



uOttawa

L'Université canadienne
Canada's university

**FACULTÉ DES ÉTUDES SUPÉRIEURES
ET POSTDOCTORALES**



**FACULTY OF GRADUATE AND
POSTDOCTORAL STUDIES**

Nathalie Loye

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Education)

GRADE / DEGREE

Faculty of Education

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Conditions d'élaboration de la matrice Q des modèles cognitifs et impact sur sa validité et sa fidélité

TITRE DE LA THÈSE / TITLE OF THESIS

Dany Laveault

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Bruno Zumbo

Marielle Simon

André Rupp

David Trumpower

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Conditions d'élaboration de la matrice Q des modèles cognitifs
et impact sur sa validité et sa fidélité

Nathalie Loye

Faculté d'Éducation, Université d'Ottawa



Library and
Archives Canada

Bibliothèque et
Archives Canada

Published Heritage
Branch

Direction du
Patrimoine de l'édition

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 978-0-494-50741-4

Our file Notre référence

ISBN: 978-0-494-50741-4

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Résumé

Cette recherche exploratoire compare les matrices Q élaborées par des experts dans une approche inductive pour deux questionnaires à choix multiple en mathématiques, dans trois conditions d'accès à des informations sur les items. Dans la première condition, les experts ne disposent que des énoncés des items; dans la seconde, les items leur sont fournis regroupés selon les résultats d'une analyse factorielle ; la troisième condition met à leur disposition les paramètres de difficulté et de discrimination de chaque item ainsi que les résultats de l'analyse des leurres.

L'application du modèle cognitif RUM (*Reparametrized Unified Model*) aux données permet de comparer la structure cognitive des matrices et de définir si certaines sont plus fortes que d'autres en basant le jugement sur la forme des matrices obtenues, sur la concordance du jugement des experts et sur les valeurs des paramètres estimés par le modèle RUM. Les résultats tendent à montrer que les matrices de la condition 2 ont une structure cognitive plus forte pour les deux tests qui font l'objet de cette étude, mais la différence entre les conditions 1 et 2 est peu importante. La condition 3 a amené les experts à perdre de vue le processus complet de réponse aux items au profit de détails et a fourni les matrices de moins bonne qualité.

Abstract

This exploratory research compares Q-matrices related to two multiple choice tests in mathematics. These Q-matrices are elaborated in three different conditions. In the first one, only the items are available. In the second one, the items are presented as set together with a factor analysis. In the third condition, the item difficulty and discrimination parameters are provided to the experts as well as the distractors' analysis results.

The cognitive model RUM (*Reparametrized Unified Model*) is used to compare the cognitive structure of the different matrices related to the same test. The form of the matrices, the agreement between the experts and the parameters estimated with the RUM are compared. The results tend to show that condition 2 cognitive structure is stronger for both tests. The difference with conditions 1 or 3 is, however, not very large. The third condition leads the experts to pay attention to non relevant aspects of the answering process and leads to the worst quality matrices.

À Thierry, Mélanie et Matthieu

Remerciements

Je tiens tout d'abord à remercier le professeur Louis DiBello, non seulement parce qu'il a rendu ce travail possible en m'aidant à avoir accès à Arpeggio, mais aussi parce qu'il m'a prodigué de nombreux conseils et encouragements et m'a fait cadeau de son amitié. Merci à ETS (*Educational Testing Service*) pour m'avoir fourni Arpeggio.

Je remercie le professeur Dany Laveault pour m'avoir guidée sur le chemin tout en me laissant la liberté dont j'avais besoin. Grâce à lui, j'ai énormément appris et je lui en suis très reconnaissante. Un grand merci également aux membres de mon comité pour leurs conseils et leurs critiques constructives.

Je veux également remercier les professeurs du département d'administration et fondements de l'éducation de l'Université de Montréal pour leur soutien quotidien et leurs encouragements.

Enfin, mes remerciements vont à tous les membres de ma famille qui m'ont épaulée et fait confiance. Merci en particulier à mon conjoint et à mes enfants dont le soutien a toujours été solide et inconditionnel dans les bons jours comme dans les moins bons.

Table des matières

1	CHAPITRE I : INTRODUCTION	1
<i>1.1</i>	<i>La problématique</i>	<i>2</i>
<i>1.2</i>	<i>La revue de la littérature</i>	<i>9</i>
1.2.1	L'évaluation diagnostique	10
1.2.2	La notion d'attribut cognitif et l'élaboration de la matrice Q	15
1.2.3	Quelques modèles.....	19
1.2.4	La validité et la fidélité de la matrice Q	24
<i>1.3</i>	<i>L'élaboration du cadre conceptuel</i>	<i>28</i>
1.3.1	Le cadre conceptuel.....	28
1.3.2	Une méthode pour élaborer les matrices Q.....	30
1.3.3	Des conditions pouvant influencer la validité et la fidélité de la matrice Q	33
1.3.4	Un modèle afin de juger de la validité et de la fidélité de la matrice Q.....	36
1.3.5	Les variables de la recherche.....	38
<i>1.4</i>	<i>Les objectifs et questions de recherche.....</i>	<i>40</i>
2	CHAPITRE II : MÉTHODOLOGIE.....	42
<i>2.1</i>	<i>Les participants</i>	<i>43</i>
<i>2.2</i>	<i>Les données.....</i>	<i>43</i>
2.2.1	Les neuf bases de données d'origine	43
2.2.2	Les deux bases de données de la recherche.....	45
2.2.3	Les caractéristiques des tests A et B.....	46
<i>2.3</i>	<i>La procédure.....</i>	<i>49</i>

2.3.1	L'élaboration des matrices	49
2.3.1.1	La notion d'attributs cognitifs.....	49
2.3.1.2	Les trois conditions expérimentales.....	49
2.3.1.3	Le déroulement	50
2.3.1.4	Les trois phases de construction des matrices synthèses.....	53
2.3.2	Le traitement des matrices Q	59
2.3.2.1	Le modèle RUM	60
2.3.2.2	La méthode d'estimation du RUM dans Arpeggio	64
2.3.2.3	La convergence et l'ajustement du modèle aux données	65
2.3.2.4	La modélisation RUM appliquée aux données et matrices de la recherche	66
3	CHAPITRE III : RÉSULTATS.....	70
3.1	L'élaboration et la description des matrices.....	71
3.1.1	Les matrices individuelles originales (phase 1 de la procédure MACB).....	71
3.1.2	Les matrices individuelles modifiées	76
3.1.2.1	Élaboration : les phases 2 et 3 de la procédure MACB.....	76
3.1.2.2	Description des matrices individuelles modifiées.....	79
3.1.3	Les matrices synthèses	87
3.2	Le traitement des matrices	91
3.2.1	L'application du modèle RUM aux matrices individuelles originales	92
3.2.2	L'application du modèle RUM aux matrices individuelles modifiées	92
3.2.3	L'application du modèle RUM aux matrices synthèses et les simplifications réalisées sur ces matrices	93
3.2.3.1	L'application du modèle RUM aux matrices synthèses.....	93
3.2.3.2	Les conséquences des choix.....	98
3.2.3.3	La description des matrices synthèses simplifiées	101
3.3	Les comparaisons des paramètres	114
3.3.1	La difficulté p_k des attributs : résultats et comparaisons.....	115

3.3.2	Les paramètres d'items π : résultats et comparaisons.....	127
3.3.3	Les paramètres d'items r : résultats et comparaisons	134
3.4	Une illustration du diagnostic des sujets.....	147
4	CHAPITRE IV : INTERPRÉTATION, DISCUSSION ET CONCLUSION.....	151
4.1	L'interprétation et la discussion	151
4.2	Les limites	163
4.3	Les recommandations.....	167
4.4	La conclusion	169

Liste de figures

Figure 1. Exemple de liens entre 5 attributs et 3 items et la matrice Q correspondante. Les formes arrondies représentent les attributs qui sont des variables latentes. Les rectangles représentent les réponses des sujets qui sont des variables observées. Dans la matrice Q, les lignes représentent les items et les colonnes représentent les attributs. La valeur est 1 lorsqu'un lien existe entre l'item et l'attribut, la valeur est 0 sinon.....	4
Figure 2. Les variables de la recherche	39
Figure 3. La difficulté des items des tests A et B	48
Figure 4. Le plan d'élaboration des 24 matrices individuelles (et des six matrices synthèses) relativement au test, à l'expert et à la condition et les dimensions selon lesquelles les comparaisons sont possibles.	50
Figure 5. Les trois phases et les tâches à accomplir lors de l'application de la version adaptée de la méthode MACB ...	58
Figure 6. Les paramètres estimés dans le modèle RUM (Hartz, 2002)	60
Figure 7. Le nombre moyen d'intitulés individuels relativement aux catégories de la synthèse pour les trois conditions.	80
Figure 8. Le nombre d'intitulés dans les matrices individuelles des experts relativement aux catégories de la synthèse..	82
Figure 9. Le comportement des experts relativement aux 6 catégories synthèses.....	87
Figure 10. Les nombres de liens pour chaque attribut dans les matrices synthèses au fil des 3 conditions.....	89
Figure 11. Les trois graphiques correspondant à la difficulté de l'attribut 5 (valeur pk_5) avec la matrice QA1	94
Figure 12. Les trois graphiques correspondant au paramètre π de l'item 9 (valeur π_9) avec la matrice QA1	94
Figure 13. Les trois graphiques correspondant au paramètre c de l'item 19 (valeur c_9) avec la matrice QA1	94
Figure 15. Le nombre d'items dans les quatre groupe en fonction du contenu, pour les tests A et B	113
Figure 16. La difficulté des attributs pour les trois conditions et les deux tests	116
Figure 17. Les distributions des difficultés des attributs pour chaque condition et chaque attribut en fonction du score total obtenu au test	122
Figure 18. Les distributions des difficultés des attributs pour chaque condition et chaque attribut en fonction du score total obtenu au test B.....	126
Figure 19. Les valeurs de π présentées pour les items des tests A et B classés en ordre de difficulté décroissante, pour les 3 conditions.....	130

Figure 20. La différence entre les valeurs de π obtenues aux conditions 2 et 3 et celles obtenues à la condition 1. Les items sont classés pour des différences entre les conditions 1 et 3 de plus en plus grandes.	131
Figure 21. Les valeurs moyennes de π pour les items appartenant ou non à un regroupement selon les analyses factorielles pour les deux tests	133
Figure 22. Les valeurs moyennes de r pour les items appartenant ou non à un regroupement selon les analyses factorielles pour les deux tests.....	136
Figure 23. Les pourcentages de valeurs r dans 4 intervalles pour les tests A et B et les 3 conditions.....	138
Figure 24. Les valeurs moyennes des paramètres r relativement aux 6 attributs pour les trois conditions et les deux tests	140
Figure 25. Le nombre d'items relativement à leur r moyen pour chaque condition et chaque test	141
Figure 26. Les valeurs moyennes de r lorsque les items sont regroupés selon leur contenu pour chaque condition et chaque test.....	142
Figure 27. Les valeurs moyennes de r lorsque les items sont regroupés selon leur appartenance à un groupe pour chaque condition et chaque test.....	143
Figure 28. Le diagnostic de huit sujets relativement aux attributs, en fonction de leur score total pour le test B et la matrice synthèse simplifiée de la condition 2	148
Figure I1. Les valeurs r pour l'attribut 1 selon les conditions et les tests A et B.....	45
Figure I2. Les valeurs r pour l'attribut 2 selon les conditions et les tests A et B.....	xlvii
Figure I3. Les valeurs r pour l'attribut 3 selon les conditions et les tests A et B.....	xlix
Figure I4. Les valeurs r pour l'attribut 4 selon les conditions et les tests A et B.....	li
Figure I5. Les valeurs r pour l'attribut 5 selon les conditions et les tests A et B.....	liv
Figure I6. Les valeurs r pour l'attribut 6 selon les conditions et les tests A et B.....	lvi

Liste de tableaux

Tableau 1.....	18
Quelques exemples de recherches empiriques de la littérature.....	18
Tableau 2.....	47
Les tableaux de spécification des tests A et B.....	47
Tableau 3.....	47
Les caractéristiques du score total des tests A et B.....	47
Tableau 4.....	54
La similarité entre les processus d'établissement de standards et d'élaboration de la matrice Q.....	54
Tableau 5.....	69
La procédure de modélisation avec Arpeggio appliquée à toutes les matrices synthèses.....	69
Tableau 6.....	72
L'opinion des experts en regard des regroupements de la condition 2.....	72
Tableau 7.....	72
L'opinion des experts en regard de l'utilité des informations de la condition 3.....	72
Tableau 8.....	74
Le nombre total de liens établis par chaque expert dans chacune des 24 matrices individuelles originales et le nombre d'attributs correspondant.....	74
Tableau 9.....	78
Les regroupements des attributs suite au MACB.....	78
Tableau 10.....	83
Le nombre total de liens établis par chaque expert dans chacune des 24 matrices individuelles modifiées et le nombre d'attributs correspondant.....	83
Tableau 11.....	86
Les 36 coefficients de contingence pour les matrices individuelles modifiées.....	86
Tableau 12.....	88
Le nombre total de liens établis dans chacune des 6 matrices synthèses et le nombre d'attributs correspondant.....	88

Tableau 13.....	91
Les 12 coefficients de contingence entre les matrices synthèses et les matrices individuelles.....	91
modifiées	91
Tableau 14.....	96
La simplification des matrices synthèses.....	96
Tableau 15.....	98
Les indices d'ajustement des modèles retenus pour les trois conditions	98
Tableau 16.....	100
Les différences entre l'analyse initiale et l'analyse retenue pour l'ensemble des matrices	100
Tableau 17.....	102
La description des six matrices synthèses simplifiées	102
Tableau 18.....	105
La mise en parallèle des changements et de la structure factorielle pour les tests A et B	105
Tableau 19.....	106
La mise en parallèle de la structure factorielle des tests A et B et du tableau de spécification	106
Tableau 20.....	108
Nombre d'items en lien avec chaque attribut pour chaque facteur et chaque condition.....	108
Tableau 21.....	109
Les items des tests A et B classés selon l'attitude des experts	109
Tableau 22.....	110
Le contenu des items de chacun des 4 groupes	110
Tableau 23.....	117
La difficulté des attributs.....	117
Tableau 24.....	118
Les nombres d'items faciles et difficiles reliés aux attributs les plus difficiles.....	118
Tableau 25.....	137
Les valeurs moyennes de r pour chaque test, chaque condition et globalement.....	137
Tableau 26.....	145
La synthèse des comparaisons des valeurs de r pour les trois conditions, par attribut et par test.....	145

