicale*, 7, p. 20-30.

Gagnon, R., Charlin, B., Coletti, M., Sauvé, E. et van der Vleuten, C. (2005), « Assessment in the context of uncertainty : How many members are needed on the panel of reference of a script concordance test ? », *Medical Education*, 39, p. 284-291.

Gerrity, M. S., Earl, J. A. L. et deVellis, R. F. (1992), « Uncertainty and professional work : Perceptions of physicians in clinical practice », *Motivation and Emotion*, 97, p. 1022-1051.

Grant J, Marsden P. (1988), « Primary knowledge, medical education and consultant expertise », *Medical Education*, 22, p. 173-179.

Hall, K. H. (2002), « Reviewing intuitive decision-making and uncertainty : the implications for medical education », *Medical Education* 36, p. 216-224.

Lambert, C. (2005), *Le test de concordance de script : étude de validation d'un nouvel outil d'évaluation du raisonnement clinique des résidents en radio-oncologie*. Mémoire de maîtrise inédit, Université de Montréal.

Meterissian, S., Zabolotny, B., Gagnon, R. et Charlin, B. (soumis), « Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills ? », *American Journal of Surgery*.

Schön, D. (1983), *The Reflective Practitioner : How Professionals Think in Action*, New York, Basic Books.

Sibert, L., Darmoni, S. J., Dahamna, B., Weber, J. et Charlin, B. (2005), « Online clinical reasoning assessment with the Script Concordance test : a feasibility study », *Biomedcentral, Medical Informatics and Decision Making*, 5, p. 18-28.

van der Vleuten, C. P. et Newble, D. I. (1995), « How can we test clinical reasoning ? », *Lancet*, 345, p. 1032-1034.

CHAPITRE 8

L'évaluation du profil psychosocial de demandeurs d'emploi, une population peu familiarisée à la technologie informatique[1]

Claude Houssemand, Raymond Meyers et Romain Martin
Université du Luxembourg

INTRODUCTION

Le testing assisté par ordinateur (TAO) est devenu au cours des dernières décennies d'utilisation courante auprès de multiples populations (Alkhader, Anderson et Clarke, 1994 ; Bartram et Bayliss, 1984 ; Butcher, Keller et Bacon, 1985). Nous utiliserons le terme de TAO dans le sens de toute passation automatisée par ordinateur d'une série d'items/questions auxquels un sujet répond individuellement. Le concept est donc plus large que la passation de « tests » au sens usuel du terme. Dans cette optique, le TAO pourra comporter la passation d'un questionnaire (voir Erdman, Klein et Greist, 1985 ; Kleinman, Leidy, Crawley, Bonomi et Schoenfeld, 2001), d'une échelle d'attitude ou de personnalité (Butcher, Keller et Bacon, 1985) ou d'un test de performance à proprement parler (par exemple Elwood et Griffin, 1972, pour une étude déjà ancienne).

Cette généralisation de l'usage des outils informatisés ne doit pas faire oublier un certain nombre de biais pouvant créer des différences artificielles entre les sujets. Dès les premières expérimentations du TAO, la familiarité avec l'outil informatique et l'anxiété envers l'ordinateur ont été mises en

1. Cette recherche a bénéficié d'un financement du Fonds National de la Recherche (FNR) luxembourgeois dans le cadre du programme VIVRE, projet n° FNR/02/05/07.

avant comme de potentielles sources de variation dans les performances enregistrées. Une foule de chercheurs s'est alors évertuée à comparer les passations papier-crayon avec les passations par ordinateur ou par Internet, afin d'examiner si ces différentes versions des mêmes tests donnaient bien des résultats équivalents. Leurs nombreuses études se sont concentrées généralement sur des populations relativement homogènes. Subsiste alors la question des personnes culturellement et cognitivement défavorisées. L'utilisation de l'outil informatique auprès de publics défavorisés et peu familiarisés avec cet outil ne risque-t-elle pas d'induire un biais substantiel en créant une difficulté supplémentaire pour ces personnes, difficulté que les personnes utilisant l'ordinateur au jour le jour ne connaîtraient pas ? Outre les écarts de résultats « intrinsèques », ne risque-t-on pas de voir s'y ajouter des différences supplémentaires dues à la non-maîtrise d'un outil nouveau, déroutant et compliqué ? L'ordinateur ne risque-t-il pas d'induire une anxiété vis-à-vis des tests (Zeidner, 1998) ou une anxiété informatique (Rosen, Sears et Weil, 1987) particulière faussant les résultats ? La conséquence serait alors une exacerbation de la « fracture numérique » (ou *digital divide*) qui opposerait ceux qui possèdent et utilisent un ordinateur et ceux qui ne les utilisent pas et les évitent, à savoir les groupes à niveaux d'éducation et de revenu bas (Wilson, Wallinet Reiser, 2003 ; Looker et Thiessen, 2003). Quels seraient alors les aménagements à éventuellement prévoir pour que l'outil informatique perde son côté anxiogène ou étranger et devienne accessible à ces publics défavorisés ?

La suite de ce chapitre est divisée en cinq parties. Dans la première partie, nous discuterons, à partir d'une revue de la littérature, la question générale de l'équivalence entre scores obtenus lors d'évaluations où les données sont récoltées sous le mode papier-crayon et lors d'évaluations où elles sont récoltées sous le mode écran-clavier. Dans la deuxième partie, nous ferons une brève revue de la question de l'anxiété vis-à-vis des tests par rapport à l'usage du TAO auprès des populations défavorisées. La troisième partie traite de l'anxiété informatique en lien avec les résultats d'évaluations faites par ordinateur, aussi bien pour les publics hétéroclites que particulièrement défavorisés. Dans la quatrième partie, nous présenterons un outil informatisé pour la prise de données et d'évaluation que nous avons utilisé dans les services publics de l'emploi (SPE) luxembourgeois avec une cohorte de 381 demandeurs d'emploi, dont la moitié n'avait aucune qualification. Finalement, au cinquième point, nous présenterons quelques données psychométriques permettant de discuter de la validité des scores obtenus avec cet outil informatisé auprès de cette population culturellement et socialement défavorisée. Dans la conclusion, nous synthétiserons les acquis de ces cinq points et formulerons quelques recommandations pour l'usage du TAO auprès de populations défavorisées.

COMPARAISON ENTRE SCORES OBTENUS EN MODE PAPIER-CRAYON ET EN MODE INFORMATISÉ

Avant d'aborder la problématique spécifique du TAO auprès de groupes défavorisés, nous allons traiter d'une première question d'importance. Les passations informatisées consistent souvent en la mise sur ordinateur d'un questionnaire, d'une échelle ou d'un test de performance qui préexistait sous une forme papier-crayon. La validation de l'outil s'est généralement faite sous cette forme, et les normes ont été également établies pour cette version. Se pose alors la question de la validité de la transposition de la version papier sur ordinateur, ainsi que de celle des normes qui lui sont afférentes. Depuis environ trois décennies, les études comparant les versions papier-crayon aux versions informatisées des mêmes outils se sont multipliées au point même de rendre difficile toute revue un peu systématique de la question. Heureusement, pour certains outils, des revues et des méta-analyses des recherches publiées ont été faites et permettent des synthèses.

Les célèbres lignes directrices de l'*American Psychological Association* (APA, 1986) mettent en avant trois séries de critères pour déterminer l'équivalence de la version papier et de la version informatisée d'un test : 1) les statistiques descriptives se rapportant aux moyennes, aux variances, aux distributions et à l'ordre de succession des scores ; 2) la validité de construit ; et 3) la fidélité. Bartram (1994) donne quatre critères de l'équivalence entre différentes formes de tests. D'après lui, deux formes sont identiques si elles ont la même fidélité, corrèlent fortement entre elles, ont des corrélations comparables avec d'autres variables, et donnent des moyennes et des écarts-types identiques. Cette conception est d'ailleurs proche de la définition de tests parallèles dans la théorie du score vrai.

Tests d'attitudes, de personnalité et d'opinions

Une première série d'études se rapporte à des questionnaires de prise d'informations, d'évaluation d'attitudes ou de personnalité. La plupart des études ont trouvé une équivalence psychométrique entre les deux versions. Par exemple, Kleinman, Leidy, Crawley, Bonomi et Schoenfeld (2001) ont montré qu'un questionnaire de qualité de vie donné à des patients souffrant de reflux gastro-oesophagien donnait les mêmes résultats avec la version informatisée qu'avec la version papier-crayon. Donovan, Drasgow et Probst (2002) ont montré qu'un questionnaire de satisfaction au travail donnait les mêmes résultats dans la version informatisée que dans la version papier-crayon. Lumsden, Sampson, Reardon, Lentz et Paterson (2004) ont testé trois versions du *Self-Directed Search* de Holland (à savoir les versions papier, ordinateur et Internet), un outil d'auto-évaluation des intérêts professionnels

qui est très largement utilisé. Ils n'ont pas trouvé de différences statistiquement significatives entre ces trois versions du même outil.

Un instrument qui est beaucoup utilisé en psychologie clinique est le *Minnesota Multiphasic Personality Inventory* (MMPI), un questionnaire recherchant un ensemble de symptômes morbides et prémorbides et comportant un grand nombre de sous-échelles. Cet inventaire a fait l'objet d'études d'équivalence particulièrement nombreuses, qui ont été synthétisées dans une méta-analyse récente (Finger et Ones, 1999). Les auteurs concluent à l'équivalence des deux versions, conclusion d'ailleurs reprise par une étude plus récente portant sur le MMPI Adolescents (Hays et McCallum, 2005).

Un problème qui a beaucoup occupé les chercheurs est celui de la désirabilité sociale et de la gestion de l'image de soi (*impression management* ou *self-presentation*) au cours des tests de personnalité et d'attitude. Comme les sujets ont généralement une idée de ce qui est mieux vu, et que rien ne les oblige à être « honnêtes », ne vont-ils pas sciemment ou inconsciemment se présenter de façon plus positive, et cela, tout particulièrement si le test est utilisé à des fins de sélection et de recrutement ? Et surtout, cette distorsion est-elle plus forte avec une version informatisée qu'avec une version papier ? Il semble que le contraire soit le cas, les études montrant que, s'il y a un effet, il y a moins de distorsion avec l'ordinateur (Martin et Nagao, 1989), les sujets ayant tendance à être plus « honnêtes » (Meinert, Vitell et Blankenship, 1998) lors de passations informatisées. Deux méta-analyses ont confirmé ces résultats à travers un large échantillon d'études (Richman, Kiesler, Weisband et Drasgow, 1999 ; Dwight et Feigelson, 2000). En raison sans doute d'une plus grande banalisation de l'usage de l'ordinateur, cet effet aurait tendance à diminuer (Dwight et Feigelson, 2000). Une méta-analyse a également été conduite sur les études comparatives portant sur la « révélation de soi » lors de passations informatisées comparées à des interviews et des questionnaires-papier (Weisband et Kiesler, 1996). Elle montre une plus forte tendance à la confidence en face de l'ordinateur, tendance également en évidence dans les études plus récentes.

Tests de performance ou d'aptitudes

Même si l'équivalence des versions informatisées et papier-crayon semble généralement acquise, qu'en est-il des tests de performance ? Si on peut comprendre que faire une croix derrière une question ou cliquer dans une bulle ne fasse pas de différence notable, qu'en est-il des épreuves qui demandent d'apporter une preuve de compétences ? On pense évidemment aux traditionnels tests d'intelligence, qui mesurent des aptitudes cognitives, mais

également aux épreuves de vitesse. Selon la nature du test, on peut distinguer les tests de puissance (difficulté progressive, sans limites de temps) et les tests de vitesse (problèmes simples, temps limité). Cette distinction est quelque peu théorique, les tests de puissance ayant toujours en pratique une limite de temps pour que la passation ne s'éternise pas. Dans le premier type de test (comme le WAIS, le WISC ou les Matrices de Raven), le sujet doit trouver la bonne réponse puis faire une croix sur le papier ou cliquer dans une bulle dessinée sur l'écran. Il est plausible que le temps de cocher ou de cliquer soit négligeable par rapport à la recherche de la bonne solution et ne fasse donc pas de différence. Par contre, un test de vitesse propose des problèmes simples avec un temps limité consistant, par exemple, à barrer certaines figures bien précises dans un alignement de figures diverses.

Très tôt, dans une des premières études traitant de la question, Elwood et Griffin (1972) avaient mis au point et testé une version complètement automatisée du WAIS qui avait montré une bonne fidélité. De très nombreuses études se sont ensuite attelées à examiner l'équivalence des versions papier et informatisée des tests de performance. Ces recherches ont été réunies dans une méta-analyse (Mead et Drasgow, 1993) montrant qu'il n'y avait pas de différences pour les tests de puissance même limités dans le temps, mais qu'il y en avait une pour les tests de vitesse (voir aussi Bodmann et Robinson, 2004, pour une étude plus récente). En effet, pour Bodmann et Robinson, ce n'est pas seulement le raisonnement, mais surtout la vitesse d'exécution qui compte, et donc de petites différences dans la présentation des items et dans la production des réponses peuvent, par un effet cumulatif sur l'ensemble de la passation, affecter les scores finaux (Mead et Drasgow, 1993). Cela dit, les tests de vitesse, jadis fréquemment utilisés dans le recrutement de personnel de bureau, ont tendance à perdre de l'importance au profit d'épreuves impliquant des processus cognitifs plus complexes.

Bilan de l'équivalence des versions

Comme nous avons pu le montrer, nous pouvons considérer aujourd'hui la question de l'équivalence entre version papier-crayon et version informatisée comme largement résolue. Les études qui ont été menées l'ont généralement été sur des populations hétéroclites, ce qui est logique, car il fallait montrer l'équivalence des deux versions pour la population générale. Toutefois, nous n'avons connaissance que de très peu d'études de ce type sur les populations défavorisées[2].

2. Une recherche déjà fort ancienne (Johnson et Mihal, 1973), donc dotée d'une technologie d'une autre époque, comparait les résultats entre une version papier et une version informatisée d'un test

Nous pouvons donc conclure que les passations par ordinateur sont largement équivalentes aux passations papier-crayon pour tout public, du moment qu'on a rendu la version informatisée semblable à la version papier. Cela vaut aussi bien pour les questionnaires, les échelles de personnalité et d'attitude que pour les épreuves de performance du type des tests d'intelligence. Il semble même que l'ordinateur induit une plus grande honnêteté, une plus grande tendance à se révéler et une distorsion moins importante due à la désirabilité sociale. Comme nous l'avons vu, des exceptions à l'équivalence des résultats sont à souligner pour certains types d'épreuves de performance impliquant surtout la vitesse. Nous n'avons pas non plus traité de la question du testing adaptatif par ordinateur (Raîche, 2004) basé sur la théorie de réponse à l'item (TRI), celui-ci impliquant une procédure uniquement informatisée qui donne le choix des items proposés aux sujets en fonction de leurs réponses antérieures (voir le chapitre 3).

L'ANXIÉTÉ EN FACE DES TESTS, ET LE TAO AUPRÈS DES POPULATIONS DÉFAVORISÉES

L'anxiété face aux tests est un phénomène très largement répandu à travers toutes les cultures et toutes les couches de la société (pour une revue très extensive, voir Zeidner, 1998). Trois questions théoriques et méthodologiques sont importantes pour notre propos, et nous tenterons d'y répondre à partir de la revue de la littérature : 1) l'anxiété face aux tests est-elle plus importante dans les groupes défavorisés de la société ?, 2) l'anxiété face aux tests biaise-t-elle les résultats aux tests chez ces populations ?, et 3) cette anxiété influence-t-elle la performance à une passation écran-clavier, comparativement à une passation papier-crayon traditionnelle ?

La revue de la littérature effectuée par Zeidner (1998) conclut effectivement à une relation inverse, mais d'amplitude assez modeste, entre le statut socio-économique et l'anxiété face aux tests (p. 272). Dans différentes cultures, on a pu montrer que les personnes issues d'un milieu socio-économique inférieur éprouvent une plus grande anxiété devant les tests que leurs compatriotes des classes moyennes et supérieures. Zeidner (1990) a également réalisé une des rares études ayant systématiquement évalué l'éventuel biais socioculturel introduit par ce niveau d'anxiété sur les performances aux tests. Aux résultats au test d'admission utilisé dans les universités

d'aptitude scolaire pour des Blancs et des Noirs aux États-Unis. Les Noirs ont généralement un niveau socio-économique moins élevé que les Blancs. Les auteurs ont trouvé que la passation informatisée améliorait les scores des Noirs dans le sous-test verbal.

israéliennes, aucun effet d'interaction n'est apparu pour les variables démographiques testées (sexe, classe sociale et groupe ethnique).

Finalement, l'anxiété vis-à-vis des tests influence-t-elle les performances lors d'une passation informatisée ? Les résultats de la littérature sont contradictoires. Llabre et ses collaborateurs (1987) avaient trouvé un effet plus important de l'anxiété sur les performances lors des tests d'intelligence informatisés que ceux de la version papier, et cela, pour des sujets ayant peu d'expérience de l'ordinateur par rapport à ceux qui en avaient beaucoup. Une étude plus récente de Powers (2004) a par contre trouvé le même niveau d'anxiété et les mêmes performances avec une version informatisée qu'avec une version papier du *Graduate Record Examination – General Test*, une épreuve d'admission couramment utilisée dans les universités américaines.

La question de l'anxiété aux tests a été récemment relancée par les recherches en psychologie sociale qui portent sur la menace du stéréotype (*stereotype threat*), à la suite de l'étude séminale de Steele et Aronson (1995). Quand une personne se trouve dans une situation où un stéréotype qui concerne son groupe d'appartenance est activé, la personne développe des préoccupations concernant ce stéréotype qui diminuent sa performance. C'est par exemple le cas des femmes dans un test de mathématiques, ou des Noirs américains lors d'une épreuve d'intelligence. Au moins deux conditions doivent être remplies pour que l'effet apparaisse. D'abord, la personne doit appartenir à un groupe socialement déterminé vis-à-vis duquel le stéréotype a cours. Ensuite, la personne doit se sentir concernée et identifiée par rapport à l'enjeu du test en question (par exemple, les Noirs américains par rapport à un test d'aptitude cognitive mais pas par rapport à un test de performance athlétique, les femmes par rapport à un examen de mathématiques mais pas par rapport à une épreuve de langues). Jusqu'à maintenant, le phénomène a été exclusivement démontré dans des situations de laboratoire, et des tests portant sur des données issues de situations de la vie courante n'ont pas montré de généralisation de cet effet à ces contextes. Il se pourrait que l'effet soit dû, en partie du moins, aux particularités des études en laboratoire (Cullen, Hardison et Sackett, 2004). Il faudra être attentif à ces développements en psychologie sociale et en tenir éventuellement compte quand des groupes « stéréotypisés » sont concernés, mais de toute façon, il n'y a jusqu'à présent aucune donnée permettant de dire que l'effet soit plus ou moins important lors de passations informatisées.

En conclusion, l'anxiété face aux tests ne semble pas être un facteur qui empêche l'utilisation du TAO auprès de populations défavorisées ou qui introduirait nécessairement des biais faussant les résultats.

ANXIÉTÉ INFORMATIQUE ET ÉVALUATIONS PAR ORDINATEUR

Avec la mise au point de différentes variantes du TAO, la question de l'anxiété informatique n'a pas manqué de se poser. L'anxiété informatique consiste en des émotions négatives déclenchées par l'évocation ou l'anticipation d'un usage de l'ordinateur qui ont l'évitement comme manifestations comportementales. Cette anxiété est un phénomène largement répondu, malgré la prolifération des ordinateurs à tous les aspects de la vie économique et sociale (Bozionelos, 2004). Cette anxiété informatique a été de façon consistante associée à l'expérience avec les ordinateurs (Beckers et Schmidt, 2003), les personnes utilisant moins l'ordinateur étant aussi celles qui éprouvent le plus d'anxiété informatique.

L'anxiété informatique affecte-t-elle les résultats au TAO ?

Pour répondre à cette question, nous devons de nouveau distinguer épreuves d'attitude et de personnalité et épreuves de performance. Pour les premières, la réponse doit être nuancée. En effet, des études ont montré que certaines attitudes par rapport à l'ordinateur sont influencées par l'anxiété informatique, c'est le cas notamment de la perception qu'ont les sujets quant à leurs compétences informatiques (Beckers, Rikers et Schmidt, 2006). Dans ce cas, les sujets anxieux ont des scores plus faibles sur leur perception de compétence. Par ailleurs, même des traits de personnalité tout à fait généraux, comme c'est le cas des *Big Five,* sont corrélés à l'anxiété informatique. Ainsi, des personnes anxieuses ont tendance à avoir un niveau plus élevé de névrosisme et un niveau moins élevé d'ouverture et d'agréabilité (Korukonda, 2007). Cependant, ces résultats différentiels selon les attitudes ou les traits de personnalité ne sont pas à considérer comme des biais, mais plutôt comme des corrélations entre différents traits de personnalité, parmi lesquels l'anxiété informatique.

Qu'en est-il des tests de performance et d'un biais potentiel dû à l'anxiété informatique ? Si Beckers, Rikers et Schmidt (2006) n'ont pas trouvé d'effet, même sur des tâches informatisées complexes, il est vrai que l'étude a été réalisée avec des étudiants universitaires, et que ceux-ci sont déjà habitués à des tâches plus complexes. Gaudron (1998) n'a pas trouvé d'effet significatif, ni de la plus ou moins grande pratique de l'ordinateur, ni de l'anxiété informatique, sur des résultats à deux tests d'intelligence, l'un passé en version papier, l'autre en version informatisée. Le temps d'exécution a été influencé par l'expérience antérieure de l'ordinateur, mais pas par l'anxiété informatique.

Anxiété informatique et publics défavorisés

Comme les populations défavorisées sont moins utilisatrices de l'informatique, et que le manque d'expérience est lié à l'anxiété informatique, ces personnes pourraient avoir un niveau d'anxiété plus élevé (Bozionelos, 2004). Il faut cependant noter que, dans l'étude de Bozionelos, l'effet de l'origine socio-économique sur l'anxiété informatique était très faible, de même que l'effet de l'origine sociale sur le niveau de pratique de l'ordinateur. Cela dit, Brosnan (1998) a trouvé qu'il y avait bien un effet de l'anxiété informatique sur une tâche informatisée complexe, et qu'il pourrait donc y avoir un biais par rapport aux publics défavorisés, ces derniers ayant justement souvent plus de difficultés avec ces tâches. Par contre, rien ne permet d'affirmer que dans une tâche informatisée simple, comme cliquer des réponses plutôt que faire des croix (ne pas confondre cette simplicité avec la complexité des raisonnements qui sont à faire avant de cliquer), l'ordinateur joue en défaveur des publics défavorisés et biaise leurs réponses. En conclusion, rien dans la recherche ne semble indiquer que l'utilisation du TAO soit une source de biais pour les publics culturellement et économiquement défavorisés, à condition que le type de manipulation avec l'ordinateur soit maintenu à un niveau très simple – ce qui n'exclut évidemment pas de poser des questions à contenu complexe, puisque la manipulation n'interfère pas avec le contenu.

UN OUTIL INFORMATISÉ UTILISÉ AVEC DES CHÔMEURS

Mise au point de l'outil

Après cette revue de la littérature, nous allons exposer les détails d'une expérience de TAO réalisée au Luxembourg auprès d'un groupe de demandeurs d'emploi. Un critère important d'un outil utilisable avec ces personnes est celui de la facilité de manipulation. Cette population est très hétérogène, son niveau de scolarisation est souvent faible, et on peut supposer que ses niveaux d'aptitude intellectuelle sont très variables. Les statistiques montrent que les personnes à faible niveau de scolarité sont beaucoup plus représentées parmi les chômeurs que parmi la population active. En date du 31 décembre 2006, la part de la population des demandeurs d'emploi inscrits présentant un niveau de formation scolaire inférieur (ayant uniquement accompli la scolarité obligatoire – 9 années d'études) était de 51,3 % (ministère du Travail et de l'Emploi, 2007, p. 114). Il importe donc de prévoir des items qui présentent une grande facilité d'utilisation (réponses fermées plutôt qu'ouvertes, réponses à choix multiple plutôt que réponses libres, écran tactile plutôt que clavier, etc.).

Les variables choisies pour l'étude ont été créées sous forme de questionnaires. L'outil de la passation complète comprenait en fait trois types d'items. Un premier groupe comprenait des questions d'information (Quel est votre genre ?, Quelle est votre nationalité ?) ; une deuxième catégorie était constituée d'échelles psychométriques de personnalité et d'attitude comme le *General Health Questionnaire* (GHQ12), l'échelle d'estime de soi de Rosenberg ou une échelle de stress perçu ; et finalement un test d'intelligence non verbale, le B53 de Bonnardel, a été utilisé. Les items ont été informatisés à l'aide du logiciel *Quest Net+*. Le programme résultant de ce travail permet une passation du questionnaire sur ordinateurs portables dans les locaux de l'Administration de l'emploi (ADEM). Dans la figure 1, nous avons reproduit deux exemples d'items de notre questionnaire, le premier se rapportant au GHQ12, une échelle étudiant le mal-être perçu, le deuxième étant extrait du test d'intelligence B53. On peut voir que la personne ne voit jamais apparaître qu'une seule question ou, éventuellement, un paragraphe de consigne à l'écran. Dans la plupart des cas, une seule réponse est demandée, et le programme présente automatiquement la page suivante, ce qui est le cas pour les deux items présentés à la figure 1. Pour certaines questions, plusieurs réponses peuvent être fournies, et la personne clique sur une flèche en bas de l'écran pour passer à l'item suivant. Les sujets ne pouvaient passer à l'item suivant qu'après avoir répondu au précédent. Cette procédure a également l'avantage de supprimer les réponses manquantes, ce qui est un atout majeur pour les échelles psychométriques. Pour certaines questions de renseignements ne faisant pas partie d'échelles, les sujets pouvaient répondre « Je ne sais pas » pour ne pas créer artificiellement des réponses qui n'auraient pas de sens. Le seul test de performance utilisé, le B53 a été placé à la fin de la séance, afin que l'ensemble de la passation ne soit pas interprété de manière erronée comme un test d'intelligence.

Pour la majorité des passations, nous avons eu recours à des ordinateurs portables de type tablette PC, où les personnes n'ont besoin que de cliquer avec un stylet sur l'écran tactile pour fournir leurs réponses. Il a été montré, dans des études comparatives sur l'anxiété informatique, que ce genre de dispositif n'affectait pas les scores dans le domaine de l'évaluation de l'humeur et de l'intelligence (Tseng, Tiplady, Macleod etWright, 1998). Ce point est important, car il est le garant de la comparabilité des résultats obtenus avec ceux provenant d'études pour lesquelles les questionnaires sont généralement sous une forme papier-crayon. D'autres études comparatives (Overton, Taylor, Zickar et Harms, 1996) ont montré que ce type d'outil garantit plus facilement l'équivalence des construits avec les tests développés initialement sous version papier-crayon. Certains chômeurs, plus à l'aise avec les outils informatiques, pouvaient répondre aux questionnaires sur des portables classiques (avec une souris au lieu du stylet) s'ils le souhaitaient. Cette

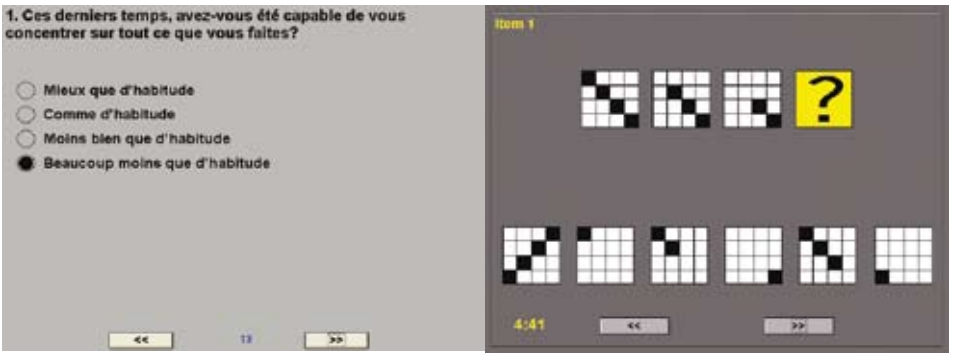


Figure 1 – Deux exemples d'items utilisés avec les chômeurs

précaution a été prise afin de ne pas infantiliser des personnes se sentant parfaitement à l'aise avec cet outil et rendant, de ce fait, leurs observations plus fiables.

Par ailleurs, le questionnaire était disponible en trois versions linguistiques différentes (français, allemand et portugais), les personnes choisissant leur langue au début de la session et pour toute la durée de celle-ci. Ce point est important, vu le multilinguisme très prononcé du Luxembourg : le français et l'allemand sont deux des trois langues officielles du Luxembourg (le luxembourgeois, étant peu lu et écrit, n'aurait été d'aucune aide, même pour les Luxembourgeois), et le portugais est la langue du principal groupe d'immigrés.

Mode de passation

La passation a eu lieu au cours de la période allant de juin 2005 à janvier 2006 auprès d'un échantillon de 381 demandeurs d'emploi nouvellement inscrits. Un premier échantillon de 33 chômeurs avait été interrogé à l'agence SPE de Luxembourg en juin 2005 afin de tester l'outil : ce test s'était avéré satisfaisant et n'avait nécessité que quelques corrections mineures des items. Notre échantillon était composé de 158 femmes et 226 hommes nouvellement inscrits comme chômeurs (Meyers et Houssemand, 2006). Les personnes étaient vues après leur premier entretien d'inscription avec leur placeur. Elles étaient volontaires pour passer le test, mais dans les faits, très peu de personnes à qui on l'a proposé ont refusé. L'âge des sujets variait de 15 à 64 ans. Le seul critère de sélection retenu était des connaissances minimales dans l'une des trois langues du questionnaire informatisé. En effet, les personnes devaient être capables de lire et de comprendre les questions, la réponse ne posant aucun problème supplémentaire, car il suffisait que la

personne clique avec un stylet sur l'écran tactile. Comme nous l'avons indiqué plus haut, certaines personnes plus familiarisées avec l'informatique répondaient grâce à la souris.

Une personne de l'Université était présente lors de toutes les sessions pour expliquer le but du questionnaire, lancer la passation, expliquer le maniement du stylet ou de la souris et répondre à d'éventuelles questions de sens. Dans ce dernier cas, l'intervention du collaborateur se limitait à donner des synonymes des termes non compris. L'anonymat était également garanti aux personnes, seul le numéro de dossier de l'ADEM ayant été noté pour pouvoir ensuite extraire anonymement des informations complémentaires de ce dossier. Ce numéro ne contient aucune information (comme la date de naissance) permettant d'identifier la personne. Les sessions duraient en règle générale entre trois quarts d'heure et une heure et quart.

Un outil utilisable avec des publics défavorisés

En conclusion, on peut noter que, même avec des personnes très peu familiarisées avec un outil de TAO, voire le découvrant, des mesures de différentes dimensions de personnalité, d'attitude et d'intelligence ont été réalisées sans rencontrer de difficultés majeures ou de refus. Il est évident qu'un certain nombre d'adaptations des instruments et du matériel ont été nécessaires, comme la mise à disposition de l'outil dans une langue familière à la personne, la simplicité des consignes et de la présentation, la non-nécessité de compétences informatiques, la présence d'une personne d'accompagnement. Mais il semble, à la lumière de nos résultats, que l'utilisation du TAO puisse être généralisée.

VALIDITÉ DES SCORES OBTENUS EN MODE INFORMATISÉ AUPRÈS D'UNE POPULATION DÉFAVORISÉE

Après la présentation de l'outil, nous allons exposer quelques résultats psychométriques afin de vérifier, sur la base des données empiriques, la qualité des réponses, du moins aux échelles psychométriques. En effet, on pourrait se demander si les personnes défavorisées n'ont pas répondu de manière aléatoire, ne voulant pas refuser la passation, mais ne comprenant néanmoins pas ce qu'on leur demandait de faire. Comme il n'y avait que des réponses fermées et un avancement automatique d'une question à l'autre, ces personnes auraient pu répondre de façon erratique.

Équivalence du GHQ12 pour les personnes sans compétence informatique

Afin de vérifier l'équivalence psychométrique des réponses reçues, nous avons comparé les résultats de deux groupes de personnes (établis sur la base du degré de familiarité avec l'ordinateur) selon une échelle qui n'est pas une échelle de performance. Le GHQ12 (Goldberg, 1972) est un outil d'autoévaluation très largement validé qui mesure le mal-être perçu d'une personne. Les analyses factorielles ont généralement montré deux ou trois facteurs (Shevlin et Adamson, 2005). Dans nos résultats, nous avons trouvé une corrélation de -0,01 (p = 0,92 ; NS) entre la mesure de l'intelligence non verbale (test B53) et les scores au GHQ12 pour l'ensemble de notre échantillon (Meyers et Houssemand, 2006), ce qui montre bien que l'échelle mesure une dimension indépendante des compétences cognitives. Nous avions également posé une question sur les compétences informatiques des demandeurs d'emploi. Les résultats sont les suivants : 24 % déclaraient ne pas savoir se servir d'un ordinateur, 29 % disaient s'en servir uniquement pour Internet et le courrier électronique, 44 % déclaraient avoir des compétences professionnelles en utilisation de l'ordinateur (traitement de texte, tableur, logiciel de comptabilité, gestion de bases de données, etc.), et 4 % se disaient informaticiens professionnels. On peut donc comparer les réponses des personnes qui n'ont aucune compétence informatique avec celles des autres sujets.

L'analyse des réponses des sujets montre qu'ils se sont appliqués et n'ont pas coché les modalités de réponses de manière aléatoire. En effet, il apparaît que seules trois personnes ont toujours donné la même modalité de réponse à l'ensemble des douze questions posées. Deux d'entre elles étaient des novices en informatique et l'autre utilisait déjà l'ordinateur. De plus, on constate que le score au GHQ12 n'est pas significativement différent pour les groupes caractérisés par les compétences informatiques ($F_{(1,376)} = 0,110$; p = 0,741 ; moyenne = 13,54 ; écart-type = 6,51). On peut donc affirmer qu'indépendamment de leur familiarité avec l'outil informatique, les sujets ont pu répondre d'une manière sensée aux questions posées.

Pour analyser la structure de l'échelle, deux facteurs ont été retenus (Politi, Piccinelli et Wilkinson, 1994). Le premier est un facteur de dépression et le second un facteur d'évitement de situations. L'analyse structurale a été réalisée avec le logiciel AMOS 6.0 en comparant un modèle comprenant l'ensemble des 381 sujets ayant répondu au questionnaire et un modèle identique, mais différenciant les sujets coutumiers de l'informatique et les novices (Arbuckle, 2005). La comparaison des valeurs des coefficients d'adéquation des modèles (une statistique du khi carré) ne permet pas de mettre en évidence de différence significative entre les deux structures (Δ khi carré = 60,9 ; Δ dl = 53 ; p>0,05). La structure obtenue pour l'ensemble des sujets est présentée dans la figure 2.

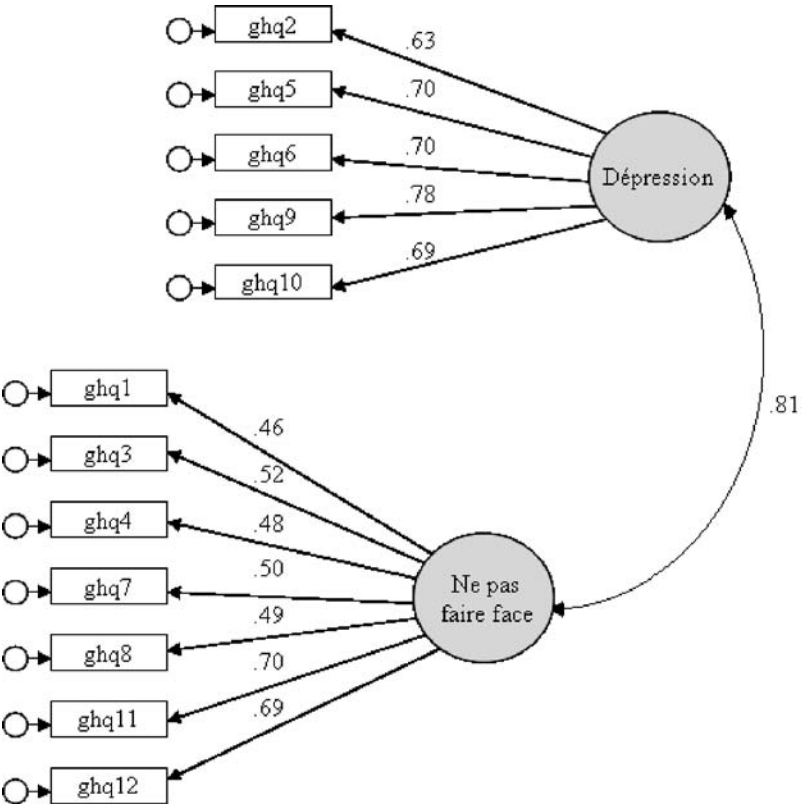


Figure 2 – Structure du GHQ12 pour l'ensemble de l'échantillon des chômeurs

On peut donc conclure, au vu de ces résultats, que les différences qui pouvaient être attendues entre personnes novices ou non en informatique ne sont mises en évidence par aucune des analyses réalisées ici. De ce fait, nous validons de manière indirecte que l'utilisation de l'informatique pour des questionnaires ou des tests ne défavorise aucune tranche des populations auxquelles ils sont proposés, à condition que les adaptations nécessaires à cette population aient été réalisées.

Équivalence du GHQ12 pour les non-qualifiés

Plutôt que de différencier les sujets selon leur niveau de familiarisation avec l'ordinateur, on peut le faire selon le niveau de qualification. Nous avons des données sur la formation initiale de notre échantillon, nous avons notamment calculé l'âge théorique qui correspond à leur niveau de formation.

En examinant la structure de l'enseignement luxembourgeois, on voit qu'aucun diplôme ne peut être acquis avant l'âge théorique de 17 ans. Nous pouvons séparer notre échantillon en deux groupes, ceux d'un âge théorique de formation de 17 ans ou moins (41 % de l'échantillon) et les autres (59 % de l'échantillon). Il faut préciser que 41 % est donc une estimation conservatrice du pourcentage de non-qualifiés, celui-ci étant forcément plus élevé. Nous pouvons alors comparer la structure du GHQ12 pour ces deux sous-populations. En procédant exactement de la même manière que précédemment avec les deux groupes selon l'expertise des sujets, on obtient des résultats absolument comparables. Aucune différence de score au GHQ12 ne peut être mise en évidence entre les sujets plus ou moins qualifiés ($F(1,339) = 0,085$; $p = 0,771$; moyenne = 13,49 ; écart-type = 6,49). Les légères différences observées ici par rapport à l'analyse précédente sont dues au fait que les informations relatives au parcours scolaire des personnes interrogées n'ont pu être obtenues que pour 341 des participants à l'enquête. De la même manière, il n'est pas possible de montrer une différence significative de structure de l'échelle du GHQ12 entre les personnes plus et moins qualifiées (Δ khi carré =20,1 ; Δ dl = 53 ; $p>0,05$). Les mêmes conclusions que précédemment peuvent donc être tirées de ces analyses complémentaires.

CONCLUSIONS ET RECOMMANDATIONS

Un faisceau convergent d'éléments nous permettent d'affirmer que le TAO est utilisable et socialement équitable auprès des populations défavorisées. D'abord, de façon générale, les versions informatisées et papier-crayon se valent, sauf pour certains tests de vitesse bien spécifiques. Une éventuelle anxiété face aux tests, même si elle est légèrement plus répandue dans les milieux défavorisés, n'affecte pas les résultats aux tests. Il en est de même pour l'anxiété informatique. Même si elle peut être plus importante chez des personnes défavorisées par manque de familiarisation avec l'outil, elle ne biaise pas les résultats aux tests. Une expérimentation pratique que nous avons réalisée avec un échantillon de chômeurs luxembourgeois n'a pas posé de problèmes pratiques, malgré le fait que presque la moitié de ces personnes n'avaient aucune qualification. L'étude psychométrique de certaines échelles n'a pas montré de différence de résultats pour les personnes ne sachant pas se servir d'un ordinateur par rapport aux autres, ni entre les personnes non qualifiées et les autres.

Néanmoins, des aménagements importants quant à la tâche informatisée doivent être réalisés. Les items doivent, si c'est possible, être présentés dans la langue maternelle ou bien dans une langue que les sujets maîtrisent bien. Les formats doivent être simples (réponses fermées plutôt qu'ouvertes, réponses à choix multiple plutôt que réponses libres, écran tactile plutôt que

clavier, etc.), un item doit être présenté après un autre afin de minimiser les distractions. Un encadrant doit être présent en permanence pour assurer que les personnes comprennent bien ce qu'on leur demande et pour intervenir en cas de problème ou de mauvaise compréhension technique. Par contre, aucune aide ne doit bien entendu être fournie pour le contenu des réponses. Par ailleurs, une limite du nombre d'items doit être prévue afin de ne pas surcharger les personnes peu familières avec ce genre de tâche, la fatigue venant d'ailleurs beaucoup plus de l'effort cognitif nécessaire pour répondre aux questions que du maniement du stylet.

Certaines limites à notre étude doivent être signalées. Nous n'avons pas testé notre type d'outil auprès d'autres populations défavorisées, et la généralisation de l'usage du TAO, même sous cette forme adaptée, ne peut être garantie auprès de populations plus fortement dysfonctionnelles (personnes analphabètes, illettrées, sans domicile fixe, ayant un long passé de déstructuration, etc.). Tout comme pour les outils papier-crayon, les outils informatisés exigent une aptitude à la lecture, ainsi qu'une aptitude à se concentrer un certain temps sur une tâche. Cependant, pour les illettrés, l'utilisation de supports auditifs peut être imaginée. Ensuite, notre type d'évaluation exclut bien évidemment un certain nombre de prises de données, en fait toutes celles qui ne peuvent pas être formatées sous forme fermée et à choix dichotomique ou multiple.

Références

Alkhader, Othman, Neil Anderson et David Clarke (1994), « Computer-based testing : A review of recent developments in research and practice », *European Work and Organizational Psychologist*, vol. 4, n° 2, p. 169-187.

APA (1986), *Guidelines for computer-based tests and interpretations*, Washington, auteur.

Arbuckle, James, L. (2005), *Amos 6.0 user's guide*, Spring House, Amos Development Corporation.

Bartram, Dave (1994). « Computer-based assessment », dans C. L. Cooper (dir.), *International Review of Industrial and Organizational Psychology,* London, Wiley.

Bartram, D. et R. Bayliss (1984), « Automated testing : Past, present and future », *Journal of Occupational Psychology*, vol. 57, n° 3, p. 221-237.

Beckers, John J. et Henk G. Schmidt (2003), « Computer experience and computer anxiety », *Computers in Human Behavior*, vol. 19, n° 6, p. 785-797.

Beckers, John J., Remy M. J. P. Rikers et Henk G. Schmidt (2006), « The influence of computer anxiety on experienced computer users while performing complex computer tasks », *Computers in Human Behavior*, vol. 22, n° 3, p. 456-466.

Bodmann, Shawn M. et Daniel H. Robinson (2004), « Speed and performance differences among computer-based and paper-pencil tests », *Journal of Educational Computing Research*, vol. 31, n° 1, p. 51-60.

Booth-Kewley, Stephanie, Jack E. Edwards et Paul Rosenfeld (1992), « Impression management, social desirability, and computer administration of attitude questionnaires : Does the computer make a difference ? », *Journal of Applied Psychology*, vol. 77, n° 4, p. 562-566.

Bozionelos, Nikos (2004), « Socio-economic background and computer use : the role of computer anxiety and computer experience in their relationship », *International Journal of Human-Computer Studies*, vol. 61, n° 5, p. 725-746.

Butcher, James N., Laura S. Keller et Steven F. Bacon (1985), « Current developments and future directions in computerized personality assessment », *Journal of Consulting and Clinical Psychology*, vol. 53, n° 6, p. 803-815.

Cullen, Michael J., Chaitra M. Hardison et Paul R. Sackett (2004), « Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory », *Journal of Applied Psychology*, vol. 89, n° 2, p. 220-230.

Donovan, Michelle A., Fritz Drasgow et Tahira M. Probst (2000), « Does computerized paper-and-pencil job attitude scales make a difference ? New IRT analyses offer insight », *Journal of Applied Psychology*, vol. 85, n° 2, p. 305-313.

Dwight, Stephen A. et Melissa E. Feigelson (2000), « A quantitative review of the effect of computerized testing on the measurement of social desirability », *Educational and Psychological Measurement*, vol. 60, n° 3, p. 340-360.

Elwood, David L. et Richard Griffin (1972), « Individual intelligence testing without the examiner : Reliability of an automated method », *Journal of Consulting and Clinical Psychology*, vol. 38, n° 1, p. 9-14.

Erdman, Harold P., Marjorie H. Klein et John H. Greist (1985), « Direct patient computer interviewing », *Journal of Consulting and Clinical Psychology*, vol. 53, n° 6, p. 760-773.

Finger, Michael S. et Deniz S. Ones (1999), « Psychometric equivalence of the computer and booklet forms of the MMPI : A meta-analysis », *Psychological Assessment*, vol. 11, n° 1, p. 58-66.

Gaudron, Jean-Philippe (1998), « Les effets de l'anxiété informatique et de la pratique de l'ordinateur sur les performances à un test automatisé d'intelligence », *Le Travail Humain*, vol. 61, n° 3, p. 263-280.

Goldberg, David (1972), *The Detection of Psychiatric Illness by Questionnaire*, London, Oxford University Press.

Greaud, Valerie A. et Bert F. Green (1986), « Equivalence of conventional and computer presentation of speed tests », *Applied Psychological Measurement*, vol. 10, n° 1, p. 23-34.

Hays, Shannon H. et Steve R. McCallum (2005), « A comparison of the pencil-and-paper and computer-administered Minnesota Multiphasic Personality Inventory – Adolescent », *Psychology in the Schools*, vol. 42, n° 6, p. 605-613.

Johnson, Douglas F. et William L. Mihal (1973), « Performance of blacks and whites in computerized versus manual testing environments », *American Psychologist*, vol. 28, n° 8, p. 694-699.

Kleinman, Leah, Nancy Kline Leidy, Joseph Crawley, Amy Bonomi et Phil Schoenfeld (2001), « A comparative trial of paper-and-pencil versus computer administration of the Quality of Life in Reflux and Dyspepsia (QOLRAD) Questionnaire », *Medical Care*, vol. 39, n° 2, p. 181-189.

Korukonda, Appa Rao (2007), « Differences that matter : A dialectical analysis of individual characteristics and personality dimensions contributing to computer anxiety », *Computers in Human Behavior*, vol. 23, n° 4, p. 1921-1942.

Lautenschlager, Gary J. et Vicky L. Flaherty (1990), « Computer administration of questions : More desirable or more social desirability ? », *Journal of Applied Psychology*, vol. 75, n° 3, p. 310-314.

Llabre, Maria M., Nancy E. Clements, Katherine B. Fitzhugh, Gary Lancellotta, Roy D. Mazzagatti et Nancy Quinones (1987), « The effect of computer-administered testing on test anxiety and performance », *Journal of Educational Computing Research*, vol. 3, n° 4, p. 429-433.

Looker, E. Dianne et Victor Thiessen (2003), « Beyond the digital divide in Canadian schools », *Social Science Computer Review*, vol. 21, n° 4, p. 475-490.

Martin, Christopher L. et Dennis H. Nagao (1989), « Some effects of computerized interviewing on job applicant responses », *Journal of Applied Psychology*, vol. 74, n° 1, p. 72-80.

Mead, Alan D. et Fritz Drasgow (1993), « Equivalence of computerized and paper-and-pencil cognitive ability tests : A meta-analysis », *Psychological Bulletin*, vol. 114, n° 3, p. 449-458.

Meyers, Raymond et Claude Houssemand (2006), *Résultats complets de la première passation du questionnaire informatisé auprès d'une cohorte de demandeurs d'emploi au Luxembourg*, Luxembourg, Université du Luxembourg.

Meinert, David, Scott J. Vitell et Ray Blankenship (1998), « Respondent « honesty » : A comparison of computer versus paper and pencil questionnaire administration », *Journal of Marketing Management*, vol. 8, n° 1, p. 34-43.

Ministère du Travail et de l'Emploi (2007), *Rapport d'activité 2006,* Luxembourg, MTE.

Overton, Randall C., Rogers L. Taylor, Michael J. Zickar et Harvey J. Harms (1996), « The pen-based computer as an alternative platform for test administration », *Personnel Psychology*, vol. 49, n° 2, p. 455-464.

Politi, P. L., M. Piccinelli et G. Wilkinson (1994), « Reliability, validity and factor structure of the 12-items General Health Questionnaire among young males in Italy », *Acta Psychiatrica Scandinavica*, vol. 90, n° 6, p. 432-437.

Powers, Donald E. (2004), « Test anxiety and test performance : Comparing paper-based and computer-adaptive versions of the Graduate Record Examinations (GRE) General Test », *Journal of Educational Computing Research*, vol. 24, n° 3, p. 249-273.

Raîche, Gilles (2004), « Le testing adaptatif », dans R. Bertrand et J.-G. Blais, *Modèles de mesure : l'apport de la théorie des réponses aux items*, Montréal, Presses de l'Université du Québec.

Richman, Wendy, Sara Kiesler, Suzanne Weisband et Fritz Drasgow (1999), « A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews », *Journal of Applied Psychology*, vol. 84, n° 5, p. 754-775.

Rosen, Larry D., Deborah C. Sears et Michelle M. Weil (1987), « Computerphobia », *Behavior Research Methods, Instruments and Computers*, vol. 19, n° 2, p. 167-179.

Shevlin, Mark et Gary Adamson (2005), « Alternative factor models and factorial invariance of the GHQ-12 : A large sample analysis using confirmatory factor analysis », *Psychological Assessment*, vol. 17, n° 2, p. 231–236.

Steele, Claude M. et Joshua Aronson (1995), « Stereotype threat and the intellectual performance of African Americans », *Journal of Personality and Social Psychology,* vol. 69, n° 5, p. 797-811.

Tseng, Hsu-Min, Brian Tiplady, Hamish A. Macleod et Peter Wright (1998), « Computer anxiety : A comparison of pen-based personal digital assistants, conventional computers, and paper assessment of mood and performance », *British Journal of Psychology*, vol. 89, n° 4, p. 599-610.

Weisband, Suzanne et Sara Kiesler (1996), « Self disclosure on computer forms : Meta-analysis and implications », dans R. Bilger, S. Guest et M. J. Tauber (Dirs), *Proceedings of the Conference on Human Factors in Computing Systems (CHI 96)*. New York, ACM Press.

Williams, John Eustis et David D. McCord (2006), « Equivalence of standard and computerized versions of the Raven Progressive Matrices Test », *Computers in Human Behavior*, vol. 22, n° 5, p. 791-800.

Wilson, Kenth R., Jennifer S. Wallin et Christa Reiser (2003), « Social stratification and the digital divide », *Social Science Computer Review*, vol. 21, n° 2, p. 133-143.

Zeidner, Moshe (1990), « Does test anxiety bias Scholastic Aptitude Test performance by gender and sociocultural group ? », *Journal of Personality Assessment*, vol. 55, n^{os} 1 et 2, p. 145-160.

Zeidner, Moshe (1998), *Test anxiety : The state of the art*, New York, Plenum.

Page laissée blanche intentionnellement

SECTION III

Les défis d'une approche adaptative et diagnostique

Page laissée blanche intentionnellement

CHAPITRE 9

Les modèles cognitifs

Nathalie Loye
Université de Montréal

INTRODUCTION

Dans une approche « classiquement » psychométrique, la création de tests sous-entend le recours à une échelle de mesure liée à un construit général (par exemple, l'habileté en mathématiques) sur laquelle les personnes peuvent être mises en ordre (Huff et Goodman, 2006). Selon cette approche, le développement de tests adaptatifs vise à améliorer la précision de l'estimation de la position de chaque sujet sur ce continuum grâce à un test sur mesure pour chacun d'eux. Le but est plutôt de l'ordre de la sélection ou de la classification des sujets, puisqu'une telle manière de les caractériser ne fournit pas d'informations sur leurs habiletés spécifiques et s'adapte assez mal à une perspective d'évaluation diagnostique ou formative (Chipman, Nichols et Brennan, 1995 ; McGlohen, 2004). Pourtant, avec la place de plus en plus importante allouée à l'évaluation formative, il existe un besoin d'outils efficaces pour aider à diagnostiquer ce qui fait défaut chez l'élève afin de développer une stratégie de remédiation appropriée. La littérature abondante sur l'amélioration de l'enseignement et de l'apprentissage en témoigne (Roussos, Dibello, Stout, Hartz, Henson et Templin, 2007).

En parallèle au développement des tests adaptatifs, les vingt dernières années ont vu se multiplier les recherches qui jumellent l'approche cognitive à des modèles psychométriques. Il en résulte aujourd'hui de nombreux modèles qui visent à fournir un diagnostic sur des variables latentes discrètes représentant des processus cognitifs (Chipman *et al.*, 1995). Les nouveaux modèles cognitifs (Loye, 2005) qui se développent modélisent la probabilité de répondre correctement à un item en fonction de la maîtrise d'un ensemble d'attributs cognitifs propres à cet item. Notons que les avancées technologiques

contribuent à la fois à l'évolution du *testing* adaptatif et à celle de ces modèles de mesure complexes.

L'état de développement de certains de ces modèles permet d'envisager la mise en place d'études visant à contribuer à l'évaluation formative et diagnostique sur le terrain. Ainsi, Roussos et ses collègues (2007) affirment que le modèle de fusion (DiBello, Stout et Roussos, 1995) « reparamétrisé » par Hartz (2002) est prêt pour un usage opérationnel. Il n'y aurait donc en principe qu'un pas à faire pour combiner les avantages d'un modèle cognitif à ceux du *testing* adaptatif. Ce pas a d'ailleurs été franchi par McGlohen qui a consacré sa thèse de doctorat à comparer plusieurs méthodes pour optimiser l'estimation de l'habileté des sujets qui passent un test, de même que l'estimation de leur niveau de maîtrise des attributs cognitifs en lien avec le test, à l'aide du modèle de fusion reparamétrisé (McGlohen, 2004).

Dans le but de vérifier si on peut utiliser un test adaptatif afin de diagnostiquer le degré de maîtrise des processus cognitifs mis en œuvre par chaque sujet pour répondre aux items, un premier défi consiste à déterminer les attributs reliés à chacun de ces items. Cela signifie qu'avant de songer à rendre le test adaptatif, il faut établir la structure cognitive de chacun des items de ce test. Ce chapitre présente les résultats d'une recherche exploratoire portant sur l'élaboration de la liste des attributs cognitifs propres à un ensemble d'items à réponse choisie en mathématiques. Les items étudiés font partie d'un prétest de mathématiques à l'entrée de l'École Polytechnique de Montréal. Cette étude vise en particulier à déterminer les meilleures conditions à mettre en place pour aider les experts à définir la structure cognitive d'un test. Les recommandations auxquelles elle aboutit constituent un premier pas vers la mise en place opérationnelle d'un modèle cognitif comme partie intégrante d'un test adaptatif.

LA MATRICE Q DES ATTRIBUTS COGNITIFS

De façon générale, les modèles cognitifs s'appuient sur l'élaboration, par des « experts », d'une liste des attributs cognitifs nécessaires et suffisants pour répondre aux items du test. Un attribut cognitif se définit par la description de connaissances procédurales ou déclaratives aboutissant à la création d'une stratégie pour répondre à un item donné. Il représente un état latent de connaissance ou une habileté, ne s'observe pas directement, et est éventuellement relié aux autres attributs selon un ordre hiérarchique (Gierl, Leighton et Hunka, 2000 ; Leighton, Gierl et Hunka, 2002 ; Milewski et Baron, 2002 ; Rupp, 2007 ; Tatsuoka, 1983). En outre, les experts peuvent déterminer au préalable les processus cognitifs à diagnostiquer et fabriquer un test approprié dans une approche déductive. Dans une approche

inductive, ils peuvent également utiliser un test existant et déterminer les processus cognitifs sous-jacents à chaque item.

Que l'approche soit déductive ou inductive, les liens entre les attributs et les items sont effectués dans ces modèles par une matrice notée Q. Les lignes de la matrice Q correspondent aux items du test alors que ses colonnes représentent les attributs cognitifs. La figure 1 propose un exemple de relations existant entre 5 attributs et 3 items mises sous la forme d'une matrice Q. Étant donné que les items sont caractérisés dans le modèle cognitif par les réponses des sujets, les variables correspondantes sont observées et donc représentées par des rectangles dans la figure 1. Par contre, les attributs ne sont pas directement observables et correspondent à des variables latentes, représentées par des formes arrondies. Pour les experts, élaborer une matrice Q signifie créer d'une part la liste des attributs cognitifs nécessaires et suffisants pour l'ensemble des items du test, et d'autre part les liens spécifiques matérialisés par les valeurs 0 et 1 dans la matrice Q

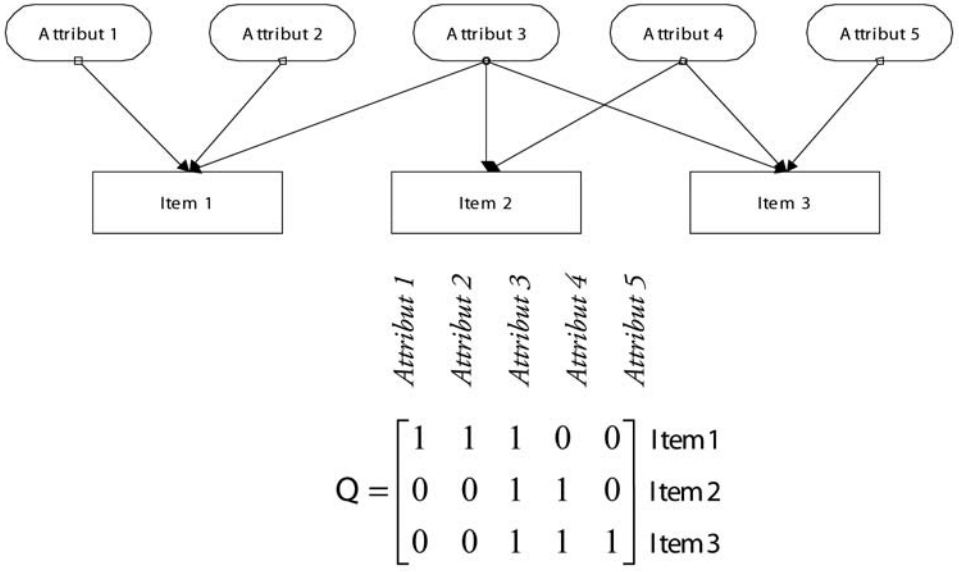


Figure 1 – Exemple de liens entre 5 attributs et 3 items et la matrice Q correspondante

Selon DiBello et ses collègues (1995), les modèles cognitifs, et la théorie qui leur est sous-jacente, devraient s'appliquer aussi bien à des tests élaborés par les enseignants qu'à des tests nationaux ou internationaux standardisés.

Dans tous les cas, la perspective de combiner les modèles cognitifs au *testing* adaptatif implique l'accès à une banque d'items calibrés. En effet, les paramètres des items doivent être connus afin que le système puisse choisir celui qui est le plus approprié après chaque estimation de l'habileté de l'élève, et éventuellement de son degré de maîtrise des attributs dans le cas d'un test adaptatif incluant un modèle cognitif.

On peut donc considérer que la matrice Q fait partie intégrante du calibrage des items, puisqu'elle les caractérise par leurs liens avec les attributs. Comme la validité et la fidélité des mesures produites par les modèles cognitifs dépendent de l'identification des attributs pertinents, des liens logiques existants entre eux, et de leurs liens avec les items, elles dépendent étroitement de la capacité des experts à fournir des jugements valides, fidèles et concordants (Gierl, 2000). Un modèle cognitif pose ainsi un problème de validité ou de fidélité dès que les attributs cognitifs identifiés ne sont pas de bonnes variables pour prédire la performance du sujet (Junker, 1999). Par exemple, le sujet peut utiliser une habileté différente de celle prévue par les experts ou encore deviner la réponse. Des attributs importants peuvent également être oubliés dans la matrice Q. La validité est alors remise en question, puisque le modèle ne diagnostiquera pas convenablement les attributs possédés ou non par le sujet. Si le sujet commet une erreur en utilisant mal un attribut qu'il maîtrise, ou s'il trouve la bonne réponse par hasard, la mesure produite pose un problème de fidélité puisqu'elle ne sera vraisemblablement pas reproductible à une autre occasion.

Le travail des experts doit donc aboutir à une matrice Q valide et fidèle. Cela implique de se questionner sur les caractéristiques que ces experts doivent posséder, sur le nombre d'experts à mobiliser, de même que sur la manière de les faire travailler. Il faut également s'interroger sur les informations à leur fournir pour les aider à comprendre et à mener à bien leur tâche. L'idéal serait que ces experts soient d'accord entre eux en ce qui concerne à la fois la liste des attributs cognitifs et les liens entre items et attributs. De plus, la matrice Q devrait permettre un diagnostic sur les caractéristiques des sujets responsables de leurs succès et de leurs difficultés, et ces caractéristiques devraient aller au-delà du contenu ou des objectifs disciplinaires du test et viser les processus mentaux utilisés par les sujets pour répondre aux items.

Jusqu'à présent, peu de recherches portant sur la validité et la fidélité de la matrice Q ont été menées. Dans les cas d'applications de modèles cognitifs à des données réelles, les chercheurs, en partenariat avec leurs experts, élaborent une première matrice Q. Ils utilisent cette matrice avec leur modèle cognitif et se basent ensuite sur certains résultats, comme le taux de maîtrise des attributs, pour déterminer les attributs plus ou moins

pertinents et modifier leur matrice dans un processus itératif (voir, par exemple, Buck et Tatsuoka, 1998 ; Dimitrov et Raykov, 2003 ; Hartz, 2002 ; Jang, 2005 ou Tatsuoka *et al.,* 1993). Une telle façon de faire leur permet d'avoir une certaine confiance dans la matrice Q finalement retenue, mais ne constitue pas l'objet même de leur recherche.

En prenant pour exemple la littérature abondante sur le travail des experts dans l'établissement de standards en éducation, la présente étude pose l'hypothèse que certains résultats d'analyses préalables pourraient aider les experts à élaborer les matrices Q. Dans le cas d'une approche inductive, l'accès aux données permet donc de se demander dans quelle mesure le fait d'analyser les données issues d'un test et de fournir les résultats aux experts peut les aider à établir une matrice Q plus valide et plus fidèle.

LES CONDITIONS EXPÉRIMENTALES

Dans cette étude, les experts doivent élaborer des matrices dans une approche inductive, ce qui implique la disponibilité des données constituées par les réponses des sujets aux items du test. On peut donc s'interroger sur les informations concernant les données qui sont susceptibles d'apporter du soutien aux experts. Trois façons différentes d'exploiter les données sont privilégiées. D'abord, étant donné que les items sont à réponse choisie, les choix des leurres ou de la bonne réponse constituent les données originales. Comme les modèles cognitifs actuels s'appliquent à des données dichotomiques, ordinales ou continues, les données doivent être dichotomisées avant que soit appliqué le modèle cognitif. L'information contenue dans les données incluant les choix de réponses par les sujets pourrait donc être analysée et fournie aux experts. Cette approche est d'autant plus intéressante que les informations sur les choix de réponses disparaissent lors de la « dichotomisation ». Une deuxième façon d'exploiter les données consiste à caractériser les items dans une approche psychométrique en calculant les valeurs de différents paramètres (la difficulté, par exemple). La troisième façon vise à exploiter le fait que l'approche cognitive cherche à établir un diagnostic dans une vision multidimensionnelle basée sur l'hypothèse que l'espace latent d'habileté peut être divisé en un certain nombre d'aspects qui constituent les attributs cognitifs. La recherche de la présence de plusieurs dimensions dans les données constitue une information qui semble donc en lien avec l'approche cognitive.

En ce qui concerne les erreurs commises par les sujets et matérialisées par leur choix de réponses, Thissen, Steinberg et Fitzpatrick (1989) affirment qu'elles sont une partie importante de la question. Un graphique peut être utilisé pour représenter chaque alternative d'une question à choix multiple

par la probabilité de choisir chaque alternative d'un item en fonction du score total des sujets (Ramsay, 1996 ; Ramsay et Silverman, 1997 ; Thissen *et al.*, 1989). Un tel portrait de chaque item montre aux experts les choix faits par les sujets. Dans certains cas, l'un des leurres peut être très attrayant pour les sujets. Chercher ce qui le différencie de la bonne réponse pourrait donc fournir une piste à suivre aux experts.

Au sujet de la caractérisation des items au sens psychométrique, la théorie classique des tests définit la difficulté d'un item par la proportion de sujets qui ont répondu convenablement à la question. Cette information peut amener les experts à rechercher ce qui rend une question difficile. Il se peut qu'elle implique un attribut qu'ils estiment difficile ; sa difficulté peut aussi venir d'un grand nombre de liens avec des attributs, ou encore de sa formulation. Ce sont autant de raisonnements qui pourraient aider les experts dans leur tâche. Le pouvoir de discrimination des items apporte une information supplémentaire en indiquant s'il y a une différence entre les sujets faibles et les sujets forts lorsqu'ils répondent à l'item. La recherche d'un attribut qui expliquerait une telle différence peut être une autre piste de réflexion. Ces paramètres d'items au sens de la théorie classique des tests peuvent facilement être calculés (par exemple, avec le progiciel SPSS).

Enfin, la possibilité d'exploiter la présence de plusieurs dimensions dans les données pour aider les experts à définir une structure dans le test et ainsi à élaborer la matrice Q est présente dans la littérature. Corter (1995) a utilisé une analyse en grappes afin de regrouper les items pour tenter de valider la structure diagnostique du test proposée par Tatsuoka (1990), mais sans y parvenir. Hartz (2002) suggère d'utiliser des analyses factorielles pour définir la structure de la matrice Q dans une approche inductive. Jang (2005) s'aide de cette approche lorsqu'elle élabore la matrice Q dans sa recherche. L'identification de plusieurs dimensions dans les données peut se faire sous la forme d'analyses factorielles exploratoires, confirmatoires, discriminantes, ou encore d'analyses en grappes. L'objectif vise alors à définir les items qui partagent une même dimension latente. Les experts peuvent avoir ensuite accès aux items regroupés lorsqu'ils élaborent leur matrice Q, et baser leur réflexion sur ce que ces items ont en commun afin d'identifier des attributs cognitifs.

Afin de pouvoir juger de la pertinence de fournir ces diverses informations aux experts, il faut que trois conditions expérimentales soient mises en place. Dans la première condition expérimentale, les experts ont seulement accès aux énoncés des items. Dans la deuxième condition expérimentale, les items leur sont fournis tels qu'ils sont regroupés selon les résultats d'une analyse factorielle. Dans la troisième condition, les experts ont accès à des

graphiques et aux valeurs des paramètres de difficulté et de discrimination de chaque item.

LE MODÈLE DE FUSION REPARAMÉTRISÉ

Le modèle de fusion (Dibello *et al.*, 1995) a été reparamétrisé par Hartz en 2002, et baptisé RUM. Le logiciel Arpeggio développé par *Educational Testing Service* vise à l'appliquer à des données. Le modèle RUM se base sur une matrice Q qui contient k attributs reliés aux items du test. Il tient également compte du fait qu'un certain nombre d'attributs non spécifiés dans Q pourraient jouer un rôle dans la performance des sujets. Le modèle inclut deux paramètres liés aux sujets, trois paramètres liés aux items et à leurs liens avec les attributs dans la matrice Q, et un paramètre lié aux attributs.

Les sujets sont caractérisés par un vecteur $\tilde{\alpha}_j$ qui représente la probabilité de maîtrise de chaque attribut présent dans la matrice Q. En outre, comme un sujet peut aussi être caractérisé par un ensemble d'habiletés résiduelles non prises en compte par la matrice Q, un paramètre θ_j correspond à la projection de l'ensemble de ces habiletés résiduelles sur une échelle unidimensionnelle (Roussos *et al.*, 2007).

Le paramètre d'item π_i^* correspond à la probabilité d'appliquer convenablement tous les attributs requis par l'item i selon la matrice Q, quand ils sont maîtrisés. Cela revient à dire que le paramètre π_i^* correspond à la difficulté de l'item i relativement aux attributs qui lui sont reliés dans la matrice Q. Il est un indicateur de la bonne utilisation des attributs reliés à l'item et est donc en lien avec la fidélité de la matrice Q. Le paramètre d'item r_{ik}^* représente la pénalité due au fait de ne pas maîtriser un attribut k relié à l'item dans la matrice Q. Ce paramètre indique si la maîtrise ou non de cet attribut prédit la réussite ou l'échec à l'item. Il est donc un indicateur de la validité de la matrice Q, tout comme le paramètre d'item c_i. Ce dernier permet de juger si la matrice Q contient ou pas tous les attributs pertinents (*completeness*). Enfin, chacun des attributs présent dans la matrice Q est caractérisé par sa difficulté, notée p_k, qui correspond à la probabilité de maîtriser l'attribut. Cette probabilité est estimée globalement pour l'ensemble des sujets relativement à chaque attribut, mais la probabilité de maîtrise de chaque attribut est également estimée pour chacun des sujets et correspond au vecteur $\tilde{\alpha}_j$.

Le modèle RUM est défini par l'équation (1) dans laquelle les indices i caractérisent les items, les indices j caractérisent les sujets et les indices k caractérisent les K attributs présents dans la matrice Q. Soient les paramètres $\pi_{ik} = P(Y_{ikj} = 1 \mid \alpha_{jk}=1)$ et $r_{ik} = P(Y_{ijk} = 1 \mid \alpha_{jk}=0)$ où Y_{ikj} fait référence à un

évènement non observé lors duquel le sujet *j* applique convenablement un attribut *k* à l'item *i*. De plus, α_{jk} prend la valeur 0 si le sujet ne maîtrise pas l'attribut *k*, et la valeur 1 s'il maîtrise l'attribut *k*. Enfin, les applications correctes ou non des différents attributs sont supposées indépendantes pour un couple (α, θ) donné, soit :

$$P(X_{ij} = 1 \mid \alpha_j, \theta_j) = \pi_i^* \prod_{k=1}^{K} r_{ik}^{*(1-\alpha_{jk})q_{ik}} P_{c_i}(\theta_j) \qquad 1$$

dans laquelle

$$P_{c_i}(\theta_j) = \textit{Fonction Logistique}\{-1{,}7\ (\theta_j - c_i)\},$$

$$\pi_i^* = \prod_{k=1}^{M} \pi_{ik}^{q_{ik}} \ ,$$

et

$$r_{ik}^* = \frac{r_{ik}}{\pi_{ik}} = \frac{P(Y_{ijk} = 1 \mid \alpha_{jk} = 0).}{P(Y_{ijk} = 1 \mid \alpha_{jk} = 1)}$$

Il existe un seul paramètre π_i^* et un seul paramètre c_i pour chaque item *i* du test. Par contre, à chaque lien établi dans la matrice Q, c'est-à-dire à chaque valeur 1 de la matrice Q, correspond une valeur estimée de r_{ik}^*. La figure 2 illustre la correspondance entre la forme de la matrice Q et les paramètres estimés par le modèle RUM. Chacun des paramètres qui caractérisent les items permet d'observer la validité ou la fidélité de la matrice Q par comparaison de la valeur estimée à des valeurs seuils définies dans la littérature.

Un paramètre π_i^* est estimé pour chaque item du test, et sa valeur varie entre 0 et 1. En ce qui concerne le paramètre π_i^*, une valeur supérieure à 0,6 signifie que les sujets ont une bonne chance de répondre correctement à l'item *i* s'ils maîtrisent les attributs qui lui sont reliés dans la matrice Q (Templin, 2004). La comparaison des paramètres estimés à 0,6 permet donc de porter un regard pertinent sur la force de la structure cognitive du test associé à sa matrice Q et ainsi sur la fidélité de chaque ligne de la matrice Q.

Un paramètre r_{ik}^* est estimé pour chaque lien établi dans la matrice Q. Les valeurs des paramètres r_{ik}^* varient entre 0 et 1. Comme le montre l'équa-

tion (1), un sujet qui ne maîtrise pas l'attribut k_0 requis par l'item i a une probabilité de répondre correctement à l'item i proportionnelle à $r^*_{ik_0}$. Chaque paramètre r^*_{ik} compare la probabilité de bien répondre à l'item i selon que l'attribut k est maîtrisé ou non. Ainsi, plus une bonne réponse à l'item dépend de la maîtrise de l'attribut k_0 (c'est-à-dire que plus la valeur de $r^*_{ik_0}$ est proche de zéro) plus la probabilité de répondre correctement à l'item est petite. Autrement dit, plus la valeur de $r^*_{ik_0}$ est proche de zéro, plus l'item est discriminant pour l'attribut k_0. Une matrice Q valide devrait établir des liens entre attributs et items pour lesquels la maîtrise ou la non-maîtrise de l'attribut permet de prévoir la réussite ou l'échec à l'item. De tels liens sont associés à une valeur de r qui est d'autant plus petite que le lien est fort. La moyenne des valeurs de r est également un indicateur de la force de la structure de la matrice pour cet item. En ce qui concerne la moyenne des r^*_{ik}, pour un item i donné, Templin (2004) suggère qu'une valeur inférieure à 0,5 implique une structure cognitive élevée pour l'item i et est un indicateur de la validité de la ligne correspondant à cet item dans la matrice Q.

Enfin, un paramètre c_i est estimé pour chaque item et quantifie la force de la dépendance aux attributs absents de la matrice Q. Il varie entre 0 et 3 (Roussos *et al.*, 2007 ; Templin, 2004). Comme il est inversement relié à cette force, plus sa valeur est grande et plus les attributs inclus dans la matrice sont complets. Templin (2004) suggère qu'un paramètre c_i supérieur à 1,5 indique que la matrice inclut les attributs nécessaires, ce qui est un indicateur de la validité de la ligne correspondant à cet item dans la matrice Q.

Attributs
1 2 3 4 5

$$Q = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

item 1 →
item 2
item 3

Items	π_i	r_{ik} (att. 1)	r_{ik} (att. 2)	r_{ik} (att. 3)	r_{ik} (att. 4)	r_{ik} (att.5)	c_i
Item 1	π_1	$r_{1,1}$	$r_{1,2}$	$r_{1,3}$	-	-	c_1
Item 2	π_2	-	-	$r_{2,3}$	$r_{2,4}$	-	c_2
Item 3	π_3	-	-	$r_{3,3}$	$r_{3,4}$	$r_{3,5}$	c_3

Figure 2 – Un exemple de lien entre la forme de Q et les paramètres estimés

L'ÉLABORATION DES MATRICES Q INDIVIDUELLES

La première action à mener pour élaborer une matrice Q consiste à recruter des experts. Les critères de sélection de cette étude ont visé à engager des personnes ayant étudié les mathématiques à l'université et enseigné les mathématiques aux niveaux secondaire, collégial ou universitaire depuis au moins 10 ans. Les quatre personnes qui ont répondu à l'appel satisfaisaient à ces critères et ont été retenues. Les données proviennent d'un prétest en mathématiques donné chaque année aux étudiants qui sont admis à l'École Polytechnique de Montréal. Les réponses des sujets à dix items à choix multiple portant sur les fonctions élémentaires et à dix items portant sur le calcul différentiel de deux épreuves forment les données. La sélection des deux cohortes d'étudiants de 1997 et de 1999 est basée sur le fait que le nombre d'étudiants est plus élevé pour ces deux années avec 426 sujets pour le test A (1997) et 478 sujets pour le test B (1999). Une analyse de variance montre en outre que le test B est statistiquement plus difficile que le test A au seuil de 0,05 ($F(1, 902) = 8{,}918$ et $p = 0{,}003$).

La méthode d'élaboration adoptée consiste d'abord en un travail individuel de chaque expert dans chacune des trois conditions expérimentales afin de tirer parti de son expertise et du fait qu'il connait bien la population d'étudiants, tout en évitant l'influence des autres experts. La tâche individuelle de chaque expert comporte : (a) l'élaboration d'une liste d'attributs, et (b) l'établissement des liens entre les attributs choisis et les items pour fabriquer chaque matrice Q individuelle. Un livret propre à chaque condition est remis à chaque expert au début de chaque condition expérimentale.

Dans la condition 1, le livret 1 fournit les consignes, des exemples d'attributs tirés de la littérature et les énoncés des items. Ces deux dernières informations font également partie des livrets 2 et 3. Pour préparer le livret de la condition 2, une analyse factorielle est réalisée sur les données des tests A et B afin de déterminer le nombre de dimensions présentes dans chaque test et les items qui se regroupent dans ces dimensions. Les analyses factorielles, qui utilisent les corrélations tétrachoriques à l'aide du logiciel MPlus (version 4) (Muthèn et Muthèn, 2004), fournissent des solutions comprenant quatre facteurs pour le test A et six facteurs pour le test B. Il y a donc quatre groupes d'items pour le test A et six pour le test B, sept items du test A et trois items du test B n'appartenant à aucun groupe. Les consignes du livret 2 précisent que les experts doivent réfléchir à ce que partagent les (ou certains) items d'un même regroupement factoriel qui pourrait avoir un lien avec les attributs cognitifs reliés à ces items. Les consignes informent également les experts qu'il est possible qu'ils ne trouvent rien de commun.

Pour préparer le livret de la condition 3, l'indice de difficulté et l'indice de discrimination, au sens de la théorie classique des tests, sont calculés pour

chaque item à l'aide de SPSS (version 14.0.1). La difficulté correspond au pourcentage de bonnes réponses de chaque item. La corrélation point-bisérial (corrélation de Pearson entre chaque item et le score total) fournit l'indice de discrimination, car les variables sont dichotomiques. En outre, le logiciel Testgraph (Ramsay, 2000) permet de représenter graphiquement le fonctionnement des leurres de chaque item grâce à une modélisation non paramétrique. Chaque item est présenté dans le livret 3 associé à ses paramètres de difficulté et de discrimination et au portrait fourni par Testgraph. Le livret 3 contient également un exemple d'utilisation de la documentation fournie, ainsi que des commentaires et suggestions pour guider les experts dans leur interprétation des informations. Les livrets 2 et 3 incluent un court questionnaire sur la perception de l'expert relativement à la documentation fournie.

Dans chacune des trois conditions, le rôle de l'expert consiste à trouver un maximum de dix attributs par matrice Q. Cette valeur correspond à plusieurs études empiriques pour lesquelles le nombre d'attributs est égal à la moitié du nombre d'items (par exemple, Buck *et al.*, 1997 ; Hartz, 2002 ; Tatsuoka, 1993). Elle oblige les experts à cibler les attributs cognitifs qu'ils jugent importants sans les limiter trop strictement. Le rôle du chercheur consiste à expliquer la tâche à chaque expert, à définir les enjeux et à estimer en quoi son expertise est essentielle. Le chercheur rencontre individuellement chaque expert au début et à la fin de chaque condition expérimentale. Lors de la première rencontre, le chercheur donne à l'expert le livret approprié qui inclut le format de remise des résultats, identique pour tous et pour chaque condition. Lors de la deuxième rencontre, le chercheur recueille les livrets remplis par les experts et qui incluent des renseignements sur les difficultés rencontrées et sur l'opinion de l'expert quant à la pertinence des informations fournies. Afin de pouvoir considérer les conditions expérimentales comme indépendantes, le chercheur s'assure de ne laisser aucune documentation aux experts entre les conditions expérimentales qu'il espace dans le temps. Lorsque les experts ont travaillé dans les trois conditions, un total de vingt-quatre matrices Q individuelles est disponible.

L'ÉLABORATION DES MATRICES Q SYNTHÈSES

Par la suite, la réalisation d'une matrice Q synthèse, représentative du travail des quatre experts pour chaque condition expérimentale, implique (a) d'obtenir une liste d'attributs qui fasse consensus, et (b) de définir une règle pour établir les liens avec les items dans la matrice Q synthèse. Étant donné que le fait de faire réfléchir les experts collectivement après un travail individuel va dans le sens d'un *feed-back* normatif utilisant le groupe comme une norme, les experts sont réunis dans une séance de travail collectif afin

de proposer une liste d'attributs cognitifs qui fasse consensus. La méthode appelée *Multi-Attribute Consensus Building* (MACB) et développée par Vanderwood, Ysseldyke et Thurlow (1993) dans le cadre du *National Center on Educational Outcome* est retenue, car elle permet de déterminer, de façon quantitative et objective, l'opinion d'un petit groupe d'individus (généralement entre 4 et 10 personnes). À la suite de l'élaboration individuelle d'une liste d'énoncés, cette méthode permet un classement selon l'importance accordée à chaque énoncé et un consensus.

Les trois phases de la méthode MACB originale sont adaptées afin d'obtenir un classement consensuel des attributs élaborés individuellement par les experts. La phase 1 correspond au travail individuel des experts. La phase 2 aboutit à un classement consensuel des attributs individuels dans des catégories définies par le groupe d'experts au début de la rencontre. À cette étape, le chercheur impose de limiter le classement à six catégories afin de simplifier par la suite l'application du modèle RUM aux données. Cette décision est basée sur des analyses préliminaires à l'aide des premières matrices Q individuelles pour lesquelles l'estimation des paramètres à l'aide d'Arpeggio présente des problèmes de convergence. En outre, l'analyse préalable du contenu des attributs suggère une possible limitation du nombre de catégories, entre autres à cause de la ressemblance de nombreux intitulés. Enfin, la phase 3 fournit les attributs qui seront utilisés dans les matrices Q synthèses sous la forme des catégories ayant servi au classement.

La construction des matrices Q synthèses n'a ensuite plus rien à voir avec la méthode MACB. Une fois que les experts ont regroupé les attributs individuels dans les six catégories, chacune de leur matrice Q peut être convertie en remplaçant chaque attribut par le nom de la catégorie à laquelle il appartient. La transformation aboutit à des matrices Q individuelles modifiées de dimension 20x6 (20 items dans chacune des 6 catégories). Les lignes représentent les vingt items, et les colonnes représentent les six catégories. Les catégories sont ordonnées de la même façon pour l'ensemble des matrices Q individuelles modifiées. Dans une matrice Q individuelle, lorsque plusieurs intitulés originaux appartiennent à la même catégorie, il suffit d'une seule valeur 1 pour attribuer la valeur 1 à la catégorie dans la matrice Q individuelle modifiée. La matrice Q synthèse a ensuite la même forme que les matrices Q individuelles après modification. Une valeur 1 est attribuée lorsque deux experts ou plus ont estimé que l'item avait un lien avec l'attribut (catégorie), sinon la valeur est 0.

Au total, six matrices Q synthèses sont produites pour les deux tests et les trois conditions. L'application du modèle RUM aux six matrices Q synthèses aboutit ensuite à la suppression de certains liens jugés inutiles dans la matrice (voir la procédure en détail dans Roussos *et al.*, 2007), et à des matrices Q synthèses simplifiées. L'ensemble des matrices Q, de même que les paramètres estimés par Arpeggio à l'aide des matrices Q synthèses simplifiées constituent le matériel analysé pour comparer les conditions.

LES RÉSULTATS LIÉS À L'ÉLABORATION DES MATRICES

La perception des experts

Lors de la remise des livrets de la condition 1, les quatre experts ont dit n'avoir pas eu de difficulté à suivre les instructions, et s'être inspirés des listes fournies dans le livret 1. L'expert 2 a souligné que le plus difficile avait consisté à comprendre la notion d'attribut cognitif. Dans le livret de la condition 2, les experts devaient consigner s'ils avaient ou non trouvé un lien entre les items des divers regroupements factoriels. Les traces laissées montrent que les experts 1 et 3 ont trouvé des liens entre les items de sept regroupements factoriels sur dix, que l'expert 4 en a trouvé dans quatre regroupements, alors que l'expert 2 n'a vu des liens entre tous les items que dans un seul de ces regroupements. Dans le livret de la condition 3, les experts devaient consigner s'ils avaient ou non réussi à tenir compte des informations fournies pour chaque item. L'expert 2 n'a pas réussi à prendre en compte les informations fournies pour huit des items du test A et pour la moitié des items du test B. En comparaison, les experts 3 et 4 ont utilisé ces informations, au moins en partie, pour la totalité des items, tandis que l'expert 1 n'a pas rempli le formulaire.

La recherche de consensus entre les experts

Au total, cent six libellés sont élaborés dans les trois conditions expérimentales. À la suite des trois phases de la méthode MACB, les experts ont identifié six catégories qui regroupent quatre-vingt-neuf intitulés. Douze intitulés sont classés dans une catégorie « autre » car ils ne sont pas directement reliés aux mathématiques, comme le fait « d'utiliser les choix de réponse pour trouver la solution ». Enfin, le manque de clarté de cinq énoncés entraîne leur suppression. Le tableau 1 présente les six catégories ainsi que le nombre d'intitulés dans chacune d'elle.

Tableau 1 – Les regroupements des attributs après le MACB

Nombre d'attributs individuels classés (total 101)	Catégories
9	**Attribut 1 : Décoder** : tout ce qui a trait au fait de lire la question correctement (symboles mathématiques, mots clés)
37	**Attribut 2 : Appliquer une technique** : calcul algébrique, arithmétique, application de règles ou d'algorithmes
8	**Attribut 3 :** Faire des **liens** entre des **notions** (niveau cognitif supérieur)
9	**Attribut 4 :** Utiliser une **visualisation graphique** (graphique, tableau, dessin ou autre pour aider à trouver la réponse)
18	**Attribut 5 :** Utiliser des **propriétés, définitions, théorèmes, éléments de base,** et faire des **liens** entre ces éléments
8	**Attribut 6 : Organiser** la solution : plusieurs étapes, stratégie
12	Autre (non utilisés d'un commun accord)
5 (éliminés)	Éliminés pour manque de clarté

La comparaison des matrices individuelles des experts montre que l'aspect technique des questions (attribut 2) fait l'objet du plus de représentations dans les trois conditions. Dans la condition 3, les experts ont fait preuve d'une plus grande précision sur cet aspect au détriment du contenu et de l'organisation (attributs 5 et 6) particulièrement. Dans la condition 2, trois des quatre experts n'ont pas pris l'attribut 1 (décodage) en considération, alors que cet attribut est relié à presque tous les items de la condition 3. Notons que cette augmentation massive du nombre de liens avec l'attribut 1 est associée à un biais de méthode, car cet attribut illustrait l'exemple présenté aux experts dans le livret 3. Le consensus a permis de modifier les vingt-quatre matrices Q individuelles originales et de créer les six matrices Q synthèses.

La concordance entre les matrices Q individuelles modifiées et la représentativité des matrices Q synthèses

Les matrices Q individuelles modifiées sur la base du consensus ont une forme semblable qui permet le calcul d'un coefficient de contingence. Pour une même condition et un même test, ce coefficient permet de comparer les matrices des experts deux à deux. Les résultats font l'objet du tableau 2.

Tableau 2 – Les 36 coefficients de contingence pour les matrices individuelles modifiées

		Experts							
		1		2		3		4	
Test A		**Coeff. c**	**Probabilité**	**Coeff. c**	**Probabilité**	**Coeff. c**	**Probabilité**	**Coeff. c**	**Probabilité**
Condition 1	Expert 1	-	-	-	-	-	-	-	-
	Expert 2	**0,212***	**0,018**	-	-	-	-	-	-
	Expert 3	0,410*	0,000	**0,176**	0,050	-	-	-	-
	Expert 4	0,345*	0,000	**0,012**	0,893	0,199*	0,026	-	-
Condition 2	Expert 1	-	-	-	-	-	-	-	-
	Expert 2	**0,158**	0,080	-	-	-	-	-	-
	Expert 3	**0,441***	**0,000**	**0,383***	**0,000**	-	-	-	-
	Expert 4	0,285*	0,001	**0,144**	0,112	0,242*	0,006	-	-
Condition 3	Expert 1	-	-	-	-	-	-	-	-
	Expert 2	**0,100**	0,272	-	-	-	-	-	-
	Expert 3	**0,180***	**0,045**	**0,364***	**0,000**	-	-	-	-
	Expert 4	**0,385***	**0,000**	**0,244***	**0,006**	**0,301***	**0,001**	-	-
Test B									
Condition 1	Expert 1	-	-	-	-	-	-	-	-
	Expert 2	**0,209***	**0,019**	-	-	-	-	-	-
	Expert 3	**0,542***	**0,000**	**0,336***	**0,000**	-	-	-	-
	Expert 4	0,349*	0,000	**0,120**	0,187	0,284*	**0,001**	-	-
Condition 2	Expert 1	-	-	-	-	-	-	-	-
	Expert 2	**0,254***	**0,004**	-	-	-	-	-	-
	Expert 3	**0,466***	**0,000**	**0,277***	**0,002**	-	-	-	-
	Expert 4	0,381*	0,000	**0,077**	0,397	**0,118**	0,194	-	-
Condition 3	Expert 1	-	-	-	-	-	-	-	-
	Expert 2	**0,079**	0,386	-	-	-	-	-	-
	Expert 3	**0,042**	0,646	0,396*	0,000	-	-	-	-
	Expert 4	**0,087**	0,337	0,425*	**0,000**	**0,254***	**0,004**	-	-

Note. Les valeurs en caractère gras montrent que le degré d'association entre les deux experts est nul, ce qui montre un désaccord entre les deux experts concernés. Celles qui sont surlignées impliquent l'expert 2.
*p<0,05

Onze des trente-six valeurs calculées montrent un manque de concordance entre deux matrices Q individuelles modifiées. Étant donné que quatre intitulés individuels de l'expert 2 ont été abandonnés en raison du consensus, plusieurs de ses matrices Q individuelles modifiées ne contiennent que trois des six attributs. Ceci explique, au moins en partie, que les matrices Q de cet expert manquent de concordance avec les autres dans huit des onze cas identifiés. Quant aux trois derniers problèmes de concordance, l'un est lié à la condition 2 et au test B, les deux autres sont liés à la condition 3 et au test B. Notons que, étant donné la valeur du coefficient de contingence observée, l'expert 1 n'est systématiquement pas d'accord avec les trois autres experts à la condition 3 pour le test B.

En ce qui concerne les matrices Q synthèses, le calcul des coefficients de contingence entre chaque matrice Q individuelle modifiée et la synthèse correspondante permet de juger leur représentativité du panel d'experts. Ces calculs montrent que la matrice Q synthèse est représentative de chaque matrice Q individuelle, et ce, pour les trois conditions. Le fait que les coefficients de contingence permettent d'identifier le comportement marginal de certains experts, mais montrent que toutes les matrices Q sont représentatives du panel d'experts, fournit une certaine validation à la procédure de calcul choisie. En outre, la représentativité des matrices est un indicateur de la pertinence de la méthode MACB pour établir la synthèse des matrices Q individuelles.

LES RÉSULTATS LIÉS À LA MODÉLISATION RUM

Les matrices Q synthèses simplifiées

Rousseau et ses collègues (2007) détaillent la procédure permettant d'appliquer le modèle RUM à des données. En particulier, ils indiquent qu'un manque de convergence des routines d'estimation est parfois observé et peut être amélioré en imposant une valeur fixe au paramètre *c*. Lors de la modélisation des données avec le logiciel Arpeggio et les matrices Q synthèses, le manque de convergence de certaines estimations associé au souci d'appliquer une procédure identique afin de pouvoir ensuite comparer les paramètres estimés a donc pour conséquence de fixer les valeurs des paramètres *c*. De plus, selon les recommandations de Rousseau *et al.* (2007), certains liens sont supprimés pour alléger le modèle sans en changer la structure, mais en simplifiant les matrices Q synthèses. Les conditions n'influencent pas le nombre de suppressions, mais les liens retirés sont essentiellement d'ordre technique pour le test B.

L'examen de l'évolution des liens dans les matrices Q synthèses simplifiées d'une condition à une autre montre que la variation est la moins

importante entre les conditions 1 et 2, et la plus importante entre les conditions 1 et 3. De plus, presque aucun item ne change seulement entre les conditions 1 et 2. Les liens des items reliés au calcul différentiel ont tendance à plus varier que ceux des items reliés aux fonctions élémentaires. Il n'est pas possible de savoir si la plus grande facilité du test A explique que plus d'items sont restés stables d'une condition à l'autre pour ce test (en particulier des items reliés aux fonctions élémentaires) ou encore si le fait que les experts ont toujours travaillé sur le test A avant le test B joue un rôle.

Toutefois, les différences entre le test A et le test B ne s'arrêtent pas là. Le regroupement factoriel du test B a un lien plus étroit avec son tableau de « spécifications » que celui du test A (voir tableau 3). En effet, deux des quatre facteurs du test A sont composés d'items reliés à trois contenus, selon le tableau de spécification du test A, alors que les facteurs du test B contiennent des items reliés à au plus deux contenus du tableau de spécification du test B. Si l'on s'intéresse par exemple au facteur 4 du test B, la condition 2 a identifié l'attribut 5 (propriétés, définitions et liens) comme partagé par les quatre items. Les paramètres *r* du modèle RUM estimés pour ces quatre liens ont en outre des valeurs inférieures à 0,5, ce qui permet de conclure à des liens très valides. La comparaison de la structure du test selon le tableau de particularités et selon les regroupements factoriels suggère donc que les experts ont plus de facilité à exploiter les regroupements d'items lorsqu'ils sont associés à des contenus semblables.

Tableau 3 – La mise en parallèle de la structure factorielle des tests A et B et du tableau de particularités

	% d'items (Fonctions élémentaires)		% d'items (Calcul différentiel)	
	Notions	**Fonctions**	**Notions**	**Technique**
Test A				
Facteur 1 : 12, 19	100 %			
Facteur 2 : 7, 9, 10	33,3 %		33,3 %	33,3 %
Facteur 3 : 2, 8, 11, 15, 18	60 %		40 %	
Facteur 4 : 14, 17, 20		33,3 %	33,3 %	33,3 %
Test B				
Facteur 1 : 1, 5, 7	33 %	67 %		
Facteur 2 : 2, 9, 10, 11, 16	60 %		40 %	
Facteur 3 : 18, 20		50 %	50 %	
Facteur 4 : 8, 12, 16, 19	25 %			75 %
Facteur 5 : 4, 15		50 %	50 %	
Facteur 6 : 3, 6				100 %

L'examen des attributs reliés aux items des regroupements factoriels des deux tests montre que la majorité des items partagent l'attribut 5 (propriétés, définitions et liens) dans huit des dix facteurs. Pour le test A, les attributs ne parviennent pas à bien différencier ce que partagent les items des regroupements factoriels. Par contre, le facteur 2 du test B contient des items reliés à tous les attributs, et les facteurs 3 et 4 ajoutent l'organisation (attribut 6) à la technique (attribut 2) et à l'attribut 5. Notons que les informations additionnelles fournies aux conditions 2 et 3 ont minimisé l'intérêt des experts pour l'attribut 6.

Les comparaisons des paramètres du RUM

Les matrices Q synthèses simplifiées ont permis de modéliser les données des deux tests à l'aide du modèle RUM. Elles ont fourni des paramètres très robustes aux trois conditions pour le test A, et à la condition 1 pour le test B. La robustesse des paramètres estimés aux conditions 2 et 3 pour le test B a toutefois été jugée acceptable lorsque les paramètres *c* ont été fixés. Les analyses ont donc permis de vérifier que le fait de fixer la valeur du paramètre *c* augmente la robustesse des estimations et améliore la structure cognitive, comme l'ont annoncé dans la littérature Jang (2004) et Roussos *et al.* (2007). L'inconvénient dans la présente recherche vient de n'avoir pas pu comparer les paramètres *c* des trois conditions. La difficulté des attributs (p_k) et les paramètres d'items π et r ont été estimés pour chacune des matrices Q synthèses simplifiées. Dans le but de comparer les conditions expérimentales, on peut effectuer la comparaison de ces trois paramètres estimés dans chacune des trois conditions.

Les premières comparaisons concernent les paramètres (p_k), même s'il n'existe pas de lien direct entre ces paramètres et la validité ou la fidélité des matrices. Les moyennes des paramètres (p_k) ne diffèrent ni d'un test à l'autre, ni d'une condition à l'autre. Le résultat intéressant lié à la difficulté des attributs vient des représentations graphiques des valeurs des paramètres $\tilde{\alpha}_j$ de chaque sujet en fonction de son score total. Ces graphiques mettent en évidence un problème de définition des attributs dans la condition 3 par le partage des nuages de points en plusieurs parties. À titre d'exemple, la figure 3 présente les graphiques des trois conditions pour l'attribut 4 et le test B. Pour chaque nuage de points, l'axe des abscisses représente le score total des sujets, et l'axe des ordonnées représente la probabilité de maîtrise de l'attribut (difficulté de l'attribut pour chaque sujet). Un manque de convergence pourrait fournir une explication à ce phénomène dans le cas du test B, mais il se produit aussi pour le test A, pour lequel la convergence est bonne. Cette information suggère donc le manque de validité des attributs à la condition 3, puisque les nuages de points des conditions 1 et 2 sont plus homogènes.

Stout (2004) rapporte l'existence de ce problème, et la solution qu'il propose consiste à chercher à diviser les attributs concernés. Comme ce phénomène ne s'observe pas aux deux autres conditions, les informations de la condition 3 semblent aboutir à des attributs au contenu moins homogène.

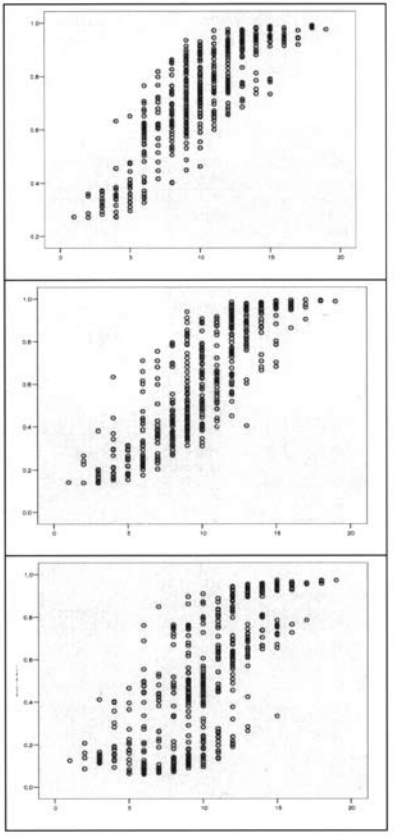


Figure 3 – Les distributions des difficultés des attributs pour chaque condition et l'attribut 4 en fonction du score total obtenu au test

Les paramètres π sont des indicateurs de la fidélité de la matrice et font l'objet des comparaisons suivantes. Les valeurs moyennes de ce paramètre sont semblables pour les deux tests. Les items qui manquent de fidélité, au sens d'une valeur de π inférieure à 0,6, sont les items les plus difficiles pour les deux tests. Même si les tests diagnostiques comportent traditionnellement des items moins difficiles que les tests de sélection, des attributs pertinents

reliés aux items difficiles sont pourtant essentiels pour que la structure de la matrice soit forte. Selon la valeur du paramètre π, les conditions ne permettent pas d'améliorer la fidélité des matrices ni d'identifier des attributs pour représenter convenablement les items les plus difficiles. D'après Roussos et ses collègues (2007), une valeur du paramètre π trop petite s'explique par des attributs qui ne prennent pas en compte la difficulté des items, parce que ces items font intervenir des attributs différents et plus difficiles à maîtriser ou, encore, plus nombreux.