Tableau 27.....	150
Les items réussis par les sujets 23 et 151 ainsi que les liens entre les items et l'attribut 5.....	150
Tableau A1.....	i
Influence du test sur la moyenne des tests A et B	i
Tableau A2	i
Influence du test sur le nombre moyen de liens dans les matrices individuelles originales (test non paramétrique de Friedman).....	i
Tableau A3	i
Influences de l'expert et de la condition sur le nombre moyen de liens dans les matrices individuelles originales.....	i
Tableau A4	ii
Influence de la condition et de l'expert sur le nombre moyen de liens dans les matrices individuelles originales (sans l'expert 2).....	ii
Tableau A5	ii
Influence du test sur le nombre moyen d'intitulés individuels reliés aux catégories du MACB (test non paramétrique de Friedman).....	ii
Tableau A6	ii
Influence de la condition, de l'expert et de l'attribut sur le nombre moyen d'intitulés individuels reliés aux catégories du MACB.....	ii
Tableau A7	iii
Influence du test sur le nombre moyen de liens dans les matrices individuelles modifiées (test de Friedman)	iii
Tableau A8	iii
Influence de la condition et de l'expert sur le nombre moyen de liens dans les matrices individuelles modifiées	iii
Tableau A9	iv
Influence de la condition sur le nombre moyen de liens dans les matrices individuelles modifiées pour chaque attribut séparément (expert est une mesure répétée).....	iv
Tableau A10	iv
Influence de l'expert et de la condition sur le nombre moyen d'attributs consensuels utilisés dans matrices individuelles modifiées.....	iv
Tableau A11	v

Influence du test sur le nombre moyen de liens (total) dans les matrices synthèses (test non paramétrique de Friedman)..	v
Tableau A12	v
Influence de la condition sur le nombre moyen de liens (total) dans les matrices synthèses	v
Tableau A13	v
Influence du test sur le nombre moyen de liens (par attribut) dans les matrices synthèses (test de Friedman)	v
Tableau A14	vi
Influence de la condition et de l'attribut sur le nombre moyen de liens (par attribut) dans les matrices synthèses	vi
Tableau A15	vi
Influence de la condition sur le nombre moyen de liens dans les matrices synthèses pour chaque attribut séparément	vi
Tableau A16	vii
Influence de test sur le nombre moyen de liens supprimés dans les matrices synthèses lors de l'application du modèle RUM	vii
Tableau A17	vii
Influence de la condition sur le nombre moyen de liens supprimés dans les matrices synthèses lors de l'application du modèle RUM (test de Friedman)	vii
Tableau A18	vii
Influence du test sur le nombre moyen de liens (par attribut) dans les matrices synthèses simplifiées (test de Friedman)	vii
Tableau A19	viii
Influence de la condition et de l'attribut sur le nombre moyen de liens (par attribut) dans les matrices synthèses	viii
Tableau A20	viii
Influence de la condition sur le nombre moyen de liens dans les matrices synthèses simplifiées pour chaque attribut séparément	viii
Tableau A21	ix
Influence de la condition sur le nombre moyen de liens par item dans les matrices synthèses simplifiées (test de Friedman)	ix
Tableau A22	ix
Influence du test sur la valeur moyenne de pk (test de Friedman)	ix
Tableau A23	ix
Influence de la condition sur la valeur moyenne de pk (attribut est une mesure répétée)	ix

Tableau A24	x
Influence du test sur la valeur moyenne de π dans les matrices synthèses simplifiées (test de Friedman).....	x
Tableau A25	x
Influence de la condition sur la valeur moyenne de π dans les matrices synthèses simplifiées (test est une mesure répétée)	x
Tableau A26	x
Influence de l'item sur la valeur moyenne de π avec les matrices synthèses simplifiées.....	x
Tableau A27	xi
Influence de la difficulté des items sur la valeur moyenne de π avec les matrices synthèses simplifiées	xi
Tableau A28	xi
Influence de l'appartenance à un groupe sur la valeur moyenne de π avec les matrices synthèses simplifiées	xi
Tableau A29	xii
Influences de la condition et de l'appartenance à un regroupement (AF) sur la valeur moyenne de π avec les matrices synthèses simplifiées.....	xii
Tableau A30	xii
Influence du test sur la valeur moyenne des paramètres r avec les matrices synthèses simplifiées.....	xii
Tableau A31	xiii
Influence de la condition sur la valeur moyenne des paramètres r pour les deux tests et pour chaque test séparément avec les matrices synthèses simplifiées (tests de Friedman)	xiii
Tableau A32	xiv
Influence de la condition sur les valeurs moyennes de r pour les items de chacun des deux tests regroupés selon leur groupe ou selon leur contenu (tests de Friedman).....	xiv
Tableau A33	xv
Influence de l'appartenance à un regroupement (AF) sur la valeur moyenne des paramètres r avec les matrices synthèses simplifiées.....	xv
Tableau A34	xv
Influence de l'item et de l'attribut sur la valeur moyenne des paramètres r avec les matrices synthèses simplifiées	xv
Tableau A35	xvi
Influence de l'appartenance à un groupe sur la valeur moyenne de r avec les matrices synthèses simplifiées	xvi

Tableau A36	xvi
Influence des conditions sur la valeur moyenne de r pour chaque attribut avec les matrices synthèses simplifiées pour le test A	xvi
Tableau A37	xvii
Influence des conditions sur la valeur moyenne de r pour chaque attribut avec les matrices synthèses simplifiées pour le test B	xvii
Tableau A38	xviii
Influence des conditions sur la valeur moyenne de r pour chaque item avec les matrices synthèses simplifiées pour le test A	xviii
Tableau A39	xix
Influence des conditions sur la valeur moyenne de r pour chaque item avec les matrices synthèses simplifiées pour le test B.....	xix
Tableau F1.....	xxxvii
Les indices d'ajustement des analyses factorielles exploratoires et confirmatoires pour les tests A et B	xxxvii
Tableau F2.....	xxxviii
Le rapport des analyses factorielles tel que fourni aux experts lors de la condition 2.....	xxxviii
Tableau I1.....	xlii
Les valeurs de r pour le test A et les 3 conditions	xlii
Tableau I2.....	xliii
Les valeurs de r pour le test B et les 3 conditions.....	xliii

Liste des annexes

ANNEXE A.....	i
Résultats des analyses de variance et des tests non paramétriques. Tous les effets sont fixes.	i
ANNEXE B.....	xxi
Test de signification des valeurs alpha de deux tests indépendants.....	xxi
ANNEXE C.....	xxii
Définitions et exemples d'attributs tirés de la littérature.....	xxii
ANNEXE D.....	xxv
Les 106 énoncés d'attributs.....	xxv
ANNEXE E.....	xxx
Les 106 attributs regroupés en six catégories tels qu'obtenus par consensus entre les experts.....	xxx
ANNEXE F.....	xxxv
La description des analyses factorielles exploratoires et confirmatoires dont les résultats sont fournis aux experts lors de la condition 2	xxxv
ANNEXE G.....	xxxix
Un exemple de la documentation fournie aux experts à la condition 3.	xxxix
ANNEXE H.....	xl
Les équations du modèle RUM	xl
ANNEXE I.....	xlii
Les valeurs de r pour chacun des liens existant dans les matrices synthèses simplifiées.	xlii
ANNEXE J.....	lvii
Les attestations d'approbation déontologique	lvii

Légende

Liste des acronymes :

DINA: *Deterministic Inputs, Noisy "and" gate model*

GEN : *Théorie de la généralisabilité*

GLTM: *General component Latent Trait Model*

LLTM: *Linear Logistic Test Model*

MACB: *Multi Attribute Consensus Building*

MAU: *Multi-Attribute Utility*

MLTM: *Multicomponent Latent Trait Model*

QCM: *Questionnaire à choix multiple*

RUM: *Reparameterized Unified Model*

TCS : *Théorie classique des scores*

TRI : *Théorie de réponse à l'item*

1 Chapitre I : introduction

Le mot diagnostique est issu du grec *diagnosis* dans lequel *gnosis* signifie connaissance et *dia* signifie à travers. Un diagnostic médical vise à identifier une maladie à travers des symptômes et parfois des investigations complémentaires. L'évaluation diagnostique en éducation cherche à identifier les particularités des élèves à travers leurs réponses aux questions d'un test. Selon Scallon (1988), le rôle de l'évaluation diagnostique va au-delà « du dépistage des élèves en difficulté, mais doit permettre de découvrir les forces et les faiblesses ainsi que le degré de préparation des élèves avant que ceux-ci n'entreprennent une séquence importante d'apprentissage ».

L'enjeu auquel est confronté le médecin consiste à découvrir la maladie dont souffre son patient. Il fournit ensuite les remèdes et traitements appropriés. L'enjeu d'une évaluation diagnostique en éducation vise à apprécier les « caractéristiques individuelles du sujet » (Legendre, 2005) en lien avec ses savoirs ou ses savoirs faire. Le but est, par la suite, de prendre des décisions pour procurer à chaque sujet en difficulté une remédiation appropriée.

Dans des situations d'apprentissage et d'évaluation complexes et authentiques, l'observation des sujets dans leur résolution de problèmes ou la trace de leurs démarches sont autant de façons d'inférer les processus cognitifs mis en œuvre par chacun d'eux. Toutefois une approche individualisée nécessite beaucoup de temps de la part de l'enseignant. Un test diagnostique à correction objective, comme un examen à choix multiple, est plus rapide à administrer et à corriger. Le problème qui se pose alors consiste à cibler les informations individuelles qui ont une influence positive ou négative sur le cheminement de chaque étudiant et à les déduire à partir des seules réponses des sujets, sans trace de leur démarche.

Un médecin définit ce qu'il doit observer chez son patient pour poser son diagnostic, mais aussi comment interpréter ses observations. Dans la phase d'élaboration d'un test diagnostique, il faut déterminer ce qui doit être évalué, la forme que doit prendre le test, mais également et surtout la façon dont les résultats seront interprétés. Les modèles modernes de la mesure sur lesquels se penche cette recherche unissent les approches psychométriques et cognitives. Ils permettent d'inférer, dans un but diagnostique, les processus cognitifs maîtrisés ou non par chaque sujet même lorsque les seules données accessibles sont la réussite ou l'échec à chaque item. Ces modèles offrent donc d'exploiter un questionnaire à choix multiple dans la perspective d'établir un diagnostic cognitif individuel.

1.1 La problématique

L'objectif d'une évaluation devrait être clairement défini car il est relié à des variables latentes de natures différentes: continues, catégorielles, dichotomiques. Le score vrai du sujet est une variable latente continue dans le cas de la théorie classique des tests, tout comme l'est le score univers dans la théorie de la généralisabilité ou l'habileté dans le cas de la théorie de réponse à l'item. Tous trois sont fonction du score total obtenu par le sujet, sont attachés à une échelle continue unidimensionnelle et permettent traditionnellement de classer les étudiants.

Lorsque l'objectif de l'évaluation consiste en un diagnostic relativement à un certain nombre de caractéristiques possédées ou non par les sujets, les variables latentes sont les niveaux discrets de classification pour chacune de ces caractéristiques (DiBello, Roussos, et Stout, 2006). L'hypothèse posée est que l'espace latent d'habileté sous une forme unidimensionnelle peut être scindé en un certain nombre de composantes (Jang, 2005). Par exemple, la mesure à interprétation critériée offre un profil de maîtrise sous forme de sous-scores en lien avec un ensemble d'objectifs clairement

explicités dans le tableau de spécification du test (Scallan, 2000). Les variables latentes correspondent alors aux habiletés des sujets pour chaque objectif. Dans un profil de scores, la performance de l'élève est simplement décrite par la fréquence des erreurs systématiques qu'il a commises et de ses bonnes réponses. On pourrait voir dans la propension à commettre chaque type d'erreur une variable latente.

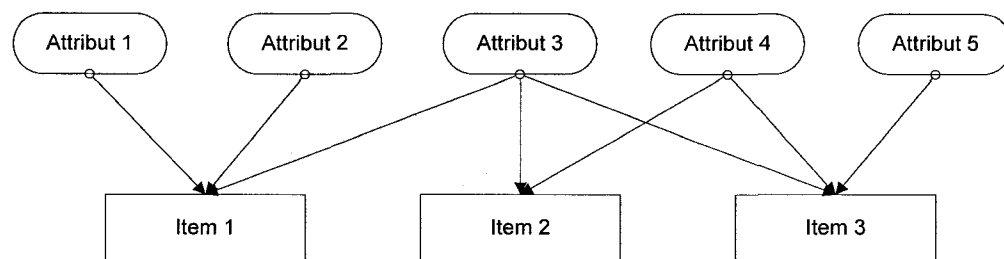
Depuis le milieu des années 1980, des recherches qui allient approches cognitives et psychométriques ont commencé à apparaître. Elles visent à fournir un diagnostic en lien avec des variables latentes discrètes représentant des processus cognitifs (Nichols, Chipman, et Brennan, 1995). Les nouveaux modèles qui se développent sont souvent appelés modèles cognitifs. Ils sont en pleine évolution grâce aux avancées en psychologie cognitive et en informatique (Rupp, 2007). Ils modélisent la probabilité de répondre correctement à un item en fonction de la maîtrise d'un ensemble d'attributs cognitifs spécifiques à cet item.

Un attribut cognitif est défini par la description de connaissances procédurales ou déclaratives aboutissant à la création d'une stratégie pour répondre à la question. Il représente un état latent de connaissance ou encore une habileté (Leighton, Gierl, et Hunka, 2002). Les modèles cognitifs se basent sur le postulat que la performance des sujets au test dépend d'un ensemble d'attributs cognitifs. Ceux-ci ne sont pas directement observables et ils sont éventuellement reliés les uns aux autres selon un ordre hiérarchique (Gierl, Leighton, et Hunka, 2000 ; Milewski, et Baron, 2002; Rupp, 2007; Rupp, et Watermann, 2005 ; Tatsuoka, 1983).

Ces modèles s'appuient sur l'élaboration, par des experts, d'une liste des attributs cognitifs nécessaires et suffisants pour répondre aux items du test. Les experts peuvent identifier au préalable les processus cognitifs à diagnostiquer et fabriquer un test approprié. Ils peuvent également utiliser un test existant et déterminer les processus cognitifs sous-jacents à chaque item. Dans les deux cas, le

modèle définit si chaque sujet maîtrise ou non chaque attribut ou classe les sujets dans un état de connaissance matérialisé par la possession ou non possession des divers attributs.

Les liens entre les attributs et les items sont opérationnalisés dans ces modèles par une matrice notée Q . La figure 1 présente un exemple de relations existant entre 5 attributs et 3 items.



$$Q = \begin{matrix} & \begin{matrix} \textit{Attribut 1} \\ \textit{Attribut 2} \\ \textit{Attribut 3} \\ \textit{Attribut 4} \\ \textit{Attribut 5} \end{matrix} & \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} & \begin{matrix} \textit{Item 1} \\ \textit{Item 2} \\ \textit{Item 3} \end{matrix} \end{matrix}$$

Figure 1. Exemple de liens entre 5 attributs et 3 items et la matrice Q correspondante. Les formes arrondies représentent les attributs qui sont des variables latentes. Les rectangles représentent les réponses des sujets qui sont des variables observées. Dans la matrice Q , les lignes représentent les items et les colonnes représentent les attributs. La valeur est 1 lorsqu'un lien existe entre l'item et l'attribut, la valeur est 0 sinon.

Une telle matrice peut être obtenue déductivement par les experts lorsque le test est élaboré à partir de la liste des attributs cognitifs. Dans ce cas, un cadre de référence contenant la liste des attributs cognitifs permet de générer les items qui sont fabriqués, ou choisis dans une banque, en même temps que la matrice. La priorité vise alors à satisfaire l'objectif diagnostique de l'évaluation.

La matrice Q peut également être obtenue par induction après la création du test. Les experts doivent identifier les attributs cognitifs et leurs liens avec les items avec la contrainte d'utiliser des items déjà existants. Ils peuvent utiliser une liste préalable d'attributs cognitifs pour établir les liens ou élaborer eux-mêmes cette liste en plus d'établir les liens. L'objectif est d'améliorer le pouvoir diagnostique d'un test ou d'exploiter l'information diagnostique d'un examen non initialement élaboré à cette fin. 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. L'approche implique d'extraire une information plus riche que celle obtenue à partir d'une échelle unidimensionnelle (DiBello et al., 2006).

Aussi bien dans l'approche déductive que dans l'approche inductive, le jugement des experts est essentiel. La validité et la fidélité des mesures produites par les modèles dépendent de l'identification des attributs, des liens logiques existants entre eux et de leurs liens avec les items. Elles dépendent étroitement de la capacité des experts de fournir des jugements valides, fidèles et concordants (Gierl et al., 2000). La concordance entre les jugements des experts est un gage de la fidélité des matrices et est essentielle pour assurer la validité et la fidélité des mesures produites par les modèles cognitifs.

La validité d'une mesure est traditionnellement définie en psychométrie par le fait de réellement mesurer ce que l'on prétend mesurer. Elle n'est toutefois pas une propriété du test lui-même, mais des inférences faites à partir des scores (Laveault, et Grégoire, 1997). En particulier, la validité dépend de la relation qu'il y a entre la réponse fournie à l'item et le processus qui aboutit à la

réponse. Par suite, la validité est menacée dès qu'une partie du processus est oubliée ou encore dès qu'un processus non prévu ou non souhaitable (comme deviner la réponse par exemple) intervient (Messick, 1995).

La fidélité d'une mesure est reliée à la notion de stabilité. Avoir de la chance ou commettre une erreur d'étourderie par exemple menacent la fidélité puisque ces comportements ne sont probablement pas reproductibles et se produisent au hasard. Dans le cas des matrices Q, la fidélité concerne à la fois les mesures produites par le modèle et la concordance entre le jugement des experts. Dans le but de simplifier la lecture, le terme de fidélité sera consacré aux mesures estimées par les modèles. Le terme de concordance sera quant à lui associé à la fidélité du jugement des experts.

Un modèle cognitif pose un problème s'apparentant à un problème de validité ou de fidélité dès que l'ensemble des attributs cognitifs identifiés ne sont pas des bons prédicteurs de la performance du sujet (Junker, 1999). Par exemple, le sujet peut utiliser une habileté différente de celle prévue par les experts ou deviner la réponse. Des attributs importants peuvent être oubliés dans le modèle. La validité est alors remise en question puisque ce dernier ne diagnostiquera pas convenablement les attributs possédés ou non par le sujet. De la validité de la matrice Q découlera la validité des mesures aboutissant au diagnostic des sujets relativement à chacun des attributs.

De la même façon, 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. En effet, elle ne sera vraisemblablement pas reproductible à une autre occasion. Une matrice pour laquelle de nombreuses erreurs de ce type sont décelées pose un problème de fidélité des mesures produites pour les sujets par le modèle. Toutefois, un tel problème en est éventuellement aussi un de validité. Si un item n'est pas relié aux attributs appropriés, une bonne réponse inattendue de la part du sujet pourrait seulement montrer qu'il a utilisé d'autres attributs pour répondre convenablement. Il est donc difficile de bien faire la distinction entre validité et fidélité lorsqu'il s'agit de la matrice Q.

L'élaboration d'une matrice valide et fidèle revient aux experts. Ceci implique de se questionner sur les caractéristiques que ces derniers doivent posséder, sur leur nombre de même que sur la manière de les faire travailler. De plus, la façon de fabriquer la matrice varie selon que l'approche est inductive ou déductive. Il faut s'interroger sur les informations à fournir pour aider les experts à comprendre et à mener à bien leur tâche. L'idéal serait que ces experts soient d'accord entre eux, non seulement en ce qui concerne la liste des attributs cognitifs, mais sur les liens entre items et attributs. Enfin la matrice qui résulte du travail des experts devrait permettre un diagnostic valide et fidèle qui informe des caractéristiques des sujets responsables de leurs succès et de leurs difficultés.

Jusqu'à présent, peu de recherches portant sur la validité et la fidélité de la matrice Q ont été menées. Dans plusieurs cas d'applications de modèles cognitifs à des données réelles, les chercheurs améliorent la validité et la fidélité de la matrice initiale au cours de la recherche. Pour ce faire, ils se basent sur les résultats obtenus après l'application du modèle (voir par exemple : Buck, et Tatsuoka (1998), Dimitrov, et Raykov (2003), Hartz (2002), Jang (2005) ou Tatsuoka et al.(1993)). Dans ces études, ils ne créent pas plusieurs matrices qu'ils comparent, mais font évoluer une matrice unique. Rupp et Templin (2006) se servent de paramètres estimés par un modèle cognitif pour comparer différentes matrices. Ils s'intéressent à l'impact des changements opérés dans la matrice Q sur l'estimation des paramètres, mais ne visent pas l'étude de la validité et de la fidélité de Q. De plus, leurs données sont simulées, ils n'ont donc pas recours à des experts.

Le processus d'élaboration de la matrice Q par les experts est le premier pas vers des inférences valides et fidèles concernant le diagnostic des sujets. Cette étape du processus est donc cruciale. Si la littérature ne fournit pas de recherches visant à soutenir la tâche des experts lors de l'élaboration de la matrice Q, elle abonde pourtant d'études pour les assister dans d'autres tâches.

Par exemple, les erreurs de mesure aléatoires et systématiques dues au jugement humain sont beaucoup étudiées dans le cadre de l'accord entre les correcteurs (Lipkins, Jones, et Halkitis, 1996; Lunz, Stahl, et Wright, 1994; Murphy, et Davidshofer, 1988 ; Thorndike, 1920 ; Wolfe, et Kao, 1996). Différentes solutions ont été proposées afin de minimiser les facteurs d'erreurs dues aux juges dans un processus de correction. Un entraînement avant le début de la correction implique une meilleure utilisation des échelles et une meilleure compréhension des consignes. Un entraînement en cours de correction à la suite de l'estimation de la sévérité des juges guide les juges. Une multiple correction de chaque copie donne la possibilité de détecter les scores biaisés par l'effet de juge et de les corriger (Wilson, et Case, 2000; Loe, 2002).

De même, lors de l'établissement de standards et de seuils de compétence, les experts doivent à la fois établir le score minimal acceptable et la définition des connaissances et habiletés correspondantes (Kane, 2001; Sadesky, et Gushta, 2004; Zieky, 2001). Si la sagesse et l'expérience sont indispensables, les associer à un entraînement et à des informations statistiques permet d'assurer une meilleure validité, une meilleure fidélité ou simplement un gain de temps appréciable (Jaeger, 1996; Lunz, 2000; Sadesky, 2003; Sadesky et al., 2004).

La procédure itérative fournit à chaque expert des renseignements basés sur des régressions multiples. Ces informations donnent la possibilité à chaque expert de comparer son comportement à ceux des autres experts. L'expert peut ensuite réviser son jugement dans un processus interactif (Jaeger, 1996). La méthode des signets (*bookmark method*) allège la tâche des experts en leur fournissant a priori une aide basée sur l'estimation de la difficulté des items grâce à la théorie de réponse à l'item (Buckendahl, Smith, Impara, et Plake, 2000, 2002). L'utilité d'une analyse en grappes (*cluster analysis*) préalable a été mise en évidence par Sireci, Robin et Patelis (1999) et par Sadesky (2003). Des analyses de classes latentes (Brown, 2000) ou une variation du modèle *rule space* de Tatsuoaka (1983), le modèle à attribut hiérarchique (*Attribute Hierarchy Model*) (Leighton et

al., 2002; Sadesky et al., 2004) donnent des informations aux experts sur l'appartenance des sujets à une classe latente particulière.

Différents résultats issus de l'analyse des données fournissent donc un appui aux experts. Ils doivent malgré tout garder un esprit critique car leur jugement reste l'élément essentiel de la démarche (Sadesky, 2003). La fidélité des scores attribués par des correcteurs ou la validité et la fidélité des standards et seuils de réussite établis par des experts peuvent donc parfois être améliorées par un entraînement des participants ou grâce à l'accès aux résultats de diverses analyses des données.

Soutenir le travail d'experts dans diverses circonstances est l'objet de nombreuses recherches, c'est aussi le sujet de la présente étude. Elle pose l'hypothèse que certains résultats d'analyses préalables aideront les experts à élaborer les matrices Q. Dans le cas d'une approche inductive, l'accès aux données permet de se demander dans quelle mesure le fait d'analyser les données issues d'un questionnaire à choix multiple et de fournir les résultats aux experts peut les aider à établir une matrice Q plus valide et plus fidèle.

1.2 La revue de la littérature

La recension des écrits fait état de la situation de l'évaluation diagnostique. Étant donné que la notion d'attribut cognitif dans la matrice Q constitue la particularité de l'approche cognitive, une partie de la revue de littérature lui est consacrée. Par la suite, quelques modèles cognitifs sont présentés. La description de la façon dont ces modèles permettent de juger de la validité et de la fidélité des matrices Q clôture la revue de la littérature.

1.2.1 L'évaluation diagnostique

Il est naturel de se demander jusqu'à quel point les modèles cognitifs permettent un diagnostic plus approfondi et pertinent que les modélisations plus anciennes, mieux connues et éventuellement mieux maîtrisées. Il faut faire le point sur les approches diagnostiques existantes afin de voir comment ces nouveaux modèles viennent éventuellement les compléter.

D'après Legendre (2005), un test diagnostique recherche les forces et faiblesses des sujets dans leur apprentissage. Par la suite, la démarche d'évaluation aboutit à des décisions pour remédier aux problèmes rencontrés. Il est possible d'organiser des groupes d'étudiants avec des lacunes semblables ou des forces complémentaires. Dans le cas d'une évaluation à grande échelle, les forces et faiblesses de différentes populations d'étudiants peuvent être comparées. Dans tous les cas, le portrait individuel de l'apprentissage de chacun des sujets est l'objet premier.

La création de groupes d'étudiants est une intention possible de l'évaluation diagnostique et une analyse en grappes (*cluster*) répond à ce besoin. En effet, l'analyse en grappe consiste en une technique statistique qui vise à regrouper des sujets de profil semblable relativement à un ensemble de caractéristiques. Ainsi, une telle analyse donne la possibilité de créer des groupes de sujets en fonction de leur schéma de réponses. Néanmoins, elle ne permet pas directement de comprendre les éventuelles difficultés que partagent les sujets d'un même groupe. Cette approche ne fournit pas d'indications sur les caractéristiques de chaque individu ou de chaque regroupement. Elle renseigne seulement sur le fait que ses membres partagent une façon de répondre aux questions relativement semblable.

La représentation des forces et faiblesses des sujets sert à fournir une remédiation. Afin de guider la réflexion sur les différentes approches, les caractéristiques individuelles des sujets sur lesquelles baser le diagnostic sont définies ici de trois façons différentes : (1) la position de l'habileté du sujet sur un continuum, (2) les erreurs commises et (3) la structure du test. Ces trois définitions

permettent de classer les approches existantes, y compris les modèles cognitifs, et d'en définir les avantages et inconvénients.

1. La position de l'habileté du sujet sur un continuum

La théorie classique des scores (TCS), la théorie de la généralisabilité GEN) ou la théorie de réponse à l'item (TRI) ont été élaborées pour déterminer les items qui sont informatifs et qui discriminent bien les sujets. Elle sert à fabriquer des tests pour lesquels ces sujets peuvent être placés sur une échelle de mesure de leur habileté en fonction de leur score total individuel.

Il s'agit de mettre les étudiants en rang et le plus souvent de les comparer les uns aux autres. Le score total, le score vrai, le score univers ou l'habileté au sens de la théorie de réponse à l'item peuvent parfois être interprétés par comparaison à un seuil de maîtrise. Dans une perspective diagnostique, le mieux qui peut être fait est de prédire que certains s'en sortiront mieux que d'autres dans les activités qui suivront (Chipman, Nichols, , et Brennan, 1995). Pas de trace ici d'informations précises sur ce qui influence, en bien ou en mal, le cheminement de l'élève. Cette approche n'explicite pas les problèmes qu'il a rencontrés en répondant aux questions. Deux étudiants avec un schéma de réponses et des habiletés différents peuvent se retrouver au même endroit sur l'échelle d'habileté. Difficile donc de savoir sur quoi faire travailler chacun pour améliorer sa performance, ni comment constituer des groupes d'étudiants ayant des lacunes semblables ou des forces complémentaires. Les approches psychométriques traditionnelles sont basées sur le concept d'unidimensionnalité qui décrit difficilement la complexité de la cognition humaine (Gitomer et al., 1993).

2. Les erreurs commises pour caractériser le sujet

Pour Scallon (1988), une démarche de diagnostic peut passer par la détection systématique des erreurs par le correcteur qui code chaque réponse de chaque sujet. Une telle approche permet de décrire la performance de l'élève par l'intermédiaire d'un profil de scores. Le bénéfice est le

dépistage des élèves en difficulté auxquels une aide appropriée peut être apportée. Cette approche offre la possibilité de faire un portrait de chaque élève d'un groupe.

Une difficulté vient du fait que la procédure doit être faite individuellement pour chaque sujet, ce qui limite la quantité de sujets pouvant être traités lorsqu'elle n'est pas informatisée. Toutefois, dans le cas d'un test à choix multiple (QCM) les leurres peuvent être codés au préalable. Cette méthode comporte des limites importantes. Tout d'abord, le format impose une bonne réponse et trois ou quatre leurres. Il est probable que toutes les disciplines ne se prêtent pas facilement à la création de leurres similaires pouvant être codés selon des erreurs systématiques. De plus, l'usage de stratégies différentes pour répondre à un item peut rendre difficile le codage des erreurs. Enfin, les profils de scores peuvent se révéler ardu à interpréter (Scallan, 1988). L'approche cognitive offre la possibilité de faire partager un attribut par plusieurs items et de diagnostiquer une éventuelle faiblesse du sujet relativement à cet attribut. Il n'est alors pas nécessaire de coder les erreurs et l'interprétation est simple à réaliser puisque le sujet maîtrise ou pas l'attribut.

3. La structure du test pour caractériser le sujet

Dans une approche par objectifs, le tableau de spécification d'une épreuve de maîtrise met en évidence un découpage qui permettra d'établir un profil des sujets. Ce diagnostic se fait à partir des scores partiels totalisés sur un ensemble d'items congruents à un même objectif (Scallan, 2000) ou à la facette d'un domaine (GEN). Ils sont interprétés relativement à un seuil de réussite dans une approche critériée. Ils rendent compte des acquis spécifiques à chaque élève par une identification des points forts et des points faibles de chacun. Il faut recourir à des experts pour déterminer le découpage et le seuil de réussite.

Toutefois, « attitudes, processus cognitifs, stratégies métacognitives, stratégies de résolution de problèmes, pensée critique, capacité d'intégration [, de même que notion de] transfert [et de] compétence semblent [...] échapper à la façon traditionnelle d'explicitier les objectifs » (Scallan,

2000, p 153). Dans une approche par objectifs, les items d'un même regroupement dans le tableau de spécifications partagent un contenu commun qui est différent d'un regroupement à l'autre. Si c'est ce contenu qui sert de base au diagnostic, les caractéristiques des sujets sont assujetties à ce découpage des items lié à la notion d'objectif pédagogique.