Le dernier paramètre qui peut être comparé d'une condition à l'autre est le paramètre *r* qui est un indicateur de la validité de la matrice Q. Les valeurs de *r* estimées par Arpeggio mettent en évidence la meilleure structure cognitive du test B (avec plus de petites valeurs et moins de grandes valeurs de *r*) relativement au test A (voir figure 4). Le tableau 3 présente une synthèse des informations concernant les paramètres *r* pour chaque test, attribut et condition.

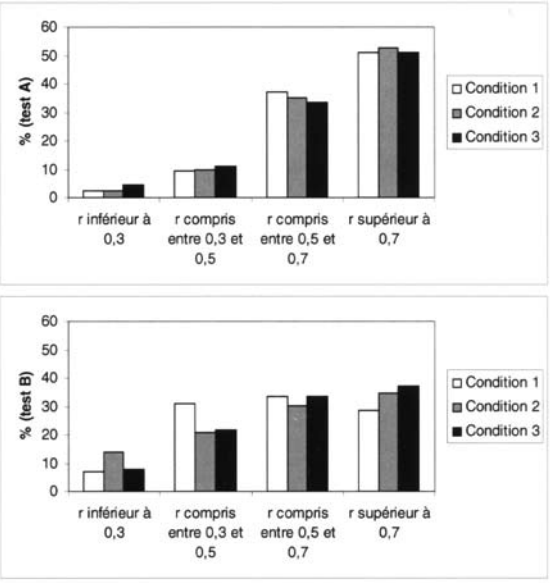


Figure 4 – Les pourcentages de valeurs *r* dans 4 intervalles pour les tests A et B et les 3 conditions

Tableau 3 – La synthèse des comparaisons des valeurs de *r* pour les trois conditions, par attribut et par test

	Critères									
	Nombre de valeurs de *r* au moins inférieures de 0,1 à une autre lorsqu'une comparaison est possible		Le nombre de valeurs de *r* inférieures à 0,4		Le nombre de valeurs de *r* inférieures à 0,6		Le nombre de liens **particuliers à une condition** pour lesquels la valeur de *r* est inférieure à 0,4		Le nombre de liens **particuliers à une condition** pour lesquels la valeur de *r* est inférieure à 0,6	
Test	A	B	A	B	A	B	A	B	A	B
Attribut 1										
Condition 1	2/2		0/4	0/2	3/4	0/2				
Condition 2										
Condition 3	0/2		0/12	4/18	0/12	8/18	0/12	4/16	0/12	7/16
Attribut 2										
Condition 1	4/6	4/8	2/16	4/14	7/16	8/14		2/2		2/2
Condition 2	1/6	2/8	1/16	1/13	6/16	5/13				
Condition 3	1/6	2/8	2/17	3/13	5/16	5/13	0/1	1/1	0/1	1/1
Attribut 3										
Condition 1	0/3	1/3	0/6	0/4	1/6	2/4	0/3		1/3	
Condition 2	**2/3**	**2/3**	0/7	**2/6**	1/7	3/6	0/4	0/2	0/4	0/2
Condition 3	1/3	0/3	0/1	0/6	1/1	**4/6**		0/1		0/1
Attribut 4										
Condition 1		0/3		0/2		2/2				
Condition 2	0/1	**2/3**	0/4	**2/5**	0/4	2/5	0/1		0/1	
Condition 3	1/1	1/3	0/6	1/4	0/6	2/4	0/3		0/3	
Attribut 5										
Condition 1	3/8	5/9	1/12	2/14	3/12	6/14	0/4	0/3	0/4	2/3
Condition 2	**4/8**	3/9	1/11	**3/15**	**3/11**	**7/15**	0/1	**2/4**	1/1	**3/4**
Condition 3	1/8	1/9	1/7	0/5	2/7	2/5		0/1		1/1
Attribut 6										
Condition 1		1/4	0/5	**2/6**	0/5	**4/6**	0/4	**2/2**	0/4	**2/2**
Condition 2		**2/4**	0/2	1/3	0/2	2/3	0/1		0/1	
Condition 3		1/4	0/1	0/5	0/1	2/5	0/1	0/1	0/1	0/1

Note. Les cases auxquelles aucune valeur ne correspond sont laissées vides. Les plus grands nombres de valeurs ainsi que la condition correspondante apparaissent en caractères gras.

Le tableau 3 illustre le fait que la condition 2 fournit globalement des valeurs de *r* plus petites pour les attributs 3, 4 et 5. C'est la condition 1 qui joue ce rôle pour les attributs 2 et 6. Les conditions 1 et 2 sont donc plus propices à une matrice plus valide, selon les paramètres *r*, que la condition 3. Même si les différences entre les conditions 1 et 2 sont souvent ténues et non statistiquement significatives, la condition 2 améliore la validité des liens des matrices Q correspondantes pour la moitié des six attributs. Dans une approche inductive, les experts tirent donc profit des informations issues de l'analyse factorielle préalable des données.

De plus, les résultats montrent que les paramètres de fidélité π (pour les deux tests : F(1, 104)=4,010 et p=0,048) et les paramètres de validité *r* (pour les deux tests : F(2, 251)=14,766 et p=0,000) sont statistiquement différents selon que les items sont ou non associés à d'autres par les analyses factorielles (voir figures 5 et 6). Les items qui partagent une dimension avec d'autres items correspondent à des liens statistiquement plus fidèles (π) et plus valides (*r*) que les items isolés dans les trois conditions. Ce résultat est valable, que les experts connaissent les liens entre les items (condition 2) ou pas (conditions 1 et 3). Pour revenir à la valeur du paramètre π, il est possible que la validité des items les plus difficiles soit moins aisée à obtenir parce qu'ils partagent moins de dimensions avec les autres items. Il paraît logique qu'un construit latent sous-jacent à un attribut cognitif se mesure mieux si plusieurs items contribuent à cette mesure. En outre, un tel résultat montre la nécessité d'élaborer un tableau de particularités dont chaque catégorie vise un contenu précis et comporte plus d'un item.

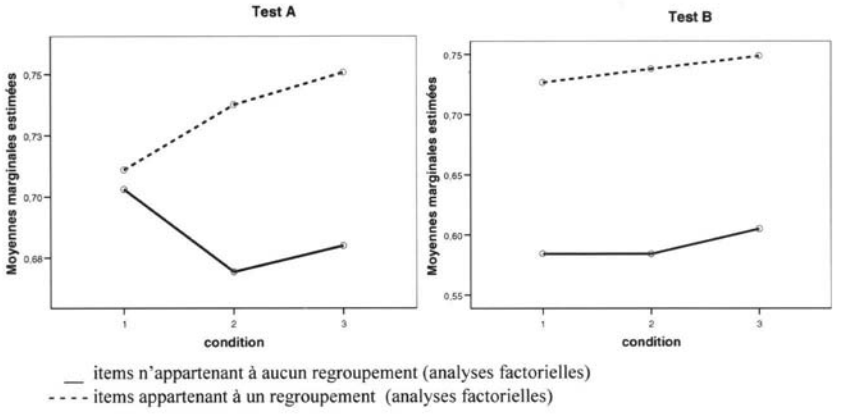


Figure 5 – Les valeurs moyennes de π pour les items appartenant ou non à un regroupement selon les analyses factorielles pour les deux tests

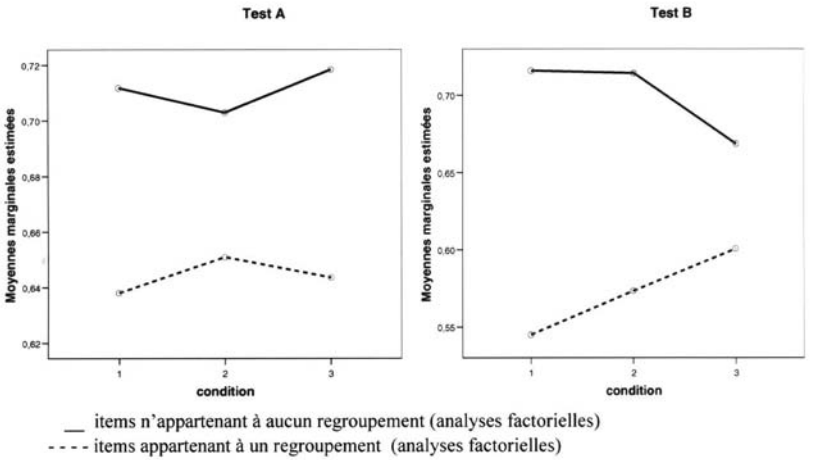


Figure 6 – Les valeurs moyennes de *r* pour les items appartenant ou non à un regroupement selon les analyses factorielles pour les deux tests

LA COMPARAISON DES TROIS CONDITIONS EXPÉRIMENTALES DE L'ÉTUDE

Dans la condition 1, les experts n'avaient accès qu'aux énoncés des items des tests A et B. L'aspect technique des items est bien cerné dans cette condition et ne l'est pas au détriment des autres aspects. Les valeurs des paramètres de difficulté montrent que les attributs sont bien définis, et les valeurs des paramètres r montrent une validité acceptable des liens présents dans les matrices Q synthèses simplifiées. Dans la condition 2, les experts devaient considérer les items tels qu'ils sont regroupés par une analyse factorielle. Les experts ont presque tous négligé l'attribut 1 concernant le décodage des questions. En revanche, les aspects liés au fait d'établir des liens entre des notions ou entre des définitions et des propriétés, et d'utiliser des représentations graphiques pour répondre aux questions ont été mieux exploités (attributs 3, 4 et 5). Les valeurs des paramètres de difficulté montrent que les attributs sont bien définis dans la condition 2, parfois mieux que dans la condition 1. Les valeurs des paramètres *r* montrent une validité acceptable des liens des matrices Q synthèses simplifiées de la condition 2.

La condition 3 a posé plusieurs problèmes : (a) elle a amené les experts à délaisser l'importance de faire des liens entre des définitions et des propriétés et d'organiser la réponse au profit de détails concernant la technique à utiliser, (b) les consignes fournies ont influencé les experts en ce qui concerne le décodage des items et biaisé les résultats, (c) les matrices Q de l'expert 1 ne

concordaient pas avec celles des trois autres experts pour le test B, (d) la convergence était moins bonne, (e) les valeurs des paramètres de difficulté des attributs montrent une mauvaise définition du contenu des attributs, et (f) les valeurs des paramètres *r* ne montrent pas d'amélioration de la validité à cette condition.

Dans cette étude, les conditions expérimentales ont été considérées comme indépendantes et traitées comme telles dans les résultats. Le contrôle mis en place pour garantir l'indépendance des conditions a consisté à laisser s'écouler un laps de temps entre chacune des conditions, à ne laisser en possession des experts aucun document d'une condition à l'autre et à leur demander de ne pas tenir compte (autant que possible) de leur travail précédent à chaque nouvelle condition. Les experts ont pourtant inévitablement gardé des souvenirs comme en témoignent plusieurs énoncés semblables issus des conditions successives. Toutefois, ils ne pouvaient pas se remémorer chacun des liens établis pour les quarante items. D'ailleurs, peu d'items sont restés totalement stables d'une condition à l'autre. S'ils ont retenu quelques informations, les experts ont visiblement refait l'exercice à la lumière des nouveaux documents fournis. On peut supposer qu'un apprentissage ou un cumul d'informations aurait pu amener la troisième condition à surpasser les autres, ce qui n'est pas le cas.

Il se peut aussi que le plan expérimental adopté ait encouragé les experts à faire un bon travail dans la condition 1, et à tenter de l'améliorer dans la condition 2 à la lumière des souvenirs qu'ils avaient gardés et des nouvelles informations dont ils disposaient. Le fait de leur demander de recommencer le travail une troisième fois peut les avoir conduits à remettre en cause les changements précédents et à en opérer de nouveaux plus importants, ou encore à bâcler le travail par lassitude. La meilleure qualité des matrices de la condition 2 serait alors le reflet de leur enthousiasme face à un travail qui n'était plus nouveau et pour lequel ils se sentaient bien préparés. La moins bonne qualité des matrices de la condition 3 serait le reflet de l'incertitude des experts face à une tâche trop répétée et devant laquelle ils étaient déstabilisés ou lassés. Toutefois, rien ne permet de valider ou d'invalider cette hypothèse.

CONCLUSION

Les résultats de cette étude constituent un premier pas vers la mise en place opérationnelle d'un modèle cognitif comme partie intégrante d'un test adaptatif, puisqu'ils présentent un certain nombre de considérations en lien avec (a) le choix des experts, (b) la méthode pour les faire travailler, et (c) les conditions à mettre en place pour favoriser l'élaboration d'une matrice

Q de qualité. Tout d'abord, les choix des experts et de leur nombre ont été faits dans une approche pragmatique dans cette étude. Il est donc difficile de savoir s'il suffit de recourir à quatre experts ou si un nombre plus important ferait une différence quant à la qualité des matrices Q obtenues. En ce qui concerne le choix de ces experts, les résultats ont toutefois mis en évidence le comportement marginal de l'expert 2. Cette différence a été rapidement décelée à partir de sa perception de l'utilité des informations fournies dans les conditions 2 et 3 et de l'élimination de quatre de ses attributs après le consensus. L'expert 2 est le plus diplômé, mais son diplôme de troisième cycle est un diplôme d'ingénieur dans lequel les mathématiques sont probablement plus appliquées que pures. En outre, son expérience d'enseignement concerne peu les niveaux secondaire et collégial, et il n'a passé que six de ses vingt années d'enseignement à enseigner les mathématiques. En comparaison, deux des autres experts avaient une expérience d'enseignement exclusivement au secondaire alors que le dernier enseigne à l'École Polytechnique et connaît donc parfaitement la population visée par le test. Comme cet expert est le seul à n'être pas diplômé en mathématiques pures et à n'avoir pas enseigné que les mathématiques, cette étude tend à suggérer de favoriser le recrutement d'experts dont la formation et l'expérience sont strictement reliées à la discipline du test.

Ensuite, en ce qui concerne la méthode pour faire travailler les experts, les modifications apportées à la méthode MACB ont permis d'obtenir une liste d'attributs cognitifs qui fasse consensus auprès des experts. En plus d'être conseillée pour fournir la liste des attributs, il apparaît que cette méthode aurait également pu servir à placer les liens dans les matrices Q synthèses. Par exemple, en fournissant aux experts les matrices Q individuelles modifiées à l'issue de la phase 2, il aurait été possible d'appliquer à nouveau la méthode MACB pour décider des liens pertinents ou non, en cherchant à obtenir un consensus. Une telle façon de faire n'impliquerait plus de vérifier la représentativité de la matrice Q synthèse, puisque le consensus interviendrait non seulement au niveau des listes d'attributs, mais également au niveau des liens qu'elle contient. Le seul inconvénient à cette façon de faire est le temps de travail supplémentaire demandé aux experts, mais elle aurait l'avantage de fournir des matrices Q synthèses encore plus représentatives du panel d'experts que celles utilisées dans cette recherche.

Enfin, en ce qui concerne les conditions à mettre en place, les résultats tendent à montrer que l'accès aux items regroupés par une analyse factorielle aide les experts à élaborer une matrice Q avec une structure cognitive un peu plus forte. Ces résultats sont issus de l'analyse de deux cas particuliers, et ne peuvent pas être généralisés sans que d'autres recherches soient menées. En outre, les différences entre la condition 1 et la condition 2 sont ténues. Toutefois, il est rassurant de voir qu'une autre étude visant à utiliser des

analyses factorielles pour identifier les attributs cognitifs a été menée en même temps que celle-ci (Leighton *et al.*, 2007) et qu'elle montre, elle aussi, la pertinence de cette approche.

En plus d'informer les experts des items qui se regroupent, il semble que les items qui partagent une dimension avec d'autres items soient caractérisés de manière plus valide et plus fidèle dans la matrice Q. Un soin particulier doit donc être porté à l'élaboration du tableau de spécification du test afin que sa structure soutienne l'interprétation des regroupements factoriels des items. Plus le lien entre le tableau de « spécificités » et les facteurs est étroit, mieux les experts semblent en mesure d'identifier les attributs que partagent les items. Ainsi, l'élaboration d'un tableau de « spécificités » du test aussi précis que possible devrait précéder l'analyse factorielle et la fabrication de la matrice Q.

Enfin, les résultats montrent que la condition 3 est peu favorable à l'obtention de matrices de qualité. Il semble donc logique de déconseiller de fournir les valeurs de la difficulté et du pouvoir discriminant des items en même temps que des informations sur le fonctionnement de leurres aux experts pour les aider dans leur tâche d'élaboration de la matrice. Il se peut que le fait de mettre plusieurs informations à la disposition des experts en même temps ne soit pas pertinent. Il aurait peut-être mieux valu présenter ces informations dans des conditions différentes. Il est également possible que ce type d'informations n'aide pas les experts à élaborer les attributs requis par chaque item. En outre, l'exemple qui illustrait les consignes a clairement biaisé les résultats dans la condition 3. Il convient donc soit de ne pas utiliser d'exemple, soit de définir un autre type d'exemple.

Ainsi, cette recherche exploratoire propose plusieurs pistes qui peuvent guider les chercheurs dans l'élaboration de la liste des attributs cognitifs en lien avec des items appartenant à une banque. Ces recommandations pourront être exploitées pour combiner un *testing* adaptatif à une approche cognitive, et utiliser la matrice Q pour le calibrage des items. Toutefois d'autres défis devront être relevés avant que la procédure ne soit opérationnelle.

Références

Buck, G. et Tatsuoka, K. K. (1998), « Application of the rule space procedure to language testing : examining attributes of a free response listening test », *Language testing*, *15*(2), p. 119-157.

Chipman, S. F., Nichols, P. D. et Brennan, R. L. (1995), « Introduction », dans P. D. Nichols, S. F. Chipman et R. L. Brennan (Ed.), *Cognitively diagnostic assessment,* Hillsdale, NJ, Lawrence Erlbaum Associates, p. 1-18.

Corter, J.E. (1995), « Using clustering methods to explore the structure of diagnostic tests », dans P.D. Nichols, S.F. Chipman et R.L. Brennan (Ed.), *Cognitively diagnostic assessment,* Hillsdale, NJ, Erlbaum, p. 305-326.

DiBello, L.V., Stout, W. F. et Roussos, L. A. (1995), « Unified cognitive/psychometric diagnostic assessment likeklihood-based classification techniques », dans P. D. Nichols, S. F. Chipman et R. L. Brennan (Ed.), *Cognitively diagnostic assessment,* Hillsdale, NJ, Lawrence Erlbaum Associates, p. 361-389.

DiBello, L., Roussos, L. A. et Stout, W.F. (2006), *Special topics section chapter : Review of cognitively diagnostic assessment and a summary of psychometric models*, Draft.

Dimitrov, D. M. et Raykov, T. (2003), « Validation of cognitive structures : a structural equation modeling approach », *Multivariate Behavioral Research, 38*(1), p. 1-23.

Gierl, M. J., Leighton, J. P. et Hunka, S. M. (2000), « Exploring the logic of Tatsuoka's Rule-Space model for test development and analysis », *Educational Measurement : Issues and Practices, 19*(3), p. 34-44.

Hartz, S. (2002), *A Bayesian framework for the unified model for assessing cognitive abilities : Blending theory with practicality*, Thèse de doctorat inédite, Champaign, IL, University of Illinois.

Huff, K. et Goodman, D. P. (2006), « The demand for cognitive diagnostic assessment », dans J. P. Leighton et M. J. Gierl (Ed), *Cognitive diagnostic assessment for education : Theory and applications*,New York, NY, Cambridge University Press, p. 275-318.

Jang, E. E. (2005), *A validity narrative : Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*, Thèse de doctorat non publiée, Champaign, IL, University of Illinois.

Junker, B. (1999), *Some statistical models and computational methods that may be useful for cognitively-relevant assessment*, Rapport de recherche préparé pour le Committee on the Foundations of Assessment.

Leighton, J.P., Gierl, M.J. et Hunka, S.M. (2002), T*he attribute hierarchy model for cognitive assessment,* Paper presented at the annual meeting of the National Council for Measurement in Education, Nouvelle Orléans, April.

Leighton, J. P., Goikert, R. J. et Cui, Y. (2007), « Using exploratory and confirmatory methods to identify the cognitive dimensions in a large-scale science assessment », *International Journal of Testing, 7*(2), p. 141-189.

Loye, N. (2005), « Quelques modèles modernes de mesure », *Mesure et évaluation en éducation, 28*(3), p. 51-68.

McGlohen. M. K.(2004), *The application of cognitive diagnosis and computerized adaptive testing to a large scale assessment*, Thèse de doctorat non publiée, Austin, University of Texas.

Messick, S. (1995), « Validity of psychological assessment », *American Psychologist, 50*, p. 741-749.

Milewski, G. B. et Baron, P. A. (2002), *Extending DIF methods to inform aggregate reports on cognitive skills*, Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April.

Muthèn, B. et Muthèn, L. (2004), *Mplus user's guide*.

Ramsey, J. O. (1996), « A geometrical approach to item response theory », *Behaviormetrika, 23*, p. 3-17.

Ramsey, J. O. (2000), « Testgraph [computer software and manual] », obtenu à l'adresse URL ftp://ego.psych.mcgill.ca/pub/ramsay/testgraf/

Ramsey, J. O. et Silverman, B. W. (1997), *Functional data analysis*. New York, Springer.

Roussos, L., DiBello, L., Stout, W.F., Hartz, S., Henson, R. A. et Templin, J. (2007), « The fusion model skills diagnosis system », dans J. P. Leighton et M. J. Gierl (Ed), *Cognitive diagnostic assessment for education : Theory and applications*, New York, NY, Cambridge University Press, p. 275-318.

Rupp, A. A. (2007), « The answer is in the question : A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models », *International Journal of Testing, 7*(2), p. 95-125.

Stout, W.F. (2004), *The Reparameterized unified model skills diagnostic system*, Conference presented at the Spearman diagnostics conference, Philadelphia, PA.

Tatsuoka, K. K. (1983), « Rule-space : an approach for dealing with misconceptions based on item response theory », *Journal of educational measurement*, *20*, p. 345-354.

Tatsuoka, K. K. (1990), « Toward an integration of item-response theory and cognitive error diagnosis », dans N. Frederiksen, R. Glaser, A. Lesgold et M. G. Shafto (Eds), *Diagnostic monitoring of skill and knowledge acquisition*, Hillsdale, NJ, Lawrence Erlbaum Associates, p. 453-488.

Tatsuoka, K. K., Birenbaum, M., Lewis, C. et Sheehan, K.M. (1993), « Proficiency scaling based on conditional probability functions for attributes », *Educational Testing Service*, Princeton, NJ.

Templin, J. L. (2004), *Generalized linear mixed proficiency models for cognitive diagnosis*, Thèse de doctorat inédite, Urbana-Champaign, University of Illinois.

Thissen, D., Steinberg, L. et Fitzpatrick, A.R. (1989), « Multiple-choice models : The distractors are also part of the item », *Journal of Educational Measurement, 26*(2), p. 161-176.

Vanderwood, M., Ysseldyke, J. et Thurlow, M. (1993), *Consensus building : A process for selecting educational outcomes and indicators*, National Center on Educational Outcomes, Report 2, University of Minnesota.

CHAPITRE 10

Les modèles graphiques

Michel C. Desmarais, Xiaoming Pu
École Polytechnique de Montréal

INTRODUCTION

Les tests adaptatifs peuvent être considérés comme une des premières applications d'un domaine aujourd'hui en pleine effervescence, celui des interfaces adaptatives et personnalisées. Le principe des interfaces adaptatives consiste à construire un modèle de l'utilisateur d'une application et à adapter le comportement de l'application à ce modèle. C'est précisément ce qu'un test adaptatif effectue. Il construit un modèle de la connaissance du répondant et adapte les questions qu'il propose en fonction de ce modèle, généralement dans le but précis d'en arriver à un diagnostic de connaissance avec un minimum de questions. Même si le domaine des interfaces adaptatives inclut une grande gamme de types d'adaptation, qui vont des préférences aux intentions de l'utilisateur (McTear, 1993), il demeure que les tests adaptatifs sont peut-être la toute première application à grande échelle du principe de base de ces applications : construire un modèle de l'utilisateur et adapter le comportement de l'application en fonction de ce modèle.

Curieusement, les deux domaines ont évolué plus ou moins en vases clos, chacun ignorant les développements de l'autre. Le domaine des interfaces adaptatives, et en particulier celui des environnements d'apprentissages adaptatifs, a donné lieu à plusieurs modèles de la connaissance d'un apprenant et à plusieurs techniques pour l'évaluer (Self, 1988). Alors que la Théorie des Réponses aux Items (TRI) était en plein développement dans le domaine des tests adaptatifs, le domaine des tutoriels intelligents développait ses propres approches pour représenter et diagnostiquer les compétences pour des environnements d'apprentissage adaptatifs (Carr et Goldstein, 1977). La majorité de ces efforts étaient basés sur des systèmes à base de règles et visaient à fournir un diagnostic très détaillé de la connaissance de l'apprenant.

La caractéristique principale de ces modèles était d'en arriver à un diagnostic précis qui relevait non seulement les concepts précis maîtrisés, mais aussi les conceptions erronées, ou les *mal-rules* (Payne et Squibb, 1990). Ces modèles avaient l'avantage d'offrir un haut niveau de *granularité*, c'est-à-dire qu'ils pouvaient offrir un diagnostic très précis des connaissances/compétences acquises ou manquantes, mais n'intégraient aucune notion de l'incertitude qui est inhérente à la modélisation de la connaissance.

À l'inverse, les travaux en psychométrie et les modèles de la TRI incorporaient d'emblée la notion d'incertitude et se consacraient en grande partie à estimer la fiabilité des modèles et les intervalles de confiance qui permettent de porter un diagnostic avec un degré de certitude déterminé. Cependant, la granularité des modèles basés sur la TRI demeure encore faible et se limite généralement à une dimension ou, dans le cas des travaux plus récents sur le TRIM (Théorie des Réponses aux Items Multidimentionnelle), à quelques dimensions simultanées, bien loin du niveau de granularité que l'on peut retrouver avec les modèles à base de règles des environnements de tutoriels intelligents.

Ces divergences s'expliquent facilement du fait que, dans le cas du domaine de la psychométrie, les exigences les plus fréquentes proviennent du contexte de l'évaluation sommative et consistent à déterminer si le répondant va réussir un test ou y échouer. Les exigences des environnements de tutoriels intelligents visent plutôt à déterminer les problèmes d'apprentissage de l'apprenant par le choix de capsules très spécifiques de contenu pédagogique visant à corriger des conceptions erronées ou à le guider vers un contenu plus avancé. Les exigences respectives des deux domaines sont donc fort différentes, ce qui explique en bonne partie le peu d'influence qu'ils ont eue l'un sur l'autre.

Le lien entre les deux domaines a émergé avec les travaux sur les modèles graphiques de la connaissance, parmi lesquels on peut mentionner ceux de Almond et Mislevy (1999), de Mislevy et Gitomer (1995), et le survol de Jameson (1995). En plus de démontrer que le modèle de la TRI pouvait être considéré comme un modèle graphique, ces travaux ont mis en valeur l'importance, pour tout diagnostic de l'apprenant, d'incorporer des mesures de l'incertitude d'une part et, d'autre part, de faciliter un diagnostic précis comme l'exigent les environnements informatisés d'apprentissage avancés.

Le présent chapitre fournit une comparaison de différentes approches propres à chaque domaine. Ces approches sont décrites dans les premières sections, puis nous rapportons une série d'expérimentations qui offrent une comparaison quantitative. Finalement, nous proposons une discussion autour des résultats observés et des critères qualitatifs qui permettent d'apprécier les avantages et inconvénients de chaque approche.

LES MODÈLES GRAPHIQUES ET L'ÉVALUATION DE LA CONNAISSANCE

Les modèles graphiques sont depuis longtemps utilisés pour représenter un domaine de connaissances (voir, par exemple, Findler, 1979). Ils offrent à la fois la visualisation d'un domaine de connaissances intuitives et un formalisme à partir duquel on peut construire des modèles rigoureux et même mathématiques. On les utilise maintenant de façon très courante dans le domaine de la représentation et du diagnostic des connaissances et compétences acquises.

Concepts, items et nœuds cachés

Les nœuds d'un modèle graphique se catégorisent typiquement en deux catégories :

- les concepts ;
- les items.

Les nœuds « items » représentent des questions, des exercices, ou toute autre manifestation observable d'une compétence de l'individu. Les nœuds « concepts » représentent des compétences « latentes ». Ils ne sont pas directement observables. Ainsi, on classifie les nœuds d'une structure graphiques selon qu'ils sont observables ou non observables, et on les nomme aussi des nœuds « cachés ».

La distinction entre un nœud observable et un nœud caché est fondamentale non seulement dans une perspective épistémologique, mais aussi par rapport aux techniques que l'on utilise pour construire automatiquement ces structures à partir de données et pour effectuer le diagnostic de connaissances.

La figure 1 illustre un modèle graphique comprenant des nœuds cachés, les concepts, et des nœuds observables, les items. Les concepts sont en fait divisés en deux groupes :

- les compétences, nommées *concepts* dans la figure ;
- les compétences fausses, nommées *concepts erronés*.

Alors que les compétences déterminent le succès aux items, les fausses compétences, ou concepts erronés, sont plutôt source d'échecs. Ce sont des erreurs typiques que l'on trouve fréquemment dans un domaine de connaissances.

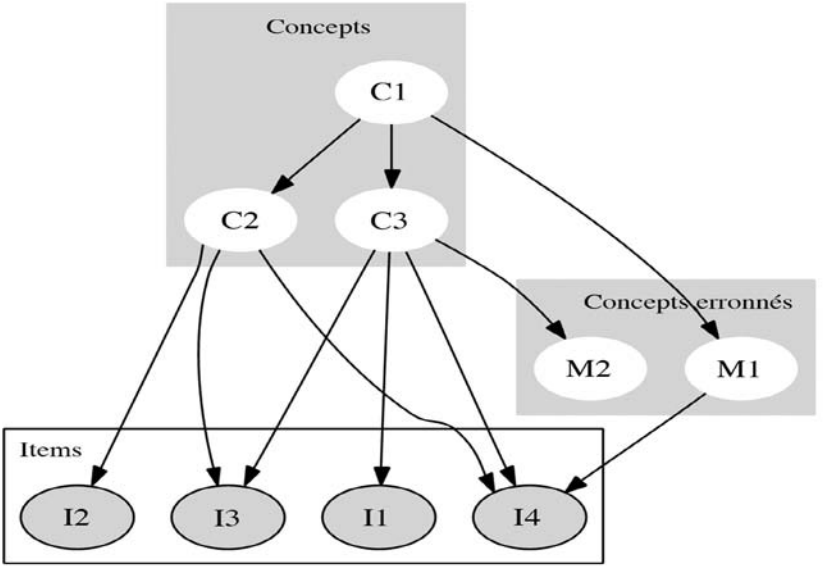


Figure 1 – Exemple d'un modèle graphique

Par exemple, lors de l'apprentissage des opérations sur les fractions, on trouvera souvent, pour une addition de deux fractions, l'erreur consistant à additionner les numérateurs et dénominateurs ensemble, et dénotée par la règle ici :

$$\frac{a}{b} + \frac{c}{d} \overset{e_1}{\rightarrow} \frac{a + c}{b + d} .$$

Évidemment, la bonne façon de procéder consiste à transformer les fractions afin d'obtenir le dénominateur commun. On pourrait décomposer cette compétence, correspondant à la règle :

$$\frac{a}{b} + \frac{c}{d} \overset{c_1}{\rightarrow} \frac{ad + cb}{bd} ,$$

en deux sous-compétences, correspondant aux règles et :

$$\frac{a}{b} + \frac{c}{d} \xrightarrow{c_2} \frac{ad}{bd} + \frac{cb}{bd} \xrightarrow{c_3} \frac{ad + cb}{bd} .$$

La notion « d'additionner avec dénominateur commun » sera donc considérée comme un concept, ou une compétence précise que l'élève devra maîtriser pour l'addition de fractions, et on la reliera à deux autres concepts plus précis.

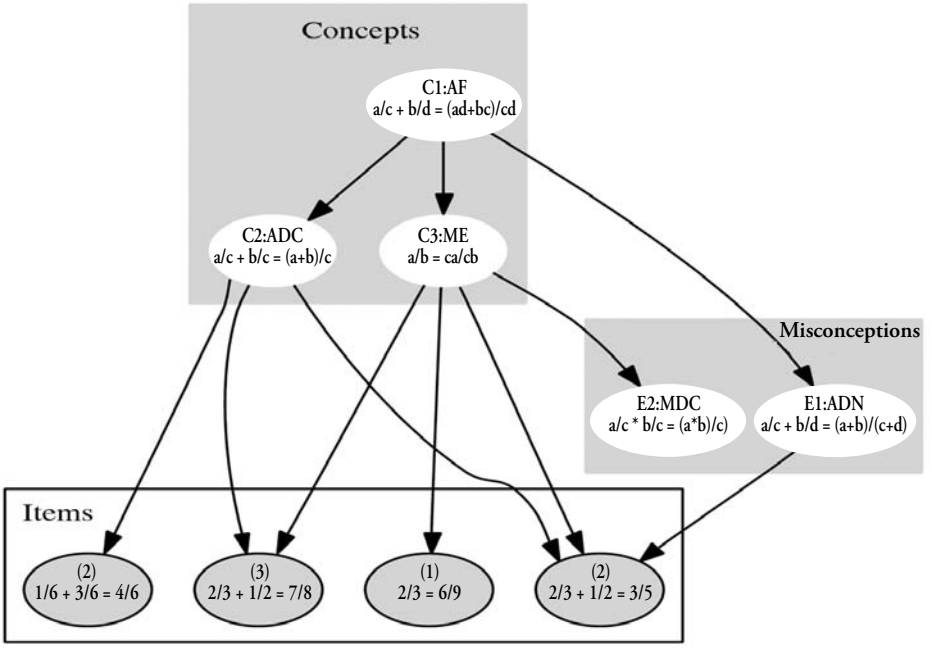


Figure 2 – Exemple d'un modèle graphique avec des concepts et items du domaine de l'arithmétique des fractions

La figure 2 fournit un modèle graphique qui illustre ces relations entre des items, des concepts et des concepts erronés pour le domaine de l'arithmétique des fractions. On y retrouve quatre items, dont le numéro [4] constitue une réponse erronée. Il est ainsi lié au concept erroné E1:ADN, qui représente l'erreur de la règle . Les concepts, ou compétences C2:ADC (addition avec dénominateur commun) et C2:ME (multiplication par un entier) représentent respectivement des notions de transformations algébriques qui sont impliquées dans la maîtrise des items auxquels ils sont liés dans la figure. Le concept de plus haut niveau, C1:AF (*addition de fractions*) est quant à lui relié aux deux concepts de niveau inférieur C2:ADC et C3:ME.

Comme on peut le constater, les modèles graphiques ont un haut niveau d'expressivité grâce à la représentation des liens complexes, des compétences précises et des compétences de plus haut niveau, et des concepts erronés expliquant des erreurs particulières que l'élève pourra effectuer sur des items bien précis.

Notons toutefois que la valeur cognitive d'un modèle graphique dépend essentiellement du choix des nœuds, des relations, et de la sémantique que l'on accorde à une relation entre les nœuds. Ainsi, le modèle de la figure 2 n'est fondé que sur une intuition des compétences latentes impliquées dans la résolution des quatre items et non pas sur un modèle cognitif démontré. La signification des relations n'est pas clairement définie et demeure intuitive elle aussi. Elle pourrait en fait signifier « est relié à » de façon générale.

Nous verrons toutefois que la sémantique des liens diffère radicalement d'un modèle graphique à un autre. Particulièrement lorsqu'on utilise un modèle graphique probabiliste dans une perspective prédictive ou de diagnostic. C'est notamment le cas des réseaux bayésiens que nous verrons plus loin, et dont la sémantique est très particulière. Il importe donc de ne pas oublier que tous les modèles graphiques ne sont pas nécessairement fondés sur une base cognitive solide, et que la sémantique de leurs liens diffère d'un modèle à un autre.

L'approche de la TRI dans la perspective graphique

La Théorie des Réponses aux Items (TRI) est un modèle qui peut justement être représenté graphiquement par un modèle graphique dont la sémantique est celle des réseaux bayésiens. La figure 3 illustre le principe.

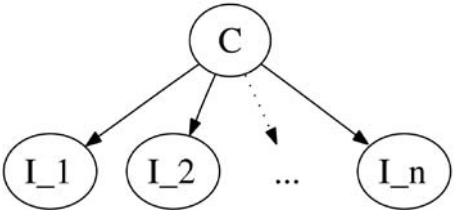


Figure 3 – Le modèle de la TRI représenté graphiquement

On retrouve, dans le modèle TRI de base, une seule compétence représentée par le concept C dans le graphique et à laquelle on réfère par θ dans le cadre de la TRI. La compétence θ représente essentiellement le fait qu'un test mesure une dimension, et que le succès à tous les items dépend de cette

dimension. De plus, en stipulant que la figure 3 se conforme à une représentation de réseau bayésien, on en conclut alors que, étant donné un niveau de compétence θ déterminé, le succès à un item est indépendant du succès à un autre item. En termes bayésiens, on peut donc conclure que :

$$P(I_i, I_j / \theta) = P(I_i / \theta)\ P(I_j / \theta).$$

C'est à dire que $P(I_i|\theta)$ et $P(I_j|\theta)$ sont indépendants l'un de l'autre. Cette conclusion correspond en fait à une des hypothèses fondamentales de la TRI : le succès à un item étant donné un niveau de compétence précis n'influence pas le succès à un autre item.

Cette définition correspond à ce que l'on nomme un réseau bayésien dit « naïf ». Compte tenu de l'hypothèse d'indépendance mentionnée ci-dessus, il en découle que :

$$P(I_1, I_2, \ldots, I_n / \theta) = P(I_1 / \theta)\ P(I_2 / \theta) \ldots P(I_n).$$

Cette équation correspond à l'hypothèse simplificatrice des réseaux bayésiens naïfs. Il existe donc un lien direct entre la TRI et la représentation graphique des réseaux bayésiens naïfs (Almond et Mislevy, 1999). La différence fondamentale entre un réseau bayésien naïf et la TRI réside dans le fait que le lien entre la probabilité de succès à un item, et un niveau de compétence, θ, est déterminé par une fonction sigmoïde qui, dans sa déclinaison à trois paramètres, correspond à :

$$P(I_i / \theta) = c_i + \frac{1}{1 + e^{-a_i (\theta - b_i)}},$$

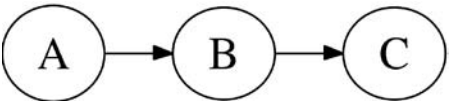
où

a_i est la capacité de discrimination d'un item i ; b_i le niveau de difficulté de l'item i, et c_i le facteur de chance (aussi appelé pseudochance) qui détermine une probabilité minimale de succès.

Le formalisme des réseaux bayésiens

Les réseaux bayésiens ne se limitent pas à une représentation graphique. Ils sont en fait une représentation sémantiquement très précise des relations de dépendance et d'indépendance entre les nœuds du réseau. Nous n'entrerons pas dans les détails de la sémantique du formalisme, mais en présentons néanmoins les principes fondamentaux. Le lecteur peut consulter Neapolitan (2004) pour une description plus complète.

Le formalisme repose sur la notion de « couverture de Markov » (*Markov blanket*), qui stipule des relations d'indépendance. Par exemple, dans la figure suivante :



le nœud *B* représente une couverture de Markov car, s'il est observé, alors les nœuds *A* et *C* seront indépendants l'un de l'autre, alors que, s'il ne l'est pas, *A* et *C* sont dépendants à travers *B*.

Les réseaux bayésiens ont la qualité à la fois d'avoir cette définition formelle de leur sémantique et d'être une représentation intuitive des relations entre des événements. Par exemple, si :

(A) Jean est un fumeur ;

(B) Jean a le cancer ;

(C) la radiographie pulmonaire de Jean indique une « tache »,

on comprend d'emblée que le graphique ci-dessus représente des relations de causalité. En outre, si on sait que *A* est une affirmation vraie, alors on aura tendance à croire que les événements *B* et *C* sont plus probables. Inversement, si on observe que *C* est vraie, alors on aura aussi tendance à croire que *A* est plus probable. Cependant, si l'on sait que *B* est vraie, alors le fait d'observer *A* n'augmente pas la probabilité de *C,* puisqu'on sait que c'est le cancer, *B*, qui est la cause de *C*. Finalement, c'est aussi le cas dans l'autre direction des flèches : sachant que *B* est vraie, l'observation de *C* n'apporte pas d'information supplémentaire quant à *A* et n'influe plus sur sa probabilité[1]. On retrouve donc la notion de couverture de Markov à travers cet exemple un peu intuitif des relations de cause à effet.

L'APPROCHE POKS ET LA THÉORIE DES ESPACES DE CONNAISSANCES

Une approche graphique qui s'est dernièrement démarquée dans le domaine des tests adaptatifs est celle des structures de connaissance avec ordre partiel, ou POKS (*Partial Order Knowledge Structures*). Elle s'inspire d'une part des travaux de Doignon et Falmagne (1999) sur les espaces de

1. On doit cependant présumer que la cigarette ne peut pas causer de tache pulmonaire autrement qu'à travers le cancer. Autrement, il y aurait alors un lien A → C, indiquant que le cancer n'est pas le seul mécanisme à travers lequel la cigarette peut entraîner une telle radiographie.

connaissance pour la représentation des connaissances, et d'autre part, des réseaux bayésiens naïfs pour l'inférence de la connaissance. Nous décrivons d'abord la théorie des espaces de connaissances, puis l'approche POKS.

La théorie des espaces de connaissances

La théorie des espaces de connaissances représente un domaine par un ensemble « d'items de connaissances ». Les items sont des manifestations de compétences comme la réussite à un exercice particulier ou la réponse à une question de connaissances. On ne réfère donc pas à des notions de compétences latentes ou de concepts dans cette théorie, mais exclusivement à des manifestations « observables » de ces compétences ou concepts.

L'état de la connaissance d'un élève est représenté par un sous-ensemble de ce domaine, les items que l'élève maîtrise. Il s'agit donc d'une façon très simple de représenter l'état de la connaissance d'un individu. Seuls des items dont la maîtrise par le répondant est observable sont impliqués dans un premier temps pour modéliser les compétences.

Dans un second temps, la maîtrise d'une compétence latente peut se faire en précisant quels items sont des manifestations de quelles compétences. On peut se reposer sur une sommation, potentiellement pondérée, des réussites/échecs attendus à ces items comme mesure de la compétence latente. C'est d'ailleurs ce que l'enseignant effectue couramment en découpant un examen en sections correspondant à différents sujets d'une matière. Cette approche contraste avec celle de la TRI qui représente explicitement la compétence du répondant dans le modèle par une dimension latente, θ.

La théorie des espaces de connaissances stipule en outre que les items d'un domaine de connaissances sont maîtrisés dans un ordre contraint. Prenons l'exemple de la figure 4 ci-dessous en présumant que l'élève doit trouver la valeur du terme de droite de chaque équation. On constate d'emblée que l'item *a* est plus difficile que les trois autres items. En ce sens, la réussite à *a* implique la réussite aux trois autres items. À l'inverse, l'item *d* est, quant à lui, plus élémentaire que les autres, et un échec à cet item implique un échec aux autres items. Par contre, il est plus difficile de se prononcer quant à un ordre donné entre les items *b* et *c*.

La figure 4 illustre ainsi un ordre partiel qui reflète des relations d'implications entre les items. Elle contraint l'ordre dans lequel les items peuvent être maîtrisés et, par conséquent, les différents états de connaissance à un ensemble spécifique de sous-ensembles :

$$\{\varnothing,\{d\},\{c,d\},\{b,d\},\{b,c,d\},\{a,b,c,d\}\}$$

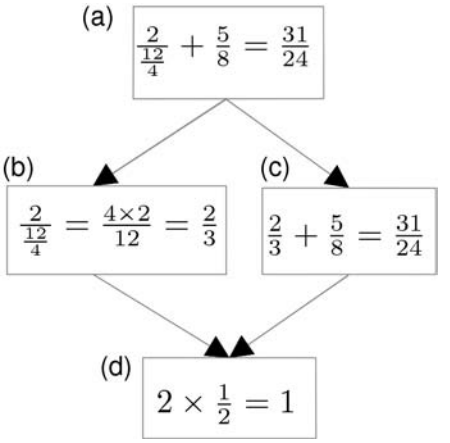


Figure 4 – Ordre partiel de la réussite à quatre items de connaissances {*a*,*b*,*c*,*d*}

Cette représentation des contraintes d'ordonnancement par un ordre partiel correspond en fait à la représentation utilisée dans l'approche POKS. Cependant, la théorie des espaces de connaissances comporte une nuance importante qui n'est pas intégrée dans le graphe dirigé de la figure 4 et qui porte sur les ordres alternatifs pour acquérir certaines compétences.

Prenons un exemple d'un ordre alternatif en revenant sur les trois concepts de la figure 2, C1 :AF, C2 :ADC, C3 :ME. Même si on présume que les transformations C2 :ADC et C3 :ME sont impliquées dans le calcul pour C1 :AF comme l'équation ci-dessous l'exprime, il est tout aussi plausible qu'un étudiant apprenne en fait la règle plus spécifique :

$$\frac{a}{b} + \frac{c}{d} \overset{c_1}{\rightarrow} \frac{ad + cb}{bd}$$

et qu'il n'ait en fait jamais acquis les notions plus élémentaires correspondant à C2 : ADC et C3 : ME (les règles et). Cette situation correspond à un apprentissage basé sur la mémorisation de règles pour résoudre des problèmes toujours de plus en plus spécialisés, plutôt qu'un apprentissage basé sur une résolution de problème fondée sur des règles plus générales et que l'on combine pour résoudre des problèmes plus complexes.

La théorie des espaces de connaissance repose donc non pas sur un ordre partiel, comme celui de la figure 4, mais bien sur un graphe qui permet la représentation d'alternatives. Nous n'irons pas davantage dans les détails de la théorie des espaces de connaissance, mais référons plutôt le lecteur à

Falmagne et *al.* (1990) et à Doignon et Falmagne (1999) pour une description complète de la théorie. De plus, il est intéressant de noter que cette théorie a donné lieu à une application de guide d'étude commerciale nommé ALEKS (www.aleks.com) et qui est aussi décrit dans Falmagne et *al.* (2006).

L'induction le la structure

Les structures graphiques donnent la représentation et l'inférence des compétences d'un individu, comme l'exemple de la figure 4 le montre. Une telle structure peut être construite par un spécialiste du domaine capable de juger de l'ordonnancement. Ceci est faisable lorsque le nombre d'items demeure faible. Cependant, lorsque le nombre augmente, la tâche devient rapidement insurmontable, étant donné le nombre de liens à considérer. Il est donc indispensable d'élaborer une approche qui permet de construire automatiquement cette structure à partir de données.

La construction de la structure se fait à partir de réponses d'un échantillon de sujets à un test comportant des items de connaissances qui représentent les nœuds de la structure. Le processus consiste en une comparaison de chaque paire d'items afin de déterminer si une relation entre les deux existe.

Par exemple, prenons la distribution de répondants à un test pour deux items, X_a et X_b telle qu'elle pourrait être déduite du tableau 1. Si on a une relation $X_a \rightarrow X_b$ comme on la retrouve à la figure 4, alors on peut s'attendre à trouver des répondants dans chacune des trois premières conditions ($\{x_{a,b}, x_{\neg a,\neg b}, x_{\neg a,b}\}$) mais pas pour la quatrième ($x_{a\neg b}$). En effet, étant donné la relation $X_a \rightarrow X_b$, on ne s'attend pas à trouver un répondant qui réussit X_a mais qui échoue à l'item normalement plus facile, X_b.

Tableau 1 – Tableau de contingence des possibilités de combinaisons des items *et* X_a *et* X_b

Condition		X_a	X_b	Répondants
(1)	x_{ab}	Réussi	réussi	oui
(2)	$x_{\neg a\neg b}$	Échoué	échoué	oui
(3)	$x_{\neg ab}$	Échoué	réussi	oui
(4)	$x_{a\neg b}$	Réussi	échoué	non

Cependant, il y a un élément probabiliste dont il faut tenir compte. La relation $X_a \rightarrow X_b$ peut exister, mais ne pas être très forte, ou encore il peut y avoir du bruit dans les données, comme des items réussis par chance ou

ratés par mégarde. Ainsi, pour déterminer si un lien $X_a \rightarrow X_b$ existe entre les deux items X_a et X_b, les trois conditions suivantes doivent être vraies :

$$P([P(X_b|X_a) \geq p_c] \mid D) > (1\text{-}\alpha_c) \quad (1)$$

$$P([P(\neg X_a|\neg X_b) \geq pc] \mid D) > (1\text{-}\alpha_c) \quad (2)$$

$$P(X_b \mid X_a) \neq P(X_b) \qquad (p < \alpha_i) \quad (3)$$

- p_c est la probabilité conditionnelle minimale pour et ; une seule valeur est retenue pour le test de toutes les relations du réseau (0,5 en général)
- α_c est l'erreur alpha des tests de probabilité conditionnelle ; cette erreur détermine la proportion tolérée de relations dont la probabilité conditionnelle de la population est en réalité sous le seuil ; les valeurs usuelles sont entre 0,2 and 0,5.
- $p < \alpha_i$ correspond à l'erreur alpha de tolérance pour le test d'interaction.
- D est la distribution de fréquence conjointe de X_a et X_b dans un échantillon de données de test. Cette distribution est un tableau de contingence 2×2 tel qu'on peut l'observer au tableau 1.