Une analyse factorielle, en grappes ou discriminante peut déceler une structure dans le test . Corter (1995) a appliqué une analyse en grappes (*cluster*) aux items d'un test afin d'en explorer la structure cognitive. Selon lui, les tests diagnostiques visent à évaluer des états d'habiletés cognitives très précises et donc souvent dichotomiques par nature : l'habileté est possédée ou pas. Comme rien ne garantit l'utilisation d'une méthode plutôt qu'une autre dans la résolution des items, il est difficile d'élaborer la structure théorique sans en référer aux données. De plus, les habiletés sont probablement reliées selon une certaine hiérarchie temporelle ou logique.

Corter a posé l'hypothèse que les groupes d'items identifiés par une analyse en grappes se trouvent reliés à cause des habiletés qui leur sont particulières. Même s'il n'a pas réussi à valider son hypothèse, cette approche offre la possibilité de définir une structure dans un test sans passer par les objectifs. Les experts doivent s'appuyer sur les résultats des analyses pour identifier ce que partagent les items regroupés. Il est alors théoriquement possible de calculer des scores partiels pour établir un profil des sujets.

Dans le cas des modèles cognitifs, la matrice Q fournit une structure au test. Elle offre l'avantage d'une grande variété des attributs cognitifs utilisables relativement au profil de scores lié à une épreuve de maîtrise. Les experts ne sont pas limités par un tableau de spécification. Ils ne sont pas non plus contraints par la définition stricte d'objectif pédagogique. La liberté apportée par la matrice Q relativement à un tableau de spécifications vient donc de la souplesse de la définition des attributs, mais aussi des liens items-attributs.

Le découpage des items selon des objectifs est exclusif. La matrice Q peut relier les items à un nombre variable d'attributs. Il en résulte une structure beaucoup plus souple. Chaque attribut crée un lien entre les items qui lui sont associés. Un item est en lien avec certains items selon un attribut et avec d'autres selon un autre attribut. Cette approche s'accorde avec la tendance actuelle qui vise l'intégration des matières. Elle permet de dépasser une vision atomisée pour laquelle la validation de contenu impose à un item de n'appartenir qu'à un seul groupe. Le calcul d'un indice de congruence item-objectif, tel que celui proposé par Hambleton (Crocker, et Algina, 1986), ne permet donc pas d'assurer la validation de Q.

Même si ces approches diagnostiques semblent différentes, une constante se retrouve. Le recours à des experts est essentiel. Qu'ils codent les erreurs, déterminent des items congruents à un même objectif, définissent un contenu latent à un regroupement d'items, établissent un seuil de réussite ou élaborent une matrice Q, leur travail pose les bases du diagnostic.

L'avantage de la matrice Q vient de la liberté qu'elle laisse aux experts dans son élaboration. Cette liberté pose toutefois problème lorsque vient le moment d'élaborer la matrice Q. Sans balises strictes, obtenir une matrice qui fasse consensus auprès d'un panel d'experts devient difficile. De plus, on peut s'interroger sur la façon d'obtenir cette matrice. Il est possible de fournir une liste d'attributs aux experts, ils n'ont ensuite qu'à établir les liens avec les items. Il est possible de leur demander de fabriquer eux-mêmes cette liste. L'étude des erreurs commises, les dimensions latentes contenues dans les items sont autant d'informations qui peuvent jouer un rôle dans le diagnostic. Pour offrir une remédiation à un étudiant en difficulté, il faut cibler a priori sur quelles dimensions on peut la lui offrir avant de déterminer s'il en a ou non besoin. Il faut définir ces caractéristiques pour qu'elles expliquent le cheminement du sujet.

S'il est difficile d'élaborer la matrice Q, elle constitue tout de même l'avantage premier des modèles cognitifs sur les autres approches. Elle peut inclure des attributs en liens avec des processus

cognitifs, des savoirs, des savoir-faire, des attitudes, des stratégies, etc. Elle n'oblige pas à créer une partition des items. Un deuxième avantage tient dans la forme du diagnostic puisque chaque sujet maîtrise ou non les attributs de la matrice. Le diagnostic est facile à interpréter si les attributs sont clairement définis et en nombre limité. La richesse d'un diagnostic relativement à une liste d'attributs cognitifs habilement élaborée justifie l'importance et la pertinence de cette recherche. Il serait facile d'en faire la démonstration à l'aide d'un exemple fictif dès à présent, mais il paraît plus intéressant de revenir à la fin de cette recherche sur la forme que prendra le diagnostic une fois que les données choisies auront été modélisées. De cette façon, les résultats de cette recherche fourniront des exemples de diagnostic (voir 3.4) qui illustreront les avantages de cette approche relativement aux autres approches.

L'approche cognitive vise donc à identifier et inférer les différences de structure des individus dans les processus mis en œuvre pour répondre aux questions. Elle se base sur l'identification préalable des attributs cognitifs ciblés et sur les réponses des sujets au test. Elle présente une approche diagnostique plus riche que les méthodes traditionnelles. Même si la manière d'élaborer la matrice Q pose un défi, les modèles cognitifs constituent un apport important à l'évaluation diagnostique.

1.2.2 La notion d'attribut cognitif et l'élaboration de la matrice Q

Les attributs cognitifs représentent des savoirs, des savoir-faire, des habiletés qui peuvent être combinés afin d'obtenir la réponse correcte à l'item. La perspective d'une interprétation diagnostique implique de retenir des attributs simples et pas trop nombreux (Chipman, Nichols, et Brennan, 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 non maîtrise de ces attributs.

Gierl (1997) définit un attribut par « la description de procédures, d'habiletés, de stratégies et de connaissances qu'un sujet doit posséder pour résoudre un test. Les attributs ne sont pas observables directement, ils doivent être inférés 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 réfère à une notion ou à un micro niveau d'habileté en plus d'inclure des variables plus générales liées au contexte, à la forme de la question ou au processus de réponse.

Un ou plusieurs experts peuvent déterminer les attributs. Dans une approche inductive, ils se basent sur l'analyse de la tâche et/ou sur les réponses écrites ou orales des étudiants. Les recherches de Tatsuoka (1995), Tatsuoka, Birenbaum, Lewis et Sheelan (1993), Tatsuoka et Tatsuoka (1997), Gierl (1997), Milewsky et Baron (2002) fournissent des exemples de listes d'attributs reliés à des items en mathématiques pour les niveaux primaire ou secondaire. Tatsuoka et ses collègues (1993) en particulier donnent les conditions de l'élaboration des attributs cognitifs reliés à une épreuve du *scholastic aptitude test* (SAT). À partir des différents types de contenus ou de questions, des différents formats de réponses ou des ressources nécessaires, ils proposent puis affinent une première liste d'attributs basée sur la littérature.

Les recherches de Buck, Tatsuoka et Kostin (1997) et Buck et Tatsuoka (1998) donnent des exemples de listes d'attributs relatives à la lecture ou à la compréhension d'une langue seconde. Elles utilisent une recension des taxonomies de la littérature et quelques études empiriques. Le processus mis en œuvre pour la création de la matrice Q est très détaillé dans l'article de Buck et Tatsuoka (1998). Il fournit au lecteur les références pour aboutir à une première liste d'attributs. Leur nombre est ensuite réduit. Les attributs reliés à un seul item, qui ont une faible corrélation avec la difficulté de l'item ou qui sont fortement corrélés aux autres attributs sont supprimés.

Jang (2005) décrit l'élaboration d'une liste d'attributs cognitifs reliés à la lecture en anglais langue seconde et au test NG TOEFL (*New Generation Test of English as a Foreign Language*). L'analyse du texte, la dimensionnalité du test, les caractéristiques de la tâche et les stratégies verbalisées par les sujets lui permettent d'élaborer cette liste.

Certaines recherches utilisent des tests et des listes d'attributs provenant d'autres études (par exemple, Dimitrov, et Raykov, 2003). Rupp (2007) propose une liste d'attributs basée sur la littérature à dix experts et leur demande d'estimer l'importance de chaque attribut relativement à chaque item sur une échelle de 0 à 3 (dans laquelle 0 représente le fait que l'attribut n'est aucunement lié à l'item et 3 le fait que l'attribut est indispensable à la solution de l'item). Parfois, l'intérêt des psychométriciens s'axe plus sur les modèles eux-mêmes que sur les méthodes d'élaboration de la matrice Q. Ils utilisent alors des données simulées (par exemple, Hai Jiang, 1996).

Les modèles cognitifs peuvent servir à élaborer un diagnostic individuel même lorsque ce n'est pas le but premier de l'évaluation. Par exemple, des recherches rapportent l'utilisation des données des grandes enquêtes internationales comme la troisième évaluation internationale sur les mathématiques et les sciences (TEIMS) ou le programme international pour le suivi des acquis des élèves (PISA) ou encore le *preliminary aptitude scholastic test* (PSAT) (Hartz, 2002; Stout, 2002). Dans un tel contexte, un diagnostic individuel n'est pas une pratique courante, mais est source d'intérêt (Birenbaum, Tatsuoka, et Yamada, 2004; Gitomer, et Rock, 1993). L'élaboration de la liste des attributs cognitifs part alors d'items déjà existants dans une approche inductive.

Un point important à considérer est le nombre d'attributs à utiliser. Le tableau 1 donne quelques exemples de nombres d'attributs relativement au nombre d'items. Ces recherches empiriques appliquent deux modèles cognitifs à des données réelles ou simulées.

Les nombres de sujets, d'items et d'attributs varient beaucoup d'une étude à l'autre. Il y a entre deux et quatre fois plus d'items que d'attributs. Plusieurs études ont un nombre d'items deux fois plus grand que le nombre d'attributs. Le recours à au moins trois ou quatre centaines de sujets semble nécessaire pour les deux modèles présents dans le tableau 1.

Tableau 1.

Quelques exemples de recherches empiriques de la littérature

Références	Nombre de sujets	Discipline	Nombre d'items	Nombre d'attributs	Rapport approx. (it./att.)	Modèle
Hartz, 2002	585	Mathématiques (PSAT)	40	16	Entre 2 et 3	RUM
Yan, Almond, & Mislevy, 2003	325	Mathématiques	15	5	3	RUM
Hai Jiang, 1996	1000	Simulation	10	5	2	RUM
Tatsuoka, 1993	2335	Mathématiques (SAT)	60	14	4	Rule Space
Buck, Tatsuoka, & Kostin, 1997	5000	Lecture	40	24	2	Rule Space
Buck et Tatsuoka, 1998	412	Compréhension (langue)	35	17	2	Rule Space

On peut se demander qui a l'expertise appropriée pour élaborer une matrice Q. La littérature sur les modèles cognitifs décrit peu les experts, mais celle sur l'établissement de seuils de réussite fournit des descriptions intéressantes. Les experts doivent être familiers avec les étudiants concernés par le test (Reckase, 2000). D'après Raymond et Reid (2001), les experts doivent posséder les connaissances et compétences liées au contenu du test. Ils doivent donc être des spécialistes du domaine évalué et avoir une formation dans ce domaine. Ils doivent avoir une bonne connaissance de la population de sujets qui passe le test et une bonne compréhension des habiletés nécessaires pour répondre correctement aux items. Toutefois, les enseignants ont souvent une expérience limitée du type d'items sur lesquels ils doivent travailler, de même qu'en psychologie cognitive (McGinty,

1996). Des spécialistes du domaine évalué et enseignant à des étudiants comparables à ceux évalués par le test devraient pouvoir élaborer les matrices Q. Il faut donc s'assurer d'une formation et d'une expérience adéquates ainsi que d'une bonne conception de ce qui est attendu, tant au niveau du contenu que du mécanisme de collecte de l'information.

Se pose également la question du nombre d'experts qu'il faut recruter. La littérature sur les modèles cognitifs donne là encore peu d'informations à ce sujet. Jang (2006) a, par exemple, élaboré ses attributs seule, mais les a faits valider par cinq experts. Les références dans la littérature concernant l'établissement de standards font mention de groupes d'experts de tailles variables allant d'un peu plus de dix (par exemple, Mc Ginty, 1996; Plake, 2006) à des nombres plus importants comme vingt quatre pour Plake (1997; 1998) et même soixante-dix pour Buckendahl (2000).

En résumé, malgré le peu de recherches sur l'étude de la validité et la fidélité des matrices Q, la littérature fournit quelques recherches empiriques sur les modèles cognitifs. Elle offre des indications sur le nombre d'items et d'attributs à retenir. Elle procure des explications sur la notion d'attributs cognitifs et des exemples de listes, particulièrement en mathématiques.

1.2.3 Quelques modèles

Fischer (1973) a utilisé en premier une décomposition en attributs de base pour estimer la difficulté des items d'un test. Son *Linear Logistic Test Model* (LLTM) est une extension du modèle unidimensionnel de Rasch (Fischer, 1973; Embretson, 1985a; 1985b; 1999; Embretson, et Reise, 2000; DiBello, Stout, et Rousseau, 1995; Bechger, Verstralen, et Verhelst, 2000; Stout, 2002). Ce modèle postule que le paramètre de difficulté d'un item peut être exprimé par une combinaison linéaire d'attributs cognitifs. Il suppose que tous les sujets utilisent les mêmes processus cognitifs pour résoudre tous les items (Fischer, 1973; Sijtsma, et Verweij, 1999). Le modèle LLTM caractérise les sujets par la position de leur habileté sur un continuum tout comme le modèle de Rasch de la TRI

dont il est issu. Il estime la valeur de leur habileté sous forme d'une variable latente continue, notée θ .

Embretson (1985) a développé une série de modèles qui prennent en compte les attributs cognitifs. Le *Multicomponent Latent Trait Model* (MLTM) opérationnalise le concept selon lequel il faut avoir la bonne information sur plusieurs attributs cognitifs pour répondre correctement à un item (Embretson, 1985b; Embretson et al., 2000). Combinant cette approche et celle de Fischer, le modèle *General component Latent Trait Model* (GLTM) tient compte des difficultés des attributs. Ce modèle fournit un outil performant dans une perspective diagnostique (Embretson, 1993, 1999; Embretson et al., 2000). Il a pour but l'estimation de la difficulté de l'item, mais aussi de l'habileté du sujet relativement à chaque attribut. Les modèles MLTM et GLTM caractérisent les sujets sous la forme d'un vecteur d'habiletés, variables latentes continues, pour chacun des attributs de la liste de départ. Ces modèles offrent un diagnostic pour chaque sujet en fonction de chacun des attributs. Ils font intervenir de nombreux paramètres difficiles à estimer et à interpréter (Hartz, 2002).

Le modèle *rule space* de Tatsuoka permet de classer les sujets dans des classes latentes qui reflètent leur état de connaissance (Tatsuoka, 1983; Embretson et al., 2000). Le sujet ne peut répondre correctement à un item que s'il maîtrise tous les attributs cognitifs qui lui sont reliés. Le modèle suppose qu'un sujet ayant un schéma particulier d'attributs maîtrisés va les utiliser pour répondre et non pas deviner ou faire des fautes d'étourderie (Tatsuoka, 1983, 1995). Il est défini pour deux dimensions : le paramètre de compétence du sujet noté θ et un indice représentant la façon de répondre noté ζ (Tatsuoka, 1983, 1995; Embretson et al., 2000). À partir de la liste des attributs établie par les spécialistes, l'algèbre de Boole permet de déterminer tous les états de connaissance possibles selon la maîtrise ou non de chacun des attributs. Le fait que certains attributs soient pré-requis à d'autres minimise le nombre d'états de connaissance idéaux, sur la base du jugement des experts et grâce à de nombreux théorèmes (Tatsuoka et al., 1993). Dans un plan cartésien, il est alors

possible de représenter graphiquement les couples (θ, ζ) de chaque sujet et de chaque état idéal. La distance de chaque sujet à chaque état idéal permet de décider du classement de chaque sujet dans un état particulier. Si le point caractérisant le sujet est trop éloigné de tous les points idéaux, le sujet correspondant peut ne pas être classé (Tatsuoka, 1983; Buck et al., 1998). L'état de connaissance est représenté par un vecteur discret de la forme $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$, dans lequel k est le nombre d'attributs et α_i peut prendre la valeur 0 ou 1. La particularité vient du fait que le vecteur est obtenu par classement plutôt que par estimation. L'interprétation est facilitée par la visualisation graphique de l'espace contenant les points représentant les sujets et les états de connaissance idéaux. Le diagnostic individuel peut se faire de façon qualitative en énonçant quels attributs sont maîtrisés ou non par chaque sujet. Le modèle *rule space* présente l'avantage de ne pas avoir de nombreux paramètres à estimer. Il peut être appliqué à des données dès qu'une liste d'attributs cognitifs liés aux items peut être définie (Rupp, 2007).

Le modèle DINA (*deterministic inputs, noisy "and" gate model*) (Junker, et Sijtsma, 2001; De la Torre, et Douglas, 2004) est un modèle conjonctif à classes latentes restreintes qui présente l'avantage de sa simplicité. S'il pose l'hypothèse que tous les attributs liés à l'item dans la matrice Q sont requis pour répondre correctement à l'item (modèle non compensatoire), il prend en considération le fait qu'un individu maîtrisant tous ces attributs peut obtenir une mauvaise réponse à l'item. Ainsi ce modèle permet l'estimation de deux paramètres par item correspondant aux facteurs de pseudo chance et d'étourderie. Ces deux paramètres rendent possible l'estimation de la sensibilité de la matrice Q au manque de fidélité des attributs cognitifs utilisés par les sujets. Il offre un diagnostic des sujets relativement aux attributs de la matrice Q sous la forme d'un vecteur $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$ dans lequel α_i est dichotomique.

Le modèle de fusion (DiBello et al., 1995) ajoute la modélisation détaillée de la source des déviations systématiques par rapport au schéma idéal. Un sujet est caractérisé, selon une approche

psychométrique, par un trait latent continu θ , qui peut être unidimensionnel ou multidimensionnel. De plus, l'approche cognitive, suppose qu'il possède ou non chacun des attributs, ce qui revient à le caractériser par un vecteur $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$ dans lequel α_i est dichotomique. La probabilité de dévier des schémas établis par la matrice Q est modélisée en fonction de θ , trait latent continu, et de α , vecteur latent discret, là où le *rule space* n'utilise que θ (DiBello et al., 1995; Hartz, 2002; Stout, 2002).

Ce modèle non compensatoire pose le postulat de l'existence d'une structure possédant un certain nombre (inconnu) de classes latentes. L'idée centrale est que dans la réalité, le comportement des sujets sera différent de ce qui a été choisi au départ dans la matrice Q . Quatre sources de variations sont incluses dans le modèle : (a) la stratégie (*strategy*): choix d'une autre stratégie que celle de Q , (b) le manque d'attributs (*completeness*): d'autres attributs que ceux de Q peuvent aider à répondre, (c) le fait qu'un attribut soit mal utilisé (*positivity*): attribut non possédé, mais utilisé et attribut possédé, mais mal utilisé, et d) l'erreur aléatoire telle une erreur de transcription ou d'attention. Ces quatre paramètres sont en lien avec la validité et la fidélité de la matrice Q .

Il est important de comprendre que le fait qu'un attribut soit mal utilisé est plus subtil que de simplement deviner ou de faire une erreur d'étourderie (Stout, 2002). En effet le partage des sujets en deux catégories, caractérisées par α_i qui peut prendre la valeur 0 ou 1, amène à considérer que certains sujets peuvent utiliser de façon routinière et avec succès un attribut sans nécessairement le maîtriser ($\alpha_i=0$) ou maîtriser un attribut ($\alpha_i=1$), mais ne pas répondre correctement, par exemple parce que la question les décourage. L'estimation de ces nombreux paramètres peut s'avérer difficile à obtenir lors de l'application du modèle à des données (Hartz, 2002). Chaque sujet est caractérisé par la valeur de son habileté estimée sur une échelle continue, comme dans les modèles de la TRI,

mais aussi par un vecteur $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$ qui est alors estimé. Le diagnostic peut se faire par la liste des attributs maîtrisés et non par chaque sujet lorsque les paramètres sont estimables.

La méthode des réseaux de Bayes a été développée dans la perspective du diagnostic médical (Mislevy, 1995; Buck et al., 1997). Tous les modèles à variables latentes qui précèdent partent du principe que la valeur de l'habileté du sujet implique une distribution de probabilités des scores qui peut se formaliser par $P(X|\theta)$. Le théorème de Bayes donne l'opportunité de calculer $P(\theta|X)$ grâce à la relation $P(\theta|X=x) \approx P(X=x|\theta).P(\theta)$. Il s'agit de définir les distributions de probabilités de toutes les quantités qui interviennent dans le modèle, qu'elles soient observables ou non. Chaque sujet est caractérisé par son état de connaissance sous la forme d'un vecteur latent θ et d'un vecteur de ses scores observés X . Il est possible théoriquement d'estimer $P(\theta|X)$, toutefois les calculs peuvent très rapidement devenir complexes. Ils risquent de ne pas aboutir à l'estimation des paramètres si l'on ne simplifie pas les relations entre les attributs et entre les attributs et les items (Levy, et Mislevy, 2004; Mislevy, 1993, 1995, 2004; Mislevy, Almond, Yan, et Steinberg, 2000). La méthode consiste à relier les attributs et les items selon un réseau tenant compte des relations qui interviennent dans la résolution des items, des relations d'ordre qui existent entre les attributs, des combinaisons d'attributs nécessaires, des diverses stratégies possibles etc. Le modèle des réseaux de Bayes permet de caractériser chaque sujet par l'estimation d'un vecteur $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$ pour un diagnostic prenant la forme d'une liste d'attributs maîtrisés et non maîtrisés par chaque sujet.

Des sept modèles présentés, seul le modèle LLTM de Fischer ne vise pas le diagnostic individuel des sujets en fonction des attributs cognitifs. Les six autres modèles permettent, au moins théoriquement, l'estimation de la maîtrise ou non maîtrise des attributs cognitifs pour chaque sujet. Le modèle le plus facile à appliquer à des données est le modèle *rule-space* qui procède par classement là où les autres modèles procèdent par estimation. Les modèles d'Embretson sont des

modèles complexes pour lesquels il est difficile d'estimer et d'interpréter les paramètres et qui se basent sur l'hypothèse que chaque attribut est requis par chaque item avec une matrice Q ne comportant que des 1 (Hartz, 2002). Le modèle DINA contient peu de paramètres à estimer et ceux-ci sont aisés à interpréter. Le modèle de fusion contient de nombreux paramètres à estimer, mais dont l'interprétation éventuelle est facile. Enfin l'élaboration du réseau est coûteuse puisqu'elle nécessite la collaboration des experts et des psychométriciens dans le cas du modèle des réseaux de Bayes (Yan, Almon, et Mislavy, 2003).

Les modèles *rule space*, DINA et le modèle de fusion posent l'hypothèse que le sujet doit posséder tous les attributs reliés à l'item pour répondre correctement (modèles non compensatoires). Une force sur un attribut ne compense donc pas une faiblesse sur un autre attribut pour un item. La forme du diagnostic est toutefois la même pour ces six derniers modèles puisqu'ils décrivent le sujet par sa maîtrise ou non maîtrise de chaque attribut cognitif de la matrice Q. Par contre ces modèles diffèrent considérablement lorsqu'il s'agit de jeter un regard critique sur la matrice et de porter un jugement sur sa validité et sa fidélité.

1.2.4 La validité et la fidélité de la matrice Q

Les modèles cognitifs permettent d'inférer les processus cognitifs mis en œuvre par les sujets pour répondre aux items à l'aide d'une matrice Q qui lie items et attributs. C'est d'ailleurs ce qui les différencie des autres approches diagnostiques. La matrice est fabriquée par des experts et le choix des attributs ainsi que des liens items-attributs peut varier d'un expert à l'autre.

La structure de la matrice Q limite la forme du diagnostic puisqu'elle impose de faire un choix lorsqu'un item peut être résolu de diverses façons. Autrement dit, une seule combinaison d'attributs apparaît dans la matrice pour chaque item. Ceci implique que la possibilité de réussir un item au moyen de plus d'une stratégie remet en cause la validité du modèle si les sujets utilisent une

autre stratégie que celle prévue. Le choix fait dans la matrice a un impact sur la validité des mesures produites et la qualité du diagnostic individuel.

Dans le cas du modèle LLTM (Fischer, 1973), la matrice Q ne joue un rôle que dans la description de la difficulté des items car il ne relie pas les attributs aux sujets. Il ne fournit aucun indicateur de validité ou de fidélité de la matrice. Dimitrov et Raykov (2003) proposent deux méthodes externes de validation de la matrice Q pour ce modèle. La première consiste à comparer la difficulté des items obtenue grâce au modèle de Rasch à la difficulté des items obtenue par le modèle LLTM. La deuxième méthode consiste à établir un sentier causal entre les items (*path diagram*). Celui-ci est basé sur la matrice Q et sur l'hypothèse qu'un item faisant appel à certains attributs sera relié de façon unidirectionnelle à un autre item faisant appel aux mêmes attributs associés à un ou plusieurs autres (Dimitrov, 1994). De cette façon, tous les items peuvent former un réseau hiérarchique. Il peut être testé par une analyse d'équations structurelles pour données dichotomiques en utilisant la matrice de corrélations tétrachoriques des réponses aux items. L'ajustement des données au modèle permet alors de valider la matrice Q. Enfin, les items devraient avoir une difficulté croissante au fur et à mesure de la progression le long du réseau hiérarchique (Dimitrov, 1994). La vérification de cette hypothèse aboutit également à une validation des attributs et de leurs liens avec les items.

Les modèles MLTM et GLTM caractérisent les sujets et les items avec l'hypothèse que chaque attribut est requis par chaque item. Ils ne permettent pas de juger de la pertinence des liens entre les items et les attributs puisque tous les liens possibles existent dans la matrice.

Le modèle *rule space* classe les sujets dans des classes latentes discrètes. Ce modèle n'offre pas d'indicateurs pour s'assurer de la validité et de la fidélité de la matrice Q. Toute déviation par rapport au modèle est classée comme erreur aléatoire. Toutefois, le taux de classification des sujets dans les états idéaux est un indicateur de la qualité de la matrice (Buck et al., 1997, 1998). Ainsi en

procédant à plusieurs analyses successives, il est possible de raffiner la liste des attributs. Il faut aboutir à une matrice qui permet de classer un pourcentage de sujets jugé satisfaisant, par exemple 80 % et plus selon Buck et ses collègues. De ce point de vue, même si le modèle ne permet pas directement de porter un jugement sur Q, le fait de l'appliquer plusieurs fois de suite avec des matrices améliorées à chaque condition, donne une certaine confiance dans le choix de Q. De plus, Tatsuoka (2005) suggère d'appliquer a posteriori diverses analyses statistiques, telles des analyses factorielles, en composantes principales ou des analyses de corrélation afin d'étudier la dimensionnalité du test, le regroupement des attributs et le pouvoir discriminant des attributs. Ces diverses analyses ont pour but une amélioration et une validation de la liste des attributs finalement retenus.

Le modèle DINA estime deux paramètres par item : un paramètre de pseudo chance et un paramètre lié aux erreurs d'étourderie. En pratique, une valeur élevée du paramètre de pseudo chance indique que de nombreux sujets ne maîtrisant pas les attributs reliés à l'item donnent une bonne réponse. De la même façon, une valeur élevée du paramètre lié aux erreurs d'étourderie indique que de nombreux sujets qui maîtrisent les attributs reliés à l'item ont une réponse fausse. Dans les deux cas, une grande valeur montre que les liens établis dans la matrice entre cet item et les attributs sont de mauvaise qualité (Templin, 2004). Si de nombreux sujets trouvent la bonne réponse par hasard ou se trompent sans raison, on peut penser que ce résultat ne sera pas reproductible à une autre occasion. En même temps, un taux élevé de tels sujets indique peut-être que ce ne sont pas les bons attributs qui décrivent l'item. Dans ce cas, le fait de maîtriser ou pas l'attribut ne prédit pas convenablement le succès ou l'échec à la question. Ces deux indicateurs donnent donc à la fois une information sur la validité et sur la fidélité de la matrice Q.

Le modèle de fusion explicite quatre sources de variation par rapport à la matrice Q. La validité de la matrice Q dépend de la relation entre la réponse fournie à l'item et le processus pour y

arriver. Le fait qu'un sujet puisse utiliser une autre stratégie que celle proposée par Q ou que des attributs ne soient pas présents dans la matrice peut être, au moins en partie, attribué à un manque de validité du modèle (Junker, 1999). Les paramètres concernant (a) la stratégie (*strategy*) et (b) le manque d'attributs (*completeness*) sont des indicateurs de la validité de la matrice Q. D'un autre côté, le fait que l'utilisation d'un attribut manque de stabilité est associé à la notion de fidélité (Junker, 1999). Les paramètres concernant (c) le fait qu'un attribut soit mal utilisé (*positivity*) et (d) l'erreur aléatoire sont plutôt des indicateurs de la fidélité de la matrice Q. Le modèle de fusion fournit donc deux indicateurs de la validité et deux indicateurs de la fidélité de la matrice Q.

Dans le cas des réseaux de Bayes, le réseau est réalisé par les experts à fois théoriquement et empiriquement (Mislevy, Steinberg, et Almond, 2003). Il relie les items aux attributs, mais aussi les items et les attributs entre eux. Le travail que demande sa fabrication prend appui sur l'expertise de ses créateurs, mais aussi sur les distributions de probabilités relatives à chaque nœud du réseau. Cette façon de faire fournit une validation de la matrice Q. Aucun paramètre estimé n'en est toutefois un indicateur.

Juger de la validité et de la fidélité des attributs cognitifs proposés par les experts et reliés à un ensemble d'items est essentiel afin que les modèles cognitifs fournissent des résultats significatifs et utiles (Corter, 1995). Les méthodes de validation proposées dans la littérature sont le plus souvent itératives et appliquées a posteriori. En effet pour valider la matrice Q lors de l'utilisation des modèles LLTM ou *rule space*, il est nécessaire d'analyser les résultats obtenus après application du modèle à l'aide d'outils qui n'ont rien à voir avec le modèle lui-même. La méthode des réseaux de Bayes valide chaque étape au fur et à mesure, sans fournir d'indicateurs de la validité et de la fidélité de Q. Le modèle DINA offre deux paramètres pour juger à la fois de la validité et de la fidélité de Q. Le modèle de fusion est le seul à fournir quatre indicateurs de la validité et de la fidélité.