La première condition (l'inégalité 1) stipule que la probabilité conditionnelle d'un succès pour l'item X_b étant donné un succès pour X_a doit être supérieure à un seuil p_c, et que nous pouvons conclure cela à partir d'un échantillon de réponses aux items D, avec un taux d'erreur plus petit que α_c.

La seconde condition (l'inégalité 2) est analogue à la première et stipule que la probabilité d'un échec à l'item X_a étant donné un échec pour X_b doit être supérieure à p_c, avec un taux d'erreur maximal de α_c étant donné la distribution D des réponses.

Ces deux conditions sont calculées à partir d'une distribution binomiale cumulative. Dans l'inégalité 1, la valeur de $P([P(X_b|X_a]|D)$ est obtenue par la sommation de la fonction binomiale pour toutes les distributions où $x_{a\neg b}$ est plus petit que la fréquence observée en D, c'est-à-dire :

$$P([P(X_b|X_a]|D) = \mathrm{P}(x \leq x_{a\neg b}|X_a)$$

$$= \sum_{i=0}^{x_{a\neg b}} \mathrm{Bp}(i, x_a, p_c)$$

$$= \sum_{i=0}^{x_{a\neg b}} \binom{x_a}{i} p_c^{(x_a - i)} (1 - p_c)^i,$$

où $x_a = x_{ab} + x_{a\neg b}$.

La probabilité conditionnelle de la seconde condition repose sur la même fonction, mais utilise $Bp(i, x_{\neg b}, p_c)$ plutôt que $Bp(i, x_a, p_c)$. La troisième condition (inégalité 3) représente un test d'indépendance, et il est vérifié par un test χ^2 avec la distribution D, le tableau de contingence 2×2 : $P(\chi^2)<\alpha_c$. Pour de petits échantillons, ce test peut aussi être remplacé par le test exact de Fisher.

Le choix de la valeur pour p_c détermine la force de la relation d'implication entre deux items. Par exemple, si on a $X_a \rightarrow X_b$ et que l'ordre dans lequel on maîtrise ces deux items est fortement contraint, alors la valeur de $P(B|A)$, telle que déterminée par la distribution des fréquences D, sera très près de 1. La valeur de p_c représente donc la limite inférieure à laquelle on accepte de retenir une relation. Le choix est relativement arbitraire, mais il ne doit logiquement pas être sous $p_c = 0{,}5$.

Les deux valeurs α_c et α_i représentent les erreurs alpha que nous sommes prêts à tolérer pour conclure que la condition est remplie. Pour de très petits échantillons, ces valeurs peuvent être aussi élevées que 0,5 afin que soient conservées le plus grand nombre de relations possible. Dans les expériences décrites plus loin, les valeurs varient entre 0,2 et 0,1.

L'inférence de la connaissance

Une fois la structure construite, on l'utilise pour choisir des items à présenter au répondant à un test et pour ensuite recalculer la probabilité de succès aux autres items à la suite d'un succès ou d'un échec. Comme nous l'avons vu dans les sections précédentes, POKS utilise un modèle de sous-ensemble d'items maîtrisés ou non maîtrisés pour modéliser l'état de connaissance du répondant, à l'instar de la théorie des espaces de connaissances. C'est donc sur la base du sous-ensemble des items estimés réussis, donc ayant une probabilité de 0,5 ou plus d'être réussis, que l'on peut ainsi évaluer les compétences acquises. Le calcul de la mise à jour des probabilités après l'observation d'un succès ou d'un échec est expliqué dans cette section.

La mise à jour des probabilités, ou l'inférence de la connaissance, peut se réaliser selon deux versions. Dans la première, on propage l'information de l'observation d'un succès ou d'un échec de façon transitive à travers le réseau. Par exemple, si l'on a $X_a \rightarrow X_b \rightarrow X_c$ et que l'on observe un succès à X_a, alors les probabilités de X_b et de X_c seront affectées. Pour la seconde version de cet algorithme, on ne propage que pour X_b, et la valeur de X_c n'est pas affectée, à moins que, lors de la construction du réseau, le lien transitif $X_a \rightarrow X_b$ ne soit aussi été dérivé. Nous ne décrivons ici que cette dernière

version, mais les détails de la première sont décrits dans Desmarais, Maluf et Lui (1996) et Desmarais et Pu (2005).

La mise à jour des probabilités dans le cadre de POKS reste essentiellement basée sur le calcul des probabilités postérieures tout en faisant l'hypothèse d'indépendance locale : c'est-à-dire que, pour deux items, X_1 et X_2, leur probabilité conjointe et conditionnelle étant donné un troisième, X_3 est indépendante, et, par définition, on a :

$$P(X1, X2|X3)=P(X1|X3)\ P(X2|X3).$$

Cette hypothèse est nécessaire pour simplifier les calculs non seulement de la construction de la structure et de l'inférence, mais surtout parce qu'autrement, il nous faudrait une quantité de données beaucoup plus importante. En effet, cette hypothèse nous permet de présumer que seules des relations binaires peuvent être évaluées alors que, dans les autres cas, il faut valider des relations à trois variables et plus. Il nous faut alors recueillir des données dans un tableau comportant n entrées, où n correspond au nombre de variables pouvant être en interaction.

Sous l'hypothèse de l'indépendance locale mentionnée, le calcul de la probabilité d'un item après l'observation d'une série d'items correspond à un enchaînement de calcul de probabilités postérieures, comme celui décrit ci-dessous.

Supposons premièrement une relation $X_a \rightarrow X_b$, la probabilité postérieure de X_b étant donné l'observation d'un succès ou d'un échec à X_a est calculée sur la base du théorème de Bayes dans sa version du rapport de vraisemblance :

$$O(X_b / X_a) = O(X_b) \frac{P(X_a / X_b)}{P(X_a / \overline{X}_b)} ,$$

$$O(X_b / \overline{X}_a) = O(X_b) \frac{P(\overline{X}_a / X_b)}{P(\overline{X}_a / \overline{X}_b)} ,$$

où $O(X_b)$ est la vraisemblance initiale, et $O(X_b| X_a)$ représente la vraisemblance de X_b étant donné l'observation de X_a. La vraisemblance est ici définie dans sa forme habituelle :

$$O(X_b / X_a) = \frac{P(X_b / X_a)}{1 - P(X_b / X_a)} .$$

Afin de combiner les observations provenant de plusieurs items, le processus d'inférence se base sur l'hypothèse d'indépendance locale pour simplifier le calcul. Ainsi, la mise à jour après l'observation de plusieurs items $X_1, X_2, \ldots, X_n$ de l'item, est le simple produit des rapports de vraisemblances. Par exemple, si on suppose que l'on a n relations de la forme , alors :

$$P(X_1, X_2, \ldots, X_n / X_b) = \prod_{i}^{n} P(X_i / X_b) .$$

Il découle de cette équation que la nouvelle probabilité de étant donné $X_1, \ldots, X_n$ peut être réécrite dans la forme de ratio de vraisemblance par :

$$O(X_b / X_1, X_2, \ldots, X_n) = O(X_b) \prod_{i}^{n} \frac{P(X_i / X_b)}{P(X_i / \overline{X}_b)} .$$

Si l'observation correspond à un échec pour X_i, alors le rapport $\frac{P(\overline{X}_i \mid X_b)}{P(\overline{X}_i \mid \overline{X}_b)}$ est utilisé.

L'hypothèse d'indépendance locale sur laquelle ces calculs reposent est très forte et caractéristique des approches bayésiennes dites naïves. Elle simplifie grandement les calculs et la quantité de données de calibration nécessaire, comme nous l'avons mentionné. Bien que cette hypothèse soit souvent incorrecte dans plusieurs situations, il a été démontré que, malgré tout, les résultats obtenus sont relativement robustes (voir Domingos et Pazzani, 1997 ; Rish, 2001 ; Friedman, Geiger et Goldszmidt, 1997).

D'un ordre partiel à un ensemble de réseaux à un niveau

Comme nous l'avons précisé dans la section précédente, la propagation d'une observation ne se poursuit pas sur la base de la transitivité des relations, contrairement aux études de Desmarais et Pu (2005), et de Desmarais *et al.* (1996). Par exemple, si l'on a A → B et B → C, la probabilité de *C* demeure intacte après l'observation de *A*, à moins que le lien A → C soit explicitement dérivé des données. Cependant, s'il existe un ordonnancement fort du type A→B→C, alors on s'attend à trouver aussi A→C.

Ce principe est illustré à la figure 5. Un réseau POKS est ici transformé en trois réseaux plus simples ne comportant qu'un niveau. La ligne pointillée de l'ordre partiel est normalement dérivée des données si les nœuds du réseau sont fortement ordonnancés, c'est-à-dire si les trois conditions (1), (2), et (3) s'approchent des valeurs maximales.

En comparaison avec la variante de l'algorithme de Desmarais *et al.* (1996), cette version a l'avantage de ne pas nécessiter de schème pour la propagation dite « partielle », à travers des relations transitives, de *A* à *C* dans une structure comme A→B→C, par exemple. Étant donné qu'on peut s'attendre à ce que les relations transitives soient elles aussi dérivées par l'algorithme de construction de la structure, c'est-à-dire que l'algorithme inférera aussi la relation A→B→C à partir des mêmes données, les résultats devraient donc être similaires. Ce résultat s'est confirmé dans nos expériences exploratoires pour comparer les deux versions.

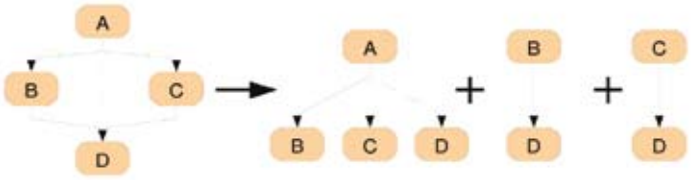


Figure 5 – Correspondance entre un graphe dirigé et un ensemble de réseaux à une couche utilisé pour l'inférence d'une observation positive (P=1)

Relations symétriques

Si les items se conformaient entièrement avec un ordonnancement partiel et qu'aucun item n'était équivalent, alors la structure obtenue à partir d'un échantillon de données serait effectivement un ordre partiel, ou un graphe acyclique dirigé. Cependant, si deux items étaient équivalents en termes de préalable et de difficulté, alors ces items seraient liés entre eux par des relations symétriques, et le réseau contiendrait alors des cycles. On pourrait donc simplement regrouper les items symétriques en un seul nœud pour reproduire un ordre partiel sans cycle. Le succès ou l'échec à un seul des items du groupe serait alors suffisant pour conclure sur le succès ou l'échec à tous les autres items.

Malheureusement, la réalité n'est pas aussi simple. En pratique, plusieurs nœuds ont des relations symétriques dérivées selon l'algorithme de la section sur l'induction de la structure, mais ils diffèrent en termes de préalable et de niveau de difficulté. Plus les valeurs retenues pour p_c et α_i sont tolérantes autour des conditions (1), (2) et (3), plus grand sera le nombre de relations symétriques dont les items ne sont pas équivalents. Cela se reflètera par des relations symétriques, mais comportant des valeurs respectives de $O(A/B)$ et

$O(B/A)$ fort différentes. Par exemple, prenons la relation symétrique A↔B. Si les deux items sont équivalents, alors $O(A|B) \approx O(B|A)$. Mais s'ils ne sont pas équivalents, une relation symétrique pourra néanmoins être dérivée bien que $O(A|B) \neq O(B|A)$,; tout dépend de la tolérance aux erreurs et de la « force » minimale des liens, toutes deux déterminées respectivement par les paramètres p_c et α_i.

La conséquence d'avoir des relations symétriques de la sorte est que la structure dérivée par l'algorithme d'induction n'est pas conforme à un ordre partiel. Cependant, avec l'algorithme d'inférence des connaissances utilisé, les cycles introduits par les relations symétriques n'ont pas d'impact sur la propagation d'inférences dans la mesure où l'algorithme ne propage pas les observations de façon transitive. Cependant, avec l'algorithme de Desmarais et Pu (2005) et de Desmarais *et al.* (1996), qui est récursif et suit les relations transitives, des mesures doivent être prises pour éviter de tomber dans des boucles infinies. Le principe le plus simple consiste à s'arrêter dès qu'un nœud a déjà été visité pour une même observation et à propager les inférences en « largeur » plutôt qu'en « profondeur ».

Il faut toutefois noter que la symétrie est la seule source possible de cycles dans une structure POKS. Car un cycle A→B, B→C et C→A ne pourrait se produire autrement qu'en ayant des relations symétriques (Desmarais *et al.,* 2006, pour une preuve formelle).

COMPARAISON DES PRÉVISIONS DES DIFFÉRENTS MODÈLES

Nous avons vu que les modèles de la TRI peuvent être considérés comme des modèles graphiques qui s'apparentent aux réseaux bayésiens naïfs. Le modèle repose sur une hypothèse d'indépendance locale et sur une relation entre les items et une compétence basées sur une fonction sigmoïde. Le modèle POKS s'apparente, lui aussi, aux réseaux bayésiens naïfs et repose sur un calcul de probabilité postérieur et une structure induite sur la base de tests statistiques pour inférer l'ordre d'apprentissage entre les items. À côté de ces deux modèles, on a aussi le modèle des réseaux bayésiens généraux qui sont aussi utilisés pour modéliser et inférer les compétences. Comment ces modèles se comparent-ils quant à leur capacité de prédire les résultats à un test de compétences ?

Nous rapportons ci-dessous une série d'expérimentations qui visent à déterminer comment ces approches se comparent entre elles en termes de performance. La première de ces expériences est décrite dans la prochaine section et porte sur la comparaison entre la TRI et POKS. Elle est suivie d'une comparaison entre l'approche POKS et les réseaux bayésiens.

Comparaison de POKS avec la TRI

Desmarais, Fu et Pu (2005) ont réalisé des simulations afin d'évaluer les prédictions des modèles POKS et TRI. Ces simulations portent essentiellement sur la capacité respective de chaque modèle de prédire le classement d'un répondant quant à sa réussite globale au test, ainsi que sur leur capacité à prédire individuellement le succès à chaque item.

Les expérimentations reposent sur la simulation d'un processus de questions-réponses. Pour chaque répondant, le processus est simulé avec des données complètes de test déjà collectées. Ainsi, les prédictions des approches TRI et POKS sont comparées aux données réelles.

Différentes variantes sont utilisées pour explorer le comportement des modèles. Ainsi, deux méthodes pour le choix de la prochaine question sont utilisées : l'information de Fisher et la réduction de l'entropie. L'information de Fisher est très répandue dans les applications de la TRI (voir Eggen, 1998, pour une comparaison des différentes stratégies de sélection d'items ave la TRI), alors que la réduction de l'entropie est une approche répandue dans le domaine des modèles probabilistes et déjà utilisée avec l'approche POKS (Desmarais et *al.*, 1996a).

La comparaison est effectuée sur deux corpus de données : un test de 160 questions de la connaissance du français utilisé dans la fonction publique canadienne et un test de la connaissance des commandes Unix qui comporte 34 questions.

Prédiction du succès au test

Les résultats de Desmarais *et al.* (2005) démontrent que les deux approches sont très efficaces pour classifier les répondants, avec un taux au-delà de 90 % de répondants bien classifiés après seulement 10 % des questions posées, dans une proportion d'environ la moitié des cas des simulations. Cependant, les résultats varient selon différents paramètres, notamment le score de césure et l'approche pour le choix des questions. Dans l'ensemble, on remarque que, pour l'approche du choix de questions basée sur l'information de Fisher, les deux approches ont des scores semblables (avec un score moyen global de 0,73 pour les deux et lorsque le paramètre α de POKS est optimal – ce à quoi on s'attend une fois le calibrage effectué). Lorsque l'approche du choix de questions est basée sur le gain d'information, POKS

réussit alors mieux que la TRI, avec un score moyen global de 0,79 comparativement à un score moyen de 0,73 pour la TRI[2].

Prédiction du succès aux items

D'autres simulations ont été effectuées pour déterminer leur capacité de prédire le succès individuel aux items. Cet objectif est évidemment plus difficile, mais il offre la possibilité de fournir un diagnostic très précis dans la mesure où certains items font appel à des compétences plus précises et que l'on peut poser un diagnostic avec une meilleure précision.

Les résultats démontrent une grande différence entre les corpus, ceux du test Unix affichant une performance nettement meilleure que ceux du test de français, et pour les deux modèles. Cependant, le modèle POKS affiche une meilleure performance dans les deux cas.

Ce résultat n'est vraisemblablement pas surprenant dans la mesure où l'approche TRI n'est pas conçue pour cette tâche et qu'elle ne peut pas traiter de plusieurs dimensions simultanément, donc distinguer les diverses compétences impliquées dans un test. Au contraire, le modèle POKS possède cette qualité, qui émane du fait que le formalisme utilisé est un ordre partiel plutôt qu'un ordre linéaire (par exemple,) qui, lui, ne pourrait modéliser adéquatement qu'une seule habileté. Il serait donc pertinent de vérifier si un modèle de TRI multidimensionnel peut réussir aussi bien que POKS.

Comparaison de POKS avec un réseau bayésien

Si la TRI simple ne peut gérer plusieurs dimensions à la fois, ce n'est certainement pas le cas des réseaux bayésiens, comme nous l'avons vu dans la section probabilité minimale de succès. **2.3** ; Desmarais *et al.* (2006) ont donc effectué une comparaison de la capacité d'un réseau bayésien de prédire le succès aux items individuellement. Un réseau bayésien est construit à partir des données avec deux méthodes courantes : l'algorithme PC (Spirtes, Glymour et Scheines, 2000) et l'algorithme K2 (Cooper et Herskovits, 1992). Les inférences avec les réseaux sont elles aussi basées sur des algorithmes courants dans le domaine. Des simulations de questions réponses sont réalisées à l'instar de la méthodologie utilisée pour la comparaison avec la TRI. Les mêmes données sont aussi utilisées, et les données d'un troisième test d'une vingtaine de questions en arithmétique sont ajoutées.

2. L'environnement de simulation ne permettait pas de tester le modèle de la TRI avec l'approche du gain d'information.

Les résultats démontrent que le modèle POKS réussit mieux que les réseaux bayésiens. Non seulement les prédictions sont plus justes, avec une réduction de l'erreur relative de l'ordre d'environ 5 % à 20 % en général, mais l'écart type des prédictions sur plusieurs simulations est beaucoup plus faible, dénotant une plus grande stabilité que le modèle basé sur un réseau bayésien. Tout comme pour la comparaison avec la TRI, les différences varient en ampleur d'un test à l'autre, mais la tendance est systématique.

Combiner les approches

On conclut donc qu'en ce qui concerne la prédiction aux items individuels, le modèle POKS semble avoir un avantage par rapport aux modèles basés sur la TRI et les réseaux bayésiens. Cela le rend donc bien indiqué pour la tâche de poser un diagnostic précis à partir d'un seul test.

Cependant, on peut toutefois se demander si les réseaux bayésiens ne sont pas mieux outillés pour exploiter les relations entre les compétences et les concepts de plus haut niveau. En effet, POKS ne représente pas ces compétences dans son modèle constitué uniquement d'items, c'est-à-dire de manifestations observables de compétences. Cette question a aussi été précisée par l'étude de Desmarais *et al.* (2006).

Les auteurs ont tenté d'améliorer le diagnostic des compétences de plus haut niveau par une technique qui consiste à augmenter les observations initiales avec POKS. Par exemple, si les items X_a et X_b sont posés au répondant et que POKS estime que l'item X_c a une chance de succès au-delà d'un certain seuil, alors les trois items seront considérés comme observés et fournis en entrée au réseau bayésien pour qu'il effectue son diagnostic de compétences.

Une expérience avec les données du test d'arithmétique a été faite avec cette technique. Les compétences de plus haut niveau en arithmétique de chaque répondant avaient été préalablement estimées de façon indépendante par des experts[3], ce qui permettait de valider les prédictions faites par le modèle.

Les résultats de l'expérimentation demeurent mitigés. Il a été démontré que POKS permet effectivement d'améliorer les prédictions aux items individuels et même de surpasser les prédictions faites avec POKS uniquement, bien que la différence ne soit pas statistiquement significative à $p<0{,}05$. Cependant, la prédiction du succès aux compétences de plus haut niveau n'a pu être améliorée. On peut en conclure que les inférences de POKS sont

3. Les données proviennent de Vomlel, 2004.

déjà exploitées par les relations du réseau bayésien qui expriment les liens entre les compétences. L'augmentation de l'ensemble d'items observés n'ajouterait ainsi rien au diagnostic. Néanmoins, le fait que la prédiction aux items soit améliorée contredit cette hypothèse dans une certaine mesure. Il est aussi possible que les évaluations indépendantes des compétences comportent trop de bruit, ou encore que l'amélioration des prédictions aux items ne soit qu'un fruit du hasard. D'autres études sont nécessaires pour élucider ces questions.

CONCLUSION

Ce chapitre présente une approche graphique et bayésienne pour la modélisation des connaissances, l'approche POKS. L'approche est inspirée à la fois par une théorie de l'apprentissage bien reconnue en psychologie cognitive, la théorie des espaces de connaissances (Doignon et Falmagne, 1999), de même que des réseaux bayésiens graphiques pour gérer le volet probabiliste. Nous avons illustré, dans un premier temps, que l'approche pouvait classifier des répondants comme « maîtres » avec un taux généralement aussi bon ou meilleur que l'approche classique du TRI à deux paramètres. D'autre part, comme un des avantages principaux de l'approche POKS est de fournir une prédiction pour la maîtrise de chaque item individuellement, nous avons aussi comparé les prédictions au niveau des items eux-mêmes. La comparaison est faite avec un modèle basé sur des réseaux bayésiens, car ils permettent aussi la prédiction de la maîtrise au niveau des items individuels. La comparaison révèle que l'approche POKS prédit la maîtrise aux items de façon plus précise que le modèle basé sur les réseaux bayésiens.

L'approche POKS semble donc offrir un modèle efficace pour la prédiction de résultats à un test de compétence, à la fois au niveau global du test comme la comparaison avec un modèle TRI le démontre, et au niveau précis des items individuellement comme le démontre la comparaison avec un modèle basé sur les réseaux bayésiens. Cependant, malgré ces avantages apparents, plusieurs incertitudes et questions importantes demeurent.

L'approche POKS s'est avérée généralement plus performante que le modèle TRI à deux paramètres, mais la question demeure quant à sa comparaison avec les développements plus récents du côté de la TRI, notamment les modèles multidimensionnels. Si un test était construit avec des items qui se conforment parfaitement à un nombre déterminé de dimension, l'approche POKS afficherait-elle toujours l'avantage que l'on observe ? Les résultats rapportés ici concernent des tests qui portent vraisemblablement sur plusieurs dimensions, comme le sont la plupart des tests, et ne sont donc probablement

pas optimaux pour un modèle unidimensionnel. Il est donc possible que les résultats soient différents pour une approche TRI multidimensionnelle.

Une seconde question est celle qui a trait à l'évaluation de la marge d'erreur et de la fiabilité des prédictions de POKS. Nous ne connaissons pas la distribution du taux de précision des prédictions de POKS, autrement qu'en le dérivant de simulations comme celles effectuées pour les expériences décrites ici. Or, dans certains contextes, il est essentiel de pouvoir estimer la marge d'erreur des prédictions. Le cas le plus évident est celui d'un test visant l'accréditation où l'on doit décider d'arrêter ou de continuer à présenter des questions au répondant. Il importe donc de mieux connaître les conditions qui influencent la fiabilité des prédictions du modèle POKS et, ultimement, de pouvoir dériver une marge d'erreur quant à ces prédictions.

De plus, on notera que les expériences réalisées jusqu'ici portent sur des tests de quelques dizaines jusqu'à cent soixante questions. Or, plusieurs contextes d'utilisation de tests adaptatifs nécessitent plusieurs centaines, voire des milliers de questions. Ces contextes posent une difficulté supplémentaire, celle de devoir construire un modèle POKS à partir de données partielles. Comme il n'est pas réaliste de faire passer un test de plusieurs centaines de questions à un répondant, le modèle doit être construit à partir de données incomplètes pour chaque individu. Le modèle de construction et d'inférence de la connaissance doit être adapté en conséquence. Enfin, ces contextes soulèvent aussi la question de savoir si le modèle demeure robuste avec des tests d'une telle ampleur. Cette démonstration reste à faire.

Références

Almond, R. G. et Mislevy, R. J. (1999), « Graphical models and computerized adaptive testing », *Applied Psychological Measurement*, *23*(3), p. 223–237.

Carr, B. et Goldstein, I. (1977), *Overlays : A theory of modelling for computer aided instruction* (Technical report), AI Memo 406, MIT, Cambridge, MA.

Cooper, G. F. et Herskovits, E. (1992), « A Bayesian method for the induction of probabilistic networks from data », *Machine Learning*, *9*, p. 309–347.

Desmarais, M. C., Fu, S. et Pu, X. (2005). « Tradeoff analysis between knowledge assessment approaches », dans C. Looi, G. McCalla, B. Bredeweg et J. Breuker (Ed.), *Proceedings of the 12th International Conference on Artificial Intelligence in Education, AEID'2005,* Amsterdam, IOS Press, p. 209–216.

Desmarais, M. C., Maluf, A.et Liu, J. (1996), « User-expertise modeling with empirically derived probabilistic implication networks », *User Modeling and User-Adapted Interaction*, *5*(3-4), p. 283–315.

Desmarais, M. C., Meshkinfam, P.et Gagnon, M. (2006), « Learned student models with item to item knowledge structures », *User Modeling and User-Adapted Interaction*, *16*(5), p. 403–434.

Desmarais, M. C. et Pu, X. (2005), « A bayesian inference adaptive testing framework and its comparison with Item Response Theory », *International Journal of Artificial Intelligence in Education*, *15*, p. 291–323.

Doignon, J.-P. et Falmagne, J.-C. (1999), *Knowledge spaces*. Berlin, Springer-Verlag.

Domingos, P. et Pazzani, M. (1997), « On the optimality of the simple Bayesian classifier under zero-one loss », *Machine Learning*, *29*, p. 103–130.

Duda, R. et Hart, P. (1973), *Pattern classification and scene analysis*, Wiley, New York.

Eggen, T. J. (1998), « Item selection in adaptive testing with the sequential probability ratio test », *Applied Psychological Measurement*, *23*, p. 249–261.

Falmagne, J.-C., Cosyn, E., Doignon, J.-P. et Thiéry, N. (2006), « The assessment of knowledge, in theory and in practice », dans R. Missaoui et J. Schmid (Ed.), *ICFCA*, vol. 3874 of *Lecture Notes in Computer Science*), Springer, p. 61–79.

Falmagne, J.-C., Koppen, M., Villano, M., Doignon, J.-P. et Johannesen, L. (1990), « Introduction to knowledge spaces : How to build test and search them », *Psychological Review*, *97*, p. 201–224.

Findler, N. V. (Ed.) (1979), *Associative networks : Representation and use of knowledge by computers*, Orlando, Academic Press.

Friedman, N., Geiger, D. et Goldszmidt, M. (1997), « Bayesian network classifiers », *Machine Learning*, *29* (2-3), p. 131–163.

Jameson, A. (1995), « Numerical uncertainty management in user and student modeling : An overview of systems and issues », *User Modeling and User-Adapted Interaction*, *5*(3-4), p. 193–251.

McTear, M. F. (1993), « User modelling for adaptive computer systems : a survey of recent developments », *Artificial Intelligence Review*, *7*, p. 157–184.

Mislevy, R. J. et Gitomer, D. (1995), « The role of probability-based inference in an intelligent tutoring system », *User Modeling and User-Adapted Interaction*, *42*(5), p. 253–282.

Neapolitan, R. E. (2004), *Learning Bayesian networks*, New Jersey, Prentice Hall.

Payne, S. J. et Squibb, H. R. (1990), « Algebra mal-rules and cognitive accounts of error », *Cognitive Science*, *14*(3), p. 445–481.

Rish, I. (2001), « An empirical study of the naive bayes classifier », *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, p. 41–46.

Self, J. (1988), « Bypassing the intractable problem of student modelling », *Proceedings of Intelligent Tutoring Systems, ITS'88*, p. 18–24.

Spirtes, P., Glymour, C. et Scheines, R. (2000), *Causation, prediction, and search*, Cambridge, Massachusetts, The MIT Press, 2e édition.

Vomlel, J. (2004), « Bayesian networks in educational testing », *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems*, *12* (Supplementary Issue 1), p. 83–100.

CHAPITRE 11

La médiatisation de récits d'expérience professionnelle

Kathye Bélanger et Michel Laurier
Université de Montréal

INTRODUCTION

L'étude présentée dans ce chapitre se penche sur le Récit d'Expérience Professionnelle Médiatisé (REPM[1]) comme stratégie de régulation des apprentissages dans le développement des compétences professionnelles chez les futurs praticiens (Bélanger, 2003). La formation des maîtres et, plus spécifiquement l'évaluation des apprentissages au primaire et au secondaire, est le domaine professionnel qui a été privilégié dans le cadre de ces travaux. Afin de répondre aux particularités d'une formation professionnalisante, nous avons choisi de mettre en scène des situations professionnelles authentiques et de placer le futur praticien en situation de résoudre un problème pour solliciter ses connaissances stratégiques. Ce faisant, nous cherchons à améliorer sa capacité à réfléchir sur l'action et à développer son jugement. Le REPM s'est avéré une activité favorisant l'individualisation de l'apprentissage et l'adaptabilité du dispositif interactif. La contribution des résultats discutés dans le présent chapitre trace des pistes d'améliorations et des tendances pour la continuité des travaux.

1. REPM est prononcé répèm.

RÉFLÉCHIR SUR L'ACTION POUR DÉVELOPPER LE JUGEMENT

Au Québec, le développement des compétences est une préoccupation dans plusieurs programmes de formation professionnelle à tous les ordres d'enseignement (secondaire, collégial, universitaire). Ces choix posent de nombreux défis dans le domaine de l'évaluation et peuvent avoir un impact sur les méthodes d'évaluation des apprentissages. Ainsi, la nature et la fonction du diagnostic pédagogique prennent différentes formes selon le type d'habiletés et le niveau de maîtrise visés par les apprentissages (Dassa et Laurier, 2003). Selon cette approche, nous nous pencherons non seulement sur les résultats des apprenants, mais aussi sur les stratégies mises en action pour apprendre.

Dans le but d'offrir un soutien à la régulation des apprentissages et de faciliter le suivi dans le cadre du cours *Méthodes d'évaluation des apprentissages* du programme de formation des maîtres de l'Université de Montréal, un dispositif médiatisé d'évaluation formative visant le développement de la compétence à *Évaluer les élèves* chez le futur enseignant a été développé et mis à l'essai. Dans ce contexte, « pour avoir une action sur le niveau d'expertise, il est nécessaire d'engager très rapidement l'apprenant dans des activités similaires à celles déployées par un expert dans un contexte relevant de son expertise » (Depover, Giardina et Marton, 1998).

Pour mettre en jeu les compétences spécifiques à la pratique enseignante et au niveau d'études visé par la formation des maîtres, le dispositif utilisé par les apprenants dans le cours *Méthodes d'évaluation des apprentissages* doit satisfaire à trois grandes exigences : mettre en scène des situations d'expérience professionnelle authentiques, placer l'apprenant en situation de résolution de problèmes pour activer ses connaissances stratégiques et, finalement, stimuler la réflexion du professionnel sur les jugements qu'il pose et la décision qui s'ensuit. D'une part, chaque praticien novice ou expert a un ensemble d'expériences de pratique qui lui est propre et a toujours une idée de ce qui aurait pu ou aurait dû être fait. D'autre part, en ayant accès à un ensemble de situations professionnelles, l'apprenant peut réfléchir et exercer son jugement sur les évènements exposés pour proposer à son tour des solutions. Le récit d'expérience professionnelle médiatisé permet l'exploration de cas mettant en scène des situations de vie authentiques reliées à un domaine de pratique et autorise l'actualisation de l'expérience professionnelle du futur

praticien (Lévy, 1995). Mettant à profit ses connaissances stratégiques, on lui demande de poser un jugement sur des décisions prises dans l'action et on l'amène ainsi à améliorer son autonomie professionnelle.

L'utilisation d'un dispositif peut faciliter ce processus d'évaluation formative dans un contexte de résolution de problèmes. Les situations de résolution de problèmes étant fréquentes dans le monde de la pratique, la formation professionnelle peut réactualiser des situations de pratique par le biais de récits d'expérience professionnelle. De fait, la médiatisation de ces récits d'expérience professionnelle facilite la contextualisation et la recontextualisation des apprentissages (Tardif, 1999). Grâce au dispositif, l'apprenant réalisant une performance complexe fait appel à ses connaissances stratégiques, cognitives et métacognitives, de façon à interagir adéquatement.

Pour mieux cerner les limites et la portée de nos travaux, un cadre conceptuel s'articulant autour des concepts suivants a été établi (voir Figure 1) :

- la flexibilité cognitive de l'apprenant mise à contribution dans le développement des compétences professionnelles (Spiro *et al.*, 1988) ;
- le savoir issu du domaine de la pratique et véhiculé par le discours professionnel (Wenger, 1998) ;
- le récit d'expérience professionnelle comme acte de sens (Ricoeur, 1975) et lieu d'actualisation des connaissances stratégiques des praticiens (Lévy, 1995) ;
- la régulation des apprentissages essentielle à l'acquisition de ces connaissances stratégiques dans le cadre de la formation professionnelle (Scallon, 2000) ;
- finalement, la résolution de problèmes, concept central et commun à chacun de ces éléments (Brien, 1994).

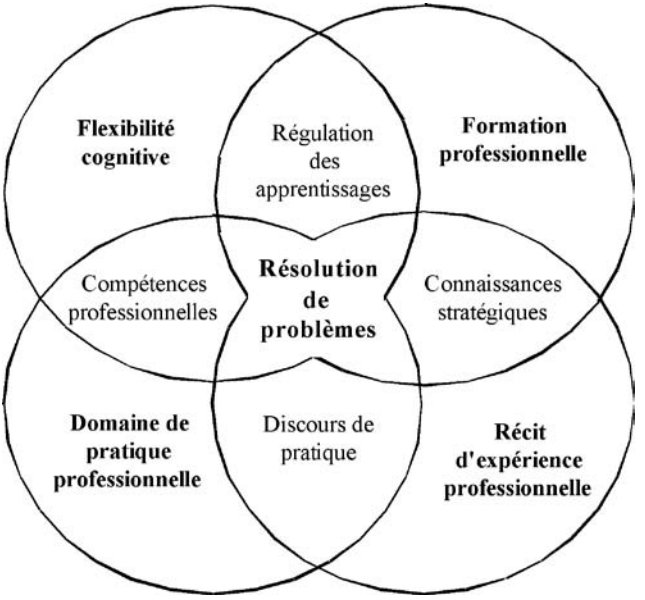


Figure 1 – Le cadre conceptuel

Sur cette base, l'objet de recherche qui nous préoccupe dans la présente étude est le récit et, plus particulièrement, celui-ci comme véhicule de l'expérience professionnelle. Depuis le milieu des années 70, Roger Schank et Robert Abelson ont mené des travaux sur le récit d'expérience comme véhicule du savoir humain. Ces travaux les ont amenés à concevoir la connaissance humaine comme issue de l'expérience et construite à partir de milliers de récits vécus, entendus ou composés par l'individu.

> […] Nous soutenons que les récits portant sur nos expériences ou celles des autres sont les constituantes fondamentales de la mémoire, de la connaissance et de la communication humaine. (Schank et Abelson, 1995, p.1 – traduction libre[2])

De son côté, Jonassen présente le raisonnement par cas dans ses travaux sur l'analyse de la tâche et suggère que, pour analyser ce que les individus savent, on peut simplement analyser leurs histoires.

> Nous nous rappelons vraiment très bien ce que nous savons sous forme de récits. Les récits sont des formalismes riches et puissants pour conserver et décrire nos souvenirs. Ainsi, une façon d'analyser ce que les gens savent est

2. […] we argue that stories about one's experiences and the experiences of others are the fundamental constituents of human memory, knowledge, and social communication.

d'analyser leurs récits. (Jonassen et Tessmer, 1999, p.147 – traduction libre[3])

Dans le contexte actuel, les changements en éducation entraînent inévitablement une redéfinition du savoir. Ce dernier revêt une dimension sociale très importante et son rapport à la pratique devient incontournable. Dans cette perspective, Wenger (1998) propose une théorie sociale de l'apprentissage. Pour cet auteur, le savoir est issu de la pratique, et le sens qui lui est conféré est vécu comme une production sociale. La négociation du sens a lieu dans la participation et la réification menées par l'individu dans sa communauté de pratique. Dans cette optique, l'éducation doit changer ses visées formatives pour des visées transformatives.

Pour l'enseignant du primaire et du secondaire, l'évaluation des apprentissages constitue une activité professionnelle d'une grande complexité. Reposant sur un processus de prise de décision, ce champ de compétence de la pratique enseignante sollicite le jugement de l'enseignant dans un contexte de résolution de problèmes. Pour répondre aux exigences de ce niveau de complexité, nous misons sur la flexibilité cognitive des futurs maîtres qui est sollicitée dans la régulation et le transfert des apprentissages. La flexibilité cognitive convient bien d'ailleurs à la multiplicité de représentations, contextes et perspectives spécifiques à la résolution de problèmes (Spiro *et al.*, 1990). Il appert que le récit d'expérience basé sur des situations de pratique authentiques est un terrain propice au déploiement du jugement chez le futur enseignant et l'encouragera dans le développement de sa compétence à *Évaluer les élèves*. Le récit d'expérience se limite ici à l'expérience professionnelle. Les situations d'expérience de vie relatent des faits réels ayant cours dans des contextes précis qui mettent en scène des acteurs et offrent de multiples points de vue. Le récit d'expérience correspond à une suite logique d'actions relatant des évènements ou des situations de pratique authentiques qui ont cours dans un contexte de vie professionnelle et auxquels participent des personnages en interaction. Ces expériences de vie issues de la pratique sont structurées et délimitées par un registre de compétences, un contexte professionnel ainsi que des rôles conférés dans le cadre de cette pratique, et peuvent être articulées en fonction d'un ensemble de connaissances stratégiques particulières (Wenger, 1998).

La médiatisation du récit d'expérience a pour enjeu l'échange entre un donateur et un destinataire (Barthes, 1966). Que l'on pense à une situation

3. « We remember so much of what we know in the form of stories. Stories are rich and powerful formalisms for storing and describing memories. So, one way to analyze what people know is to analyze their stories ».

de vie ou à un récit, un discours est tenu, et on peut se demander quel est le sens porté par ce discours. La sémiotique considère le mot comme à l'origine du sens du discours. La sémantique y oppose la phrase dans son contexte textuel. L'herméneutique, de son côté, pose l'ensemble du texte et son contexte évènementiel comme porteurs du sens du discours. Dans son ouvrage théorique sur l'analyse du discours, Ricoeur (1975) recadre chacune de ces composantes comme des éléments constitutifs du discours. Pour lui, le sens du discours a plutôt son origine dans la signification que lui accorde la personne qui le manipule. Le sens du discours se pose dans son caractère statique en terme de métaphore morte portant le sens visé par l'auteur, et dans son caractère dynamique en terme de métaphore vive portant le sens autorisé par le lecteur. Le REPM prendra un sens dans l'interactivité de l'apprenant avec le dispositif. Comme véhicule de l'expérience professionnelle, il abritera le sens visé par le praticien-narrateur et le sens autorisé par l'apprenant-lecteur. Ces derniers contribuant au sens porté par le discours, la dimension socio-affective y revêt un caractère participatif « en ce sens que le *mood* ainsi créé est une sorte de modèle pour *voir comme* et *sentir comme* […]. Or le sentiment est ontologique d'une autre manière que le rapport à distance, il fait participer à la chose » (Ricoeur, 1975, p. 309).

Un autre aspect à considérer pour entrevoir le potentiel du récit d'expérience est de l'envisager comme un lieu mettant en scène une situation de résolution de problèmes. En verbalisant, l'apprenant peut imaginer plusieurs solutions possibles. Cette verbalisation comporte une fonction heuristique. L'utilisation du langage plutôt que la conduite d'une activité concrète place l'apprenant dans une position réflexive et l'amène à penser à plusieurs façons de résoudre le problème. De cette façon, l'apprenant met à contribution des habiletés cognitives de haut niveau favorisant un niveau d'abstraction qui stimule la réflexion et permet d'aborder le problème selon plusieurs approches différentes, ce qui débouche sur un éventail de solutions plus élaborées et nuancées. Résoudre un problème avec des mots ou dans l'agir sont deux approches fondamentalement différentes. En passant par le langage plutôt que par l'agir, l'apprenant entre inévitablement dans un processus d'observation et d'analyse qui aboutit à une tentative de compréhension du monde. « Les mots analysent le monde, les mots sont notre médium principal pour l'analyse » (Vygotsky, 1934, p. 202 – traduction libre[4]).

À ce point-ci, on peut se demander quels sont exactement les processus cognitifs qui s'activeront chez l'apprenant lorsqu'il utilisera le dispositif médiatisé. L'apprenant pourra tenter de déterminer ce qui ne va pas dans l'histoire et envisager des solutions de rechange. Pour ce faire, il pourra

4. The word analyzes the world, the word is the first means of analysis.

aborder le problème à résoudre selon plusieurs points de vue, considérer le problème comme un tout, diviser le problème à résoudre en plusieurs petites parties et considérer les liens qui les structurent entre elles, se concentrer sur un aspect en particulier, examiner un éventail de possibilités (De Bono, 1967). Sous cet angle, le REPM stimule la réflexion sur l'action et exige d'avoir recours à des connaissances stratégiques. Il peut favoriser le développement du jugement et accroître les possibilités du transfert des apprentissages (Spiro *et al.*, 1990).

Le REPM orchestre une multitude de situations de pratique professionnelle. À partir d'un réseau de récits d'expérience professionnelle, l'apprenant s'engage dans un processus de résolution de problèmes. Sur la base d'un récit initial inadéquat, il proposera un récit d'expérience révisé ou reconstruit qui, selon lui, s'avérera amélioré. Pour cela, le futur praticien réfléchit à la situation professionnelle présentée dans le REPM. Sa réflexion peut l'amener à faire des ajustements au récit d'expérience. Les ajustements anticipés motivent la révision du récit. Après révision, une évaluation du récit permet de diagnostiquer d'autres aspects pouvant faire défaut. Ces aspects problématiques provoquent une nouvelle réflexion, et le cycle d'échange continue (Giardina, 1992). Dans la répétition de ce cycle, une régulation des apprentissages a cours pour favoriser le développement des compétences et le transfert des apprentissages (Giardina et Laurier, 1999). C'est ainsi que l'expertise du futur praticien progresse (Lave et Wenger, 1991). De cette façon, le REPM permet de transmettre le savoir issu d'une pratique professionnelle et de le mettre à profit dans le développement des compétences chez les futurs professionnels. Dans ce cycle d'échange de l'expérience professionnelle d'une communauté de pratique, le REPM apparaît comme un élément dynamique (Depover, Giardina et Marton, 1998) et prend son sens à partir des différents acteurs engagés dans la transmission du savoir professionnel et le développement de leur niveau d'expertise. Chacune des personnes engagées dans ce cycle d'échange professionnel possède son propre niveau d'expertise. L'expertise de chaque praticien se développe en continuum selon son implication dans le domaine professionnel. Tel que le précise Tochon (1993, p. 134), « l'expert est un professionnel qui réfléchit, sans cesse, sur la base des images de cas intériorisées par l'expérience. » De ses travaux, il ressortit que l'expérience ne suffit pas pour développer une expertise, et que cette dernière repose d'abord et avant tout sur des activités métacognitives autorégulatoires. Pour le futur praticien, la résolution de problèmes contextualisés donne accès à des simulations de situations professionnelles réelles qui facilitent une activité de réflexion orientée sur le recadrage et la remise en question des évènements relatés. Le partage de l'expérience et des réflexions de pratique d'un pair alimente la réflexion du

futur praticien. Si l'expertise n'est pas transmissible, la réflexion menant à l'expertise, elle, peut être stimulée.

Dans cette recherche, le REPM a donc été exploré comme une stratégie pour stimuler cette réflexion chez les futurs praticiens. Par le biais du REPM, on veut mettre à l'épreuve le jugement de l'apprenant en l'incitant à réfléchir sur l'action.

LE REPM, UN VÉHICULE POUR L'EXPÉRIENCE PROFESSIONNELLE

Nous abordons dans cette section le modèle descriptif du REPM qui a servi de base à la réalisation du dispositif d'évaluation formative. Cette représentation de l'objet de recherche a été conçue en termes de structuration du récit et d'organisation de ses composantes. Son élaboration n'en est qu'au stade d'ébauche et ne doit pas être considérée comme la formalisation de l'objet de recherche. Notre but étant d'explorer et de mieux comprendre la portée du REPM comme stratégie potentielle pour développer les compétences professionnelles, cette représentation nous permet de délimiter l'objet de recherche et d'encadrer le « prototypage » tout en tenant compte des conditions et des contraintes posées par le cadre conceptuel. Aussi, le modèle descriptif présenté ci-après s'appuie sur les prémisses suivantes :

- La connaissance humaine est en partie une agrégation, un raffinement et une transformation de milliers de récits vécus, entendus ou composés par l'individu (Schank et Abelson, 1995).
- La flexibilité cognitive constitue le mécanisme principal activé lors de performances complexes (Spiro *et al.*, 1988).
- Les pratiques professionnelles mettent à contribution des connaissances stratégiques (Wenger, 1998).
- Et, finalement, le sens porté par le discours provient de l'individu qui le manipule (Ricoeur, 1975).

Le REPM prend forme autour de la notion de récit et de savoirs issus de la pratique professionnelle. Tel que l'illustre la Figure 2, le REPM peut être abordé selon trois perspectives emboîtées l'une dans l'autre : A- le domaine de pratique professionnelle, B- l'individu en résolution de problèmes, et C- le récit. Chaque domaine de pratique regroupe plusieurs individus en situation de résolution de problèmes ayant chacun à leur actif plus d'un récit d'expérience professionnelle. Sur un autre plan, le REPM comporte quatre dimensions : I- la dimension sociale, II- la dimension affective, III- la dimension cognitive, et IV- la dimension de la régulation. Chacune des perspectives mentionnées auparavant peut être envisagée selon ces quatre dimensions. Ainsi, on peut voir qu'un domaine de pratique professionnelle est caractérisé par une communauté de pratique, des membres adhérant à cette communauté de pratique, un registre de compétences spécifiques et un éventail d'expériences professionnelles. À un autre niveau, on peut voir au cœur de cette pratique un individu en résolution de problèmes qui se définit par : son appartenance et sa participation à la communauté de pratique ; son identité professionnelle et les problèmes qu'il est appelé à résoudre ; le problème dans sa structure évènementielle et la compétence qu'il sollicite ; l'autoréflexion qui se concrétise dans la performance atteinte. Finalement, chaque résolution de problèmes menée par l'individu constitue un récit qui véhicule des situations de vie authentiques et est caractérisé par son type narratif, par sa forme narrative, par sa structure narrative et par l'acte narratif lui-même.

La Figure 2 montre aussi que le REPM peut être envisagé selon ces quatre dimensions différentes à la lumière de chacune des trois perspectives.

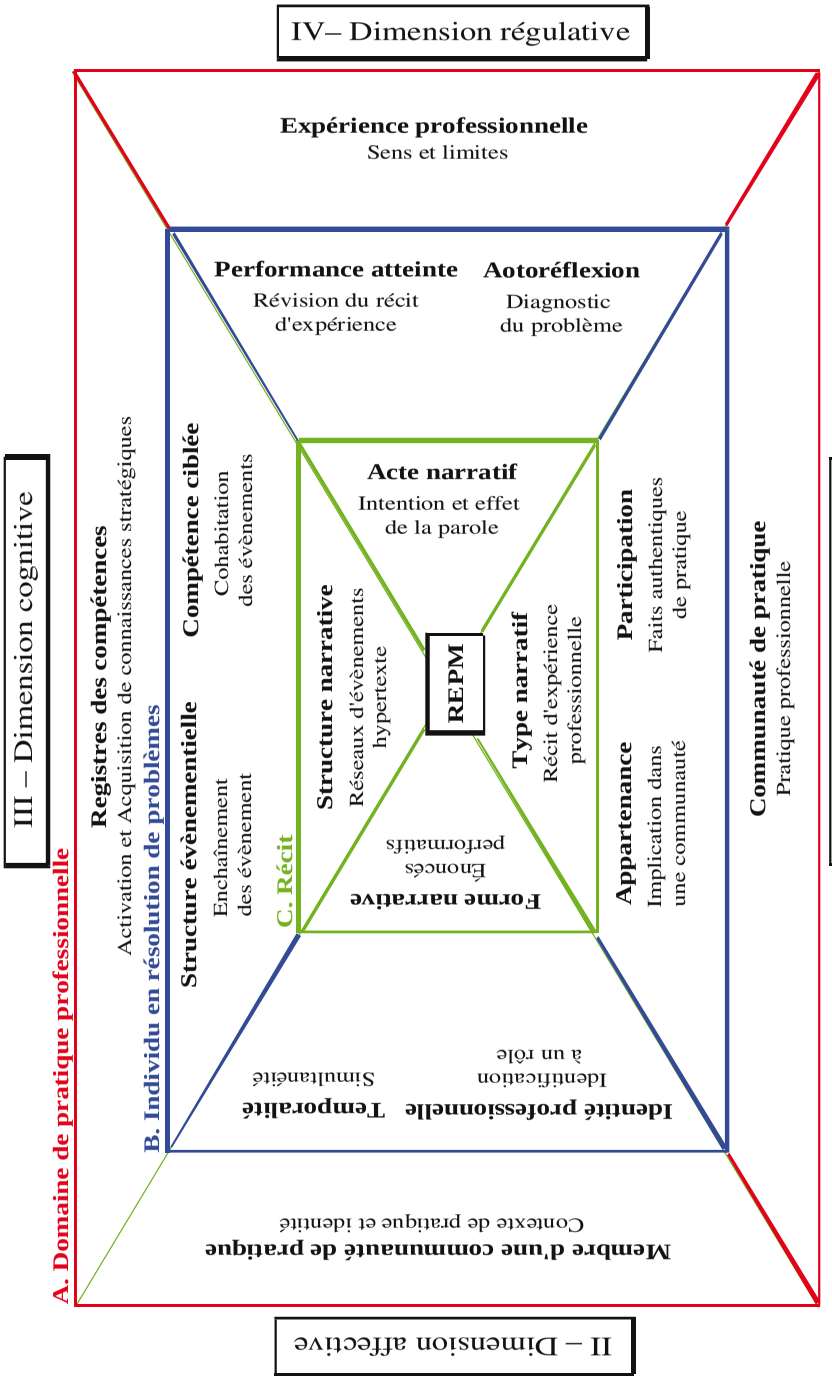


FIGURE 2 – Le modèle descriptif du REPM

De plus, en regard de :

- la dimension sociale, le discours de pratique prend forme dans un récit d'expérience professionnelle décrivant des faits authentiques au cœur desquels le praticien en résolution de problèmes démontre son implication dans sa communauté de pratique professionnelle.
- la dimension affective, le REPM est principalement composé d'énoncés performatifs c'est-à-dire écrits à la première personne du présent de l'indicatif. Le praticien-narrateur nous livre ainsi une information de première main et recrée un contexte de pratique adapté à une démarche d'autoréflexion. Ce caractère authentique du récit favorise chez le lecteur l'identification à son rôle en tant que membre de la communauté de pratique.
- la dimension cognitive, le REPM se structure en un réseau d'évènements hypertextes. Les évènements qui y sont relatés s'enchaînent et cohabitent sur la base de la compétence ciblée par la situation de pratique mise en scène. Cette situation de pratique reflète le registre des compétences du domaine professionnel et vise à l'activation et à l'acquisition de connaissances stratégiques par le futur praticien.
- la dimension de la régulation, le REPM comme acte de parole porte l'intention derrière l'acte professionnel relaté et produit un effet sur le lecteur. Le REPM présente une situation qui pose problème et pour laquelle le futur praticien doit proposer une solution, selon lui, plus adéquate. Poser un diagnostic et réviser la situation de pratique amènent l'apprenant à réfléchir sur l'action mettant à contribution sa compétence pour produire un récit révisé portant sa propre intention. Que ce soit celle du praticien au cœur du récit d'expérience ou celle de l'apprenant en interaction avec le REPM, ces expériences professionnelles portent le sens et les limites de leur pratique professionnelle. L'acte narratif en terme d'intention et d'effet de la parole donne le pouvoir d'effectuer quelque chose et non pas simplement de dire quelque chose. « Dès qu'on a saisi que l'objet à étudier, ce n'est pas la phrase, mais la production d'une énonciation dans la situation de discours, on ne peut plus guère manquer de remarquer ceci : affirmer, c'est exécuter [perform] un acte. » (Austin, 1970, p.143)

L'objet de recherche tel qu'il est esquissé précédemment offre un cadre en vue de la mise en opération du REPM menant au développement d'un dispositif d'évaluation formative destiné à un domaine de pratique professionnelle spécifique. Les multiples dimensions du REPM devraient permettre de simuler des performances complexes pour stimuler la flexibilité cognitive de l'apprenant en situation de résolution de problèmes.

REGULUM[5], LA SOLUTION PROTOTYPÉE

Dans le but de mieux comprendre les forces et les faiblesses de l'objet décrit précédemment et d'en délimiter la portée, le REPM a été mis en place pour les besoins de la formation des maîtres en « prototypant » un dispositif d'évaluation formative sur l'évaluation des apprentissages au primaire. La démarche méthodologique de recherche qui a été suivie s'inspire globalement de la méthodologie de développement EVR – LVR / *Expert Vérification et Révision – Learner Verification & Revision* telle que la décrivent Stolovitch et Keeps (1995). Cette démarche s'est déroulée en trois grandes étapes (voir **Figure 3**) : l'analyse de la situation pédagogique et des solutions envisageables, le « prototypage » de la solution retenue, les quatre mises à l'essai fonctionnelles auprès des experts et des apprenants. En alternance, chaque mise à l'essai a donné lieu à une révision du prototype RegulUM.

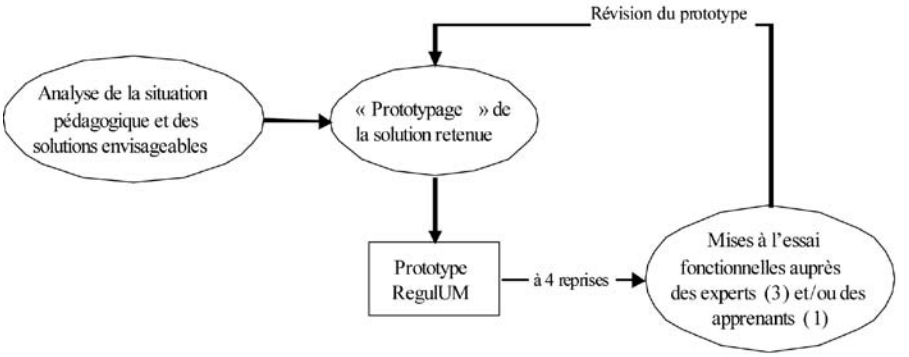


Figure 3 La démarche méthodologique

Les systèmes d'apprentissage multimédia interactif sont les choix d'implantation de la solution retenue, car « ils offrent des possibilités intéressantes lesquelles contribuent au renouvellement des stratégies pédagogiques faisant naître de nouvelles théories de l'apprentissage basées sur l'interactivité » (Depover, Giardina et Marton *et al.*, 1998). Dans la perspective de Giardina (1992), l'interactivité immédiate des environnements virtuels offre un milieu pédagogique « sans distance » permettant de répondre aux besoins de l'apprenant au bon moment et selon son rythme. Ces environnements peuvent redonner le contrôle de son apprentissage à l'apprenant et le placent dans une relation pédagogique un à un avec le système (Schank, 1988). De plus,

5. L'appellation RegulUM est composée de Regul pour régulation et UM pour Université de Montréal

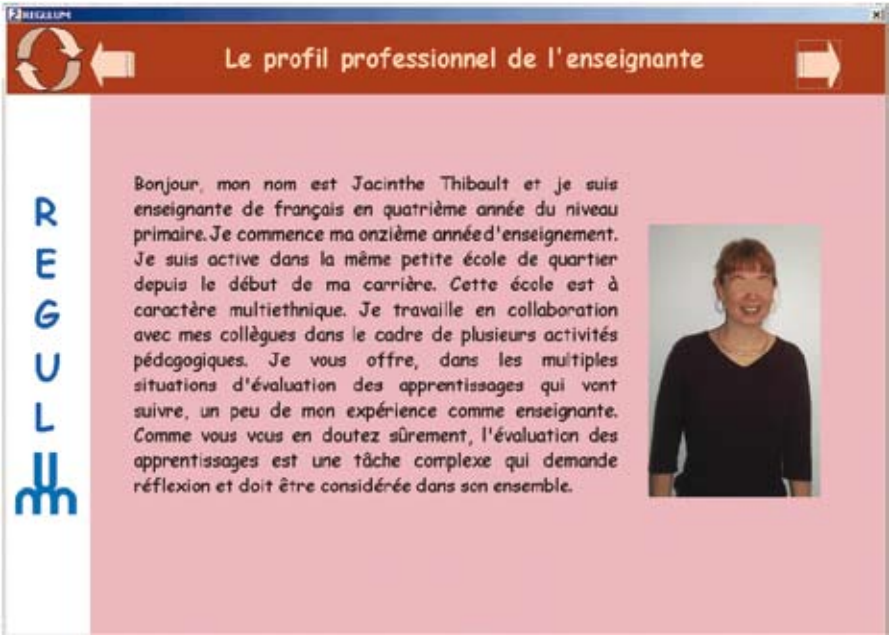


Figure 4 – Page écran – Mise en contexte de l'enseignante

les mises en situation de résolution de problèmes sollicitant les savoir-faire de l'apprenant sont très bien adaptées aux possibilités offertes par le multimédia interactif (Depover, Giardina et Marton, 1998). Mettant en scène des situations de vie authentiques et sollicitant chez l'apprenant des stratégies cognitives et métacognitives telles que les présentent West, Farmer et Wolff (1991), plusieurs dispositifs multimédias interactifs ont été développés. En effet, les environnements multimédias, micromondes de la simulation, augmentent l'interactivité apprenant-système et stimulent la créativité chez l'apprenant (Jonassen, 1996). Plus spécifiquement, les hypertextes impliquent l'apprenant en tant que lecteur ayant accès à de l'information par la consultation et la navigation ou l'engagent dans l'organisation et la structuration des ressources disponibles (Bruillard, 1997). Dans cette dernière approche, l'apprenant est impliqué à titre d'auteur dans un processus de création qui n'est pas directement mu par le questionnement et qui implique un processus de construction plutôt que de recherche (Jonassen, 1996). Ces environnements d'apprentissage ouverts, permettant de simuler des micromondes sur la base d'une interactivité significative, constituent une solution appropriée pour encourager le déploiement de stratégies cognitives et métacognitives chez l'apprenant dans un contexte de mise en situation authentique,

et ils favorisent ainsi la régulation des apprentissages (West, Farmer et Wolff, 1991). RegulUM peut être classé comme un dispositif de type « découverte », plus spécifiquement comme un système d'apprentissage multimédia interactif (Giardina et Laurier, 1999). Le prototype a été conçu à l'aide d'un système auteur[6] destiné à l'élaboration d'environnements d'apprentissage et de formation.

La structure multimédia du REPM implantée dans RegulUM est constituée principalement de huit pages-écrans. Dans un premier temps, pour situer le cadre dans lequel se déroule la situation d'expérience relatée, une description du profil professionnel (Figure 4) de l'enseignante au cœur des situations d'évaluation mises en scène est présentée au futur maître.

Après cette mise en contexte, on accède au récit à réviser. La vision intégrée du REPM (Figure 5) offre une vue globale du récit d'expérience professionnelle. Dans un premier temps, l'apprenant y consulte le récit initial à réviser. À ce point, l'apprenant peut décider d'aller en mode révision ou de demander un diagnostic sur son récit déjà révisé en cliquant sur le bouton Réviser ou Diagnostiquer selon le cas.

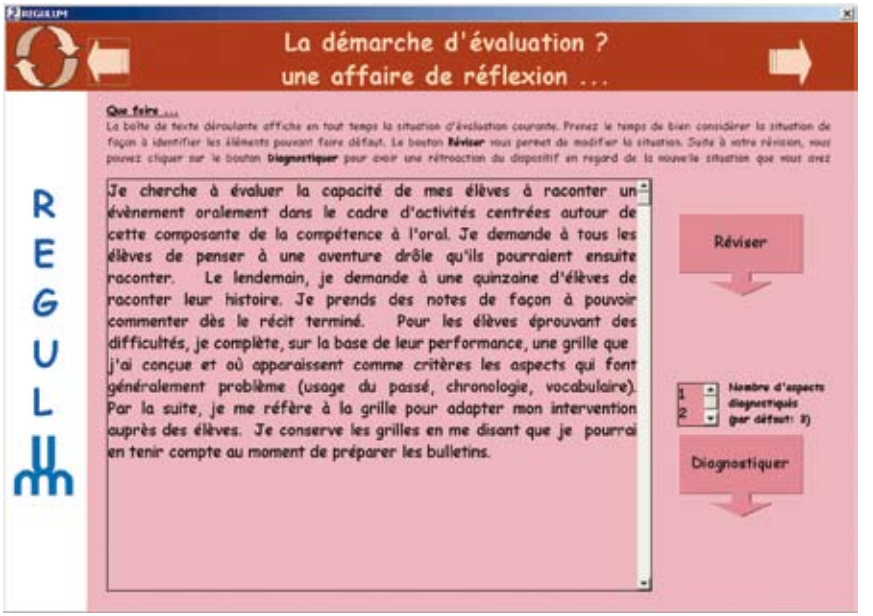


Figure 5 – Page écran – Vision intégrée du REPM

6. Dans ce cas, nous avons utilisé le système Toolbook version 8.5 de Click2learn.

En mode révision (Figure 6), le récit est scindé en cinq évènements pouvant être remplacés individuellement en cliquant sur le bouton Modifier.

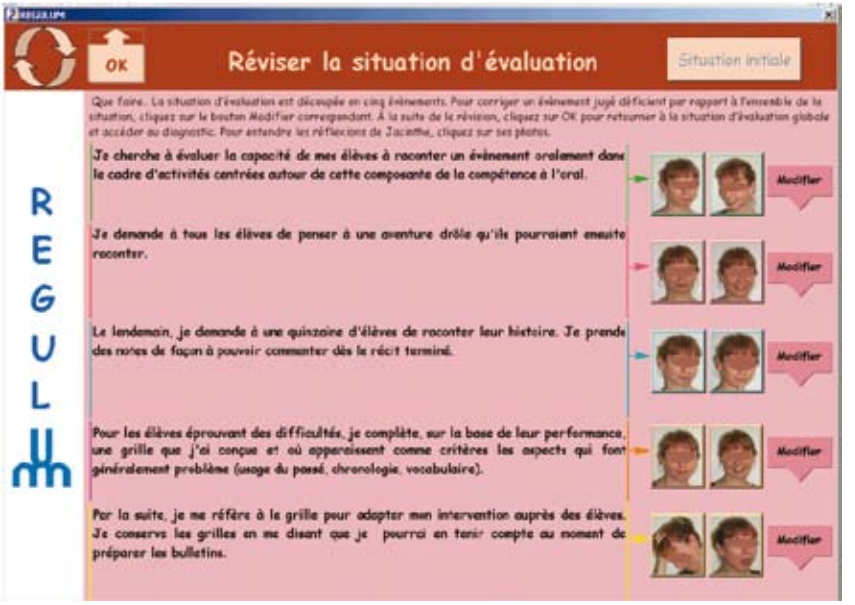


Figure 6 – Page écran – Vision parcellisée du REPM

À tout moment, l'apprenant peut accéder à une vue comparative sur le récit initial en cliquant sur le bouton Situation initiale. À chaque évènement du REPM sont associées des réflexions de pratique livrant les convictions et les incertitudes de l'enseignante au cœur de situations d'évaluations menées auprès de ses élèves de 4e année du primaire. Ces réflexions peuvent être entendues en cliquant sur les images correspondantes. À chacun des cinq évènements se greffe un ensemble de trois évènements substituts (Figure 7) pouvant être sélectionnés comme alternative à l'évènement potentiellement déficient, ici en l'occurrence le premier évènement du récit. Sur l'ensemble des combinaisons possibles d'évènements, 1024 récits au total peuvent être construits par l'apprenant.

Figure 7 – Page écran – Vision focalisée du REPM

En tout temps, l'apprenant peut retourner au niveau précédent du récit en cliquant sur le bouton OK.

LA COLLECTE DES DONNÉES

Avec RegulUM, nous avons voulu tenter d'identifier les éléments importants pour construire un dispositif efficace basé sur le REPM et qui serait utile à la formation des maîtres. Lors du développement d'un environnement d'apprentissage multimédia interactif, les apprenants constituent une source d'informations extrêmement importante pour augmenter l'efficacité d'un tel dispositif (Depover, Giardina et Marton, 1998). C'est la raison pour laquelle la collecte des données a eu lieu en trois temps (voir Figure 8) : avant la conception du dispositif, pendant et après la quatrième mise à l'essai. Les informations recueillies à ces moments précis, à partir de sources d'information différentes et utilisant des techniques variées ont permis la triangulation des données lors de l'analyse diminuant les biais et augmentant la validité de l'étude (Huberman et Miles, 1991).

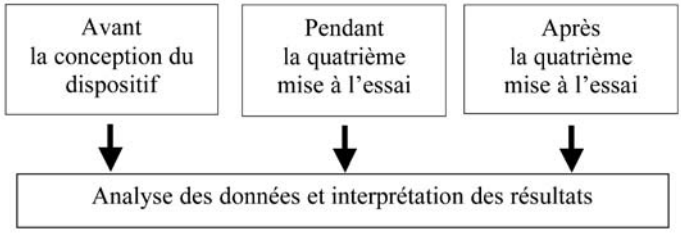


Figure 8 – La collecte des données

Avant la conception du dispositif, une collecte de données auprès du public cible a permis de brosser un portrait de l'utilisateur type confirmant la disposition des apprenants à utiliser un dispositif d'évaluation formative médiatisé. De plus, cette analyse a permis d'orienter notre approche en tant que concepteurs nous aidant à concevoir un dispositif mieux adapté aux besoins de ces étudiants et du contexte pédagogique de la formation des maîtres.

Lors de la quatrième mise à l'essai, le professeur participant a choisi d'arrimer le dispositif RegulUM à son cours comme préparation pour un examen. Quinze apprenants y ont participé sur une base volontaire : trois étudiants et douze étudiantes. Cette répartition du genre des informateurs correspond à celle des inscriptions à la formation des maîtres. La collecte de données pour cette mise à l'essai auprès des apprenants a eu lieu pendant et après la séance de travail prévue à cet effet. La technique de collecte des données utilisée lors de la quatrième mise à l'essai a été l'observation. Deux types d'observation ont permis de recueillir l'information : l'observation électronique et l'observation à l'aide d'une grille. L'observation électronique consistait en un traçage automatique de la navigation de l'apprenant pendant l'utilisation du dispositif, ce qui a permis de suivre chaque action effectuée et de mesurer le temps associé à chacune de ces actions. Ce type d'observation avait pour but d'identifier comment l'apprenant mène la tâche et de consigner le récit révisé qu'on lui demande de produire accompagné de ses commentaires. Parallèlement, l'observation à l'aide de la grille visait essentiellement à repérer les comportements de l'apprenant en regard du dispositif et de l'activité en infirmant ou en confirmant le traçage de l'observation électronique. L'observation électronique a permis de retracer les actions, tandis que l'observation à l'aide de la grille a permis de préciser et de nuancer le déroulement de ces actions. Nous avons tenté de déterminer ce qui a été utile à l'apprenant lors de l'activité en portant une attention particulière au parcours de la navigation et à la manière dont le problème a été résolu. Ces éléments ont été ciblés dans l'intention de pouvoir induire l'engagement du

futur praticien et sa manière de mener l'action dans la situation professionnelle mise en scène, ainsi que le sens qu'il accorde aux évènements dans le contexte du récit présenté. Plus précisément, l'analyse qualitative des données recueillies visait à répondre aux interrogations suivantes : les apprenants ont-ils complété l'activité ? Ont-ils tenté de déterminer ce qui n'allait pas ? Ont-ils construit un nouveau récit ou modifié le récit déficient ? Ont-ils consulté les réflexions de pratique ? Ont-ils effectué des ajustements au récit révisé ? Rappelons ici que la trace électronique à elle seule ne suffit pas pour répondre à ces interrogations ; toutefois elle donne de bonnes indications qui gagnent à être confrontées aux autres données recueillies.

À la suite de cette mise à l'essai, une entrevue de groupe a donné l'occasion de recueillir les réactions des apprenants concernant le dispositif RegulUM par rapport à l'interface, au guidage, à l'utilisation d'un dispositif d'évaluation formative et à la révision du récit d'expérience professionnelle. L'ensemble des données recueillies a permis de valider l'interface du prototype et de déterminer les ajustements nécessaires pour mettre au point un environnement d'apprentissage multimédia interactif mieux adapté à ce contexte pédagogique et aux besoins des apprenants

LES PREMIERS RÉSULTATS

Notre discussion sur la pertinence d'un dispositif tel que RegulUM basé sur le REPM s'appuie sur les facteurs pédagogiques des environnements d'apprentissage multimédia interactif identifiés par Depover, Giardina et Marton (1998). Dans l'ensemble, nos propos gardent en toile de fond les considérations suivantes :

- la participation et la motivation de l'apprenant ;
- le rythme individuel de l'apprenant ;
- le guidage de l'apprenant ;
- la perception et l'organisation des messages ;
- l'application des connaissances acquises et le choix des méthodes pédagogiques ;
- la connaissance immédiate des résultats (la rétroaction) ;
- la structuration du contenu, la stratégie de l'organisation des ressources et l'interaction avec l'apprenant.

Nous tenons à rappeler que les résultats présentés ici ne peuvent que servir à déterminer les éléments importants pour construire un REPM et ne prétendent en aucun cas à l'efficacité de ce dernier pour développer le jugement professionnel des futurs enseignants à la formation des maîtres. À cette étape, il nous est tout au plus possible d'entrevoir certains facteurs pouvant

alimenter la réflexion sur l'action. Nous avons choisi de mettre principalement en lumière les données recueillies par le biais de la trace électronique et de l'entrevue de groupe. Comme le montre la Figure 9, nous avons retenu six contrastes jugés plus pertinents pour dégager les éléments importants du REPM :

- les utilisateurs masculins/féminins ;
- les récits adéquats/inadéquats ;
- les parcours de navigation séquentiels/ciblés (selon l'ordre préétabli ou non) ;
- les parcours récurrents/non récurrents (retour sur les évènements modifiés) ;
- les réflexions de pratiques consultées/non consultées ;
- la vue comparative utilisée/non utilisée.

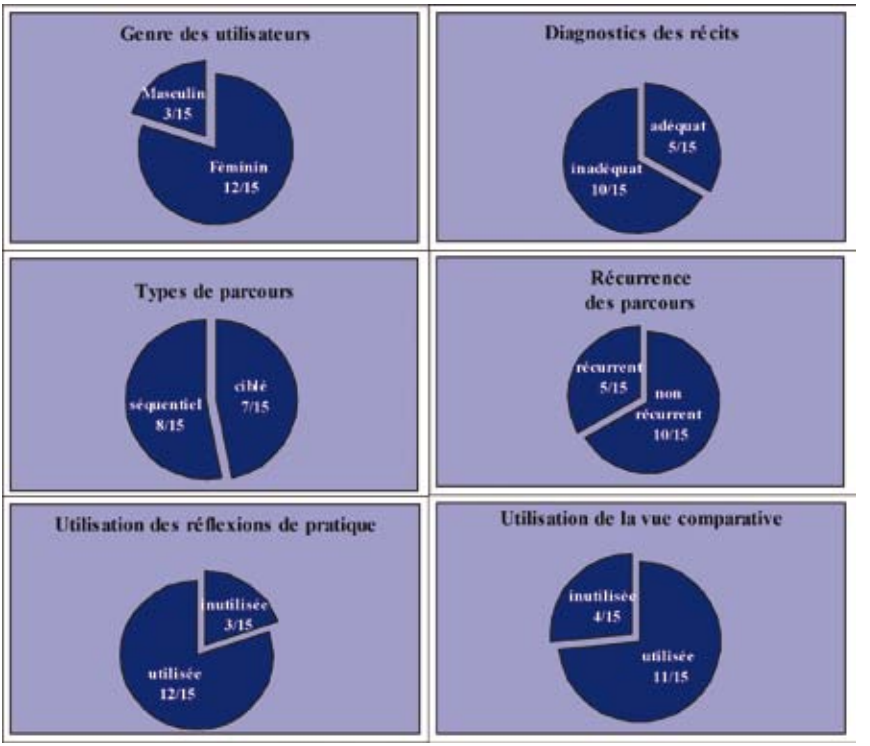


Figure 9 – Répartition des informations retenues

Ces informations sont traitées ici individuellement et sont mises en rapport entre elles par la suite pour donner un autre éclairage à nos observations. À la lumière de notre analyse, le cheminement individualisé, le compagnonnage cognitif ainsi que la résolution de problèmes sont les thèmes qui appellent certaines remarques.

Un cheminement individualisé

Il semble que le REPM encourage une participation active de l'apprenant et qu'il soit un environnement d'apprentissage stimulant. Les apprenants ont mentionné l'importance du respect du rythme individuel de chacun. De plus, les multiples récits potentiels prévus par le dispositif semblent pouvoir répondre aux particularités des profils d'apprenants et leur permettre de produire un résultat individualisé. De façon générale, les apprenants ont aimé utiliser le dispositif. Ils ont noté la facilité d'utilisation en précisant avoir aisément découvert ce qu'il fallait faire. Ils ont apprécié la possibilité de naviguer à leur propre rythme, de produire un récit selon leur profil individuel et de ne pas arriver tous au même résultat. Ils ont jugé le dispositif comme une façon différente d'aborder la matière vue en classe. Certains apprenants ont construit une histoire complètement nouvelle ; d'autres ont cherché ce qui n'allait pas pour ensuite adapter la situation de départ. Étant tous d'accord sur le fait que le choix d'évènements guide l'apprenant et lui donne des idées pour s'orienter, ils auraient tout de même aimé pouvoir écrire eux-mêmes une partie du texte. Ils ont aussi mentionné que les réflexions de l'enseignante pour chaque évènement leur permettaient de mieux comprendre la situation. Les étudiants ont précisé que les ensembles d'évènements substituts leur donnait des idées, des pistes pour s'orienter et les guidaient un peu. Il leur a été utile d'avoir l'opinion de l'enseignante sur chaque aspect, car cela leur permettait de bien comprendre. Ils ont apprécié le fait de pouvoir prendre leur temps et d'évaluer les différentes possibilités selon leur profil et leur orientation. Par un effet insoupçonné, certains étudiants avouant éprouver de la difficulté en français ont souligné que RegulUM leur avait permis d'écrire un texte rapidement. La séance a été perçue comme un exercice pouvant les aider à mettre en application les thèmes abordés en classe. Le contexte de résolution de problèmes, la mise en situation professionnelle authentique, l'organisation du contenu, les réflexions sur la pratique et les vues comparatives sont autant d'éléments qui contribuent à l'adaptabilité du dispositif RegulUM et encouragent le cheminement individualisé.

Le compagnonnage cognitif

En mettant en scène dans le dispositif RegulUM des situations de pratique authentiques, on veut, par la découverte, faire évoluer l'apprenant d'un état de compétence initial vers un état de compétence plus avancé tout en s'appuyant sur la « grande liberté laissée à l'apprenant dans le contrôle de ce qu'il va apprendre et du rythme auquel il va apprendre » (Giardina et Laurier, 1999, p. 46). Par ce compagnonnage cognitif, on veut stimuler chez le futur maître le réflexe de réfléchir sur l'action pour assurer le développement de son jugement maintenant et par la suite. On a dans ce cas une visée transformative de son savoir-faire.

Dans un premier temps, l'autoévaluation et l'autorégulation menées par l'étudiant dans une pédagogie de situation et de résolution de problèmes entrent en jeu lorsque l'individu doit travailler seul à un moment donné et déployer des stratégies métacognitives. En évaluant sa propre performance dans un contexte formatif, l'apprenant fait l'acquisition de critères pour évaluer de mieux en mieux la qualité de sa réflexion et de son action (Aylwin, 1996). L'apprenant en situation de résolution de problèmes doit faire appel à son jugement. Cette situation d'évaluation engendre une tâche complexe puisque le contexte nécessite l'usage de stratégies cognitives et métacognitives de la part de l'apprenant (Tardif, 1993). L'apprenant procède dans une telle démarche à une autoévaluation continuelle de ce qu'il produit (Scallon, 2000). Dans un deuxième temps, un des critères pédagogiques pour qu'un environnement d'apprentissage multimédia interactif soit efficace est la connaissance immédiate des résultats pour la tâche menée par l'apprenant (Giardina, 1992). La nature dynamique du REPM est favorisée par sa structure organisationnelle et le contexte de situation authentique. Celle-ci, jumelée aux réflexions de pratique et aux vues comparatives, offre une rétroaction interactive. Toutefois, le mécanisme pour le diagnostic de l'expert offert dans cette première version du prototype à la suite de la révision du récit d'expérience par l'apprenant ne s'avère pas optimal en termes de rétroaction pour stimuler la réflexion de l'apprenant et favoriser la régulation des apprentissages. Le diagnostic du récit révisé retransmis ultérieurement par le professeur exclut un retour immédiat par l'apprenant sur son récit, ce qui interrompt le cycle d'échange essentiel à la régulation (voir Tableau 1).

Tableau 1 – Caractéristiques de la navigation selon le diagnostic des récits

Quantité de récits diagnostiqués	Caractéristiques de la navigation			
	Parcours récurrent	Accès aux réflexions	Vue comparative	Parcours ciblé
Adéquats – 5	0 %	80 %	60 %	80 %
Inadéquats – 10	50 %	80 %	80 %	30 %

De surcroît, le professeur a mentionné que l'ensemble des diagnostics qu'il a produits pour tous les récits reçus a eu un impact non négligeable sur sa charge de travail. Il est donc ressorti que le REPM alimenterait mieux la réflexion de l'apprenant en jumelant au contexte de situation authentique une rétroaction globale immédiate de l'expert par le biais d'un diagnostic automatisé. Cette rétroaction accessible en tout temps et au besoin stimulerait l'apprenant à revoir et à nuancer sa propre analyse, l'encouragerait à effectuer des ajustements à plus d'une reprise et favoriserait d'autant plus le développement de son jugement. Une nouvelle version du prototype incluant le diagnostic automatisé a été réalisée à ce jour (Bélanger, Blais, Giardina et Laurier, 2004). Dans cette nouvelle version, trois sources de rétroaction alimentent la réflexion de l'apprenant. Sa propre verbalisation sur la situation professionnelle, les réflexions de pratique livrées par le praticien-narrateur pour chaque évènement du récit et le diagnostic automatique de l'expert à chaque récit révisé. Nous pouvons maintenant dire que, dans son ensemble, le REPM offre une « interactivité significative [qui] place l'apprenant au centre des échanges en lui permettant de manipuler ou de transformer des objets, des situations ou des activités que le concepteur aura mis à sa disposition dans un environnement médiatisé » (Depover, Giardina et Marton, 1998, p.144).

Finalement, il semble que l'utilisateur doive se sentir interpellé par le profil du praticien qu'il consulte et auquel il s'identifie. En fait, les réflexions sur la pratique n'ont pratiquement pas été consultées par les apprenants de sexe masculin (voir Tableau 2). À première vue, ces apprenants ne semblent pas s'être sentis concernés par l'enseignante qui livrait ses réflexions sur sa pratique. Toutefois, compte tenu du nombre limité d'apprenants masculins ayant participé, cet aspect demeure incertain. Un éventail plus varié de praticiens-narrateurs pourrait être mis à la disposition du futur praticien dans sa réflexion, lui laissant le choix de son accompagnateur.

Tableau 2 – Caractéristiques de la navigation selon le type d'utilisateurs

Genre	Caractéristiques de la navigation				
	Accès aux réflexions	Vue comparative	Parcours ciblé	Parcours récurrent	Récits adéquats
Masculin – 3	33 %	33 %	100 %	0 %	67 %
Féminin – 12	92 %	83 %	33 %	42 %	25 %

Diagnostiquer pour mieux réviser

L'immédiateté de la rétroaction et le diagnostic sur le récit révisé sont ressortis comme deux aspects semblant influencer le déroulement de la tâche menée par l'apprenant. Sur la base de la trace électronique de la navigation effectuée lors de la révision du récit, nous avons tenté d'identifier des indices révélateurs de la démarche de résolution de problèmes menée par les apprenants. Ainsi, deux types de parcours de navigation se sont dégagés : séquentiel et ciblé. Le parcours de type séquentiel est un parcours qui correspond à l'ordre préétabli par l'enchaînement des évènements composant le récit. Dans le parcours de type ciblé, l'apprenant accède aux différents ensembles d'évènements sans tenir compte de leur ordre d'apparition dans le récit. Un autre indice révélateur retrace le nombre de retours faits par l'apprenant dans chacun des ensembles d'évènements substituts. On pointe ici la récurrence du parcours. Lorsqu'un ensemble d'évènements a été visité à plus d'une reprise, le parcours est qualifié de récurrent ; dans le cas contraire, il est qualifié de non récurrent.

La structure de la navigation nous porte donc à croire qu'un parcours ciblé indique vraisemblablement un diagnostic du récit déficient avant d'amorcer la révision, et un parcours séquentiel suppose une approche plus linéaire lors de la révision du récit. Voici une description des parcours (voir Figure 10 et Figure 11) parmi les ensembles d'évènements substituts, ceux-ci étant notés A, B, C, D et E.

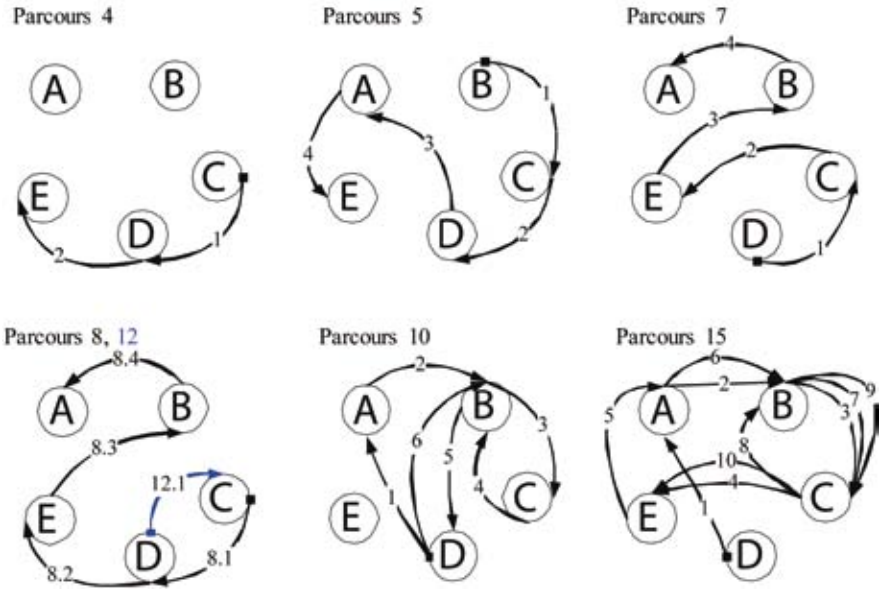


Figure 10 – Représentation des parcours ciblés

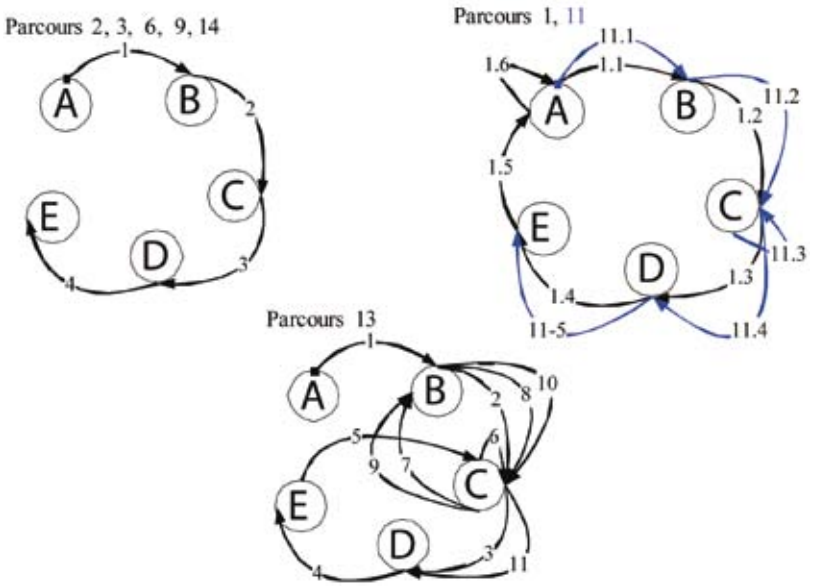


Figure 11 – Représentation des parcours séquentiels

Lors de l'entrevue de groupe, les réactions des apprenants par rapport à la révision du récit d'expérience ont été variées et confirment l'approche adoptée pour l'analyse des parcours en regard de la résolution de problèmes. Certains ont mentionné qu'ils avaient construit une histoire complètement nouvelle parce qu'ils n'arrivaient pas à déterminer ce qui n'allait pas dans le récit initial. Ils ont précisé que le choix donné dans les ensembles de substituts leur a permis d'identifier des évènements pour construire un récit plus adéquat. D'autres apprenants ont précisé de leur côté qu'ils ont lu et trouvé ce qui n'allait pas dans le récit et ont confirmé avoir consulté les alternatives pour adapter la situation. Les apprenants étaient d'accord pour dire qu'il y avait beaucoup de solutions possibles et d'autres combinaisons valides. Ces commentaires recueillis auprès des apprenants viennent soutenir les types de parcours de navigation que nous avons identifiés.

Les utilisateurs masculins ont emprunté un parcours ciblé dans tous les cas, et les utilisateurs féminins pour la grande majorité un parcours séquentiel (voir Tableau 2). Nous avons noté un pourcentage plus élevé de récits adéquats parmi les parcours ciblés par rapport aux parcours séquentiels. La consultation des réflexions de pratique a été plus fréquente dans le parcours séquentiel (voir Tableau 3). Le type de parcours emprunté a-t-il eu une incidence sur le succès du récit révisé et l'accès aux réflexions de pratique ? Est-il un symptôme d'une plus grande difficulté à ajuster adéquatement le récit initial ou une façon différente d'aborder le problème ? Ces pistes gagneraient à être explorées plus à fond pour mieux comprendre la démarche de résolution de problèmes menée dans ce type d'environnement d'apprentissage. Si nous poussons la réflexion plus loin, la consultation des réflexions de pratique pourrait-elle constituer par effet indirect une distraction lors du diagnostic mené par l'apprenant, ou est-elle simplement plus fréquente lorsque la révision du récit s'avère plus ardue ?

Tableau 3 – Caractéristiques de la navigation selon le type de parcours

Type de parcours	Caractéristiques de la navigation			
	Accès aux réflexions	Vue comparative	Parcours récurrent	Récits adéquats
Ciblé – 7	57 %	71 %	29 %	57 %
Séquentiel – 8	100 %	75 %	38 %	13 %

Nous ne pouvons conclure à ce sujet, mais voilà autant de questionnements autour du REPM qui gagneraient à être explorés.

CONCLUSION

Cette recherche a permis de tracer les traits importants du REPM et d'entrevoir les enjeux relatifs à la conception d'un tel dispositif. Sur la base des travaux que nous avons menés, nous pouvons dire que le REPM – Récit d'Expérience Professionnelle Médiatisé s'avère une stratégie d'apprentissage qui mérite d'être plus exploitée. Les résultats obtenus suggèrent que le REPM offre un environnement riche et adaptable pouvant développer les compétences du futur praticien. Le REPM semble donc une avenue prometteuse pour stimuler la réflexion sur l'action et développer le jugement et l'autonomie de l'apprenant. La participation active de l'apprenant et les multiples récits complétés soutiennent la construction et l'utilisation du REPM pour la régulation des apprentissages.

Certains aspects relevés lors de notre étude jettent une autre lumière sur le REPM et font poindre des faiblesses que nous ne saurions ignorer. Dans sa première version, le REPM comportait des limites pour ce qui est de l'ajustement et de la régulation des apprentissages. Des corrections y ont été apportées. L'automatisation du diagnostic a été réalisée pour mieux aider l'apprenant et augmenter son autonomie dans la régulation de ses apprentissages. Nous rappelons que la présente recherche comporte des limites empêchant toute généralisation et ne nous permettant que de dégager des tendances. Le REPM vise le développement des compétences dans le cadre d'un domaine de pratique professionnelle et pour les étudiants de premier cycle universitaire. Même si la taille de l'échantillon a permis de recueillir une quantité suffisante d'informations, nous croyons qu'il serait pertinent d'examiner des applications avec des apprenants provenant de disciplines différentes. Quoique la répartition garçon-filles lors de la quatrième mise à l'essai corresponde approximativement à celle des inscrits à la formation des maîtres, il serait important de prévoir des ratios plus équilibrés pour explorer les pistes relatives au genre. Afin que nous comprenions mieux les interactions entre l'apprenant et le dispositif, d'autres moyens de collecte de données sont à prévoir. Une vidéo offrant une vision globale externe sur les comportements et attitudes de l'apprenant ainsi qu'un enregistrement de la verbalisation à voix haute des raisons et des réflexions l'ayant guidé dans sa tâche devraient être prévus dans le but de mieux cerner les processus de traitement activés chez l'apprenant (Ericsson, 1984). La verbalisation cible-rait plus adéquatement le processus de métacognition chez l'apprenant, et l'observation électronique jumelée à l'observation vidéo pourrait permettre de nuancer les actions de l'apprenant. De plus, à la suite de l'utilisation du dispositif, une entrevue individuelle semi-structurée permettrait de recueillir les commentaires de l'apprenant, toujours dans le but de retracer et de décrire le phénomène de métacognition. Cette mise à l'essai auprès des apprenants

dans une optique évaluative devrait nous autoriser à statuer sur l'efficacité du REPM comme stratégie pour le développement des compétences au premier cycle universitaire. Si cette stratégie s'avère encourager la capacité à réfléchir sur l'action, elle peut vraisemblablement conduire au développement du jugement chez le professionnel.

Références

Austin, J.L. (1970), *Quand dire c'est faire*, Paris, Éditions Du Seuil.

Aylwin, Ulrich (1996), *La différence qui fait la différence…,* Montréal, AQPC.

Barthes, Roland (1966), « Introduction à l'analyse structurale du récit », dans *L'analyse structurale du récit – Communications 8*, Paris, Éditions du Seuil, 1981, p. 7-33.

Bélanger, K. (2003), *Le récit d'expérience professionnelle médiatisé, une stratégie pour le développement des compétences au premier cycle universitaire.* Mémoire de maîtrise inédit, Université de Montréal.

Bélanger, K., Jean-Guy Blais, Max Giardina et Michel Laurier (2004), « Le récit d'expérience professionnelle médiatisé, une stratégie interactive pour le développement des compétences », *RES Academica*, vol. 22, p. 197-214.

Brien, Robert (1994), *Science cognitive et formation,* Sainte-Foy, PUQ.

Bruillard, Éric (1997), *Les machines à enseigner*, Paris, Éditions Hermès.

Dassa, Clément et Michel Laurier (2003), « Le diagnostic pédagogique comme moyen d'informer », dans Michel Laurier (dir.) *Évaluation et communication – De l'éducation formative à l'éducation informative,* Outremont(Québec), Éditions Québecor, p. 103-130.

De Bono, Edward (1967), *The use of lateral thinking,* Harmondsworth, Penguin Books.

Depover, Christian, Max Giardina et Philippe Marton (1998), *Les environnements d'apprentissage multimédia – Analyse et conception,* Paris, Éditions L'Harmattan.

Ericsson, Anders (1984), *Protocol analysis : verbal reports as data,* Cambridge, MIT Press.

Giardina, Max (1992), « L'interactivité dans un environnement d'apprentissage multimédia », *La revue des sciences de l'éducation*, vol. XVIII, nº 1, p. 43-66.

Giardina, Max et Michel Laurier (1999), « Modélisation de l'apprenant et interactivité », *La revue des sciences de l'éducation*, vol. XXV, nº 1, p. 35-59.

Huberman, M. et Matthew Miles(1991), *Analyse des données qualitatives,* Bruxelles, De Boeck-Wesmael.

Jonassen, David (1996), *Computers in the classroom : Mindtools for critical thinking*, New Jersey, Prentice Hall.

Jonassen, David et M. Tessmer (1999), *Task analysis methods for instructional design,* New Jersey, Lawrence Erlbaum.

Lave, Jean et Etienne Wenger (1991), *Situated Learning : Legitimate Peripheral Participation*. New York, Cambridge Press.

Lévy, Pierre (1995), *Qu'est-ce que le virtuel ?,* Paris, Éditions La Découverte

Ricoeur, Paul (1975), *La métaphore vive*, Paris, Éditions Du Seuil.

Scallon, Gérard (2000), *L'évaluation formative*, Saint-Laurent, Éditions du Renouveau Pédagogique.

Schank, Roger (1988), *The creative attitude : learning to ask and answer the right questions*, New York, Macmillan Publishing company.

Schank, Roger et Robert P. Abelson (1995), « Knowledge and memory : The Real Story », dans Robert S. Wyer, Jr, *Knowledge and Memory : The Real Story* New Jersey, LEA Advances in Social Cognition, vol. VIII, p. 1-86.

Spiro, Rand.J., P.J. Feltovitch, M.J. Jacobson et R.L. Coulson (1992), « Cognitive Flexibility, Constructivism and Hypertext : Random Access Instruction for Advanced Knowledge Acquisition in Ill-Structured Domains », dans T.Duffy et D. Jonassen (Ed.), *Constructivism and the Technology of Instruction.* Hillsdale, NJ, Erlbaum, p. 57-75.

Spiro, R.J., P.J. Feltovitch, M.J. Jacobson et R.L. Coulson (1992), « Knowledge Representation, Content Specification, and the Development of Skill in Situation-Specific Knowledge Assembly : Some Constructivist Issues as They Relate to Cognitive Flexibility Theory and Hypertext », dans T.Duffy et D. Jonassen (Ed.), *Constructivism and the Technology of Instruction,* Hillsdale, NJ, Erlbaum, p. 121-128.

Spiro, Rand J. et Jihn-Chang Jehng (1990), « Cognitive Flexibility and Hypertext : Theory and Technology for the Nonlinear and Multidimensional Traversal of Complex Subject Matter », dans Don Nix et Rand Spiro, L., *Cognition, Education, Multimedia : Exploring Ideas in High Technology,* Hillsdale, N.J., Erlbaum, p. 163-205.

Spiro, R.J., R.L. Coulson, P.J. Feltovitch et D. Anderson (1988), « Cognitive Flexibility Theory : Advanced Knowledge Acquisition in Ill-Structured Domains », dans *Program of the Tenth Annual Conference of the Cognitive Science Society*, Hillsdale, N.J., Erlbaum, p. 375-383.

Stolovitch, Harold D. et Keeps, Erica J. (2003), Engineering Effective Learning Toolkit.

New Jersey : Pfeiffer, John Wiley & Sons, 400 pages

ISBN : 978-0-7879-6536-5 Loose-leaf

http ://www.pfeiffer.com/WileyCDA/Section/id-101711.html

Tardif, Jacques (1993), « L'évaluation dans le paradigme constructiviste », dans René Hivon (dir.), *L'évaluation des apprentissages,* Sherbrooke, Éditions du CRP, p. 27-57.

Tardif, Jacques (1999), *Le transfert des apprentissages*, Montréal, Éditions Logiques.

Tochon, François-V. (1993), *L'enseignante experte, l'enseignant expert*, Paris, Éditions Nathan.

Vygotsky, L.S. (1997), *The collected works of L.S. Vygotsky (1896-1934) – The History of the Development of Higher Mental Functions,* Plenum Press.

Wenger, Etienne (1998), *Communities of practice : Learning, meaning and identity*, New York, Cambridge University Press.

West, C., J. Farmer et Philipp Wolff (1991), *Instructional design-implications from cognitive science*, USA, Prentice Hall.

Page laissée blanche intentionnellement

SECTION IV

Regards sur des dimensions techniques avancées

Page laissée blanche intentionnellement

CHAPITRE 12

Des normes techniques pour les systèmes TIC associés à l'évaluation des apprentissages[1]

Gilles Raîche, Martin Riopel,
Martin Lesage et Komi Sodoké
Université du Québec à Montréal

INTRODUCTION

Une importante dimension du développement de systèmes pour l'évaluation des apprentissages à l'aide des technologies de l'information et des communications est certainement celle rattachée à la mise en place de normes techniques pour l'élaboration de systèmes d'apprentissage par ordinateur, c'est-à-dire le *e-Learning* (Friesen, 2004). En effet, on assiste depuis plusieurs années à la multiplication des plateformes informatiques pour soutenir la certification, entre autres, dans les secteurs de l'aéronautique et de l'informatique. De plus, on fait face à la diversification des formats de données et à leur disponibilité croissance sur support informatisé pour soutenir l'apprentissage. Ces développements ont donc engendré des travaux visant à permettre une meilleure communication entre les plateformes informatisées de soutien à l'apprentissage (*interoperability*), le partage et la réutilisation des ressources (*reusability*) (Wiley, 2000) et l'adoption d'un langage commun (*standardisation*) (Hodgins et Conner, 2000).

1. La production de ce chapitre a pu être réalisée grâce à l'obtention de subventions auprès du Conseil de recherche en sciences humaines du Canada (CRSH), du Fonds québécois de la recherche sur la société et la culture (FQRSC), de la Fondation pour l'Innovation du Nouveau-Brunswick (FINB) et du Programme d'aide aux chercheurs et chercheuses de l'Université du Québec à Montréal (PAFACC).