1.3 L'élaboration du cadre conceptuel

1.3.1 Le cadre conceptuel

Cette recherche s'intéresse à la fabrication de la matrice Q par des experts dans une approche inductive. Cette élaboration peut être influencée par les facteurs contextuels suivants : (a) les données: le domaine évalué et la forme du test, (b) la méthode de travail choisie, (c) la tâche demandée et (d) les conditions dans lesquelles les experts travaillent. La figure 2 décrit l'univers des possibilités.

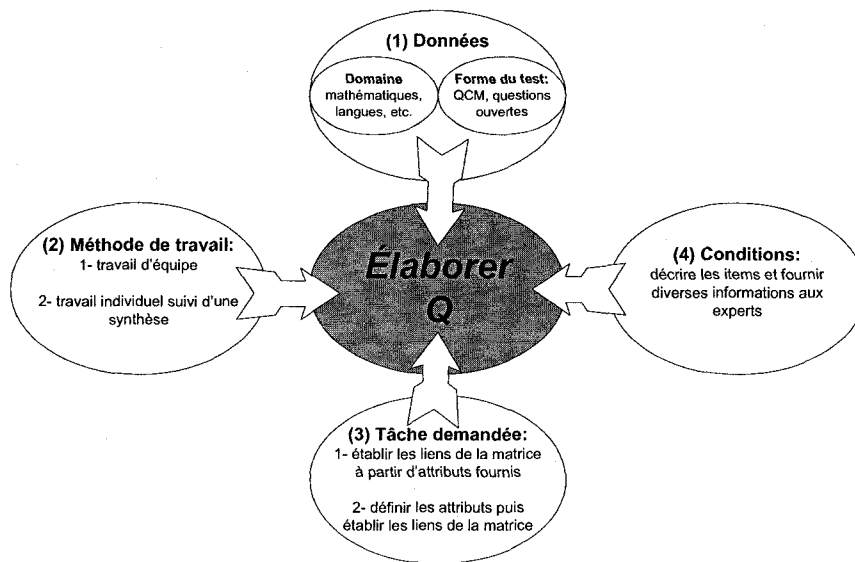


Figure 2. Les facteurs susceptibles d'influencer l'élaboration de la matrice Q

Chacune de ces quatre sources d'influence est abordée à la lumière de la littérature.

1. En ce qui concerne les données, certaines disciplines se prêtent probablement mieux que d'autres à une structure telle que celle imposée par la matrice Q. Les domaines de la lecture ou de la compréhension d'une langue seconde fournissent quelques exemples de recherches empiriques (par exemple, Buck et al., 1997; 1998; Jang, 2005). C'est toutefois en mathématiques que la

littérature offre le plus d'exemples d'application de modèles cognitifs (par exemple, Dimitrov et al., 2003; Gierl, 1997; Gitomer et al., 1993; Tatsuoka, 1983; 1995; Tatsuoka et al., 1993; 1997).

La littérature décrit la compétence en mathématiques et ses liens avec les processus cognitifs (par exemple, Boaler, 2002; Garofalo, et Lester, 1985; Kilpatrick, 2001; Noss, Healy, et Hoyle, 1997).

2. La méthode d'élaboration de la matrice Q que l'on retrouve dans la littérature consiste à demander aux experts d'ébaucher ensemble une première matrice. Le chercheur et ses experts la font ensuite évoluer à la lumière de différents résultats. Ceux-ci peuvent provenir du modèle cognitif ou être externes au modèle. Les experts peuvent également travailler individuellement avant de regrouper leurs travaux.
3. La tâche demandée peut consister à fournir les attributs aux experts et à ne leur demander que d'établir les liens dans la matrice. Une autre approche peut les amener à définir eux-mêmes les attributs avant de les relier aux items.
4. Il est possible de poser l'hypothèse d'une influence des conditions dans lesquelles les experts travaillent sur l'élaboration de la matrice Q. Les recherches empiriques de la littérature n'abordent pas cet aspect. Une approche inductive donne accès aux données du test et la possibilité de fournir aux experts divers résultats liés aux analyses de ces données. C'est la littérature sur l'établissement de standards qui suggère que l'accès à ces résultats pourrait apporter un support aux experts dans leur tâche d'élaboration de la matrice Q (par exemple, Buckendahl et al. (2002) ou Sadesky (2003)).

Cette dernière source d'influence est l'objet de la présente étude. La mise en place et le contrôle de plusieurs conditions d'élaboration de la matrice Q visent à déterminer si certaines d'entre elles influencent positivement ou négativement l'élaboration de la matrice Q. Étant donné que cette approche est originale, elle suggère de choisir des données issues d'un domaine qui a fait l'objet de

nombreuses recherches empiriques. C'est la raison du choix d'un test à choix multiple en mathématiques.

Étant donné que cette étude vise à comparer des matrices élaborées dans plusieurs conditions, l'approche retenue consiste à faire travailler les experts individuellement pour ensuite synthétiser leur travail en cherchant à obtenir un consensus. Cette approche vise à exploiter l'expertise de chacun avant de le soumettre à l'influence du groupe. De cette façon, les matrices individuelles ainsi que les matrices synthèses seront disponibles. Pour finir, étant donné le peu de balises fournies par la littérature, le choix est fait d'interférer le moins possible dans le processus et de demander aux experts de déterminer les attributs et les liens et de documenter la démarche.

Il faut définir les conditions expérimentales à mettre en place. Le choix de faire travailler les experts individuellement avant de regrouper leurs travaux implique de déterminer théoriquement un moyen d'obtenir un consensus entre les experts à l'issue de leur travail individuel. Quant au modèle cognitif utilisé, il doit permettre de comparer la validité et la fidélité des matrices (individuelles et synthèses) propres aux conditions expérimentales.

Les paragraphes qui suivent présentent une méthode pour élaborer les matrices individuellement puis pour les regrouper en obtenant un consensus de la part des experts. Les conditions expérimentales qui peuvent être mises en place sont ensuite identifiées à partir de la littérature suivie par la présentation du modèle cognitif retenu.

1.3.2 Une méthode pour élaborer les matrices Q

Il suffit de penser aux méthodes pour établir les standards ou les seuils de réussite pour reconnaître la nécessité d'un panel d'experts. Dans un premier temps, les experts travaillent individuellement à l'élaboration de leurs matrices qui contiennent des attributs libellés différemment par chacun et en quantité éventuellement différente.

L'établissement de standards requiert la synthèse du travail individuel des participants. Comme le souligne Reckase (2001), la littérature est habituellement peu prolixe sur les méthodes concrètement utilisées et appliquées. Elle offre tout de même quelques pistes pour réaliser la synthèse des matrices. Les experts qui participent à l'établissement de standards doivent produire des résultats appropriés et opérationnalisables. Le fait de faire réfléchir les experts collectivement suite à un travail individuel va dans le sens d'un *feed-back* normatif utilisant le groupe comme une norme. Cette approche est courante en établissement de standards (Reckase, 2000).

Il faut déterminer une procédure analogue pour réaliser une matrice synthèse représentative. Elle comporte deux niveaux lorsque les experts élaborent eux-mêmes leurs attributs : (a) obtenir une liste d'attributs qui fasse consensus et (b) définir une règle pour établir les liens avec les items.

Dans le cadre du *National Center on Educational Outcomes*, Vanderwood, Ysseldyke et Thurlow (1993) ont développé une méthode pour déterminer, de façon quantitative et objective, l'opinion d'un petit groupe d'individus (généralement entre 4 et 10 personnes). Suite à l'élaboration individuelle d'une liste d'énoncés, cette méthode permet un classement selon l'importance accordée à chaque énoncé et d'arriver à un consensus. La méthode, appelée *Multi-Attribute Consensus Building* (MACB) est dérivée d'un modèle d'évaluation lui-même appelé *Multi-Attribute Utility* (MAU). Ce dernier vise à faciliter la prise de décisions face à de nombreuses options ou alternatives (Lewis, Erickson, Johnson, et Bruininks, 1991).

Le processus implique trois phases dans la construction du consensus : (a) la génération des énoncés, (b) la session de travail visant le consensus et enfin (c) la synthèse. Après avoir sélectionné les participants, l'objectif de la phase 1 consiste à leur faire élaborer individuellement une liste d'indicateurs selon un cadre de référence commun. Ces listes sont ensuite regroupées dans un même document afin que chaque énoncé apparaisse une seule fois. Ce document sert de base de la session

de travail de la phase 2 qui peut durer entre une demi-journée et quatre jours selon l'ampleur de la tâche. Chacun des membres du groupe doit bien comprendre l'origine des différents énoncés et que l'objectif de la rencontre consiste à identifier les indicateurs les plus importants. Pour finir, la phase 3 vise à faire la synthèse de la phase 2 et à décider de l'importance accordée aux énoncés de manière consensuelle par les participants.

Une adaptation de la méthode MACB devrait permettre de classer les attributs. La différence avec la méthode originale vient du fait que les attributs ne doivent pas être classés par importance, mais regroupés selon leur contenu. Il est impossible de savoir à l'avance si les attributs énoncés par les différents experts différeront beaucoup ou pas. Il reviendra aux experts de définir des contenus semblables qui deviendront les attributs consensuels. C'est seulement lorsqu'elle aura été adaptée et utilisée que la pertinence de la méthode sera démontrée. Il reste ensuite à définir une règle pour établir un lien entre les attributs consensuels et les items. Elle doit représenter adéquatement l'opinion de tous les experts dans la matrice synthèse.

L'élaboration de la matrice Q constitue le point de départ à l'utilisation des modèles cognitifs. Elle offre une structure au test sur laquelle s'appuie le diagnostic des sujets. Cette élaboration ne fait pas l'objet d'une méthodologie précise dans la littérature. Elle se présente souvent comme un processus d'essais et d'erreurs réalisé conjointement par le chercheur et les experts. Dans la majorité des recherches empiriques impliquant des données non simulées, une liste d'attributs de départ est fournie. Une seule matrice est ensuite ébauchée, utilisée puis améliorée.

Dans le but de documenter l'élaboration d'une matrice Q individuellement, puis collectivement par des experts, la présente recherche vise à exploiter plusieurs matrices pour comparer leurs conditions d'élaboration. Il faut donc contrôler la méthode d'élaboration. Cette dernière doit être semblable dans toutes les conditions afin de ne pas interférer dans les résultats. Si cela ne pose pas trop de problèmes lorsque les attributs cognitifs sont imposés aux experts, il en va

autrement dès qu'ils doivent les élaborer. La méthode MACB devrait permettre d'obtenir une liste d'attributs qui fasse consensus auprès des experts. Tirer parti de l'expertise des experts en leur laissant la liberté d'élaborer eux-mêmes les attributs cognitifs au lieu de les leur imposer semble alors réalisable tout en gardant le contrôle sur le processus.

1.3.3 Des conditions pouvant influencer la validité et la fidélité de la matrice Q

La littérature sur l'établissement des standards montre que les résultats de l'analyse des données offrent souvent une aide aux experts. La procédure itérative se sert de régressions multiples pour informer l'expert de son comportement par rapport à celui des autres. La méthode des signets (*bookmark method*) aide les experts en leur fournissant la difficulté des items estimée par la TRI. Une analyse en grappes et des analyses de classes latentes informent sur l'appartenance des sujets à une classe latente.

Les experts de cette recherche doivent élaborer des matrices dans une approche inductive, impliquant la disponibilité des données constituées par les réponses des sujets aux items du test. Dans le cas d'un test à choix multiple, les données originales contiennent le choix fait par chaque sujet (leurre ou bonne réponse). Elles doivent ensuite être recodées sous forme dichotomique (0 pour une mauvaise réponse et 1 pour la bonne réponse) afin d'appliquer un modèle cognitif. Ces données peuvent être analysées sous les deux formes et les résultats fournis aux experts.

Il reste à définir les informations concernant les données susceptibles d'apporter un soutien dans la tâche. Pour guider la réflexion, rappelons que les approches diagnostiques identifiées dans la revue de la littérature offrent plusieurs façons de caractériser les sujets. Les profils de scores illustrent l'une d'elles. Dans ce cas, les caractéristiques individuelles des sujets sont basées sur les erreurs systématiques. Ce point de vue consiste à s'intéresser aux erreurs commises pour fournir un diagnostic. Thissen, Steinberg et Fitzpatrick (1989) affirment d'ailleurs que les choix de réponses des items à choix multiples sont une partie importante de la question. Les modèles cognitifs actuels

s'appliquent à des données dichotomiques, ordinales ou continues. Comme ils ne peuvent pas modéliser les choix de leurres, ils ne tiennent pas compte de cette information. Elle pourrait être utilisée pour aider les experts.

Un graphique réalisé à l'aide du logiciel Testgraph (Ramsay, 2000) peut représenter chaque alternative d'une question à choix multiple (Ramsay, 1996; Ramsay, et Silverman, 1997; Thissen et al., 1989 ; Wang, 1998). En complément, les paramètres d'items au sens de la théorie classique des tests peuvent être calculés (par exemple avec SPSS). L'accès à des informations au sujet de la difficulté des items, de leur pouvoir discriminant, de l'intérêt porté à certains leurres par les sujets peut aider à élaborer la matrice Q.

Testgraph représente graphiquement la probabilité de choisir chaque leurre d'un item en fonction du score total des sujets. La visualisation du portrait de chaque item montre aux experts comment les leurres sont utilisés. Dans certains cas, l'un des leurres peut être très attrayant pour les sujets. Chercher ce qui le différencie de la bonne réponse peut fournir une piste à suivre par les experts.

La difficulté d'un item correspond à la proportion de sujets qui a répondu convenablement à la question. Cette information nuance le portrait de l'item. Elle peut amener les experts à rechercher ce qui rend une question difficile. Il se peut qu'elle implique un attribut qu'ils considèrent comme compliqué. Sa difficulté peut venir d'un nombre important de liens avec des attributs. Sa formulation peut éventuellement expliquer la difficulté. Ce sont autant de raisonnements qui pourraient aider les experts dans leur tâche. Le pouvoir de discrimination des items apporte une information supplémentaire. Celui-ci détermine s'il y a une différence entre les sujets faibles et les sujets forts lorsqu'ils répondent à la question. La recherche d'un attribut qui expliquerait une telle différence peut être une autre piste de réflexion.

Un autre point de vue utilise la structure du test pour établir le diagnostic. Le découpage d'une épreuve de maîtrise selon son tableau de spécification illustre cette structure. Il aboutit à un profil des sujets relativement à un ensemble d'objectifs. Les modèles cognitifs cherchent toutefois à dépasser la notion de structure imposée par l'opérationnalisation de la notion d'objectifs. Ils basent le diagnostic sur les processus cognitifs. L'hypothèse sous jacente est que l'espace latent d'habileté peut être scindé en un certain nombre d'aspects qui constituent les attributs cognitifs.

La multidimensionnalité des données donne une information sur la structure du test qui peut avoir un lien avec les attributs cognitifs. Hartz (2002) suggère d'ailleurs comme adéquat 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. Corter (1995) a utilisé une analyse en grappes afin de regrouper les items dans le but d'identifier la structure diagnostique du test.

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 est de définir les items qui partagent une même dimension latente. Les experts peuvent avoir accès aux items regroupés lorsqu'ils élaborent la liste des attributs cognitifs et les relient aux items. Ils peuvent baser leur réflexion sur ce que ces items ont en commun afin d'identifier des attributs cognitifs.