C'est dans ce contexte que des normes d'apprentissage ont été développées dans le domaine des sciences informatiques. On retrouve d'ailleurs de plus en plus de plateformes informatisées conçues autour de ces normes. À titre d'exemple, on peut citer les plateformes Ganesha, WebCT et Moodle. Toutefois, puisque ces normes ont été élaborées, en bonne partie, en dehors du domaine formel de l'éducation, elles ne prennent pas toujours en compte les diverses contributions scientifiques des sciences de l'éducation. À l'inverse, les chercheurs en sciences de l'éducation connaissent généralement assez peu ces normes. Il y a ici une très belle opportunité en éducation pour soutenir l'évaluation des apprentissages à l'aide des technologies de l'information et des communications en tenant compte des normes d'apprentissages développées en informatique (Baxter et Mislevy, 2005 ; Casadei et Magnani, 2007 ; Fletcher, Tobias et Wisher, 2007 ; Giotopoulos, Alexakos, Beligiannis et Likothanassis, 2005 ; Mislevy, 2002 ; Seibert, Hamel, Haynie, Mislevy et Bao, 2006 ; Sodoké, 2007).

Ce chapitre a pour but de décrire les six normes d'apprentissage par ordinateur les plus utilisées et de porter un jugement sur leur capacité à prendre en compte l'évaluation des apprentissages. Plus spécifiquement, il s'agira des conventions du noyau de Dublin, de CMI, de LOM, d'IMS, de SCORM et de MINE SCORM. Soulignons qu'il s'agit uniquement des normes qui semblent être les plus utilisées et connaître la croissance la plus importante actuellement. Ensuite, au regard des préoccupations et des contributions des praticiens et des chercheurs liées à l'évaluation des apprentissages, des pistes d'amélioration des normes d'apprentissages actuelles seront suggérées de façon telle que nous puissions, dans le futur, tenir compte aussi de normes pour l'évaluation des apprentissages à l'aide des TIC.

LES CONVENTIONS DU NOYAU DE DUBLIN *(DUBLIN CORE)*

Le noyau de Dublin (Heery, 1996) fournit un ensemble normalisé de conventions pour décrire des ressources en ligne (qualifiées d'objets d'apprentissage – *learning objects*) de manière à faciliter leur recherche, leur utilisation et leur partage. Le noyau de Dublin est employé couramment pour décrire les fichiers numériques vidéo, audio, graphiques, textuels, ou mixtes, telle une page Web. Au départ, l'utilisation principale du noyau de Dublin s'est effectuée principalement dans le domaine de la bibliothéconomie.

Cette norme est constituée d'un noyau à deux niveaux, soit le noyau simple, qui comprend des informations sur le contenu, la propriété intellectuelle et les instances particulières des objets d'apprentissage, et le noyau qualifié, qui donne des informations sur l'assistance, la provenance et les droits d'auteurs de ces objets.

La mise en application du noyau de Dublin se fait généralement sous forme de fichiers XML (*eXtensible Markup Language*). XML est un langage constitué de balises (éléments décrivant l'information – *tags*) comme l'est le langage HTML servant à créer des pages Web. Dans le langage XML, contrairement au langage HTML, les balises sont définies par le programmeur. La figure 1 permet de visualiser un exemple de fichier complet où est appliqué le langage XML dans le but de décrire des métadonnées (*metadata*) du noyau de Dublin.[2] On remarquera la première ligne qui indique tout simplement quelle version du langage XML est utilisée, soit la 1.0. Les six lignes suivantes indiquent les liens web où trouver les informations sur les normes utilisées, dont celle du noyau de Dublin.

Toutes les autres lignes constituent les métadonnées qui permettent de décrire l'objet d'apprentissage. Par exemple, on remarquera l'ouverture de la balise <dc:creator> qui indique que le champ *creator* de la norme du noyau de Dublin (dc: signifiant *Dublin Core*) se verra attribué la valeur « Blais, J.-G. et Raîche, G. ». On termine ce champ par la balise de fin </dc:creator>. En langage XML, chaque balise d'ouverture doit avoir une balise de contrepartie finale débutant par le caractère /. Toutes les balises suivantes sont utilisées de la même façon. Nous ne les expliquerons donc pas davantage, leur signification étant suffisamment explicite.

2. Ce type d'exemple est donné une seule fois, les autres normes pouvant être toutes représentées de la même manière.

```
?xml version= »1.0 »?>

<metadata
xmlns= »http ://example.org/myapp/ »
xmlns :xsi= »http ://www.w3.org/2001/XMLSchema-instance »
xsi :schemaLocation= »http ://example.org/myapp/
http ://example.org/myapp/schema.xsd »
xmlns :dc= »http ://purl.org/dc/elements/1.1/ »>

<dc :creator>
Blais, J.-G.
</dc :creator>

<dc :date>
2007.12.25
</dc :date>

<dc :title>
Évaluation des apprentissages et technologies de l'information
et de la communication
</dc :title>

<dc :description>
L'utilisation des technologies de l'information pour soutenir
l'administration, la construction et la correction d'instruments
de mesure dans le cadre de l'évaluation des apprentissages a connu
d'importants développements depuis ses débuts. Ces développements
seront analysés à l'intérieur de ce volume.
</dc :description>

<dc :publisher>
Presses de l'Université Laval
</dc :publisher>
</metadata>
```

Figure 1 – Exemple de fichier en langage XML selon la norme du noyau de Dublin

La norme du noyau de Dublin est limitée à des informations assez simples qui permettent de retracer des « objets » à l'intérieur du réseau Internet. Les banques de données bibliographiques comme ERIC utilisent ce type de normes pour retracer des documents. Cette norme offre assez peu d'informations formelles quant à des objets d'évaluation des apprentissages. Cependant, si, dans certaines situations, on peut se satisfaire d'être limité à considérer un objet d'évaluation uniquement comme un objet bibliographique, elle permet tout de même de connaître le ou les auteurs d'un item ou d'une modalité d'évaluation, sa description, sa date de création, etc. Le contenu de l'objet d'évaluation, les réponses admissibles, les consignes de

réalisation de la tâche, la façon de le passer, le résultat de l'élève, etc., ne sont toutefois pas pris en compte. Une autre norme devra prendre ces informations en charge.

LA NORME CMI (*COURSEWARE MANAGEMENT INTERFACE*)

La norme CMI a été développée par l'AICC (*Aviation Industry Computer based training Committee*). L'AICC est un regroupement de constructeurs aéronautiques, d'entreprises du domaine de l'aviation, d'agences gouvernementales, d'unités militaires ainsi que de firmes de développement informatique (McDonald, Hyde et Montgomery, 2004). L'AICC élabore des normes pour l'industrie de l'aviation visant le développement, la livraison et l'échange de matériel destiné à l'apprentissage par ordinateur. Le but principal de cet organisme est de standardiser le matériel et les interfaces utilisés par les ordinateurs portatifs et de promouvoir l'interopérabilité des systèmes d'apprentissage informatisé à travers divers organismes et industries. Les innovations et les travaux de l'AICC au regard des systèmes d'apprentissage sont d'aspect plutôt général, et ils sont largement utilisés par des organismes externes au domaine de l'aviation.

Les normes de l'AICC/CMI sont divisées selon les neuf catégories suivantes : 1) postes de travail destinés à l'apprentissage par ordinateur (*courseware delivery stations*) ; 2) numérisation et traitement des sons (*digital audio*) ; 3) systèmes d'exploitation et d'affichage de fenêtres d'interface graphiques (*operating/windowing system*) ; 4) périphériques nécessaires à l'apprentissage par ordinateur (*computer based training peripherical devices*) ; 5) systèmes de traitement de fichiers (*file-based CMI systems*) ; 6) interopérabilité et partage de cours sous format électronique (*courseware interchange*) ; 7) numérisation et traitement des images (*digital video*) ; 8) normalisation des icônes des interfaces usagers (*icon standards : user interface*) et 9) systèmes de gestion sur le « web » (*web-based CMI systems*).

À l'intérieur de ces neuf catégories de normes, on trouve des caractéristiques qui s'intéressent plus directement à l'évaluation des apprentissages. Cependant, ces caractéristiques sont formalisées pour un cours entier et n'ont donc pas une granularité assez fine pour caractériser la diversité des objets de l'évaluation des apprentissages. Elles permettent principalement de caractériser le résultat global à un cours ainsi que les objectifs d'apprentissage (d'enseignement) associés à un cours.

On remarque ainsi que la norme CMI s'intéresse bien plus à la présentation des objets d'apprentissages et à la gestion des médias associés qu'à ce qui touche les différentes dimensions de l'évaluation. Toutefois, il ne faut

perdre de vue que ces métadonnées sont importantes dans le développement de plateformes informatisées d'évaluation des apprentissages. Par exemple, la présentation des tâches d'évaluation à un élève influence le niveau de difficulté de cette tâche. Il devient donc important de pouvoir contrôler cette présentation. Un autre exemple est celui de l'adaptation de la présentation des tâches d'évaluation selon les besoins de clientèles spéciales, comme ceux des élèves qui ont un handicap visuel.

LA NORME LOM (LEARNING OBJECT METADATA)

La norme LOM est une extension des conventions du noyau de Dublin destinée à la description de n'importe quel objet d'apprentissage. Elle est elle aussi généralement encodée sous la forme de fichiers XML.

Les objectifs de la norme LOM peuvent se résumer à trois points. Premièrement, elle vise la création de descriptions bien structurées des objets d'apprentissage, ce qui devrait faciliter la recherche, la localisation, l'évaluation et l'acquisition de ces objets par des étudiants, des professeurs ou des environnements informatiques automatisés. Deuxièmement, elle vise le partage de la description des objets d'apprentissage entre diverses plateformes pour permettre une réduction du coût de création et pour fournir des services basés sur des descriptions de qualité des ressources utilisées. Troisièmement, la norme LOM a pour objectif la description de ressources pour convenir aux besoins spécialisés de différentes communautés d'utilisateurs. Ce dernier objectif implique le choix d'un vocabulaire approprié à chacune de ces communautés.

La version actuelle de la norme LOM ne possède pas de modèle formel permettant la conception et la production de tâches ainsi que la gestion de l'interopérabilité des objets d'évaluation. Elle tient pour acquis que les données d'évaluation sont intégrées aux contenus d'apprentissage. Cette norme comprend 78 éléments répartis en neuf catégories : 1) général, 2) cycle de vie, 3) métadonnées, 4) technique, 5) pédagogie, 6) droits, 7) relations, 8) commentaires, et 9) classification.

La catégorie « pédagogie » est le cœur de la norme LOM. C'est ce qui la rend spécifique par rapport à d'autres normes. Elle accomplit une fonction de description pédagogique de l'objet d'apprentissage. S'y retrouvent, pour ne citer que certains de ses éléments, le contexte d'utilisation (scolaire, secondaire, supérieur, formation continue, etc.), le type de public cible (enseignant, auteur, apprenant, gestionnaire), la difficulté, l'âge des apprenants auquel la ressource est destinée, le temps d'apprentissage type, la langue des apprenants, le niveau et le type d'interactivité. C'est dans cette catégorie qu'on peut définir le type d'objet d'évaluation. Cet attribut peut prendre les

valeurs *Questionnaire*, *Exam*, *Exercise*, *ProblemStatement* et *SelfAssessment*. Toutefois, c'est surtout dans la catégorie « classification » qu'on trouve une préoccupation pour l'évaluation des apprentissages. La norme LOM prend ainsi en compte, entre autres, des aspects qui permettent de définir le domaine de généralisation de l'objet d'apprentissage (qui peut aussi être une tâche d'évaluation) : discipline, objectifs pédagogiques et compétences. Cette norme peut ainsi soutenir la définition des objectifs d'évaluation ainsi que certaines modalités assez simples d'évaluation des apprentissages.

LA NORME IMS (INSTRUCTIONAL MANAGEMENT SYSTEM)

Le *Global Learning Consortium* est un groupe de travail fondé en 1997 qui est composé de membres faisant partie des domaines de l'éducation, de l'entreprise privée et d'organismes gouvernementaux. Ce groupe a proposé la norme *Instructional Management Systems* (IMS), qui a pour objectif principal la définition de caractéristiques d'interopérabilité des applications et des services d'enseignement distribués sur l'Internet. Ces caractéristiques sont fondées sur cinq principes fondamentaux : l'interopérabilité, l'accessibilité, la réutilisation, la pérennité, l'indépendance et la portabilité.

La norme IMS est en fait plutôt une série de caractéristiques qui sont en continuel développement. Dans plusieurs de celles-ci, on trouve des éléments utiles à la gestion des objets d'évaluation des apprentissages. Par exemple, la QTI-IMS (*Question and Test Interoperatibility*) permet de définir des modalités d'évaluation de type classique, principalement des examens. La RDCEO-IMS (*Reusable Definition of Competency of Educational Objectives*), pour sa part, permet la définition des objectifs d'évaluation. La ePortfolio-IMS facilite la gestion d'un dossier d'apprentissage. Les autres caractéristiques ne sont pas propres aux divers aspects de l'évaluation des apprentissages. On peut toutefois souligner que la LIP-IMS (*Learner Information Packaging*) offre l'opportunité d'organiser le profil d'un élève, dont ses résultats d'évaluation, tandis que la spécification GLC (*Instructional Management System/ Global Learning Consortium*) structure l'organisation de la séquence des objets d'apprentissages, dont ceux qui ont trait à l'évaluation.

Actuellement, la spécification sous IMS est probablement celle qui offre le plus de particularités sur l'évaluation des apprentissages. Il est aussi assez simple pour la communauté scientifique de contribuer au développement de nouvelles caractéristiques qui seraient associées à la norme IMS. Par exemple, les membres du groupe CAMRI de l'UQAM tentent actuellement de développer trois normes additionnelles sous IMS, qui fourniraient des caractéristiques pour les paramètres de chacun des items selon les

modélisations de la théorie de la réponse à l'item, de la génération automatisée d'items et des caractéristiques de leur présentation matérielle adaptative, ainsi que de caractéristiques propres à l'administration des tests adaptatifs (Lesage, Raîche, Riopel et Sodoké, 2007).

LA NORME SCORM (*SHARABLE CONTENT OBJECT REFERENCE MODEL*)

La norme SCORM consiste aussi, comme la norme IMS, en un agrégat de spécifications. Elle récupère le meilleur des normes AICC/CMI et IMS (*Secretary of defence 2000*). Elle s'impose aujourd'hui comme le standard en matière de conception de cours et de plateformes d'apprentissages informatisées. Un cours basé sur la norme SCORM est composé d'objets à contenu partagé (*Sharable Content Object*, SCO). Un SCO est une unité de contenu d'un niveau de granularité variable qui a une signification pédagogique, qui peut être réutilisée dans un autre cours et qui sera reconnaissable par toute plateforme d'apprentissage informatisée utilisant la norme SCORM. Un SCO pourra être composé de pages HTML, d'animations Flash, de dessins, de vidéos, etc. Plusieurs SCO pourront former un objet d'apprentissage, et un ou plusieurs objets d'apprentissage pourront former un cours. Un cours respectant la norme SCORM est dit RAID, ce qui signifie que ce cours est :

- Réutilisable : peut être utilisé et facilement modifié à partir de différents outils de développement informatique ;
- Accessible : peut être soumis à une recherche et offert aussi bien à des apprenants qu'à des développeurs ;
- Interopérable : peut fonctionner sur une grande variété d'environnements, plateformes informatisées, systèmes d'exploitation, navigateurs Web, etc. ;
- Durable : ne requiert pas d'importantes modifications lors de nouvelles versions des logiciels et des plateformes informatisées.

La norme SCORM présente trois aspects : 1) l'agrégation du contenu (*Content Packaging*), 2) l'environnement d'exécution (*runtime*) et 3) les métadonnées (*metadata*). On remarquera ainsi que cette norme ne s'intéresse pas particulièrement à l'évaluation des apprentissages. Toutefois, les SCO qui constituent un cours peuvent être spécifiés selon d'autres normes d'apprentissage, dont celles issues d'IMS, de CMI ou de LOM. Il est donc possible d'intégrer les caractéristiques d'évaluation de ces normes à l'intérieur d'un cours dit RAID.

LA NORME *MINE SCORM*

Pour pallier les limites de la norme SCORM au regard des objets d'évaluation des apprentissages, Chang, Hsu, Smith et Wang (2004) proposent une nouvelle norme, soit MINE SCORM. Celle-ci vise aussi à pallier certaines contraintes de la norme QTI-IMS qui, selon ces auteurs est, entre autres, très complexe et limitée à un nombre restreint de types d'items (huit), et ne comporte pas de particularités au regard des processus cognitifs impliqués dans une démarche d'évaluation.

La norme MINE SCORM, pour sa part, est plus souple quant aux types de questions admissibles (quoique limitée à notre avis), intègre une taxonomie qui s'apparente de près à celle de Bloom (1956), associe un domaine de connaissances à chacun des items, indique leurs niveaux de difficulté et de discrimination selon la théorie classique des tests ainsi que le temps moyen de réponse.

PROBLÈMES AVEC LA NORMALISATION DES OBJETS D'ÉVALUATION

La majorité des métadonnées prises en compte à l'intérieur de ces normes peuvent être classifiées en tant que métadonnées subjectives (Sodoké, 2007). En effet, mis à part les données physiques (la date de création, le type, le format, l'identifiant, les niveaux de difficulté et de discrimination, etc.), elles sont issues en majorité de l'appréciation subjective du concepteur. C'est le cas, par exemple, du sujet et de la description des objets d'apprentissage (ou d'évaluation). Le niveau de confiance envers ces métadonnées est ainsi limité. Dans ce contexte, la description des objets d'apprentissages ou d'évaluation est réalisée par des personnes ou des organismes, tandis que ces objets sont utilisés dans la plupart des cas par d'autres personnes ou organismes. Toute la responsabilité de la qualité des métadonnées revient donc à leurs concepteurs qui doivent alors être motivés pour fournir la description la plus précise possible. L'expérience dans des domaines similaires nous démontre toutefois qu'il faut se méfier de la qualité de la définition de ces métadonnées par leurs concepteurs.

L'exemple concret est celui du standard HTML qui offre aux concepteurs l'option d'inclure dans leur document des métadonnées relatives au contenu d'une page web. Cela se fait à travers la balise <meta> dans l'en-tête du document. Le but de cette balise était (et reste) identique au but des métadonnées utilisées par les normes d'apprentissages. L'utilisation de cette balise a été un échec, car beaucoup de moteurs de recherche qui s'appuient sur elle obtiennent de piètres résultats. Cette situation s'explique par le fait que ces

métadonnées sont volontairement ou involontairement mal conçues au départ, ou bien sont plus promotionnelles que descriptives.

Un autre problème consiste en l'utilisation intensive des métadonnées pour décrire les objets d'apprentissage. Généralement, le grand nombre de ces métadonnées, quel que soit le standard utilisé, fait en sorte que beaucoup de temps et d'efforts de la part des concepteurs sont exigés pour élaborer et mettre en application les objets d'apprentissage et d'évaluation. Ainsi, dans une activité d'apprentissage ou de formation, divers contextes peuvent se présenter, ce qui engendre rapidement une augmentation exponentielle des métadonnées nécessaires à la description d'un cours.

Outre ces considérations techniques, qui relèvent bien plus de préoccupations propres aux informaticiens, qui les ont d'ailleurs élaborées, les normes actuellement proposées présentent des lacunes importantes au regard des besoins en évaluation des apprentissages. Ces besoins concernent principalement la prise en compte formelle 1) des diverses fonctions de l'évaluation, 2) des avancées de la psychologie cognitive, 3) des contributions psychométriques au regard de la théorie de la réponse à l'item, 4) de la passation des tests adaptatifs, et 5) de la production automatisée de tâches d'évaluation. Ces besoins seront maintenant décrits plus en détail, quoique brièvement.

Ainsi, au regard de la prise en compte des fonctions de l'évaluation, les normes devront permettre la spécification des fonctions usuelles reconnues de l'évaluation des apprentissages : formative, sommative (certificative), diagnostique, de classement et de placement, etc. À ce sujet, cette spécification devra toutefois être ouverte, car les fonctions de l'évaluation ne sont pas nécessairement limitées à celles-ci.

La prise en compte des avancées en psychologie cognitive, couplée avec la tendance actuelle dans le mode de l'éducation d'une transition vers une fonction plus formative de l'évaluation des apprentissages, est aussi une préoccupation importante par rapport à l'utilisation des technologies de l'information et des communications pour soutenir l'évaluation. Plusieurs auteurs soulignent que la prise en compte des avancées de la psychologie cognitive est primordiale pour porter un jugement éclairé sur des apprentissages réalisés par un élève (Glaser et Nitko, 1971 ; Myslevy, 1995 ; Nitko, 1989 ; Snow et Lohman, 1989). Pour obtenir des mesures dites intelligentes, il faut tenir compte de la structure cognitive de l'élève et en rendre compte par le biais des environnements informatisés d'évaluation. À ce titre, les normes devront offrir des caractéristiques relatives aux diverses taxonomies des objectifs d'évaluation ainsi qu'à la caractérisation des objectifs vus sous l'angle des compétences. Des particularités devront aussi être proposées pour tenir compte des connaissances déclaratives, procédurales et métacognitives.

Les contributions psychométriques, surtout au regard de la théorie de la réponse à l'item (van der Linden et Hambleton, 1997), qui va au-delà des niveaux de difficulté et de discrimination de la théorie classique des tests (Cronbach, 1990), devront aussi pouvoir être utilisées à l'intérieur des plateformes éducatives qui appliquent les diverses normes. À cette fin, les paramètres d'items ou de sujets associés aux différentes modélisations issues de la théorie de la réponse à l'item devront pouvoir être formellement spécifiés, tout comme la modélisation cible impliquée. Il sera aussi nécessaire de pouvoir spécifier plus d'une modélisation à chacun des items, ainsi que des modélisations différentes entre ceux-ci.

Les normes actuelles ne permettent pas encore de supporter la passation de tests adaptatifs par ordinateur. Cette situation est quelque peu étrange si on considère l'important développement qu'a connu la passation de ce type de tests depuis quelques années. Il faudra alors offrir des particularités pour prendre en charge les règles de départ, de suite et d'arrêt qui sont utilisées dans les tests adaptatifs. La règle qui nécessitera qu'on lui apporte le plus de soins est sûrement celle de suite. C'est d'ailleurs sur elle que les plus importants travaux de recherche se font actuellement, puisque c'est à ce niveau que l'on juge qu'on peut résoudre une grande partie des problèmes de sécurité associés aux tests adaptatifs par ordinateur.

Enfin, la production automatisée de tâches d'évaluation nécessitera au cours des prochaines années qu'on se préoccupe d'élaborer des normes qui permettent de spécifier les caractéristiques qui affectent les divers paramètres des items. De cette façon, il sera possible de simplifier les travaux de création de banques d'items dans l'application des tests adaptatifs et de tenter de produire des évaluations qui tendent vers l'équivalence des échelles de mesure en jeu.

CONCLUSION

En conclusion, pour répondre à ces préoccupations propres aux besoins en évaluation des apprentissages, nous sommes d'avis qu'une série de nouvelles normes devront être formellement développées. Elles devront être complémentaires à celles qui existent actuellement pour éviter d'adopter une attitude de concurrence improductive dans leur utilisation. Enfin, il sera plus prometteur d'ajouter graduellement cette série de normes à la norme IMS, norme à laquelle la communauté des chercheurs et praticiens qui utilisent et développent les plateformes éducatives informatisées semble de plus en plus s'intéresser.

Références

Baxter, G. P. et Mislevy, R. J. (2005), *The case for an integrated design framework for assessing science inquiry*. PADI technical report 5, Menlo Park, CA, SRI International.

Bloom, B. S. (1956), *Taxonomy of educational objectives : the classification of educational goals – Cognitive domain*, New York, Longman.

Casadei, G. et Magnani, M. (2007), *Assessment strategies of an intelligent learning management system*, Bologna, Italy, University of Bologna.

Chang, W.C., Hsu, H.-H., Smith, T. K. et Wang, C.-C. (2004), « Enhancing SCORM metadata for assessment authoring in e-Learning », *Journal of Computer Assisted Learning,* vol. 20, p. 305-316.

Cronbach, L. J. (1990), *Essential psychological testing*, New York, Harper.

Fletcher, J. D., Tobias, S. et Wisher, R. A. (2007), « Learning anytime, anywhere : advanced distributed learning and the changing face of education », *Educational Researcher,* vol. 36, nº 2, p. 96-102.

Friesen, N. (2004), « Three objections to learning objects and E-learning standards ». Dans R. McGreal (dir.), *Online education using learning objects*, London, UK, Routledge, p. 59-70.

Giotopoulos, K. C., Alexanos, C. E., Beligiannis, G. N. et Likothanassis, S. D. (2005), « Computational intelligence techniques and agents' technology in E-learning environments », *International Journal of Information Technology,* vol. 2 nº 2, p. 147-156.

Heery, R. M. (1996), « Review of metadata formats », *Program,* vol. 30, nº 4, p. 345-373.

Hodgins, W. et Conner, M. (2000), « Everything you ever wanted to know about learning standards but were afraid to ask », *Linezine*,http ://linezine.com/2.1/features/wheyewtkls.htm.

Lesage, M., Raîche, G, Riopel, M. et Sodoké, K. (2007), *Interfaces Modelization in Computerized Adaptive Testing (CAT) Context for Learning Optimization*, Proceedings of LORNET annual Meeting, Montréal.

McDonald, W. A. Hyde, J. et Montgomery, A. (2004), *CMI guidelines for interoperability,* Sugar City, ID, AICC.

Mislevy, R. J., Steinberg, L. S. et Almond, R. G. (2002), « On the structure of educational assessments », *Measurement : Interdisciplinary Research and Perspectives,* vol. 1, nº 1, p. 3-66.

Secretary of defence (2000), *Department of defence implementation plan for advanced distributed learning*, Washington, DC, Office of the deputy under secretary of defence.

Seibert, G., Hamel, L., Haynie, K., Mislevy. R. J. et Bao, H. (2006), *Mystery powders : an application of the PADI design system using the four-process delivery system*, PADI technical report 17, Menlo Park, CA, SRI International.

Sodoké, K. (2007), *Conception et implémentation d'une plate-forme d'évaluation adaptative des apprentissages*, Mémoire de maîtrise inédit, Montréal, Université du Québec à Montréal.

van der Linden, W. J. et Hambleton, R. K. (1997), *Handbook of modern item response theory*, New York, Springer.

Wiley, D. A. (2000), « Connecting learning objects to instructional design theory : A definition, a metaphor, and a taxonomy », Dans D. A. Wiley (dir.), *The Instructional Use of Learning Objects : Online Version*, Document récupéré le 15 octobre 2007 à l'adresse URL *http ://reusability.org/read/chapters/wiley.doc*.

Page laissée blanche intentionnellement

CHAPITRE 13

La production automatisée de tâches d'évaluation[1]

Martin Riopel, Gilles Raîche,
Mélissa Pilote et Patrice Potvin
Université du Québec à Montréal

INTRODUCTION

La production automatisée de tâches d'évaluation est aussi désignée par l'expression anglaise *item generation* et consiste à utiliser un modèle formel pour prédire les caractéristiques de tâches nouvellement créées. Cette méthode permet de proposer des solutions à certains problèmes associés à la construction et à la validation de tests et qui sont présentés brièvement dans ce qui suit.

Lors de la construction d'un nouveau test, des experts ont habituellement la responsabilité de créer un ensemble de tâches adaptées au contexte de l'évaluation projetée. Dans le cas d'un test classique, toutes les tâches du test sont proposées à tous les candidats et le rôle des experts consiste à s'assurer que, pour une habileté suffisante/insuffisante d'un candidat donné, les tâches proposées sont généralement réussies/ratées. Si l'on suppose que les candidats à évaluer ont des habiletés semblables, alors un test classique comportant un seul ensemble de tâches validées pourra suffire pour obtenir une évaluation adéquate de leur habileté. Cependant, si des candidats présentent des habiletés très élevées ou très faibles, alors il sera plus difficile d'obtenir des données pertinentes pour les évaluer, car ils auront tendance à tout réussir

1. La production de ce chapitre a été rendue possible grâce à des subventions du Conseil de recherches en sciences humaines du Canada (CRSH) et du Fonds québécois de la recherche sur la société et la culture (FQRSC).

ou échouer à tout en bloc. Seule une limite inférieure ou supérieure de leur habileté pourra être obtenue.

Lorsque l'habileté varie beaucoup d'un candidat à l'autre, la création d'un seul ensemble de tâches ne suffit plus pour évaluer tous les candidats avec beaucoup de précision : de nouvelles tâches doivent être créées pour les cas extrêmes. De plus, ces nouvelles tâches ne doivent être proposées qu'aux candidats pour lesquels elles ont été créées : les tests doivent « s'adapter » aux candidats. L'adaptation optimale des tests aux candidats peut être encadrée par différents modèles, dont ceux de la théorie de la réponse à l'item (voir aussi le chapitre 10 sur des modèles adaptatifs reposant sur le concept d'espace de connaissance). Dans un contexte adaptatif, plus la variabilité des habiletés est grande, plus le nombre de tâches à créer doit augmenter. Évidemment, il devient alors de plus en plus difficile de valider ces nouvelles tâches au moyen d'expérimentations contrôlées.

Les recherches sur la production automatisée de tâches d'évaluation s'intéressent à la création d'un très grand nombre de nouvelles tâches dont les caractéristiques seraient prédites à l'aide d'un modèle formel. Dans ce contexte, le processus de validation ne ferait plus intervenir individuellement chacune des tâches créées, mais plutôt le modèle dans son ensemble. Une fois le modèle validé, de nouvelles tâches adaptées pourraient être créées automatiquement au besoin. Cette façon de procéder présenterait plusieurs avantages pour le coût du développement de nouvelles tâches (qui pourraient être produites en grand nombre rapidement) et pour la sécurité des tests produits (un grand nombre de tâches diminue significativement la possibilité de répétition de la même tâche). Bien qu'elle soit actuellement soutenue par les modélisations issues de la théorie de la réponse à l'item et qu'elle convienne bien aux approches adaptatives et informatisées, la production automatisée de tâches d'évaluation peut aussi être appliquée lors de situations plus classiques en salle de classe. La suite de ce texte présente les travaux qui ont jeté les bases théoriques de la production automatisée de tâches d'évaluation ainsi que les principales théories qui permettent de modéliser les paramètres d'items dans ce contexte. Il présente aussi quelques exemples d'application convenant au contexte d'une évaluation des apprentissages soutenue par la technologie informatique.

PREMIERS TRAVAUX

Selon Irvine (2002), les travaux sur la production automatisée de tâches d'évaluation ont permis d'envisager l'unification des deux principaux domaines de recherche en psychologie expérimentale tels que les définit Cronbach (1957). Le premier domaine s'intéresse à la mise en évidence

d'effets systématiques au moyen d'expérimentations contrôlées basées sur des modèles cognitifs explicites et un nombre de candidats relativement restreint. Le second domaine s'intéresse plutôt à la modélisation des différences individuelles observées dans les résultats pour différents contextes et, par conséquent, nécessite un nombre de candidats significativement plus élevé. Ces deux domaines sont aujourd'hui associés aux deux principaux courants concernant la production automatisée de tâches d'évaluation présentés par Bejar, Morley et Bennett (2003), soit le courant théorique fort qui se base sur des modèles cognitifs explicites pour produire des tâches d'évaluations standardisées, et le courant théorique faible, ou empirique, qui se base plutôt sur l'analyse des résultats de banques de tâches validées pour synthétiser des modèles.

Embretson (1999) et Bejar, Morley et Bennett (2003) s'entendent pour rapporter que les premiers travaux sur la production automatisée de tâches d'évaluation remontent à la fin des années 1960, où Hively, Patterson et Page (1968) ont proposé un système basé sur la combinaison d'éléments fixes (des phrases trouées) et d'éléments variables (des nombres ou des mots) pour produire des problèmes équivalents en mathématiques. On peut en voir une illustration dans l'exemple fictif suivant, où les éléments variables sont en caractères italiques :

Claude a acheté *3 concombres* à *2,69*$, *2 laitues* à *0,49*$
et *4 tomates* à *0,99*$. Quel est le montant total de ses achats ?

D'autres travaux complémentaires ont aussi été effectués durant la même période. Par exemple, Guttman (1969) a proposé une théorie pour faciliter la production régulée d'un grand nombre de tâches d'évaluation. De même, Uttal, Rogers, Hieronymous et Pasich (1969) ont proposé l'expression « *instruction générative* » pour décrire des systèmes informatisés qui prennent en considération les problèmes rencontrés par les apprenants afin d'adapter la présentation de leur contenu. Finalement, Bormuth (1970) et Hively (1974) ont proposé des approches permettant de produire des tâches d'évaluation pour mesurer l'atteinte d'objectifs pédagogiques spécifiques. Ces approches standardisées s'opposaient à la position défendue par certains (dont Wesman, 1971) qui défendent le point de vue selon lequel la production de tâches d'évaluation est essentiellement créative et correspond à un art dont la qualité ne peut être garantie par un ensemble de formules ou de règles strictes.

Tous les travaux présentés précédemment utilisaient la théorie classique des tests. Fisher (1995) rapporte que Scheilblechner (1971) a été le premier à s'intéresser à la production automatisée de tâches d'évaluation dans le contexte des modélisations issues de la théorie de la réponse à l'item. À ce

moment, il a appliqué un modèle de régression linéaire dans le but de prédire la valeur du paramètre de difficulté de tâches d'évaluation dans le contexte de la modélisation logistique à un paramètre de Rasch. Il a pu ainsi prédire de façon satisfaisante la valeur du paramètre de difficulté de tâches de compréhension de propositions logiques présentées graphiquement en fonction de trois opérations cognitives, soit la négation, la disjonction et l'asymétrie.

Une des limites de la méthode précédente provient du fait que la modélisation du paramètre de difficulté des tâches n'est pas effectuée directement à l'intérieur de la modélisation retenue de la réponse à l'item. En fait, l'application de la régression multiple est réalisée une fois les paramètres d'items obtenus, et non pas simultanément. Pour pallier cette limite, Fischer (1977) a proposé une adaptation de la modélisation logistique à un paramètre, soit le modèle logistique linéaire (*Linear Logistic Test Model*, LLTM). Cette modélisation de la réponse à l'item permet de calculer de façon plus précise la valeur du paramètre de difficulté des tâches d'évaluation, et est aussi plus près de l'esprit des modélisations issues de la théorie de la réponse à l'item.

À la suite de ces travaux fondateurs, Irvine (2002) rapporte que l'essor du champ de la production automatisée de tâches d'évaluation s'est poursuivi durant les années 1980 dans plusieurs centres géographiquement éloignés. Irvine propose une liste des auteurs qui rendent compte de cette période, soit Bartman (1987), Bejar (1986a, 1986b, 1986c), Carrol (1980, 1983, 1986, 1987), Christal (1984), Dann et Irvine (1986), Embretson (1983, 1985), Hornke et Habon (1986), et Kyllonen et Christal (1989). Selon Irvine, malgré leur relatif isolement scientifique, ces travaux créatifs et originaux ont contribué à mieux cerner la définition actuelle de ce nouveau champ et à préparer son implantation à de plus grandes échelles.

TRAVAUX PLUS RÉCENTS

Il existe actuellement deux théories pour modéliser les paramètres d'items dans le contexte de la génération de tâches d'évaluation : la théorie forte et la théorie faible. La théorie forte consiste à se baser sur des principes psychologiques lors de la conception des modèles de tâches afin de contrôler finement leurs paramètres. Comme le souligne Fischer (1995), les études de ce type sont limitées, bien sûr, par la disponibilité d'hypothèses à propos des opérations cognitives nécessaires à la réussite de la tâche et sont alors propres à celle-ci. C'est pourquoi il est impératif de toujours effectuer préalablement une analyse des tâches à l'étude. Dans cet esprit, les travaux d'Embretson (1993, 1999) indiquent que la théorie forte convient à des domaines de tâches spécifiques pour lesquelles une analyse cognitive est possible.

Pour des domaines plus larges, l'analyse cognitive est plus difficile, et la théorie faible est alors généralement préférée (Bejar, Morley et Bennett, 2003). Elle consiste à se baser sur un ensemble de tâches existantes dont les paramètres d'items sont connus. À partir de cet ensemble, certaines tâches sont choisies afin de construire des modèles généraux d'items. Dans le contexte de la théorie faible, ce ne sont pas principalement les principes psychologiques, mais plutôt l'expertise disciplinaire propre au domaine qui dirige la conception des modèles de tâches qui sont ensuite validés.

Plusieurs travaux ont été entrepris en considérant la théorie forte ou la théorie faible et ont permis de prédire, avec des succès variables, les valeurs du paramètre de difficulté et de discrimination de tâches d'évaluation. D'un point de vue théorique, ces travaux doivent être séparés en trois catégories que Irvine, Dann et Evans (1987) présentent ainsi : les modélisations de type D, de type L ou de type R. Les modélisations de type D concernent des tâches d'évaluations qui sont répétées plusieurs fois durant l'apprentissage. Les modélisations de type L concernent des tâches d'évaluation dont la réussite dépend de la vitesse d'exécution. Finalement, les modélisations de type R concernent des tâches d'évaluations dichotomiques qui seront soit réussies, soit ratées. Selon Irvine (2002), parmi ces trois types de modélisations, les deux premiers (D et L) présentent des problèmes techniques qui rendent difficile l'interprétation non biaisée des corrélations. Ces difficultés font en sorte que seules les modélisations de types R sont actuellement utilisées avec un certain succès dans les recherches (et sont présentées dans le texte qui suit). Les modélisations de type D et L représentent cependant un défi pour les prochaines années.

Un recensement de travaux moins récents impliquant des modélisations de type R est aussi présenté dans Crocker et Algina (1986) ainsi que dans Wainer et Messick (1983). Les sections qui suivent présentent des exemples de modélisations proposées plus récemment.

Propriétés essentielles et propriétés accessoires

Une des applications possibles des modélisations de type R est la passation systématique de tests à des échelles nationales. Bien que ces tests soient encore généralement proposés sous la forme papier-crayon à des groupes, plusieurs versions équivalentes des tests doivent tout de même être produites et continuellement renouvelées. Ce genre de test est utilisé, par exemple, pour le recrutement de l'armée américaine (Kyllonen et Christal, 1989) ou britannique (Irvine, Dann et Anderson, 1990). Cette dernière application pour l'armée britannique a permis à Irvine (2002) de comparer les résultats obtenus pour plusieurs versions de mêmes tests dans trois contextes différents

(les recrues, l'armée régulière, l'armée navale) ainsi que pour trois catégories différentes de questions (raisonnement, orientation, mémoire). Le tableau 1 présente la moyenne des résultats obtenus (en %) ainsi que la corrélation entre deux versions équivalentes des tests pour chacune des catégories. On peut y observer que les moyennes diffèrent d'au plus 1 %, et que les corrélations sont typiquement assez élevées (entre 0,76 et 0,92). Ainsi, comme le propose Irvine, la production automatisée de tests équivalents à partir d'une banque de tâches existantes dans ce contexte peut être réalisée par l'application d'algorithmes reconnus de la théorie de la réponse à l'item lors de la conception et de la validation des tâches d'évaluation (Armstrong, Jones et Wang, 1994). Par contre, si l'on envisage la production automatisée des tâches elles-mêmes, alors la situation devient plus difficile, et il faut bien distinguer entre les propriétés essentielles (qui produiront un effet systématique sur la difficulté de la tâche) et les propriétés accessoires (qui produiront aussi un effet sur la difficulté, mais de nature plutôt aléatoire). En effet, seules les propriétés essentielles pourront vraiment servir à prédire le niveau de difficulté de tâches nouvellement créées (Carroll, Meade et Johnson, 1991).

Tableau 1 – Résultats moyens et corrélation pour deux versions équivalentes de tests pour l'armée britannique (adapté de Irvine, 2002, p. 12)

	Recrues			Armée régulière			Armée navale			Nouveau		
Version	R	O	M	R	O	M	R	O	M	R	O	M
A	54	57	37	36	30	18	51	33	46	19	23	42
B	53	58	36	39	29	19	51	34	45	19	23	43
Corrélation	0,76	0,92	0,90	0,86	0,88	0,84	0,88	0,89	0,90	0,82	0,86	0,87
N	279			240			294			3000		

R = raisonnement, O = orientation, M = mémoire.

En modélisant les propriétés essentielles de tâches concernant l'orientation de figures, plusieurs chercheurs ont obtenu des coefficients de corrélation variant entre 0,72 et 0,92 lors de la prédiction du niveau de difficulté (Irvine, Dann et Anderson, 1990 ; Mislevy, Wingersky, Irvine et Dann 1991 ; Dennis, 1993, 1995). Les propriétés considérées étaient, par exemple, le nombre de rotations, l'angle des rotations, la position initiale et la longueur du trait. Dans le cas de tâches concernant la compréhension verbale, Dennis, Collis et Dann (1995) ont obtenu un coefficient de corrélation de 0,85 en considérant le type de phrase, le degré de transformation

grammaticale et la longueur des mots. Enfin, Kyllonen et Christal (1990) ont mis en évidence le rôle central de la mémoire (à court terme) dans la réalisation de l'ensemble des tâches d'évaluation données à grande échelle dans le contexte militaire avec une corrélation globale d'environ 0,40.

Modélisation cognitive du raisonnement abstrait

Comme cela a été mentionné précédemment, la théorie forte consiste à se baser sur des principes psychologiques lors de la conception de nouvelles tâches. Cette approche convient évidemment bien à la modélisation du raisonnement. Les travaux effectués par Embretson ont appliqué les théories de la psychologie cognitive au développement de tâches d'évaluation associées au raisonnement et à la prédiction du niveau de difficulté de ces tâches. Une dizaine de types de tâches ont été considérées : les analogies et les classifications verbales (Embretson, 1984, 1985 ; Embretson et Scheider, 1989 ; Embretson, Scheider et Roth, 1985), les analogies et les classifications géométriques (Whitely et Schneider, 1981), la complétion de séries numériques (Embretson, 1985), la compréhension de paragraphes (Embretson et Wetzel, 1987), l'assemblage spatial (Embretson 1991, 1994), les problèmes mathématiques en mots (Embretson, 1995a) et la complétion de matrices (Embretson 1995b, 1998). Pour l'ensemble de ces tâches, des corrélations toujours supérieures à 0,70 ont été observées. On peut considérer l'exemple de la figure 1 qui illustre une tâche typique concernant le raisonnement abstrait et où l'on demande de compléter une matrice.

Pour ce type de tâche, un modèle cognitif a été développé par Carpenter, Just et Shell (1990), puis raffiné par Embretson (1999). Il consiste à considérer deux propriétés cognitives essentielles (le nombre de règles présentes et la correspondance abstraite), ainsi que trois propriétés perceptuelles (la superposition, la fusion et la distorsion des objets). Les corrélations obtenues sont présentées dans le tableau 2.

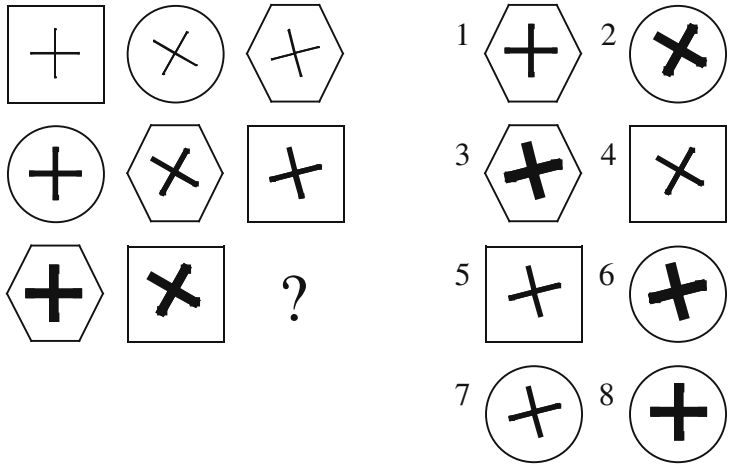


Figure 1 – Tâche d'évaluation concernant le raisonnement abstrait

Tableau 2 – Corrélation obtenue pour la complétion de matrice (globale = 0,79) (adapté de Embretson, 1999, p. 412)

Catégorie	Propriété	Corrélation
Cognitive	Nombre de règles	0,69
	Correspondance abstraite	0,71
Perceptuelle	Superposition	-0,24
	Fusion	-0,42
	Distorsion	0,31

Dans d'autres travaux concernant la modélisation du raisonnement abstrait, Newstead, Bradon, Handley et Dennis (2002) identifient cinq propriétés influençant les niveaux de difficulté observés :

1) les négations sont significativement plus difficiles que les affirmations ;
2) les règles d'assignation sont les plus faciles ;
3) les ensembles de règles qui restreignent le nombre de possibilités sont plus faciles ;
4) les leurres possibles sont plus difficiles que les leurres impossibles ;

5) le nombre de règles présentes influence considérablement le niveau de difficulté.

Ces chercheurs ont utilisé, entre autres, des tâches d'évaluation sur les orientations et les distances, dont deux exemples typiques ont été traduits librement dans le tableau 3.

Tableau 3 – Tâches d'évaluation sur les orientations et les distances

Descripteur	Énoncé de la tâche
Mouvement d'un bateau	Un bateau quitte le port, navigue 10 km Nord-Est, puis 15 km Sud-Est, puis encore 4 km Nord-Est et, finalement, 15 km Nord-Ouest. Dans quelle direction se situe maintenant le port par rapport au bateau ? a) Nord-Ouest. b) Sud-Ouest. c) Nord-Est. d) Sud-Est.
Campus	Sur un campus, le Département de chimie se trouve à 500 m au Sud-Est de la Bibliothèque, la Bibliothèque se trouve à 800 m à l'Ouest du Registrariat, et le Registrariat se trouve à 500 m au Nord-Est de la Chapelle. Dans laquelle des paires suivantes les bâtiments sont-ils les plus rapprochés ? a) Le Département de chimie et la Chapelle. b) Le Département de chimie et le Registrariat. c) La Bibliothèque et la Chapelle. d) La Bibliothèque et le Registrariat.

Pour ces deux exemples, Dennis, Handley, Evans et Newstead (2002) rapportent une valeur du coefficient de corrélation de 0,88. Leur modèle pour le premier exemple était basé sur trois propriétés essentielles :

1) la méthode utilisée pour indiquer les directions (droite/gauche ou nord/sud) ;
2) l'utilisation des cardinaux (N, S, E, O) ou des semi-cardinaux (NE, NO, SE, SO) ;
3) le type de question « Dans quelle direction… » ou « À quelle distance… ».

De façon générale, ces chercheurs insistent sur le fait qu'une mise en application réussie de la production automatisée de tâches d'évaluation repose en grande partie sur le choix des tâches à considérer. Ils proposent une approche qui utilise des tâches pour lesquelles on peut construire un ensemble complet présentant à la fois un grand éventail de niveaux de difficulté et un nombre restreint de propriétés essentielles dont on pourra étudier à fond les effets. Ils proposent aussi d'éviter l'utilisation des leurres qui multiplient le nombre de propriétés à considérer et ont tendance à complexifier grandement leur étude.

Modélisation du raisonnement quantitatif

Contrairement à la théorie forte, la théorie faible consiste à se baser sur une banque de tâches pour un domaine donné dont le niveau de difficulté est connu, et d'en choisir un sous-ensemble afin de construire un modèle. L'objectif ultime est évidemment d'étendre les sous-ensembles jusqu'à considérer toutes les tâches de la banque. Les tâches d'évaluation du raisonnement quantitatif conviennent bien à cet exercice à cause de leur construction. Les travaux de Goldin et McClintock (1984) ont porté sur les propriétés qui influencent la difficulté de tâches de résolution de problèmes en mathématiques. En plus de la syntaxe, du contenu et du contexte du problème, ils ont constaté que la structure mathématique pouvait influencer la représentation mentale d'un problème. Enright et Sheenan (2002) ont aussi choisi cette approche à partir des tâches du *Graduate Record Examinations (GRE) quantitative measure*, qui évalue le raisonnement quantitatif de façon très large. Un groupe de vingt questions à développement ont d'abord été choisies en fonction de leurs propriétés comparables. Ensuite, une analyse des solutions produites par 50 étudiants a permis d'en extraire les structures mathématiques présentes. Enfin, trois propriétés essentielles ont été identifiées :

1) la complexité mathématique : le nombre d'opérations, le nombre de contraintes, le nombre de niveau de parenthèses ;
2) le contexte dans lequel sont effectués les calculs : temps, argent, distance ;
3) l'algébricité : le fait que la réponse soit demandée sous la forme d'une expression algébrique plutôt que comme un nombre.

Dans le contexte de la théorie de la réponse à l'item, les paramètres de difficulté, de discrimination et de pseudochance ont ensuite été obtenus pour chacune des valeurs possibles des propriétés essentielles précédentes. Ces propriétés ont permis de prédire le niveau de difficulté des tâches avec un coefficient de corrélation linéaire de 0,95. Parmi les propriétés considérées, le contexte avait peu d'influence sur le niveau de difficulté.

Les chercheurs ont ensuite tenté d'étendre la modélisation à deux groupes d'environ 350 tâches représentatives de l'ensemble de la banque et classées selon les propriétés suivantes :

1) type de calcul(avec des entiers, des nombres réels ou des fractions, avec des inégalités, des probabilités ou des pourcentages) ;
2) niveau d'habileté cognitive (application de procédures, compréhension des concepts, réflexion de plus haut niveau) ;
3) format (données numériques ou variable algébrique).

Ces propriétés ont permis de prédire le niveau de difficulté pour les deux groupes de tâches avec une valeur du coefficient de corrélation de Pearson de 0,60. D'autres propriétés ont aussi été prises en considération afin d'améliorer le modèle, mais sans succès. Les valeurs des coefficients de corrélation sont demeurées comparables.

Ces derniers résultats sont cohérents avec ceux recensés par Hornke (2002), qui a comparé l'efficacité de la production de tâches d'évaluation pour différents types de tests. Le tableau 4 présente ces résultats. On peut y observer que les tâches faisant intervenir la mémoire présentent les corrélations les plus élevées.

Tableau 4 – Corrélation obtenue pour la prédiction du niveau de difficulté dans différents types de test (adapté de Hornke, 2002, p. 176)

Type de test	Description	Corrélation
Analogies verbales	Difficulté basée sur la complexité de la terminologie et de la formulation de l'énoncé	0,39
Problèmes quantitatifs	Résolution à l'aide d'opérations sur des nombres ou des variables	0,59
Rotations mentales	Pairage d'objets ayant subi des rotations en trois dimensions	0,62
Analyse visuelle	Reconnaissance de la présence d'une figure dans une autre figure plus complexe	0,63
Complétion de matrices	Détermination du contenu de la case manquante d'une matrice	0,72
Mémoire visuelle	Mémorisation de la position des éléments sur une carte	0,88
Mémoire verbale	Remise en ordre des étapes mémorisées d'un itinéraire routier	0,94

En fait, comme Hornke le suggère, plus les tâches sont produites selon un algorithme contraignant, plus les corrélations semblent élevées. Ainsi, la modélisation de l'ensemble des tâches d'une banque sera généralement moins bien corrélée que la modélisation d'un sous-ensemble de tâches choisies spécifiquement pour la production automatisée.

Modélisation de la compréhension de texte

Dans le tableau 4, les corrélations les plus modestes concernent la prédiction du niveau de difficulté relié à la formulation d'un énoncé. Ce constat est repris par Gorin et Embretson (2006), qui proposent que les modèles cognitifs associés à la compréhension de texte pourraient être beau-

coup plus complexes que ceux associés au raisonnement quantitatif. Lors de la prédiction du niveau de difficulté, par exemple, les propriétés des tâches de compréhension de texte sont souvent plus difficiles à modéliser quantitativement.

Certains travaux de recherche en intelligence artificielle (Powers, Burstein, Chodorow, Fowles et Kukich, 2000 ; Passonneau, Hemat, Plante et Sheehan, 2002) ont permis d'automatiser la caractérisation d'un extrait de texte. Cependant, pour produire automatiquement des tâches d'évaluation de difficulté connue à partir d'un extrait donné, il est nécessaire de mieux comprendre les processus cognitifs associés à la lecture. Plusieurs recherches ont été effectuées pour établir des relations entre les modèles cognitifs et les difficultés de compréhension de lecture. On peut consulter à ce sujet, par exemple, Anderson (1982), Mitchell (1983), Embretson et Wetzel (1987), Freedle et Kostin (1993), et Sheenan et Ginther (2000). Selon Gorin et Embretson (2006), dans tous ces modèles, les propriétés à considérer peuvent être séparées en trois catégories principales :

1) les propriétés associées à la représentation du texte (nombre et type de propositions, nombre d'adjectifs, nombre de verbes, proportion des mots de contenu, proportion du texte utile) ;
2) les propriétés associées à la décision de réponse (niveau du vocabulaire des réponses proposées ; confirmabilité ou réfutabilité des réponses proposées ; niveau de transformation nécessaire pour confirmer ou infirmer les réponses) ;
3) les propriétés particulières reliées à la banque de tâches d'évaluation choisie (longueur de l'extrait, format de la question).

Le tableau 5 présente les corrélations obtenues par Gorin et Embretson (2006) avec ces catégories à partir de la banque de tâches du *Graduate Record Examination – Verbal*.

Tableau 5 – Corrélation obtenue pour la prédiction du niveau de difficulté dans la compréhension de texte (adapté de Gorin et Embretson, 2006, p. 395)

Catégories	Corrélation
Représentation du texte	0,09
Représentation du texte + décision de réponse	0,57
Représentation du texte + décision de réponse + particulières	0,62

On peut y constater que ce sont les propriétés associées à la décision de réponse, plutôt que celles associées au texte lui-même, qui semblent déterminantes dans la prédiction du niveau de difficulté. Un autre résultat

intéressant rapporté par ces auteurs de même que par Riopel, Gagnon, Gagnon et Maisonneuve (2006) est que les propriétés associées aux choix de réponses proposés comme leurres ne semblent pas significatives, contrairement à celles associés aux bonnes réponses.

Modélisation dans un contexte informatisé

En plus des propriétés associées aux tâches d'évaluation elle-mêmes, Irvine (2002) propose de prendre aussi en considération les propriétés associées aux modalités de passation. Dans ce contexte, l'ordinateur enrichit l'espace des possibilités en permettant, entre autres, de présenter des contenus dynamiques ou interactifs. Le tableau 6 présente des propriétés supplémentaires qui pourraient contribuer significativement à prédire le niveau de difficulté dans un contexte informatisé.

Tableau 6 – Propriétés informatiques susceptibles d'influencer le niveau de difficulté (adapté de Irvine, 2002, p. 7)

Affichage	Information	Réponse
Statique	Modalité	Clavier
Dynamique	Quantité	Console
Séquentiel	Ordre	Écran tactile
Interactif	Rythme	Souris
	Qualité	Voix

En plus d'influencer le niveau de difficulté des tâches elles-mêmes, l'ordinateur pourrait aussi être utilisé dans un contexte d'intelligence artificielle pour passer des tests en tant que candidat. Ce genre d'approche a déjà été explorée par Bejar (1986) ainsi que Irvine, Dann et Evans (1987), et pourrait être poursuivie dans les prochaines années, et plus spécialement dans le contexte de l'éducation informatisée à distance.

CONCLUSION

L'ensemble des travaux présentés dans ce texte donne un aperçu de la jeunesse, du dynamisme et de l'importance de recherches sur la production automatisée de tâches d'évaluation. Dans le contexte actuel, cette nouvelle façon de procéder présente plusieurs avantages pour le coût de développement de nouvelles tâches et la sécurité des tests produits. Cette approche permet aussi, en considérant explicitement les modèles cognitifs, de contribuer à leur développement formel. Enfin, l'application de cette approche dans un contexte informatisé ouvre la porte à des évaluations de plus en plus

personnalisées dans un contexte éducatif où l'on favorise actuellement des évaluations plus authentiques du cheminement d'apprentissage.

Il est important de mentionner en terminant que la plupart des travaux effectués jusqu'à maintenant sur la production de tâches d'évaluation ont été réalisés à partir de tâches, soit relativement simples, soit dans un contexte expérimental assez contrôlé. Ils ont un peu moins ciblé des situations d'évaluation plus généralisables aux divers contextes éducatifs. La plupart de ces travaux se sont aussi limités à la modélisation d'une seule caractéristique de la tâche, sa difficulté. Dans les travaux futurs, il pourrait être important de mieux considérer d'autres caractéristiques des tâches, telles que leur puissance de discrimination.

Bibliographie

Anderson, R. C. (1982), « How to construct achievement tests to assess comprehension », *Review of Educational Research*, nº 42, p. 145-170.

Armstrong, R. D., Jones, D. H, et Wang, Z. (1994), « Automated parallel test construction using classical test theory », *Journal of educational statistics*, nº 19, p. 73-90.

Bartman, D. (1987), « The development of an automated testing system for pilot selection : The MICROPAT project », *Applied psychology : international review*, nº 36, p. 279-298.

Bejar, I. I. (1986), *Final report : Adaptive testing of spatial abilities (ONR 150 531),* Princeton, NJ, Educational Testing Service.

Bejar, I. I. (1986a), *The psychometrics of mental rotation (RR-86-19)*, Princeton, New Jersey, Educational testing service.

Bejar, I. I. (1986b), *Analysis and generation of hidden figure items : a cognitive approach to psychometric modeling (RR-86-20)*, Princeton, New Jersey, Educational testing service.

Bejar, I. I. (1986c), *Final report : adaptive testing of spatial abilities (ONR 150 531)*, Princeton, New Jersey, Educational testing service.

Bejar, I. I., Morley, M. E. et Bennett, R. E (2003), « A feasibility study of on-the-fly item generation in adaptive testing », *The journal of technology, learning and assessment*, vol. 2, nº 3, p. 1-29.

Bormuth, J. R. (1970), *On the theory of achievement test items*, Chicago, University of Chicago press.

Carpenter, P.A., Just, M.A. et Shell, P. (1990), « What one intelligence test measures : A theoretical account of processing in the Raven's Progressive Matrices Test », *Psychological Review*, nº 97, p. 404-431.

Carrol, J. B. (1980), *Individual difference relations in psychometrics and experimental cognitive tasks (report no 163)*, Thurstone psychometric laboratory, Chapel Hill, University of North Carolina.

Carrol, J. B. (1983), « The difficulty of a test and its factor composition revisited », dans H. Wainer et S. Messick (dir.), *Principals of modern psychological measurement*, Hillsdale, New Jersey, Lawrence Erlbaum Associates.

Carrol, J. B. (1986), « Defining habilitéis through the person characteristic function », dans S. E. Newstead, S. H. Irving et P. L. Dann (dir.), *Human assessment : cognition and motivation*, Dordrecht, Netherlands, Nijhoff.

Carrol, J. B. (1987), « New perspectivas in the analisis of habilities », dans R. R. Ronning, J. A. Glover, J. C. Conoley et J. C. Witt (dir.), *The influence of cognitive psychology on testing, Hillsdale*, New Jersey, Lawrence Erlbaum Associates.

Carroll, J. B., Meade, A. et Johnson, E. S. (1991), « Test analysis with the person characteristic function : implications for defining abilities », dans R. E. Snow et D. E. Wiley (dir.), *Improving inquiry in education, psychology and social science : a book in honour of Lee J. Cronbach*, Hillsdale, NJ, Lawrence Erlbaum Associates, p. 109-143.

Christal, R. E. (1984), *New cognitive tests being evaluated by TTCP services, Report to the technical cooperation program meeting of 1984,* USAF Armstrong laboratory, San Antonio, Texas, Brooks AFB.

Crocker, L. et Algina, J. (1986), *Introduction to classical and modern test theory*, New York, Harcourt Brace Jovanovich College Publishers.

Cronbach, L. J. (1957), « The two disciplines of scientific psychology », *American psychologist*, nº 12, p. 671-684.

Dann, P. L. et Irvine, S. H. (1986), *Handbook of computer-based cognitive tasks*, Plymouth, UK, Center for computer-based assessment, University of Plymouth.

Dennis, I. (1993), *The development of an item generation test of spatial orientation closely related to test SP80a SP(N),* Report TR 307, Office of senior psychologist (naval), ministry of Defence, London.

Dennis, I. (1995), *The structure and development of numeracy and literacy tests in the navy personnel series*, technical report for the Office of senior psychologist (naval), Plymouth, UK, Human assessment laboratory, University of Plymouth.

Dennis, I., Collis, J. M. et Dann, P. L. (1995), « Extending the scope of item generation to tests of educational attainment », *Proceedings of the 37th international military testing association conference*, Toronto, Canada.

Dennis, I., Handley, P. B., Evans, J. et Newstead, S. (2002), « Approaches to modeling item-generative tests », dans S. H. Irvine et P. C. Kyllonen (dir.), *Item generation*

for test development, Mahwah, New Jersey, Lawrence Erlbaum Associates, p. 53-72.

Embretson, S. E. (1983), « Construct validity : construct representation versus nomothetic span », *Psychological Bulletin*, nº 93, p. 179-197.

Embretson, S. E. (1984), « A general multicomponent latent trait model for response processes », *Psychometrika*, nº 49, p. 175-186.

Embretson, S. E. (1985), *Test design : Developments in psychology and psychometrics*, New York, Academic Press.

Embretson, S. E. (1991), « A multidimensional latent trait model for measuring learning and change », *Psychometrika,* nº 56, p. 495-516.

Embretson, S. E. (1993), « Psychometric models for learning and cognitive processes », dans N. Frederiksen, R. J. Mislevy, et I. I. Bejar (dir.), *Test theory for a new generation of tests*, Hillsdale, NJ, Lawrence Erlbaum Associates, p. 125–150.

Embretson, S. E. (1994), « Application of cognitive design systems to test development », dans C. R. Reynolds (dir.), *Cognitive assessment : A multidisciplinary perspective,* New York, Plenum Press, p. 107-135

Embretson, S. E. (1995a), « A measurement model for linking individual change to processes and knowledge : Application to mathematical learning », *Journal of Educational Measurement*, nº 32, p. 277-294.

Embretson, S. E. (1995b), « The role of working memory capacity and general control processes in intelligence », *Intelligence*, nº 20, p. 169-190.

Embretson, S. E. (1998), « A cognitive design system approach to generating valid tests : Application to abstract reasoning », *Psychological Methods,* nº 3, p. 380-396.

Embretson, S. E. (1999), « Generating items during testing : psychometric issues and models », *Psychometrika*, vol. 64, nº 4, p. 407-433.

Embretson, S. E. et Schneider, L. M. (1989), « Cognitive models of analogical reasoning for psychometric tasks », *Learning and Individual Differences,* nº 1, p. 155-178.

Embretson, S. E. et Wetzel, D. (1987), « Component latent trait models for paragraph comprehension tests », *Applied Psychological Measurement*, nº 11, p. 175-193.

Embretson, S. E., Schneider, L. M. et Roth, D. L. (1985), « Multiple processing strategies and the construct validity verbal reasoning tests », *Journal of Educational Measurement*, nº 23, p. 13-32.

Enright, M. K. et Sheenan, K. M. (2002), « Modeling the difficulty of quantitative reasoning items : implications for item generation », dans S. H. Irvine et P. C. Kyllonen (dir.), *Item generation for test development*, Mahwah, New Jersey, Lawrence Erlbaum Associates, p. 129-158.

Fischer, G. H. (1977), « Linear logistic trait model : Theory and application », dans H. Spade et W. F. Kempf (dir.), *Structural models of thinking and learning*, Berne, Huber.

Fischer, G. H. (1995), « The linear logistic test model », dans G. H. Fischer et I. O. Molenaar (dir.), *Rasch models. Foundations, recent developments, and applications*, New York, NJ, Springer-Verlag.

Freedle, R. et Kostin, I. (1991), *The prediction of SAT reading comprehension item difficulty for expository prose passages (ETS Research Report RR 91-29)*, Princeton, NJ, ETS.

Goldin, G. A., McClintock C. E. (1984), *Task Variables in Mathematical problem solving,* Mahwah, The Franklin Institute Press.

Gorin, J. S. et Embretson, S. E. (2006), « Item difficulty of paragraph comprehension items », *Applied psychological measurement*, vol 30, nº 5, p. 394-411.

Guttman, L. (1969), « Integration of test design and analysis », *Proceedings of the 1969 invitational conference on testing problems*, Princeton, New Jersey, Educational Testing Service.

Hively, W. (1974), « Introduction to domain-reference testing », *Educational technology*, vol. 14, nº 6, p. 5-10.

Hively, W., Patterson, H. L. et Page, S. (1968), « A "universe-defined" system of arithmetic achievement tests », *Journal of educational measurement*, nº 5, p. 275-290.

Hornke, L. F. (2002), « Item-reneration models for higher cognitive function », dans S. H. Irvine et P. C. Kyllonen (dir.), *Item generation for test development*, Mahwah, New Jersey, Lawrence Erlbaum Associates, p. 159-178.

Hornke, L. F. et Habon, M. W. (1986), « Rile-based item bank construction and evaluation within the linear logistic framework », *Applied psychological measurement*, nº 10, p. 369-380.

Irvine, S. H. (2002), « The foundations of item generation for mass testing », dans S. H. Irvine et P. C. Kyllonen (dir.), *Item generation for test development*, Mahwah, New Jersey, Lawrence Erlbaum Associates, p. 3-34.

Irvine, S. H., Dann, P. L. et Anderson, J. D. (1990), « Towards a theory of algorithm-determined cognitive test construction », *British Journal of psychology*, nº 81, p. 173-195.

Irvine, S. H., Dann, P. L. et Evans, J. St. B. T. (1987), *Item generative approaches for computer-based testing : a prospectus for research,* Report for the army personnel research establishment, Plymouth, UK, Human assessment laboratory, University of Plymouth.

Kyllonen, P. C. et Christal, R. E. (1990), « Reasoning ability is (little more than) working memory capacity ?! », *Intelligence*, nº 14, p. 389-433.

Kyllonen, P. C. and Christal, R. E. (1989), « Cognitive modelling of learning abilities », dans R. Dillon et J. W. Pellegrino (dir.), *Testing : theoretical and applied issues*, San Francisco, Freeman, p. 143-173.

Mislevy, R. J, Wingersky, M. S., Irvine, S. H. et Dann, P. L. (1991), « Resolving mixtures of strategies in spatial visualisation tasks », *British journal of mathematical and statistical psychology*, nº 44, p. 265-288.

Mitchell, K. (1983), *Cognitive processing determinants of item difficulty on the verbal subtests of the Armed Services Vocational Aptitude Batters and their relationships to success in army training*, Thèse de doctorat inédite, Université Cornell.

Newstead, S., Bradon, P., Handley, J. E. et Dennis, I. (2002), « Using the psychology of reasoning to predict the difficulty of analytical reasoning problems », dans S. H. Irvine et P. C. Kyllonen (dir.), *Item generation for test development*, Mahwah, New Jersey, Lawrence Erlbaum Associates, p. 35-52.

Passonneau, R., Hemat, L., Plante, J. et Sheehan, K. M. (2002), *Electronic sources as input to GRE reading comprehension development : Source-Finder prototype evaluation (ETS Research Report RR 02-12)*, Princeton, NJ, ETS.

Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E. et Kukich, K. (2000), *Comparing the validity of automated and human essay scoring (GRE N° 98-08a, ETS RR-00-10)*, Princeton, NJ, ETS.

Riopel, M., Gagnon, A., Gagnon, F. et Maisonneuve, H. (2006), *Production et validation d'un test de lecture visant le dépistage d'élèves en difficulté d'apprentissage*, Montréal, Association de collèges privés du Québec.

Sheehan, K. M. et Ginther, A. (2000), *What do passage-based multiple-choice verbal reasoning items really measure ? An analysis of the cognitive skills underlying performance on the current TOEFL reading section*, Texte présenté à l'occasion de la rencontre annuelle du National Council of Measurement in Education, Nouvelle-Orléans, États-Unis.

Uttal, W. R., Rogers, M., Hieronymous, R., et Pasich, T. (1969), *Generative CAI in analytic geometry*, Ann Arbor, Michigan, University of Michigan.

Wainer, H. et Messick, S. J. (1983), *Principles of modern psychological measurement*, Hillsdale, NJ, Lawrence Erlbaum Associates.

Wesman, A. G. (1971), « Writing the test item », dans R. L. Thorndike (dir.), *Educational measurement*, Washington, DC, American Concil on Education, p. 81-129.

Whitely, S. E. et Schneider, L. M. (1981), « Information structure on geometric analogies : A test theory approach », *Applied Psychological Measurement*, n° 5, p. 383-397.

CHAPITRE 14

La détection des patrons de réponses problématiques dans le contexte des tests informatisés[1]

Jean-Guy Blais
Université de Montréal

Gilles Raîche
Université du Québec à Montréal

David Magis
Université de Liège

INTRODUCTION

Le résultat obtenu par une personne à un test n'est pas toujours représentatif de son niveau réel d'habileté et peut constituer une information trompeuse sur ses capacités. Parfois, les candidats manquent d'attention, de motivation ou de préparation, mais il y a également des situations où des personnes réussissent intentionnellement des tests de classement en deçà de leurs capacités en vue, par la suite, de fournir un effort moindre et qui ne correspond pas à leur vrai potentiel. À titre d'exemple, c'est la situation qu'on observe au test de classement en anglais, langue seconde à l'intérieur du réseau collégial au Québec (Raîche, 2004). Dans d'autres situations, les réponses peuvent être copiées sur celles d'une autre personne, et le résultat

1. La production de ce chapitre a pu être réalisée grâce à l'obtention de subventions auprès du Conseil de recherche en sciences humaines du Canada (CRSH), du Fonds québécois de la recherche sur la société et la culture (FQRSC), d'une bourse de stage doctoral de l'Université de Liège, et du Programme d'aide aux chercheurs et chercheuses de l'Université du Québec à Montréal (PAFACC).

obtenu est alors une mauvaise estimation du niveau d'habileté, celui-ci se trouvant surestimé. Pour chacun de ces contextes, la production d'un patron de réponses ne permet pas de connaître le niveau d'habileté réel de l'étudiant.

Dans le cas des tests conventionnels, papier-crayon ou écran-clavier, de longueur et de contenu fixes, des indices d'ajustement ont été proposés pour détecter des patrons de réponses dits « problématiques ». Ces indices s'avèrent très efficaces lorsque le niveau de difficulté des items qui composent le test varie suffisamment au dessus et au dessous du niveau d'habileté réel du candidat. Toutefois, dans le cas de la passation de tests adaptatifs par ordinateur, forme de tests sur mesure qui se modifient en fonction des réponses fournies par le candidat (voir le chapitre 3), la stratégie du « sur mesure » consiste à faire en sorte que le niveau de difficulté des items se rapproche constamment du niveau d'habileté de l'étudiant. Ainsi, le niveau de difficulté moyen des items soumis à une même personne pourrait varier très peu, et les indices de détection des patrons de réponse problématiques s'avèrent peu efficaces dans ces situations. Malgré cela, est-il possible d'utiliser des indices de détection de réponses problématiques dans le contexte des tests adaptatifs ? Avec comme objectif de contribuer à fournir des réponses à ces questions, ce chapitre présente des propositions de modélisations des paramètres associés aux personnes (plutôt qu'aux items) qui pourraient être pertinentes pour la détection des patrons de réponses problématiques dans le contexte des tests adaptatifs par ordinateur.

LE CONTEXTE

Un test adaptatif est un test où la séquence de passation des items peut se modifier en fonction des réponses obtenues à des items préalablement donnés. Les items proposés et leur nombre peuvent donc être différents d'une personne à une autre. On évite alors de donner des items trop faciles à un candidat dont le niveau d'habileté est élevé, ou des items trop difficiles lorsque le niveau d'habileté du candidat est plutôt faible. Cette stratégie permet de diminuer la longueur du test et d'augmenter la précision de la mesure du niveau d'habileté de l'étudiante ou de l'étudiant (Thissen et Mislevy 2000).

Un des problèmes rencontré avec les tests adaptatifs est la difficulté de détecter un patron de réponses qui serait problématique. En effet, la détection d'un tel patron de réponses a été bien étudié dans le cas d'un test conventionnel, et plusieurs auteurs ont proposé des stratégies et des indices à cette fin (Drasgow, Levine et Zickar, 1996 ; Molenaar et Hoijtink, 2000 ; Reise et Flannery, 1996 ; Raîche, 2002 ; Raîche et Blais, 2003). Ces indices

sont particulièrement efficaces lorsque le niveau de difficulté des items constituant le test varie considérablement. On peut ainsi détecter si une personne rate des items qui devraient être faciles pour elle, ou encore, réussit des items qui devraient être trop difficiles. Dans le cas d'un test adaptatif cependant, il est nécessaire de minimiser la variation du niveau de difficulté des items. Au fil de la passation d'un test adaptatif, le niveau de difficulté des items se rapproche constamment du niveau d'habileté de l'étudiant. Or, cette stratégie de minimisation va à l'encontre de la stratégie de maximisation de la variation du niveau de difficulté des items qui doit être utilisée pour détecter un patron de réponses qui pourrait être problématique. Comment peut-on alors tenir compte de cet obstacle dans la mise en œuvre d'un test adaptatif ?

LES INDICES DE DÉTECTION DES PATRONS DE RÉPONSES PROBLÉMATIQUES

Une des premières études publiées sur ce sujet, celle de Nering (1997), s'intéressait à deux indices, l_z et $ECI4_z$, qui ont été appliqués auparavant avec succès à des données provenant de la passation de tests conventionnels avec le format papier-crayon. Cependant, Nering soulignait tout de même le faible pouvoir de détection de ces indices et la difficulté d'interpréter les valeurs calculées dans le contexte d'un test adaptatif, étant donné que la distribution de probabilité de ces indices s'éloigne considérablement d'une loi de probabilité normale, limitant les inférences possibles sous la forme d'intervalles de confiance classiques. Il faut toutefois noter que ce constat est le même dans le contexte des tests conventionnels, mais qu'il est cependant moins sérieux. McLeod et Lewis (1999) ont également étudié l'efficacité des indices l_z et $ECI4_z$, mais l'originalité de leur travail réside dans l'utilisation d'un indice supplémentaire qui aurait pu s'avérer fort approprié au contexte d'un test adaptatif : il s'agit de l'indice Z_e. Cet indice a la particularité de subdiviser le niveau de difficulté des items proposés en trois catégories : les items faciles, les items de difficulté moyenne et les items difficiles. Toutefois, selon les résultats présentés par les auteurs, cet indice s'est avéré fort peu efficace. Van Krimpen-Stoop et Meijer (1999b), pour leur part, ont utilisé un indice de détection l^*_z qui tient compte du biais dans l'estimation du niveau d'habileté. D'ailleurs, on peut avancer que cet indice est actuellement l'indice le plus recommandé dans le contexte des tests conventionnels. À partir de cet indice, van Krimpen-Stoop et Meijer n'ont cependant pas pu arriver à des résultats plus intéressants que ceux obtenus auparavant par Nering et par McLeod et Lewis dans le cas des tests adaptatifs. Ils ont fait face aux mêmes problèmes d'interprétation et d'efficacité de détection.

Par la suite, van Krimpen-Stoop et Meijer (1999a, 2000), ainsi que Meijer (2002), ont réalisé quelques études similaires avec des indices réservés au contexte des tests adaptatifs. Ces indices sont issus des techniques de contrôle de la qualité utilisées en industrie (Page, 1954 ; Stoumbos et Raynolds, 2001). Toutefois, le problème de l'interprétation des valeurs de ces indices demeure, même si l'efficacité de la détection était meilleure que dans les tentatives précédentes. La dernière recherche recensée, celle réalisée par Meijer (2004), propose l'utilisation d'un indice non paramétrique basé sur le score total à des sous-tests, l'indice $L_{xx'}$. Encore une fois, les résultats obtenus nous amènent à conclure que cet indice est très peu efficace pour détecter des patrons de réponse problématiques dans le contexte de la passation de tests adaptatifs. Intuitivement, il nous apparaît que ce peu d'efficacité est dû à un problème d'interprétation de la distribution de probabilité de l'indice.

Raîche (2002) ainsi que Raîche et Blais (2003) ont proposé une solution au problème d'interprétation de la distribution de probabilité des indices de détection dans le contexte des tests conventionnels : ils ont utilisé la distribution empirique (ou distribution observée) des indices de détection pour en interpréter les valeurs, plutôt qu'une distribution théorique déterminée à l'avance. Cette stratégie a permis d'augmenter considérablement la puissance de détection des indices étudiés dans ces travaux, soit l_z, *Zeta*, *Infit* et *Outfit*. De prime abord, l'application de cette stratégie dans le contexte des tests adaptatifs semble prometteuse. Elle présente toutefois le désavantage d'exiger un temps de calcul considérable. D'autres solutions, où on se préoccupe du contexte d'application des indices, sont aussi à explorer. Par exemple, tenir compte du fait que les comportements qui produisent les patrons de réponses problématiques dans un test adaptatif peuvent être différents de ceux à l'œuvre dans un test conventionnel ; utiliser un estimateur du niveau d'habileté moins biaisé ; augmenter l'étendue du niveau de difficulté des items proposés ou encore utiliser des indices de détection qui conviendraient à des comportements particuliers, comme l'inattention ou les réponses données au hasard.

La suite de ce chapitre propose ainsi une approche quelque peu différente pour caractériser le degré d'adéquation du patron de réponse d'une personne, qu'il soit adaptatif ou non. Les modélisations étudiées permettraient de surmonter certaines limites des approches de détection des patrons de réponses problématiques utilisées jusqu'à maintenant. Ainsi, il sera possible d'obtenir une mesure du degré d'adéquation du patron de réponses qui n'est pas tributaire du biais quant au niveau d'habileté estimé. En fait, la plupart des indices d'adéquation sont calculés à partir d'un estimateur du niveau d'habileté qui est biaisé, et le biais est encore plus marqué lorsqu'une méthode d'estimation bayésienne est utilisée. L'estimation simultanée du degré

d'adéquation et du niveau d'habileté à l'intérieur d'une même modélisation pourrait diminuer l'impact d'un tel biais. De plus, actuellement, les différents indices d'adéquation d'un patron de réponses ne permettent pas de leur associer une erreur type, ni d'estimer leur précision. À partir des modélisations proposées dans ce chapitre, on pourra produire une mesure ou des mesures d'adéquation pour lesquelles il est possible d'obtenir directement un estimateur de la précision.

L'INTÉGRATION DE PARAMÈTRES ASSOCIÉS AUX PERSONNES

De façon générale, les modèles de la théorie des réponses à l'item intègrent des paramètres qui sont associés aux personnes et aux items. Toutefois, lorsque des paramètres sont ajoutés aux modèles plus simples pour décrire la réalité, ils ont le plus souvent comme fonction de modéliser plus adéquatement les caractéristiques des items et non celles des personnes[2]. Un bon exemple de la situation où les items sont à réponse choisie est le modèle de Rasch, qui devient le modèle logistique avec deux paramètres lorsqu'on ajoute un paramètre pour tenir compte du pouvoir de discrimination de l'item, et le modèle avec trois paramètres lorsqu'on ajoute un paramètre pour tenir compte de la possibilité qu'une personne devine la bonne réponse (paramètre de pseudochance). Dans ce dernier cas, même si c'est la personne qui devine, la modélisation considère qu'il s'agit d'une propriété inhérente à l'item, celui-ci favorisant par ses caractéristiques le choix de la réponse au hasard.

Pour illustrer le développement de modèles où on s'intéresse aux caractéristiques « métriques » des personnes plutôt qu'à celles des items, un modèle unidimensionel simple a été retenu, le modèle logistique à un paramètre pour la situation où la réponse est dichotomique, et il sera complexifié par l'ajout de paramètres « personnels » tentant de rendre compte de différents comportements des répondants. Ainsi, la modélisation d'une bonne réponse X_{ij} de la part d'un candidat j à un item i, conditionnelle au paramètre représentant la difficulté de l'item, b_i, et à un premier paramètre personnel du candidat, soit son niveau d'habileté θ_j, prend la forme de l'équation ci-dessous.

$$P_{ij} = P(X_{ij} = 1 | \theta_j, b_i) = \frac{1}{1 + e^{-(\theta_j - b_i)}}$$

2. Les modèles multidimensionnels peuvent aussi être considérés comme des modèles auxquels on ajoute des paramètres, mais ils ne seront pas abordés dans ce texte.

Par analogie avec ce qui se passe lorsque les paramètres ajoutés visent les items, de nouveaux paramètres pour les personnes, celui de discrimination α_j, celui de pseudochance χ_j et celui d'inattention δ_j, sont ajoutés simultanément au niveau d'habileté θ_j,et le modèle obtenu correspondrait à l'équation suivante :

$$P(x_{ij} = 1|\theta_j, \alpha_j, \chi_j, \delta_j, b_i) = \chi_j + \frac{\delta_j - \chi_j}{1 + e^{-\alpha_j(\theta_j - b_i)}}.$$

En nous limitant à un exemple simple comportant peu d'items, nous supposerons deux patrons de réponses hypothétiques obtenus par deux candidats ($j = 1, 2$) à 9 items ($i = 1,...,9$), que nous utiliserons pour comparer les différentes modélisations (une valeur de 1 correspondant à une bonne réponse à l'item i et une valeur de 0 à une mauvaise réponse) :

$$x_{i1} = \{1,1,1,1,1,1,0,0,0\}$$

et

$$x_{i2} = \{1,0,1,1,1,1,0,0,1\}$$

Supposons un ensemble de valeurs associées au paramètre b_j pour ces 9 items : {-2,0,0,0,1,2,3,3,3}. Les valeurs ayant été placées en ordre ascendant, il est aisé de leur faire correspondre les réponses des candidats à chacun des items. Ainsi, on note que le candidat 1 présente un patron de réponses « adéquat » au sens de Guttman. Il a donné de bonnes réponses pour les six premiers items et de mauvaises réponses pour les trois derniers items ; la hiérarchie de la difficulté des items est donc respectée. Le candidat 2, pour sa part, présente un patron de réponses où il obtient une mauvaise réponse aux items 2, 7 et 8, et une réponse correcte à l'item 9. Son patron de réponse est plutôt inusité, même s'il s'agit d'un patron de réponses réaliste.

En appliquant la méthode d'estimation par maximum de vraisemblance *a posteriori* avec le modèle logistique à un paramètre, on obtient un niveau d'habileté θ_1 égal à 1,3 pour le premier candidat. Le second candidat, quant à lui, a obtenu le même nombre de bonnes réponses, et on devrait obtenir avec ce modèle une estimation du niveau d'habileté égale à la valeur obtenue pour le premier candidat. C'est le cas, et on obtient la même valeur 1,3 pour le paramètre θ_2. La figure 1 représente la relation entre le niveau d'habileté estimé et la probabilité d'obtention du patron de réponses.

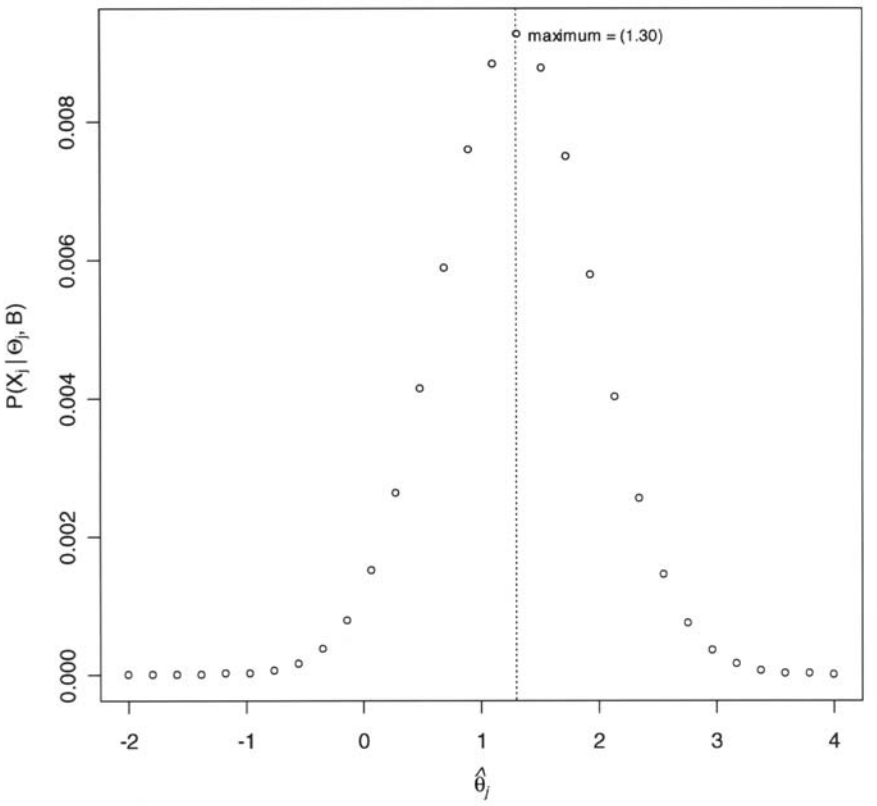


Figure 1 – Relation entre le niveau d'habileté estimé et la probabilité d'obtention du patron de réponses

Nous allons maintenant comparer ces résultats à ceux qu'on obtient avec des modélisations où d'autres paramètres pour les personnes sont ajoutés.

De la même façon qu'on produit ce qu'on appelle la courbe caractéristique d'un item (voir Bertrand et Blais 2004), il est possible de produire la courbe caractéristique d'un candidat. Plusieurs appellations ont été suggérées pour cette représentation graphique : courbe de trace (*trace line*, Weiss, 1973), courbe caractéristique de personne (*person characteristic curve*, Lumsden, 1977, 1978), courbe de réponse de personne (*person response curve*, Trabin et Weiss, 1983). Nous avons retenu la deuxième appellation et nous utiliserons donc dans ce qui suit l'expression « courbe caractéristique de la personne » (CCP).

La figure 2 offre la représentation de trois courbes caractéristiques d'une personne pour trois candidats hypothétiques (exemple adapté de Trabin et Weiss, 1983, p. 91). Supposons qu'on propose un assez grand nombre d'items à ces trois candidats, plus de 200 par exemple, et qu'on regroupe ces items en sept classes de niveaux de difficulté. Il est alors possible de calculer avec quelle fréquence relative chacun des candidats a donné une bonne réponse aux items de chacune des catégories. Pour le premier candidat, on observe une courbe caractérisant une personne qui obtient à coup sûr une bonne réponse lorsque le niveau de difficulté des items est faible, et une mauvaise réponse de façon assurée lorsque les items sont d'un niveau de difficulté élevé. Entre les items de niveau de difficulté faible et ceux de niveau de difficulté élevé, la courbe est graduellement descendante et affiche une forme en *S* typique de ce qui est attendu avec le modèle de Rasch pour des données dichotomiques.

Les deuxième et troisième candidats affichent toutefois des courbes caractéristiques atypiques. Ainsi, ces candidats ne réussissent pas à coup sûr les items les plus faciles, comme s'ils pouvaient être caractérisés par un paramètre personnel d'inattention plus faible que la valeur attendue. La valeur du paramètre personnel δ_j de ces candidats serait alors inférieure à la valeur attendue de 1. Dans le cas du deuxième candidat, on observe que, lorsque des items d'un niveau de difficulté élevé lui sont proposés, il les réussit plus fréquemment que la modélisation le prévoit. Un paramètre personnel de pseudochance χ_2 avec une valeur supérieure à 0 pourrait donc être en jeu. De plus, on observe une remontée inattendue de la fréquence relative de réponses correctes au centre de la courbe. On pourrait alors émettre deux hypothèses, qui impliquent toutes deux l'ajout de paramètres supplémentaires associés à la personne. La première hypothèse serait qu'au moins un autre niveau d'habileté serait impliqué et qu'on aurait plutôt affaire à un vecteur de niveaux d'habileté , $\theta = \{\theta_1 ... \theta_n\}$, c'est-à-dire à une situation multidimensionnelle. La seconde hypothèse serait qu'un paramètre personnel de discrimination α_j serait en jeu. Cette dernière hypothèse peut être facilement mise en rapport avec la première, car une telle modification de la discrimination dans le cas d'un item est d'ailleurs expliquée par la multidimensionalité des niveaux d'habileté (McDonald, 1981).

De plus, le paramètre de discrimination, qu'il soit associé à un item ou à un candidat, correspond à l'inverse de la racine carrée d'une variance. Ce qui veut dire que plus la valeur du paramètre de discrimination est faible, plus la variabilité du niveau de difficulté de l'item ou, comme ici, du niveau d'habileté du sujet, est grande. C'est ce qu'ont souligné Levine et Drasgow (1983) en proposant une nouvelle modélisation de la réponse à l'item qui intègre, en plus du niveau d'habileté, un paramètre qui modélise la variabi-

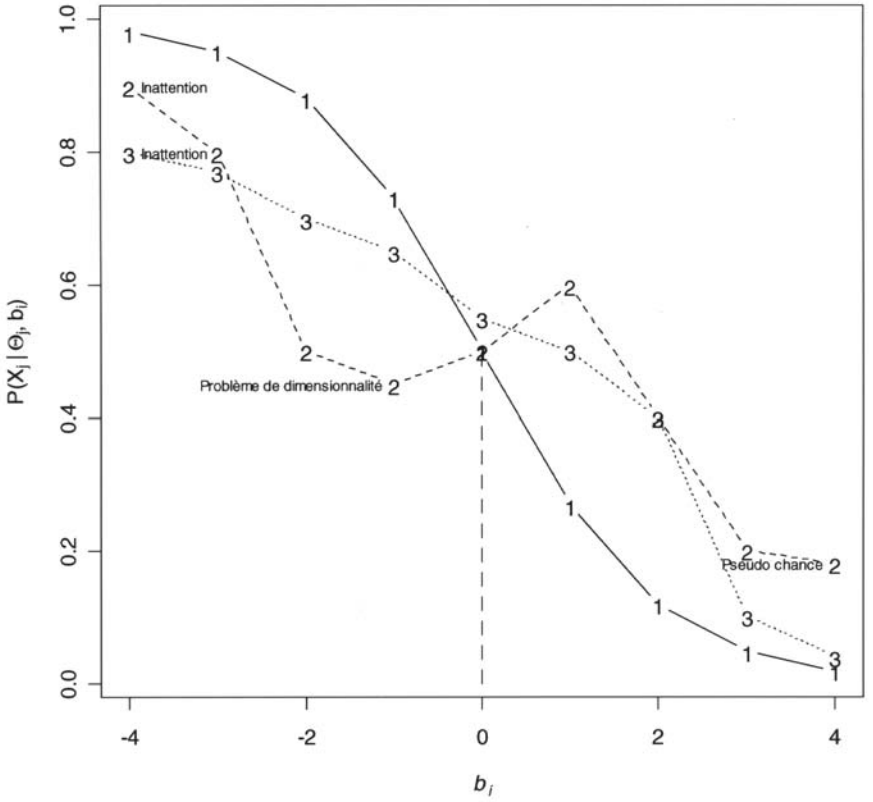


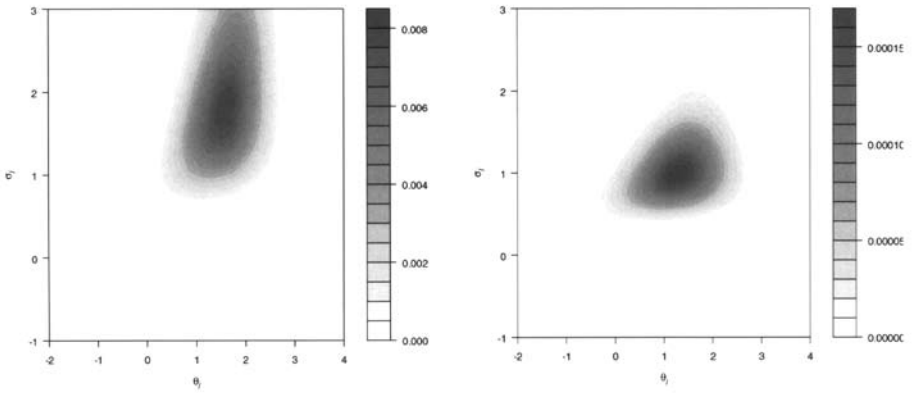
Figure 2 – Probabilité d'une réponse correcte en fonction du niveau de difficulté des items pour trois candidats

lité du niveau d'habileté du sujet à l'intérieur d'un même test. La modélisation repose sur une simple modification de la formulation du modèle logistique à un paramètre :

$$P(x_{ij} = 1|\theta_j, \alpha_j, b_i) = \frac{1}{1 + e^{-\alpha_j(\theta_j - b_i)}},$$

Un paramètre de discrimination associé à la personne, α_j, a ainsi été ajouté au modèle logistique à un paramètre. Peu de travaux ont été réalisés sur cette modélisation jusqu'à récemment. Les recherches encore exploratoires de Ferrando (2004), de Magis (2007) et de Magis et Raîche (2007) ont permis de prendre en considération de nouveau cette modélisation. En poursuivant avec l'exemple ci-dessus, la méthode du maximum de vraisemblance *a posteriori*

permet d'obtenir les valeurs suivantes : θ_1 = 1,58 ; α_1 = 1,78 ; θ_2= 1,30 ; α_2 = 0,98. Ainsi, pour le premier candidat, celui dont le patron de réponses est conforme à une échelle de Guttman, le niveau d'habileté est supérieur à celui obtenu à partir de la modélisation logistique à un paramètre, qui était de 1,30, et la discrimination personnelle devient supérieure à celle, prévue, de 1. Dans le cas du deuxième candidat, le niveau d'habileté est le même, tandis que le paramètre de discrimination personnelle est plus faible que la valeur attendue de 1. Pour cette personne, comme on pouvait s'y attendre, la discrimination est plus faible. Les figures 3a et 3b illustrent la probabilité associée à différentes valeurs des estimateurs des paramètres personnels des deux premiers candidats. On y remarque bien la correspondance avec les valeurs que nous avons obtenues à partir de la méthode d'estimation par vraisemblance maximale *a posteriori*.



a) Candidat 1 **b) Candidat 2**

Figure 3 – Représentation graphique du maximum de vraisemblance a posteriori des paramètres personnels d'habileté θ_j et de discrimination α_j

Les explications accompagnant la dérivation de cette équation par Levine et Drasgow (1983) sont instructives en ce sens qu'elles montrent que celle-ci ne peut pas malheureusement être utilisée pour corriger la valeur du niveau d'habileté malgré la valeur obtenue aux autres paramètres d'items, car on obtient plutôt un estimateur moyen du niveau d'habileté à ce test, et non pas un estimateur fixe du niveau d'habileté du sujet. Ce qu'on peut en revanche obtenir, c'est une information supplémentaire sur la variabilité de ce niveau d'habileté qu'on devrait alors décrire comme une variable aléatoire présentant une variance ($N(\theta_j, \alpha^2_j)$), et non pas simplement comme une valeur fixe du niveau d'habileté. Plus de travaux sont maintenant à réaliser sur l'interprétation des paramètres personnels dans cette modélisation.

Les deux modélisations suivantes prennent en compte un paramètre personnel de pseudochance χ_j et un paramètre personnel d'inattention δ_j (plutôt qu'un paramètre personnel additionnel de discrimination) :

$$P(x_{ij} = 1|\theta_j, \chi_j, b_i) = \chi_j + \frac{1-\chi_j}{1 + e^{-(\theta_j - b_i)}},$$

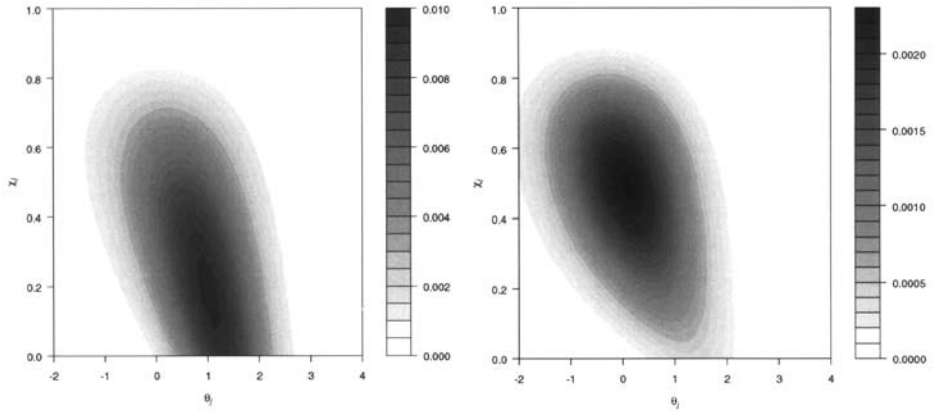
$$P(x_{ij} = 1|\theta_j, \delta_j, b_i) = \frac{\delta_j}{1 + e^{-(\theta_j - b_i)}},$$

Dans le cas de la modélisation intégrant un paramètre personnel de pseudo-chance, les estimations obtenues pour les deux candidats hypothétiques correspondent à : $\theta_1 = 1{,}02$; $\chi_1 = 0{,}18$; $\theta_2 = 0{,}08$; $\chi_2 = 0{,}49$. Dans les deux cas, le niveau d'habileté estimé est plus faible, d'autant plus que le paramètre personnel de pseudochance est élevé. Les figures 4a et 4b illustrent la vraisemblance des différentes valeurs des estimateurs des paramètres personnels des deux candidats. Il faut noter que les fonctions utilisées imposent logiquement une limite inférieure de 0 et une limite supérieure de 1 au paramètre personnel de pseudochance. Il en sera de même pour le paramètre d'inattention plus loin.

Dans ce cas, comme dans tous ceux impliquant les paramètres personnels de pseudochance ou d'inattention, la fonction de maximisation que nous avons utilisée a permis de tenir compte uniquement de la distribution de probabilité *a priori* du niveau d'habileté. Les distributions de probabilité *a priori* (distributions beta) des paramètres personnels de pseudochance ou d'inattention que nous avons expérimentées se sont avérées peu stables et donnaient des résultats imprévisibles.