Le fait de s'intéresser aux erreurs commises ou à la dimension du test offre deux possibilités d'exploiter les données pour mettre les experts dans des conditions différentes. Les renseignements sur les choix de réponses ne sont pas pris en compte par le modèle cognitif. La perspective d'en tirer parti par le biais des experts constitue une première possibilité intéressante. Les analyses suggérées sont faciles à réaliser. Les résultats de ces analyses ne demandent pas que les experts aient des connaissances particulières en statistiques. Ils peuvent être fournis dans un langage compréhensible

par tous. Fournir un graphique, la difficulté et l'indice de discrimination pour chaque item représente tout de même une documentation imposante et détaillée. Le risque est peut-être que la quantité d'information n'amène les experts à trop entrer dans le détail pour chacun des items. Il convient donc de limiter le nombre d'attributs à utiliser afin d'éviter une prolifération des attributs cognitifs.

La deuxième possibilité consiste à analyser la dimensionnalité du test et à regrouper les items dans les dimensions identifiées. Ces analyses sont assez complexes. Elles nécessitent de faire un certain nombre de choix (méthode, logiciel, nombre de facteurs, etc.). Les résultats sont toutefois facilement compréhensibles par les experts. Ils ont simplement à examiner ensemble les items d'un même regroupement. Il est possible que les items soient, parfois ou toujours, regroupés selon un construit latent qui n'a rien à voir avec la notion d'attribut cognitif. Dans ce cas, les experts risquent d'avoir beaucoup de difficulté à tirer parti de ces informations. Il est important que les consignes expliquent cette limite. Il faut éviter que les experts ne cherchent à tout prix à trouver un attribut commun et tombent dans le piège d'identifier des dimensions non pertinentes dans une perspective diagnostique au détriment de celles qui le sont.

1.3.4 Un modèle afin de juger de la validité et de la fidélité de la matrice Q

Une rapide étude comparative des sept modèles présentés dans la revue de la littérature amène à choisir celui qui fournit les indicateurs de validité et de fidélité les plus détaillés. Le LLTM ne relie pas les sujets aux attributs, les modèles MLTM, et GLTM utilisent une matrice reliant tous les attributs à tous les items (Hartz, 2002). Le modèle des réseaux de Bayes et le modèle *rule space* ne contiennent pas d'indicateur de la validité et de la fidélité de la matrice Q.

Le modèle DINA estime deux paramètres pour chaque item. Leur valeur élevée indique un problème de validité et/ou de fidélité des liens entre l'item concerné et les attributs qui lui sont reliés. Le modèle de fusion permet théoriquement de juger de la qualité de prédiction du modèle à l'aide de

quatre paramètres. Deux de ces paramètres sont des indicateurs de la validité de la matrice Q, les deux autres sont plutôt liés à sa fidélité.

Selon DiBello et ses collègues (1995) et Junker (1999), un modèle plus fin assure une meilleure fidélité et/ou validité au prix d'une faible maniabilité. Le choix d'un modèle plus fin risque de mener à des problèmes de convergence et, par exemple, à devoir limiter le nombre d'attributs. En contrepartie, un modèle plus grossier s'applique plus facilement à des données. La perspective de comparer des matrices en terme de validité et de fidélité implique d'utiliser un modèle à grain fin, en autant qu'une convergence soit possible. Les modèles de fusion (DiBello et al., 1995) et DINA (Junker et al., 2001) sont les seuls à fournir, au moins théoriquement, des indicateurs de validité et de fidélité. Le modèle DINA s'applique à des données et ses paramètres sont estimables. De plus, un programme appelé CDM (Templin, 2006) associé au logiciel Mplus existe.

Par contre, la version initiale du modèle de fusion proposée par DiBello et Stout en 1995 ne permet pas l'estimation des paramètres. Jiang (1996) propose un étalonnage du modèle qui les rend estimables grâce à l'utilisation d'un algorithme EM (espérance et maximisation). Cet algorithme itératif permet de chercher le paramètre réalisant le maximum de vraisemblance. Toutefois, la méthode proposée implique des calculs très lourds, limite la taille de la matrice Q utilisable en pratique et perd l'esprit du modèle (Templin, 2004).

Hartz (2002) fournit une solution au problème sous la forme d'un modèle de fusion reparamétrisé dans une approche bayésienne et appelé RUM (*Reparameterized Unified Model*). Il suppose que chaque sujet répond correctement à chaque item si et seulement s'il maîtrise tous les attributs reliés à cet item (modèle non compensatoire). Il estime trois des quatre paramètres du modèle initial. Deux de ces paramètres sont légèrement différents et ils sont tous trois interprétables.

Les deux nouveaux paramètres sont notés π_i^* et r_{ik}^* . Le paramètre π_i^* représente la probabilité qu'un sujet ayant maîtrisé tous les attributs requis pour l'item i par la matrice Q les ait

convenablement utilisés pour répondre à la question. Ce paramètre peut être associé à la notion de fidélité des attributs. Le paramètre r_{ik}^* compare la probabilité de bien répondre à l'item i selon que le sujet maîtrise ou non l'attribut k . Ce paramètre représente la pénalité due au fait de ne pas maîtriser l'attribut. Il peut être associé à la notion de validité puisqu'il est un indice de la pertinence des attributs. Enfin, un paramètre, noté c , est un indicateur de la validité du modèle qui est présent dans le modèle original et permet de juger si la matrice Q contient tous les attributs importants. Le RUM garde l'esprit du modèle de fusion tout en permettant son application à des données. Le logiciel Arpeggio, propriété de l'*Educational Testing Service* (Stout, 2002; Roussos, Dibello, Stout, Hartz, Henson, et Templin, 2007) permet d'estimer les paramètres. En outre, le programme CDM (Templin, 2006) applique ce modèle à des données sans l'estimation du paramètre c .

Chacun des modèles DINA et RUM offrent des indicateurs de validité et de fidélité. Ils permettent de comparer les matrices créées par des experts. Le modèle RUM est toutefois plus intéressant car il offre une interprétation plus riche. Les résultats de la modélisation des données à l'aide du RUM serviront donc à porter un jugement sur les validité et fidélité relatives des matrices Q élaborées dans cette recherche. Il fournit deux indicateurs de la validité et un indicateur de la fidélité de la matrice Q .

1.3.5 Les variables de la recherche

La recherche compte neuf variables: une variable indépendante et huit variables dépendantes (voir figure 2). Les conditions expérimentales dans lesquelles les experts travaillent composent les valeurs de la variable indépendante. Elles sont définies par les renseignements qui sont fournis aux experts.

Les matrices individuelles de chacun des experts constituent la première variable dépendante. Chaque matrice associée à un expert et à une condition est une valeur différente de cette variable.

Les matrices de tous les experts pour chaque condition expérimentale sont ensuite synthétisées pour former la deuxième variable dépendante. Ses valeurs résultent des conditions par l'intermédiaire des matrices individuelles qui ont été regroupées pour les fabriquer.

Le modèle cognitif choisi (RUM) est appliqué aux données du test avec chacune des matrices. Comme il permet l'estimation de trois paramètres, son utilisation aboutit à six nouvelles variables dépendantes des conditions expérimentales. Les valeurs des indicateurs r et les valeurs des indicateurs c forment quatre variables. Deux correspondent à la validité des matrices individuelles et les deux autres à celle des matrices synthèses. Les valeurs des indicateurs π (π_i) donnent deux variables liées à la fidélité : une pour les matrices individuelle et une pour les matrices synthèses. La figure 2 illustre l'ensemble des neuf variables et leurs relations.

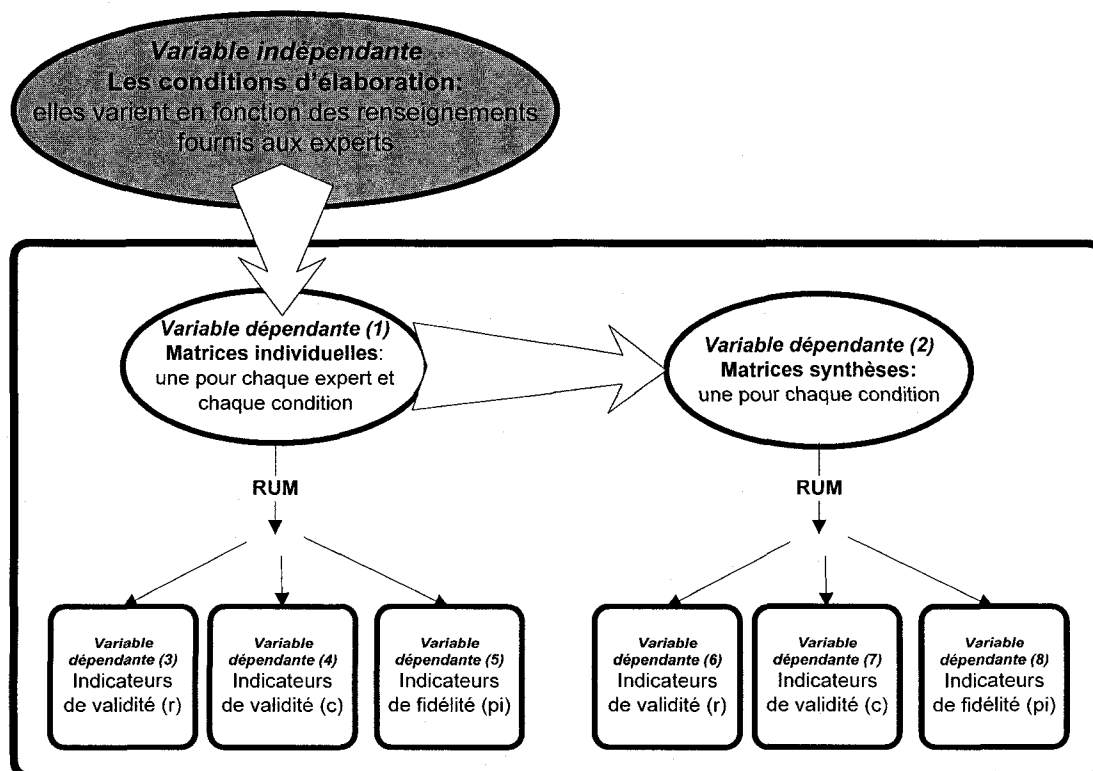


Figure 2. Les variables de la recherche

1.4 Les objectifs et questions de recherche

L'élaboration de la matrice Q est une étape cruciale pour qui veut utiliser un modèle cognitif puisque de sa validité et de sa fidélité dépendent la validité et la fidélité du diagnostic des sujets.

Dans une approche inductive, les données du test sont accessibles. Il est possible de les analyser et de fournir les résultats aux experts chargés de l'élaboration de la matrice avec pour objectif de les aider.

Cette recherche vise à identifier quelles conditions influencent la validité et la fidélité des matrices produites.

Dans le but de documenter le processus complexe d'élaboration des matrices Q, le contexte choisi est le plus ouvert possible. Dans chacune des conditions, les experts élaborent individuellement des matrices dont la validité et la fidélité sont comparées. Étant donné que l'usage diagnostique d'un modèle cognitif nécessite une seule matrice, les experts doivent ensuite parvenir à une matrice unique par condition. La validité et la fidélité de ces matrices sont alors comparées d'une condition à l'autre.

Dans cette logique, des matrices individuelles et des matrices synthèses sont produites dans chacune des conditions expérimentales contrôlées dans la recherche. Toutes ces matrices sont caractérisées par les attributs qu'elles contiennent, par leur nombre et par les liens qu'elles établissent entre les attributs et les items. Ces matrices peuvent être comparées du point de vue de leur forme pour répondre à la première question de recherche.

1. Comment les conditions expérimentales influencent-elles les matrices Q de chaque expert et les matrices synthèses?

La recherche d'un consensus dans le regroupement des matrices individuelles en matrices synthèses amène à étudier la concordance entre les matrices individuelles et la représentativité des matrices synthèses pour répondre à la deuxième question de recherche.

2. Comment les conditions expérimentales influencent-elles la concordance entre les matrices individuelles et la représentativité des matrices synthèses ?

Le modèle RUM peut modéliser les données avec chacune des matrices individuelles et synthèses. Les différents paramètres du modèle, en particulier les deux indicateurs de la validité de la matrice (r et c) et l'indicateur de la fidélité (π) de la matrice Q , sont estimés. L'étude des variations des six variables dépendantes correspondantes (voir figure 2) en fonction des conditions, dès que la robustesse des estimations est suffisante, permet de répondre à la dernière question de recherche.

3. Comment les conditions expérimentales influencent-elles la validité et la fidélité des matrices Q ?

2 Chapitre II : méthodologie

Cette recherche vise à comparer la validité et la fidélité de matrices Q élaborées dans des conditions différentes. L'objectif est de déterminer si l'une d'entre elles aboutit à la création de matrices plus valides et fidèles que les autres. Les experts fabriquent les matrices à partir d'un test en mathématiques. Dans un premier temps, les experts ainsi que les données sont présentés.

La variable indépendante du cadre conceptuel prend pour valeurs les différentes conditions d'élaboration des matrices Q. La recherche doit exercer un contrôle sur ces conditions de même que sur la procédure permettant d'élaborer les matrices individuelles et les matrices synthèses. Ce chapitre décrit les conditions ainsi que la procédure. Les comparaisons de la forme et du contenu des matrices individuelles et des matrices synthèses produites permettent de répondre à la première question de la recherche.

La concordance entre les matrices individuelles d'une même condition, de même que la représentativité de la matrice synthèse qui en découle, sont comparées d'une condition à l'autre. Ces comparaisons permettent de répondre à la deuxième question de la recherche.

Les données sont modélisées à l'aide des différentes matrices. Le modèle RUM est relativement récent, ce chapitre en inclut une description. Il décrit également les précautions à prendre avant d'interpréter les résultats. L'estimation des paramètres reliés à chaque matrice donne la possibilité de comparer la validité et la fidélité des matrices d'une condition à l'autre. La vérification de la robustesse des indicateurs estimés puis leurs comparaisons permettent de répondre à la dernière question de la recherche.

2.1 Les participants

Pour être considérés comme experts, les participants devaient à la fois être des spécialistes diplômés en mathématiques et bien connaître la population d'étudiants visés par le test. Ils devaient donc répondre aux critères suivants : (a) être diplômés universitaires en mathématiques et (b) enseigner les mathématiques aux niveaux secondaire, collégial ou universitaire depuis au moins 10 ans. Quatre personnes ont satisfait à ces critères. Considérant que les enjeux sont ici beaucoup moins élevés que ceux qui président à l'établissement de standards ou de seuils de réussite, ce nombre a été jugé suffisant en l'absence de balises claires. Les quatre experts ont fourni un gros travail qui s'est étalé sur plusieurs semaines.

2.2 Les données

2.2.1 Les neuf bases de données d'origine

Les résultats à toutes les épreuves de mathématiques administrées aux étudiants à l'entrée de l'École Polytechnique de Montréal de 1997 à 2005 constituent les neuf bases de données d'origine. Ces épreuves prennent la forme d'un questionnaire à choix multiple de 60 items puisés dans une banque de 985 questions qui couvrent les contenus mathématiques enseignés aux niveaux secondaire et collégial. Chaque épreuve est construite à partir d'un échantillonnage d'items différent, mais contient 13 items liés aux fonctions élémentaires, 7 items liés à la trigonométrie, 10 items liés à la géométrie plane et analytique, 12 items liés à l'algèbre linéaire et vectorielle, 12 items liés au calcul différentiel et 6 items liés au calcul intégral. Tous les items de la banque sont classifiés en fonction de leur appartenance à l'un ou l'autre des sept contenus visés et selon qu'ils visent une notion, un type de fonction, une application, une technique, etc., mais le nombre d'items de chaque catégorie varie d'une épreuve à l'autre.