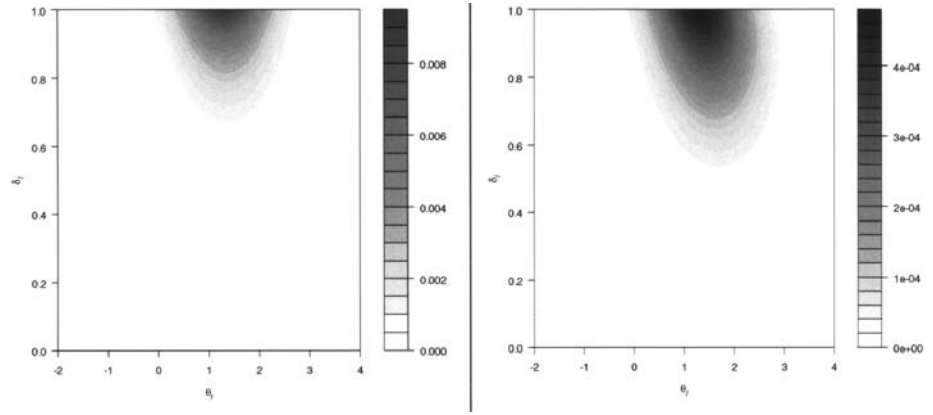
Lors de l'ajout d'un paramètre personnel relié à l'inattention, δ_j, tel que c'est attendu avec le modèle logistique à un paramètre, la valeur du paramètre d'inattention est égale à 1 pour les deux patrons de réponses considérés dans l'exemple. L'estimateur du niveau d'habileté demeure lui aussi inchangé, avec une valeur de 1,30. Les figures 5a et 5b illustrent la variation de la vraisemblance des estimations des paramètres personnels d'inattention.

Comme nous l'avons mentionné, il est aussi possible de penser à des modélisations plus complexes intégrant trois ou quatre paramètres personnels. La représentation graphique des fonctions de vraisemblance associées ne sera pas considérée à partir de maintenant, car elle est difficile à réaliser, quatre dimensions étant désormais impliquées, cinq dans le cas de du modèle qui intègre quatre paramètres associés aux personnes.



a) Sujet 1 **b) Sujet 2**

Figure 4 – Représentation graphique du maximum de vraisemblance a posteriori des paramètres personnels d'habileté θ_j et de pseudo chance χ_j



a) Sujet 1 **b) Sujet 2**

Figure 5 – Représentation graphique du maximum de vraisemblance a posteriori des paramètres personnels d'habileté θ_j et d'inattention δ_j

Avec un modèle intégrant des paramètres personnels de discrimination et d'inattention (voir ci-dessous), les valeurs obtenues sont les suivantes : $\theta_1 = 1{,}58$; $\alpha_1 = 1{,}78$; $\delta_1 = 1{,}00$; $\theta_2 = 1{,}30$; $\alpha_2 = 0{,}98$; $\delta_2 = 1{,}00$. Puisque le paramètre personnel d'inattention n'est pas affecté, nous retrouvons pour ces deux patrons de réponse précis le même résultat que celui obtenu à partir de la modélisation intégrant uniquement le paramètre personnel supplémentaire de discrimination.

$$P(x_{ij} = 1|\theta_j, \alpha_j, \delta_j, b_i) = \frac{\delta_j}{1 + e^{-\alpha_j(\theta_j - b_i)}},$$

En ce qui a trait à la modélisation intégrant simultanément les paramètres personnels de discrimination et de pseudochance (équation ci-dessous), on retrouve encore les mêmes paramètres personnels pour le premier candidat, puisque le paramètre personnel de pseudochance est nul. Toutefois, pour le deuxième candidat, on remarque une modification importante des paramètres personnels de discrimination et de pseudo-chance. Les valeurs des paramètres estimés sont les suivantes : $\theta_1 = 1{,}58$; $\alpha_1 = 1{,}78$; $\chi_1 = 0{,}00$; $\theta_2 = 0{,}00$; $\alpha_2 = 1{,}39$; $\chi_2 = 0{,}52$.

$$P(x_{ij} = 1|\theta_j, \alpha_j, \chi_j, b_i) = \chi_j + \frac{1 - \chi_j}{1 + e^{-\alpha_j(\theta_j - b_i)}},$$

Enfin, le dernier modèle intègre tous les paramètres personnels que nous avons considérés jusqu'à maintenant. Dans ce cas précis, puisque la valeur du paramètre personnel d'inattention demeure inchangée et est égale à 1, nous retrouvons exactement les mêmes valeurs pour les paramètres personnels que celles qui avaient été obtenues avec la modélisation précédente. Il est évidemment possible que d'autres patrons de réponses produisent des résultats différents.

$$P(x_{ij} = 1|\theta_j, \alpha_j, \chi_j, \delta_j, b_i) = \chi_j + \frac{\delta_j - \chi_j}{1 + e^{-(\theta_j - b_i)}},$$

CONCLUSION ET IMPLICATIONS POUR LES TESTS ADAPTATIFS

Nous avons tenté de proposer une approche assez différente pour caractériser le degré d'adéquation du patron de réponse d'un sujet à un test, qu'il soit adaptatif ou non. À cette fin, l'ajout de paramètres personnels à la modélisation logistique à un paramètre a été considéré. Ainsi, trois paramètres personnels ont été pris en compte, soit les paramètres personnels de discrimination, de pseudochance et d'inattention.

La modélisation où uniquement le paramètre personnel de discrimination est ajouté semble la plus intéressante pour la détection de patrons de réponses problématiques à l'intérieur de tests adaptatifs. L'estimation de ce paramètre personnel additionnel semble assez facile à obtenir avec une méthode bayésienne comme celle de maximisation *a posteriori*. Toutefois, comme le souligne Ferrando (2004), pour obtenir un estimateur précis du niveau d'habileté, le double du nombre d'items est requis. Cela peut entraîner le besoin de plus de 80 items. Cependant, dans un test adaptatif, on pourrait utiliser le paramètre personnel de discrimination pour optimiser le choix des prochains items à proposer et, éventuellement, obtenir un estimateur moins biaisé du niveau d'habileté. Malgré cette opportunité, il est possible que l'estimateur du niveau d'habileté n'ait plus la même signification que celui obtenu à partir de la la modélisation logistique usuelle à un paramètre. Cette remarque est valide pour toutes les modélisations alternatives que nous avons proposées ici. Il ne semble donc pas possible, malheureusement, de corriger l'estimateur du niveau d'habileté du sujet malgré le fait que son patron de réponses soit problématique.

Toutes les modélisations alternatives où des paramètres personnels additionnels de pseudochance ou d'inattention sont ajoutés, quoique potentiellement intéressantes d'un point de vue théorique, ne semblent pas très utiles pour le moment. Premièrement, les estimateurs exigent que des adaptations soient effectuées au regard des méthodes d'estimation. Actuellement, les estimateurs obtenus ne semblent pas vraiment stables. De plus, comme pour l'estimation des paramètres lorsqu'ils sont associés aux items, l'opération exige la présence de plusieurs items très faciles ou très difficiles pour que les valeurs obtenues aient un sens. Il est probable que, pour obtenir des estimateurs d'une précision satisfaisante, il soit nécessaire que le patron de réponses du sujet soit constitué de plus d'items encore que dans le cas du paramètre personnel de discrimination.

Une des applications les plus intéressantes de toutes ces modélisations sera sans contredit leur utilisation à l'intérieur de simulations de tests classiques fixes et invariables et à l'intérieur de tests adaptatifs. Il sera ainsi très facile de contrôler le degré d'inadéquation de patrons de réponses simulées

en faisant varier les paramètres personnels. Dans le cas des tests adaptatifs, cela s'avère encore plus utile, car il a toujours été difficile de simuler adéquatement des patrons de réponses inadéquats.

Une dernière application de ces travaux pourra consister à utiliser ces modélisations, surtout celle où un paramètre personnel additionnel de discrimination est ajouté, pour effectuer la calibration simultanée des paramètres d'items et des paramètres personnels. Il ne serait éventuellement plus nécessaire d'identifier dans un premier temps les sujets dont le patron de réponses est problématique ni de les retirer ensuite du processus de calibration. On pourrait ainsi tenir compte directement de leur degré d'inadéquation.

Références

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

Drasgow, F., Levine, M.V. et Zickar, M.J. (1996), « Optimal identification of mismeasured individuals », *Applied Psychological Measurement, 9*(1), p. 47-64.

Ferrando, P. J. (2004), « Person reliability in personality measurement : an item response theory analysis », *Applied Psychological Measurement, 28*(2), p. 126-140.

Levine, M. V. et Drasgow, F. (1983), « Appropriateness measurement : validating studies and variable ability models », dans D. J. Weiss (Dir.), *New horizons in testing. Latent trait test theory and computerized adaptive testing*, New York, Academic Press.

Lumsden, J. (1977), « Person reliability », *Applied Psychological Measurement, 1*, p. 477-482.

Lumsden, J. (1978), « Tests are perfectly reliable », *British Journal of Mathematical and Statistical Psychology, 31*, p. 19-26.

Magis, D. (2007), *Influence, information and item response theory in discrete data analysis,* Thèse de doctorat inédite, Université de Liège, Liège, Belgique.

Magis, D. et Raîche, G. (2007), *Two new iterative Bayesian estimators of subject ability and their properties,* Rapport nº 07.001, Département de mathématiques, Université de Liège, Liège, Belgique.

McDonald, R. P. (1981), « The dimensionality of tests and items », *British Journal of Mathematical and Statistical Psychology, 34,* p. 100-117.

McLeod, L.D. et Lewis, C. (1999), « Detecting item memorization in the CAT environment », *Applied Psychological Measurement, 23*(2), p. 147-160.

Meijer, R.B. (2002), « Outlier detection in high-stakes certification testing », *Journal of Educational Measurement, 39*(3), p. 219-233.

Meijer, R.B. (2004), « Using patterns of summed scores in paper-and-pencil tests and computer-adaptive tests to detect misfitting item score patterns », *Journal of Educational Measurement, 41*(2), p. 119-136.

Molenaar, I.W. et Hoijtink, H. (2000), « The many null distributions of person fit indices », *Psychometrika, 55*(1), p. 75-106.

Nering, M.L. (1997), « The distribution of indexes of person fit within the computerized adaptive testing environment », *Applied Psychological Measurement, 21*(2), p. 115-127.

Page, E.S. (1954), « Continuous inspection schemes », *Biometrika, 41*, p. 100-115.

Raîche, G. (2002), *Le dépistage du sous-classement aux tests de classement en anglais, langue seconde, au collégial*, Hull, Collège de l'Outaouais.

Raîche, G. (2004), « Le testing adaptatif », dans R. Bertrand et J.-G. Blais, *Modèles de mesure : l'apport de la théorie des réponses aux items*. Sainte-Foy, Presses de l'Université du Québec.

Raîche, G. et Blais, J.-G. (2003), *The distribution of person-fit indices conditional on the estimated proficiency level and the detection of underachievement at a placement test*. Communication réalisée à la rencontre annuelle de la Psychometric Society à Cagliari, Sardaigne.

Rasch, G. (1960), *Probabilistic models for some intelligence and attainment tests*, Chicago, University of Chicago Press.

Reese, S.P. et Flannery, W.P. (1996), « Assessing person-fit on measures of typical performance », *Applied Psychological Measurement, 9*(1), p. 9-26.

Stoumbos, Z.G. et Raynolds, M.R. (2001), « The state of statistical process control as we proceed into the 21th Century », *Journal of he American Statistical Association, 95*(451), p. 992-994.

Thissen, D. et Mislevy, R.J. (2000), « Testing algorithms », dans H. Wainer (Éd.) : *Computerized adaptive testing : A primer*, Mahwah, NJ, Lawrence Erlbaum Associates.

Trabin, T. E. et Weiss, D. J. (1983), « The person response curve : fit of individual to item response theory models », dans D. J. Weiss (Dir.), *New horizons in testing. Latent trait test theory and computerized adaptive testing*, New York, Academic Press.

van Krimpen-Stoop, E.M.L.A. et Meijer, R.R. (1999a), *CUSUM-based person-fit statistics for adaptive testing* (Rapport de recherche 99-05), Enschede, NETHERLANDS, University of Twente.

van Krimpen-Stoop, E. M. L. A. et Meijer, R. R. (1999b), « The null distribution of person-fit statistics for conventional and adaptive tests », *Applied Psychological Measurement, 23*(4), p. 327-345.

van Krimpen-Stoop, E.M.L.A. et Meijer, R.R. (2000), « Detecting person misfit in adaptive testing using statistical process control techniques », dans W.J. van der Linden et C.A.W. Glas (Éd.), *Computerized adaptive testing : theory and practice*, Dordrecht, Kluwer.

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

Weiss, D. J. (1973), *The stratified adaptive computerized ability test* (Rapport de recherche nº 73-3), Minneapolis, MN, University of Minnesota.

Page laissée blanche intentionnellement

CHAPITRE 15

La méthode alpha stratifiée pour le contrôle de la fréquence d'exposition des items dans les tests adaptatifs

Éric Dionne et Jean-Guy Blais
Université de Montréal

INTRODUCTION

La possibilité d'utiliser des ordinateurs dans un milieu éducatif offre différentes perspectives pour l'évaluation des apprentissages et des compétences. Ainsi, dans une perspective adaptative et diagnostique, il est possible d'offrir des séquences de tâches qui ne sont pas identiques pour tous, des tests adaptatifs plus personnalisées, en fonction du niveau de compétence de chacun des candidats (voir le chapitre 3). Il est également possible pour les candidats en formation de faire la démonstration de leurs compétences au moment qui leur convient. Le rythme d'apprentissage des personnes peut donc être mieux respecté.

Bunderson, Inouye et Olsen (1993, p. 381) définissent les tests adaptatifs ainsi :

> [...] des tests administrés par ordinateur et pour lesquels la décision d'administrer ou non un nouvelle item – ou une nouvelle tâche – est adaptative. Une tâche peut être un item ou une situation standardisée plus complexe impliquant une ou plusieurs réponses. Le caractère adaptatif signifie que la présentation d'une nouvelle tâche dépend de la performance du candidat aux tâches qui lui sont jusqu'alors administrées.

La plupart des tests adaptatifs s'appuient généralement sur l'un ou l'autre des modèles de mesure issus de la théorie des réponses aux items (TRI)[1]. Comme l'indiquent Embretson et Hershberger (1999), la TRI permet de calculer la probabilité de réussir un item en fonction du niveau d'habileté effectif du candidat et des caractéristiques métriques (par exemple la difficulté) de l'item. Les modèles de la TRI permettent de positionner l'habileté estimée du candidat et la difficulté de l'item sur une seule et même échelle. Dans le cadre de tests adaptatifs, il faut estimer le niveau d'habileté du candidat après que ce dernier a répondu à un item donné. Au fur et à mesure du cheminement du candidat selon les réponses aux items, le système adaptatif prend une décision (poursuivre ou non le test) selon des règles préétablies. En ce sens, la TRI est présentement la famille de modèles de mesure privilégiée par plusieurs pour les applications, et elle semble offrir des perspectives intéressantes pour la mise en œuvre et l'étude de l'efficacité des tests adaptatifs.

Plusieurs initiatives d'implantation de tests adaptatifs ont déjà eu lieu avec des résultats fort prometteurs (Kingsbury et Houser, 1999 ; Laurier, 1999 ; Trentin, 1997). Les tests adaptatifs offrent des avantages indéniables par rapport aux tests plus traditionnels effectués sous format papier-crayon. Parmi ces avantages, on note que les TA sont généralement moins longs, mais qu'ils améliorent la précision de la mesure ; ils présentent des items qui offrent un défi bien dosé au candidat, la correction est facilitée, et les items permettent de mettre à profit les capacités multimédias qu'offre l'informatique. Cependant, il existe encore aujourd'hui des aspects problématiques, en particulier, la sécurité associée aux items et aux banques d'items (Meijer et Nering, 1999 ; Way, 1998). Par exemple, dans certaines situations où les enjeux sont critiques, la divulgation du contenu des items peut être facilitée par le fait que, dans une situation adaptative, ils reviennent plus souvent que d'autres et ont donc une fréquence d'exposition plus élevée.

Dans le cadre de ce chapitre, nous présenterons une stratégie développée initialement par Chang et Ying (1999) visant à assurer un meilleur contrôle de la fréquence d'exposition des items dans un test adaptatif : la méthode alpha-stratifiée. Nous discuterons également de plusieurs variantes associées à cette méthode en montrant les forces et les limites de chacune d'elles.

1. Le lecteur intéressé peut consulter le chapitre 10 de ce volume pour prendre connaissance d'une autre perspective sur les algorithmes d'adaptabilité.

PROBLÉMATIQUE

Le contrôle de l'exposition des items : définitions et contexte

Mills et Stocking (1996, p. 294) définissent le contrôle de l'exposition des items par la fréquence à laquelle un item est présenté à un ou plusieurs candidats. Elle tient compte de la proportion des candidats qui voient les items ainsi que de la période de temps durant laquelle ces items sont disponibles pour les candidats. Les méthodes de contrôle de l'exposition des items agissent, comme l'illustre la figure 1, comme un filtre afin d'éviter de toujours proposer ou de toujours négliger les mêmes items de la banque. Ainsi, une interface usager recueille et présente les informations liées au candidat ou aux items. Un modèle de mesure et une méthode générale de sélection des items (méthode bayésienne, méthode du maximum d'information, etc.), permettent d'estimer alors le niveau d'habileté du candidat en fonction de la réponse qu'il vient de donner. En fonction de cette estimation, on détermine un nouvel item à donner dans la banque d'items en tenant compte des contraintes issues de la méthode de contrôle d'exposition (on peut aussi mettre un terme au test). Lorsque l'item est choisi, il est alors proposé au candidat, et le cycle se poursuit jusqu'au moment de la règle d'arrêt du test.

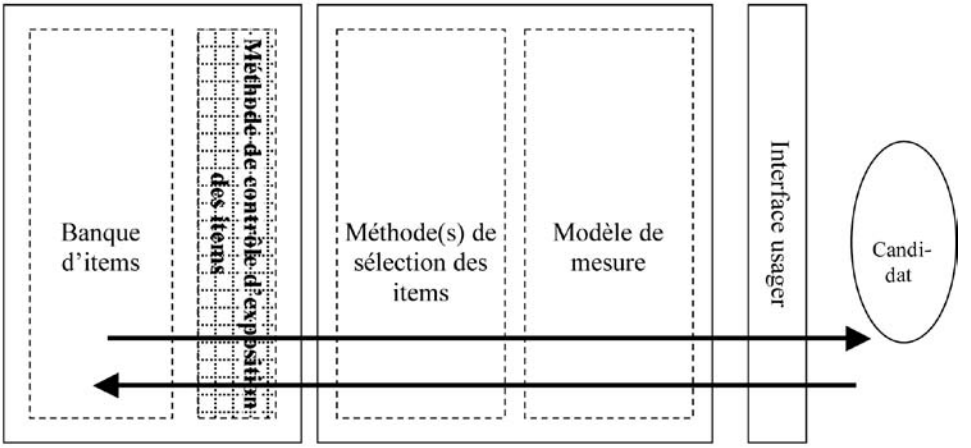


Figure 1 – Sous-systèmes d'un test adaptatif

Plusieurs recherches (Chang et Ying, 1999 ; Chen et Lei, 2005 ; Eggen, 2001 ; Hetter et Sympson, 1997 ; Revuelta et Ponsoda, 1998 ; Van der Linden, 2002) montrent que certains items donnés dans le cadre de TA – qui ne tiennent pas compte de l'exposition des items – sont exagérément proposés

ou, au contraire, ne sont pratiquement jamais exploités. Dans le cas des items surexposés, cette situation peut entraîner de sérieux problèmes, par exemple le vol des items les plus souvent présentés (Davey et Nering, 2002 ; Wainer et Eignor, 2000), et par conséquent être à la source d'un problème de validité. Les items sous-exposés entraînent une perte nette, puisque le temps et les ressources matérielles et humaines consentis à la rédaction de ces items ne sont aucunement ou faiblement récupérés. Peu importe la situation, puisque la rédaction et la validation d'items sont des opérations coûteuses et laborieuses, il y a tout intérêt à préserver l'intégrité de ces items – et plus globalement des banques d'items – le plus longtemps possible.

Dans le cadre de ce texte, nous discuterons principalement de l'efficacité de la méthode alpha-stratifiée qui est étudiée depuis plusieurs années, et qui se révèle intéressante pour mieux cerner les difficultés et les conséquences liées aux problèmes de l'exposition des items. Pour les fins de notre analyse, nous discuterons des avantages et des limites de cette méthode en comparaison avec la méthode de Sympson et Hetter (SH) dans un contexte de sélection basé sur le maximum d'informations. Cette dernière méthode représente une base de comparaison intéressante puisqu'une riche littérature lui est associée.

Les méthodes de contrôle de l'exposition des items

Il existe une vaste collection de méthodes permettant de contrôler l'exposition des items. Dans le cadre d'une recension présentant les principales méthodes de contrôle de l'exposition des items développées et étudiées au cours des vingt dernières années, Georgiadou, Triantafillou et Economides (2007) en ont comptabilisé plus d'une trentaine. Il existe quelques essais de ces classifications mais, dans le cadre de ce texte, nous utiliserons une classification en quatre catégories :

- les méthodes aléatoires ;
- les méthodes conditionnelles ;
- les méthodes alpha-stratifiées ;
- les autres méthodes.

Comme le nom l'indique, les méthodes aléatoires se basent sur le hasard afin de sélectionner l'item – fournissant normalement une information suffisante – qu'il est possible de proposer à un candidat. La méthode de McBride et Martin (1983) et celle de Kingsbury et Zara (1989) appartiennent à cette catégorie de méthodes. Les méthodes conditionnelles se basent

sur une toute autre stratégie. Il s'agit plutôt d'assigner un paramètre de contrôle d'exposition pour chacun des items afin d'en déterminer le taux de présentation. L'une des plus connues est celle de Sympson et Hetter (1985). Les différentes méthodes issues de l'approche dite alpha-stratifiées sont regroupées dans la troisième catégorie. Nous présenterons en détail cette méthode dans la section qui suit et qui lui est entièrement dédiée. Les autres méthodes composent la dernière catégorie. On peut retrouver dans cette catégorie les méthodes qui mettent l'accent sur d'autres contraintes que celles mentionnées jusqu'ici, telles la représentativité des contenus mesurés (Van der Linden et Chang, 2005) ou les approches combinant plusieurs stratégies (C.-K. Leung, Chang et Hau, 2003 ; Yi, 2002).

La méthode alpha-stratifiée (ASTR)

Hambleton et Swaminathan (1985, p. 104) ont montré que l'information était maximale lorsque l'item proposé au candidat possédait un indice de discrimination élevé et un indice de difficulté se rapprochant de la valeur de l'estimation du niveau d'habileté du candidat. Dans le contexte d'un test adaptatif, il peut donc arriver que ce soit les items les plus discriminants qui soient les plus souvent présentés, même quand cela n'est pas nécessaire, particulièrement au début du test, alors que l'estimation du niveau d'habileté du candidat peut diverger. La méthode ASTR, introduite par Chang et Ying (1999), vise à limiter cet écueil en divisant la banque d'items en n strates regroupant chacune un certain nombre d'items ayant tous un indice de discrimination relativement comparable. La figure 2 qui suit en fournit une représentation. La strate 1 regroupe les items les moins discriminants de la banque d'items. Les premiers items qui seront proposés aux candidats seront sélectionnés dans cette première strate. La strate 2, quant à elle, contient des items un peu plus discriminants que ceux contenus dans la première strate, et ainsi de suite pour l'ensemble des strates. Au fur et à mesure que le test progresse et que l'estimation de l'habileté du candidat se précise, on passe d'une strate à l'autre afin de choisir les items qui sont de plus en plus discriminants. Ce faisant, on s'assure de donner au début du test des items faiblement discriminants afin de conserver les items fortement discriminants pour la fin du test, donc au moment où ils sont normalement les plus utiles.

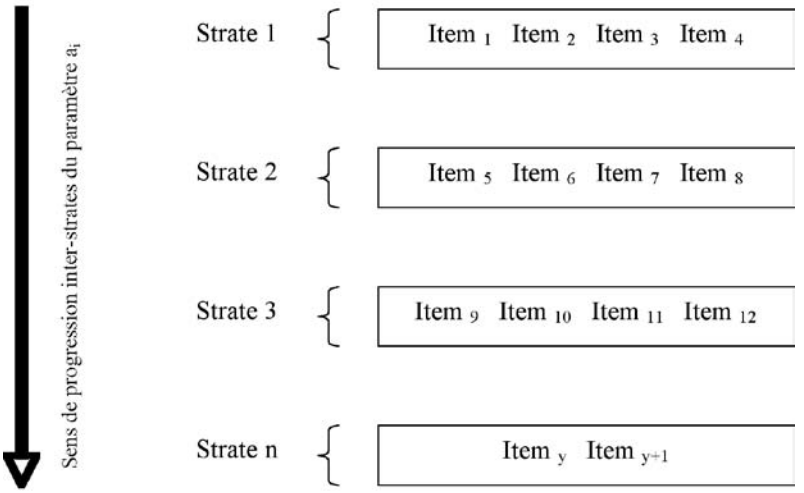


Figure 2 – Agencement des items lors de l'utilisation de la méthode ASTR

La méthode ASTR se distingue par le fait qu'elle agit d'abord sur l'architecture de la banque d'items tout en permettant d'agir sur l'algorithme de sélection d'items en y amalgamant, par exemple, une méthode aléatoire ou conditionnelle. En ce sens, les méthodes ASTR permettent d'agir à deux niveaux (sur la structure de la banque d'items et sur l'algorithme de sélection), ce qui les rend particulièrement intéressantes et efficaces.

LES STATISTIQUES POUR LA COMPARAISON DES MÉTHODES

Pour comparer l'efficacité de chacune des méthodes, nous nous sommes basés sur les statistiques qui sont les plus couramment utilisées dans la littérature : les biais, le carré moyen de l'erreur, la statistique du khi-carré, le taux de chevauchement des items, le nombre d'items sous utilisés et le nombre d'items trop utilisés. Ces statistiques permettent d'obtenir des informations sur l'efficacité de la mesure obtenue au moyen des outils d'évaluation ainsi que sur la performance de chacune des méthodes au regard de l'exposition des items.

Le biais et le carré moyen de l'erreur (CME)

Le biais et le carré moyen de l'erreur sont deux statistiques couramment utilisées pour mesurer la qualité de la mesure. Ces statistiques se calculent au moyen des équations suivantes :

$$Biais = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} (\hat{\theta}_i - \theta_i) \qquad CEM = \frac{1}{\mathrm{N}} \sum_{i=1}^{N} (\hat{\theta}_i - \theta_i)^2,$$

où $\hat{\theta}_i$ est l'estimateur de l'habileté du candidat i, et θ_i son habileté réelle.

Une méthode de contrôle qui est efficace se reconnaît quand la valeur associée soit au biais ou au carré moyen de l'erreur est faible, puisqu'alors l'estimateur du niveau d'habileté converge vers le niveau d'habileté réelle du candidat.

La statistique du khi-carré (χ^2)

Selon Deng et Chang (2001), cette statistique permet de mesurer la différence entre le taux d'exposition observé et le taux d'exposition idéal. Cette statistique permet d'obtenir de l'information sur l'utilisation de la banque d'items :

$$\chi^2 = \sum_{i=1}^{N} \frac{(er_i - L/N)^2}{L/N},$$

où er_i est le taux d'exposition observé, L la longueur du test, et N le nombre d'items dans la banque d'items.

Une valeur faible pour cette statistique signifie que l'utilisation de la banque d'items est adéquate, et qu'ainsi la méthode de sélection des items est efficace.

Le nombre d'items sur-utilisés ou sous-utilisés (TEI)

Leung, Chang et Hau (2002) définissent ces statistiques par le ratio de la fréquence de présentation d'un item aux candidats sur le nombre total de candidats à qui on fait passer le test. On considère qu'une valeur de 0,2 ou plus est un signe de sur-utilisation de l'item, alors qu'une valeur inférieure à 0,02 est le signe d'une sous-utilisation.

Le taux de chevauchement des items (TCI)

Deng et Chang (2001, p. 7) définissent le taux de chevauchement des items comme la proportion attendue d'items partagés par deux candidats tirés au hasard. Le calcul de cette statistique s'effectue en trois étapes : 1) compter le nombre d'items communs pour chacune des m(m-1)/2 paires de candidats ; 2) faire la somme de chaque valeur m(m-1)/2 obtenue ; et 3) diviser la somme par Lm(m-1)/2. Un contrôle satisfaisant des items minimise la valeur de cette statistique.

LA COMPARAISON DES RÉSULTATS DE LA RECHERCHE

Les études retenues

Pour les fins de notre analyse, nous avons aussi pris soin de croiser les données qui nous apparaissaient les plus comparables. Les résultats issus de dix simulations ont été retenus pour notre analyse, car les paramètres qui suivent étaient relativement semblables d'une simulation à l'autre :

- nature de la banque d'items (réelle ou simulée) ;
- nombre de strates ;
- caractéristiques des items (indice de difficulté (b_i), indice de discrimination (a_i), indice de pseudochance (c_i)) ;
- nombre maximum d'items employés dans le TA ;
- taille de la banque d'items.

Le tableau 1 présente une synthèse de chacune des banques d'items au regard des paramètres que nous venons de décrire. Les résultats proviennent de recherches publiées entre 1999 et 2004 et offrent une base adéquate pour comparer les résultats. Notre analyse repose sur les travaux de Chang et Ying (1999), Leung, Chang et Hau (2002), Chang et van der Linden (2003), Davis (2004) et Deng et Chang (2001).

Tableau 1 – Caractéristiques des banques d'items et des items comparés

Simulations	Auteurs	Nombre d'items du TA	Information sur la banque d'items : Nature réelle (R) ou simulée (S) Nombre d'items de la banque Nombre d'items par strate Valeur des indices a_i, b_i, et c_i[1]
1	Chang, Ying (1999)	20	R 254 (65-63-63-63)
2		40	S 400 (100-100-100-100) a_i : (0,5-1,0-1,5-2,0) b_i : N(0,1) c_i : 0
3		60	
4	Leung, Chang, Hau (2002)	24	R 254 (63-63-64-64)
5	Chang, van der Linden (2003)	30	R 360 (90-90-90-90)
6	Davis (2004)	20	R 157 (5 strates)
7	Deng, Chang (2001)	32	S 400 (100-100-100-100) a_i : (0,5-1,0-1,5-2,0) b_i : N(0,1) c_i : 0
8		40	
9	Leung, Chang, Hau (2002)	40	
10		60	

1. Certaines informations manquantes s'expliquent par le fait qu'elles n'étaient pas disponibles dans les articles ou rapports de recherche consultés.

Idéalement, nous aurions aimé comparer l'efficacité des méthodes de contrôle de l'exposition des items au moyen de simulations qui agissent sur une seule variable à la fois afin d'identifier des relations de case à effet. Puisque les recherches actuelles sont plutôt à caractère exploratoire, il n'a pas été possible de trouver des données rigoureusement comparables. Pour toutes ces raisons, nous avons opté pour la présentation des grandes tendances qui semblent se dégager des analyses que nous avons réalisées.

Résultats et discussion

Le tableau 2, qui suit, présente les résultats concernant le biais et le carré moyen de l'erreur pour les 10 simulations à l'étude. La figure 3, quant à elle, présente la distribution des valeurs associées au biais et au carré moyen de

l'erreur pour les 10 simulations comparées. On peut y observer que les valeurs associées au biais sont pratiquement identiques pour la méthode ASTR et pour la méthode MI+SH. On peut arriver à des conclusions semblables en analysant les valeurs liées au carré moyen de l'erreur. On constate que les deux méthodes de contrôle de l'exposition permettent d'obtenir une qualité de mesure tout à fait comparable.

Tableau 2 – Valeurs associées au biais et au carré moyen de l'erreur pour 10 simulations de tests adaptatifs

Simulations	Biais		Carré moyen de l'erreur	
	ASTR	MI+SH	ASTR	MI+SH
1	0,007	0,007	0,0840	0,0600
2	-0,004	-0,006	0,0220	0,0170
3	-0,001	-0,001	0,0140	0,0150
4	0,008	0,008	0,0672	0,0614
5	0,009	-0,008	0,1172	0,1245
6	-0,030	-0,030	0,1089	0,0900
7	0,003	0,004	0,0289	0,0199
8	-0,004	-0,006	0,0220	0,0170
9	-0,002	-0,004	0,0227	0,0193
10	-0,001	-0,001	0,0140	0,0150

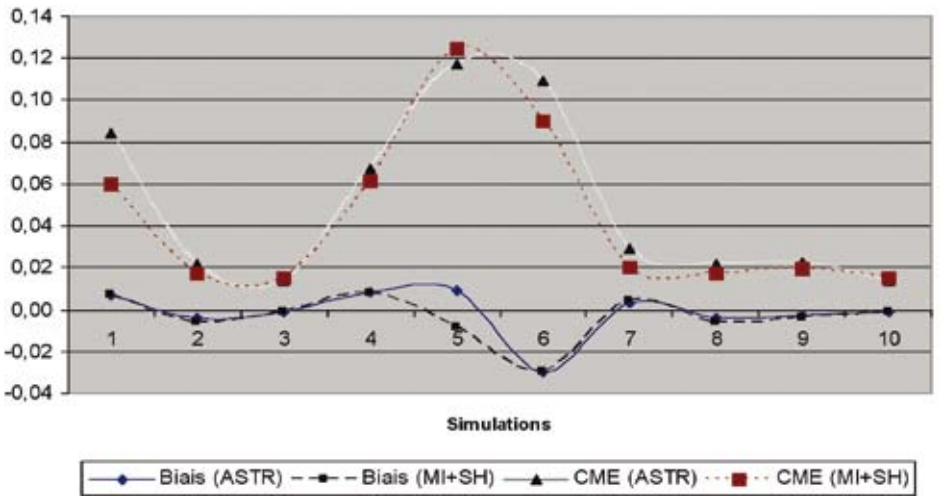


Figure 3 – Distribution des valeurs associées au biais et au carré moyen de l'erreur (CME)

La figure 4 illustre la distribution de la statistique associée au taux de chevauchement des items (TCI). À la lumière de cette figure, on peut constater que la méthode ASTR permet de diminuer le taux de chevauchement d'environ 50 % lorsqu'on la compare avec la méthode MI+SH. Ce résultat est particulièrement intéressant, puisqu'il indique que l'utilisation de la banque d'items est généralement plus uniforme quand on fait appel à la méthode ASTR que lorsqu'on fait appel à la méthode MI+SH. On constate cependant que les résultats associés aux simulations 5 et 6 ont été volontairement omis. En effet, les résultats obtenus par Chang et van der Linden (2003) ainsi que par Davis (2004) s'éloignent de ceux que nous venons de présenter. Dans le cas des premiers auteurs, ils obtiennent, avec la méthode ASTR, un TCI supérieur (4,76) à celui qui est obtenu avec la méthode MI+SH (2,98). Il en est de même avec les résultats de Davis (ASTR : 7,90, MI+SH : 6,44). Puisque celle-ci a utilisé des données dites polytomiques, une échelle de type Likert par exemple, pour réaliser ses simulations, on peut s'interroger sur la comparabilité des résultats, au regard de cette statistique, entre des simulations utilisant des items dichotomiques et d'autres utilisant des items polytomiques. Des recherches plus ciblées sur cette problématique seraient bénéfiques pour nous éclairer sur ce point. Ces résultats nous rappellent le besoin de plus en plus senti de réaliser des simulations dans un contexte plus standardisé afin de mieux cibler l'influence de chacune des variables.

On peut aussi remarquer que les gains réalisés au niveau du TCI pour la méthode ASTR semblent s'estomper au fur et à mesure de l'augmentation du nombre d'items administrés. Les simulations (1, 2 et 3) réalisées par Chang et Ying (1999) sont particulièrement intéressantes sur ce point. Ainsi, les valeurs associées au TCI augmentent (0,10 ; 0,11 ; 0,17) au fur et à mesure que le nombre d'items augmente (20, 40 et 60). Les résultats de Leung, Chang et Hau (2002) permettent de tirer des conclusions semblables quant à la valeur de la TCI (0,12 ; 0,17) lorsque le nombre d'items s'accroît de 40 à 60.

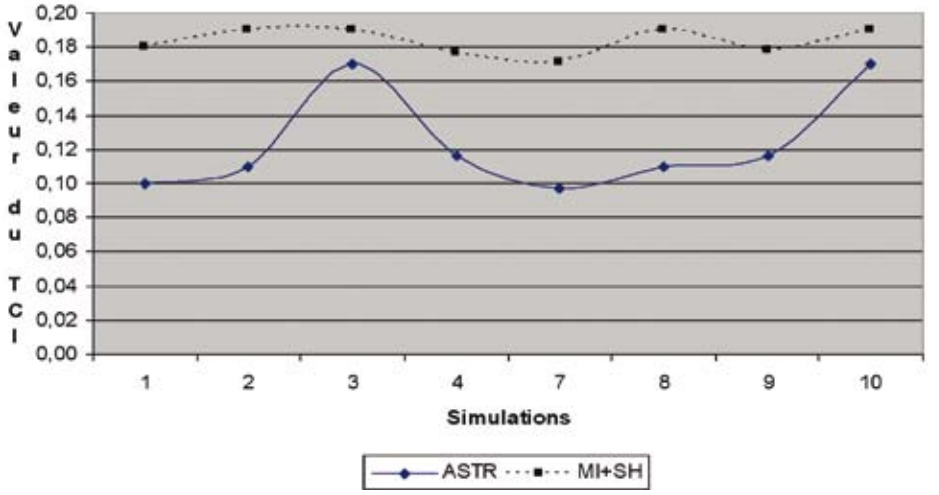


Figure 4 – Distribution des valeurs associées au taux de chevauchement des items (TCI) pour les 10 simulations sauf les simulations 5 et 6

La statistique du khi-carré montre aussi que la méthode ASTR est avantageusement comparable à la méthode MI+SH dans pratiquement toutes les simulations étudiées. La figure 5 illustre d'ailleurs ce constat. Cette statistique prend une valeur environ cinq fois moins grande dans le cas de la méthode ASTR que dans la méthode MI+SH. Encore une fois, il semble se dégager que l'écart s'amenuise au fur et à mesure de l'augmentation du nombre d'items présentés. De plus, la simulation (5) réalisée par Chang et Van Der Linden (2003) donne des résultats diamétralement opposés (MI+SH : 30,6 et ASTR : 6,85), ce qui nous laisse penser qu'il y aurait tout lieu d'approfondir les recherches sur ces résultats.

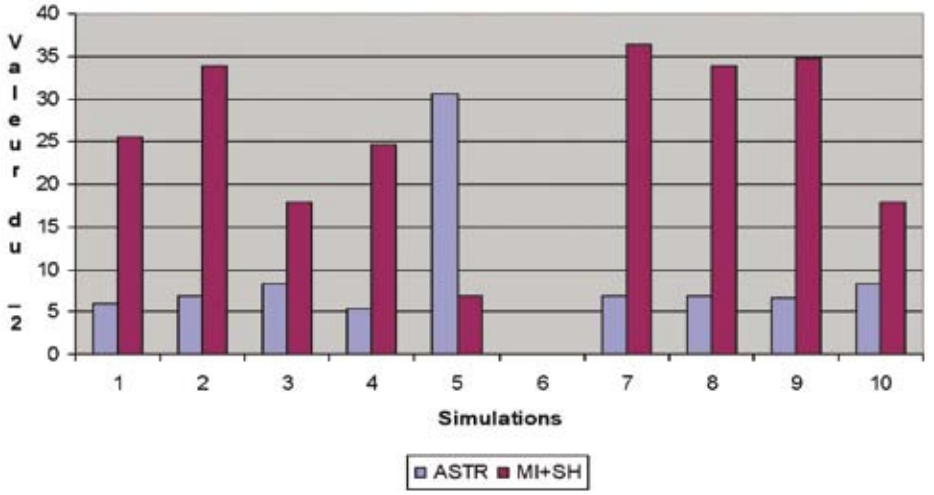


Figure 5 – Distribution des valeurs associées à la statistique du khi-carré (χ^2) pour 9 simulations

La dernière colonne du tableau 3 et la figure 6 nous présentent les résultats obtenus de la sous-exposition des items. Une brève analyse nous permet de constater que la méthode ASTR se distingue largement de la méthode MI+SH dans pratiquement toutes les simulations, sauf celle réalisée par Davis (2004) et qui correspond à la simulation 6. Comme dans le cas du TEI, on peut se questionner sur l'effet d'items polytomiques sur cet indicateur de performance. Dans toutes les autres simulations, comme on peut l'observer à la figure 6, la méthode ASTR permet de maximiser le recours à l'ensemble des items de la banque en ne négligeant que très peu d'items.

Tableau 3 – Valeurs associées au taux de chevauchement des items (TCI), au khi-carré (χ^2) et au taux d'exposition des items (TEI) pour 10 simulations de tests adaptatifs

Simulations	TCI		χ^2		TEI (≤2 %)	
	ASTR	MI+SH	ASTR	MI+SH	ASTR	MI+SH
1	0,10	0,18	5,910	25,635	5	129
2	0,11	0,19	6,931	33,923	1	169
3	0,17	0,19	8,286	17,858	0	73
4	0,12	0,18	5,362	24,700	---[2]	---
5	4,76	2,98	30,600	6,850	17	56
6	7,90	6,44	---	---	85	55
7	0,10	0,17	6,898	36,344	4	178
8	0,11	0,19	6,931	33,923	1	169
9	0,12	0,18	6,589	34,690	---	---
10	0,17	0,19	8,286	17,858	0	73

2 L'absence de certaines valeurs s'explique par le fait que certains articles ou rapports de recherche ne faisaient pas mention de ces données.

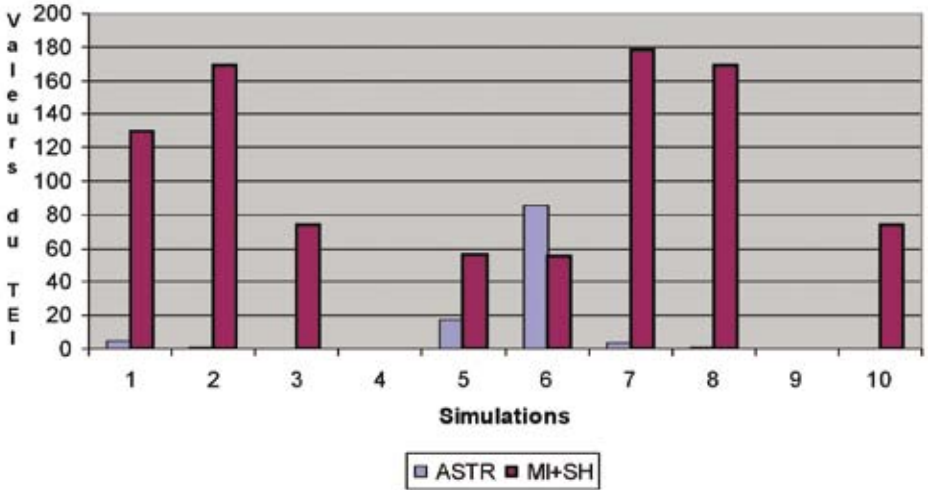


Figure 6 – Nombre d'items ayant une valeur inférieure à 2 % pour le taux d'exposition des items

CONCLUSION

Les résultats des études analysées permettent de constater que les méthodes alpha-stratifiées sont intéressantes par rapport au contrôle de l'exposition des items. Quand on les compare avec la méthode de Sympson et Hetter, on voit qu'elles permettent généralement d'améliorer la qualité de la mesure tout en favorisant un usage plus exhaustif de la banque d'items. D'autres recherches devraient être menées afin de mieux analyser l'effet de certaines variables sur l'efficacité des méthodes ASTR. Par exemple, les résultats obtenus par Chang et Van Der Linden ainsi que par Davis (2004) nous laissent croire que la nature des items utilisés lors de ces simulations peut avoir un effet direct sur les conclusions à tirer de l'efficacité des méthodes de contrôle. En ce sens, il est à souhaiter que les simulations de TA se réalisent dans l'avenir dans des conditions plus standardisées afin de mieux comparer les méthodes. Le nombre restreint de recherches comparées dans le cadre de ce texte reflète bien ce besoin de standardisation, car plusieurs résultats ont dû être omis de notre « méta-analyse » puisqu'ils s'éloignaient trop des autres résultats.

En terminant, puisqu'il semble que les tests adaptatifs seront appelés à jouer un rôle de plus en plus important en milieu scolaire et en formation, souhaitons que les recherches futures s'intéressent de plus en plus aux banques d'items réelles. De cette façon, il sera alors possible de valider les recherches théoriques (par exemple, les simulations de TA, d'items et de banques d'items) dans des contextes plus réalistes.

Références

Bunderson, C. V., Inouye, K. D. et Olsen, J. B. (1993), « The four generations of computerized educational measurement », dansans L. R. Linn (dir.), *Educational measurement* (3e ed., p. 367-407), Phoenix, National council on measurement in education, American council on education.

Chang, H.-H. et van der Linden, W. J. (2003), « Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing », *Applied Psychological Measurement, 27*(4), p. 262-274.

Chang, H.-H. et Ying, Z. (1999), « Alpha-Stratified Multistage Computerized Adaptive Testing », *Applied Psychological Measurement, 23*(3), p. 211-222.

Chen, S.-Y. et Lei, P. W. (2005), « Controlling item exposure and test overlap in computerized adaptive testing », *Applied Psychological Measurement, 29*(2), p. 204-217.

Davey, T. et Nering, M. L. (2002), « Controlling item exposure and maintainning item security », dans C. N. Mills, M. T. Potenza, J. J. Fremer et W. C. Ward (dir.), *Computer-based testing, building foundation for future assessments*, New-Jersey, Lawrence Erlbaum associates.

Deng, H. et Chang, H.-H. (2001), *A-Stratified Computerized Adaptive Testing with Unequal Item Exposure across Strata,* texte présenté dans le cadre de la rencontre annuelle de l'American Educational Research Association, Seattle, WA.

Eggen, T. J. H. M. (2001), « *Overexposure and underexposure of items in computerized adaptive testing* », Measurement and research department reports nº 2001-1. Arnhem, Citogroep.

Embretson, S. E. et Hershberger, S. L. (1999), *The new rules of measurement : what every psychologist and educator should know*, Mahwah, New-Jersey, Lawrence Erlbaum associates, 263 p.

Georgiadou, E., Triantafillou, E. et Economides, A. A. (2007), « A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005 », *The journal of technology, learning, and assessment, 5*(8), p. 1-39.

Hambleton, R. K. et Swaminathan, H. (1985), *Item response theory : Principles and applications*, Norwell, MA, Kluwer Academic Publishers.

Hetter, R. D. et Sympson, J. B. (1997), « Item exposure control in CAT-ASVAB », dans W. A. Sands, Waters, B. K., McBride, J. R. (dir.), *Computerized adaptive testing : From inquiry to operation,* Washington, American Psychological Association, p. 141-144.

Kingsbury, G. G., et Houser, R. L. (1999), « Developing computerized adaptive tests for school children »,dans F. Drasgow et J. B. Olson-Buchanan (dir.), *Innovations on computerized assessment*, New-Jersey, Lawrence Erlbaum associates.

Kingsbury, G. G. et Zara, A. R. (1989), « Procedures for Selecting Items for Computerized Adaptive Tests », *Applied Measurement in Education, 2*(4), p. 359-375.

Laurier, M. (1999), « Testing adaptatif et évaluation des processus cognitifs », dans C. Depover et B. Noël (dir.), *L'évaluation des compétences et des processus cognitifs : modèles, pratiques et contextes*, Bruxelles, DeBoeck Université, p. 255-266.

Leung, C.-K., Chang, H.-H. et Hau, K.-T. (2003), « Incorporation of Content Balancing Requirements in Stratification Designs for Computerized Adaptive Testing », *Educational & Psychological Measurement, 63*(2), p. 257-270.

Leung, C. K., Chang, H. H. et Hau, K. T. (2002), « Item selection in computerized adaptive testing : improving the a-stratified design with the Sympson-Hettter algorithm », *Applied psychological measurement, 26*(4), p. 376-392.

McBride, J. R. et Martin, J. T. (1983), « Reliability and validity of adaptive ability tests in a military setting », dans A. Press (dir.), *New horizons in testing,* New-York, p. 223-226.

Meijer, R. R. et Nering, M. L. (1999), « Computerized Adaptive Testing : Overview and Introduction », *Applied Psychological Measurement, 23*(3), p. 187-194.

Mills, C. N. et Stocking, M. L. (1996), « Practical Issues in Large-Scale Computerized Adaptive Testing », *Applied Measurement in Education, 9*(4), p. 287-304.

Revuelta, J. et Ponsoda, V. (1998), « A Comparison of Item Exposure Control Methods in Computerized Adaptive Testing », *Journal of Educational Measurement, 35*(4), p. 311-327.

Sympson, J. B. et Hetter, R. D. (1985), *Controling item exposure conditional on ability in computerized adaptive testing,* Proceedings of the 27th annual meeting of the Military Testing Association, San Diego.

Trentin, G. (1997), « Computerized Adaptive Tests and Formative Assessment », *Journal of Educational Multimedia & Hypermedia, 6*(2), p. 201-220.

Van der Linden, W. J. (2002), « *Some Alternatives to Sympson-Hetter Item-Exposure Control in Computerized Adaptive Testing. Research Report* », Research/Technical. Enschede (Netherlands) Faculty of Educational Science and Technology, Twente University.

Van der Linden, W. J. et Chang, H. H. (2005), « *Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach* », Computerized Testing Report no. 01-09, Law School Admission Council.

Wainer, H. et Eignor, D. (2000), « Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing », dans H. Wainer, (dir.), *Computerized adaptive testing : a primer* (2e ed.), Mahwah, NJ, Lawrence Erlbaum, p. 271-299.

Way, W. D. (1998), « Protecting the integrity of computerized testing item pools », *Educational Measurement : Issues & Practice, 17*(4), p. 17-27.

Yi, Q. (2002), *Incorporating the Sympson-Hetter exposure control method into the a-stratified method with content blocking,* Annual meeting of the American Educational Research Association (AERA), New Orleans, LA.