Chaque question est associée à cinq choix de réponses, le cinquième choix étant « je ne sais pas ». Comme ce choix n'a pas toujours été différencié d'une valeur manquante dans les bases de données, il a été ignoré. Un étudiant qui admet ne pas connaître la réponse ne diffère donc pas d'un étudiant qui ne répond pas.

Les neuf bases de données sont analysées de façon descriptive. Le nombre de sujets, la fréquence des choix de réponses, le nombre de valeurs manquantes, le pourcentage de bonnes réponses et le score total de chaque sujet sont calculés. Les données sont ensuite dichotomisées. Le pouvoir de discrimination de chaque item est calculé sous la forme de la corrélation entre chaque item et le score total. Les coefficients Alpha de Cronbach du score total sont également calculés. Étant donné que les résultats des différentes analyses descriptives sont semblables pour l'ensemble des neuf bases de données, le plus grand nombre de sujets permet finalement de retenir les données correspondant aux cohortes d'étudiants de 1997 et de 1999. La cohorte de 1997 est associée à la lettre A, elle est formée de 426 sujets. La cohorte de 1999 est associée à la lettre B, elle en comporte 478.

Ces bases de données sont pertinentes pour mener cette recherche à cause de leur visée diagnostique. En effet, la professeure responsable de l'élaboration de la banque d'items a cherché dès le départ à identifier parmi les étudiants admis à l'école Polytechnique ceux et celles qui avaient besoin d'un support ou d'une mise à niveau. Les résultats sont interprétés de façon critériée puisque'un cours de mise à niveau est proposé à tous les sujets dont le score total est inférieur à 70%. Même si les sous scores relatifs à un contenu visé sont fournis aux étudiants, cette épreuve vise essentiellement à identifier les plus faibles d'entre eux. Les recherches empiriques utilisent généralement les données d'évaluations à grande échelle qui privilégient souvent le format de questions à choix multiple. Les épreuves utilisées dans cette recherche sont des QCM construits par des professeurs pour leur propre usage et administrés à plusieurs centaines d'étudiants. La pertinence

de ces bases de données est donc en plus d'ordre pratique puisque le nombre de sujets est suffisant pour réaliser les analyses statistiques que demande cette étude.

2.2.2 Les deux bases de données de la recherche

Afin de donner une charge de travail acceptable à chacun des experts et de minimiser l'effet de la fatigue sur leur travail, vingt items sont sélectionnés parmi les 60 que comporte chacune des deux épreuves choisies. Le contenu visé par chaque question est clairement indiqué dans la banque d'items. La sélection des vingt items retenus s'est effectuée à partir des considérations suivantes.

Plutôt que de choisir des items dans les six contenus mathématiques des épreuves, deux d'entre eux sont d'abord sélectionnés. L'examen des items de géométrie et de trigonométrie montre qu'ils demandent souvent un niveau cognitif moins élevé que les autres et proposent parfois des leures discutables à cause de la difficulté à les élaborer (cette réflexion fait suite à une discussion avec la créatrice du test). Le nombre d'items liés au calcul intégral est petit (6), cette catégorie est éliminée. Un choix doit être fait entre les catégories restantes : fonctions élémentaires, calcul différentiel et algèbre linéaire et matricielle. Le choix arbitraire basé sur la proximité des notions de fonctions élémentaires et de calcul différentiel aboutit à retenir dix items de chacune de ces deux catégories.

Il reste à éliminer 3 des 13 items liés aux fonctions élémentaires et 2 des 12 items liés au calcul différentiel. Certaines questions sont naturellement écartées par l'intention d'éviter qu'elles ne se retrouvent dans les deux tests. Étant donné qu'un grand nombre de valeurs manquantes a un impact sur la difficulté des items, ce critère de sélection des items à supprimer est retenu pour retirer les derniers items. Les réponses des sujets à dix items portant sur les fonctions élémentaires et dix items portant sur le calcul différentiel de deux tests différents identifiés par les lettres A et B forment les données.

2.2.3 Les caractéristiques des tests A et B

Le test A et le test B ont des formes similaires puisqu'ils contiennent chacun vingt items dont dix sont liés aux fonctions élémentaires et dix sont liés au calcul différentiel. En référence à la classification des items dans la banque, deux items du test A et cinq items du test B liés aux fonctions élémentaires sont associés à la « notion » de fonction, les autres sont associés à un « type » de fonction (par exemple, polynômiale, racine carrée, etc.). En ce qui concerne les items liés au calcul différentiel, sept items du test A et cinq items du test B sont associés à la « notion » de dérivée alors que les autres sont de l'ordre de la « technique » (par exemple, produit, composition, etc.). Les tableaux de spécification des tests A et B font l'objet du tableau 2.

Le tableau 3 présente les caractéristiques du score total de chacun des deux tests et la valeur du coefficient alpha de Cronbach. Le test B comporte cinquante deux sujets de plus que le test A. Une analyse de variance montre que la moyenne des deux tests est statistiquement différente au seuil de 0,05 ($F(1, 902) = 8,918$ et $p = 0,003$). Les résultats font l'objet du tableau A1 de l'annexe A. L'hypothèse nulle disant que les scores moyens sont égaux pour les deux groupes doit être rejetée, le test B est donc plus difficile que le test A. Toutefois la dispersion des scores est semblable pour les deux tests. La statistique de Levene est égale à 0,570 avec un degré de liberté. La probabilité associée est de 0,450, elle est supérieure à 0,05, ce qui implique qu'on ne rejette pas l'hypothèse nulle disant que la variance des deux groupes est égale.

Tableau 2.

Les tableaux de spécification des tests A et B

Fonctions élémentaires					Calcul différentiel					
	Test A	Test B		Test A	Test B		Test A	Test B	Test A	Test B
Notions			Fonctions			Notions			Techniques	
définition		Item 1	puissance	Item 2	Item 3	définition	Item 3 Item 16 Item 18	Item 17	élémentaire	Item 9 Item 15
domaine		Item 11	polynomiale	Item 4		interprétation	Item 11	Item 2 Item 9	produit	Item 8
graphe		Item 10	Trigonomé- trique	Item 15		utilisation	Item 7 Item 13 Item 20	Item 6 Item 18	quotient	Item 14 Item 19
composée	Item 1		valeur absolue	Item 8	Item 20	Taylor			composition	Item 5 Item 4 Item 12
logique		Item 14	racine carrée	Item 12	Item 13					
manipulation	Item 10	item 16	exponentielle	Item 17						
			logarithmique	Item 19	Item 7					
			inéquation	Item 6	Item 5					

Tableau 3.

Les caractéristiques du score total des tests A et B

	N	Moyenne	Médiane	Mode	Écart- type	Minimum	Maximum	Alpha
Test A	426	11,5	12	12	3,25	2	19	0,659
Test B	478	10,83	11	10	3,46	2	20	0,709

Le fait d'élaborer des matrices Q relativement à des tests non équivalents, dont l'un est plus difficile que l'autre n'est pas problématique puisque les matrices ne seront pas comparées d'un test à l'autre. Par contre, il sera utile d'exploiter cette information pour se demander si une plus grande

difficulté du test rend l'élaboration de la matrice Q plus complexe ou encore si la difficulté a un lien avec la validité et/ou la fidélité des matrices.

La figure 3 présente la difficulté de chacun des items des deux tests. Elle permet de visualiser le fait que les items du test B sont souvent plus difficiles que ceux du test A.

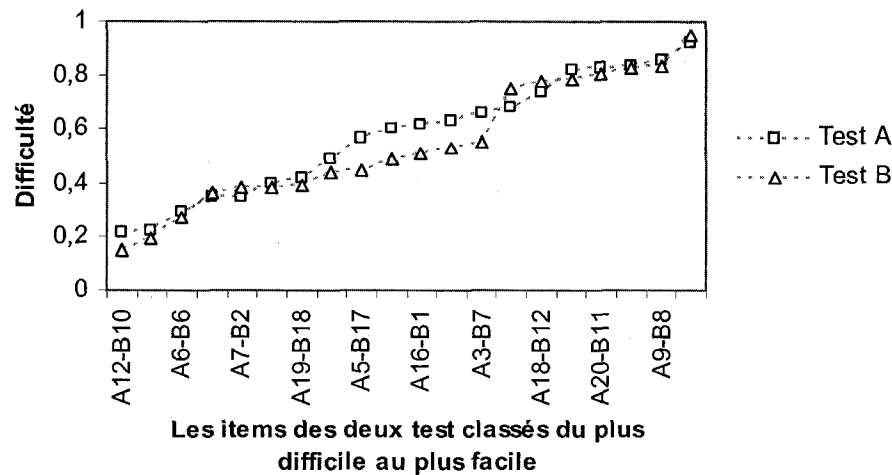


Figure 3. La difficulté des items des tests A et B

Hakstian et Whalen (1976) fournissent une méthode pour réaliser un test de signification sur les valeurs du coefficient alpha (voir annexe B). La statistique M du test est égale à 2,67. Elle est inférieure à la valeur critique du khi carré à un degré de liberté qui vaut 3,84. On ne peut donc pas rejeter l'hypothèse nulle disant que la valeur du coefficient alpha est semblable pour A et B. Si le test B est plus difficile que le test A, les deux tests ont une cohérence interne comparable.

2.3 La procédure

2.3.1 L'élaboration des matrices

2.3.1.1 La notion d'attributs cognitifs

La littérature permet de familiariser les experts avec la notion d'attribut cognitif en mathématiques. Les quatre experts ont accès à deux définitions (Buck et al., 1998; Gierl, 1997) et aux listes de Hartz (2002); Milewski (2002); Rupp, et Watermann (2005) et Tatsuoka (1996, document personnel cité dans Gierl, 1997) dans chacune des conditions (voir annexe C). Ces quatre listes sont représentatives de ce qui se trouve dans la littérature et illustrent différentes approches. Les listes de Milewski, Rupp et Hartz proposent entre sept et seize attributs formulés de façon synthétique. La liste de Tatsuoka contient seize attributs largement illustrés à l'aide d'exemples précis. La même information est disponible dans les trois conditions expérimentales afin de contrôler la compréhension par les experts de ce qu'est un attribut cognitif.

2.3.1.2 Les trois conditions expérimentales

Trois conditions expérimentales sont mises en place. Dans la condition 1, les experts n'ont accès qu'aux items. Dans la condition 2, les informations additionnelles concernent la dimensionnalité du test (résultats d'analyses factorielles). Dans la condition 3, des informations sur les erreurs commises (analyse des leurres, difficulté et pouvoir discriminant des items) remplacent celles sur la dimension du test. De cette façon, la valeur ajoutée de l'une ou l'autre source de renseignements peut être comparée à la première condition qui ne comporte que les items. Étant donné que le but visé par l'étude consiste à identifier si certaines de ces informations sont utiles aux experts, il semble suffisant, dans un premier temps, de comparer les conditions 2 et 3 à la condition 1. S'il s'avère que les deux sources d'information sont pertinentes, la condition 2 devra être comparée à la condition 3. Dans les trois conditions les attributs cognitifs sont définis de la même

manière et les exemples sont identiques. Les deux tests sont utilisés à chaque condition et chaque expert élabore une matrice. Au total, vingt quatre matrices individuelles (2 tests $\times$ 3 conditions $\times$ 4 experts = 24) sont fabriquées, comme illustré par la figure 4.

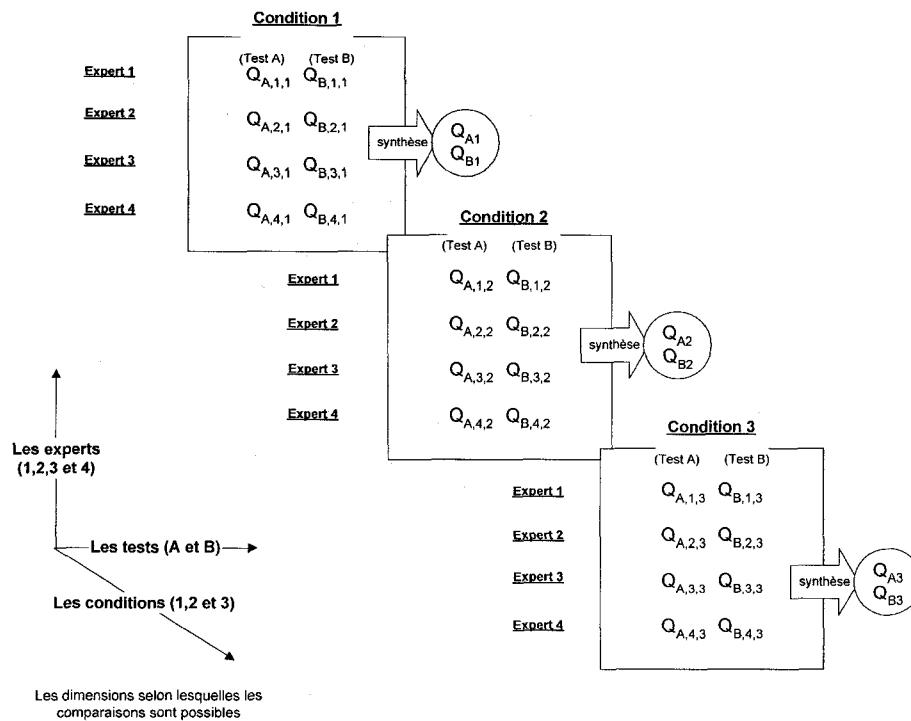


Figure 4. Le plan d'élaboration des 24 matrices individuelles (et des six matrices synthèses) relativement au test, à l'expert et à la condition et les dimensions selon lesquelles les comparaisons sont possibles.

2.3.1.3 Le déroulement

Les experts élaborent chacun de leur côté les listes d'attributs cognitifs de chaque condition. Ceci vise à tirer parti de leur expertise et du fait qu'ils connaissent bien la population d'étudiants en évitant l'influence des autres experts. Cette approche permet aux attributs de varier entre les conditions et selon les tests A et B. Dans les trois conditions, chaque expert a la possibilité

d'identifier un maximum de dix attributs par matrice. Cette valeur est en accord avec plusieurs études empiriques pour lesquelles le nombre d'attributs est égal à la moitié du nombre d'items (voir tableau 1). Elle oblige les experts à cibler les attributs cognitifs qu'ils jugent importants sans les limiter trop strictement.

Le chercheur passe en revue les définitions et exemples d'attributs avec chaque expert et répond à ses questions. Il définit les enjeux liés à la tâche demandée et en quoi son expertise est essentielle. Il les rencontre individuellement au début et à la fin de chaque condition. Lors de la première rencontre, le chercheur remet à chaque expert un livret explicatif contenant les items, les définitions, les exemples, les consignes et les informations additionnelles propres à la condition. Le format de remise des résultats fait partie de chaque livret. Il est identique pour tous et pour chaque condition.

Lors de la rencontre individuelle entre le chercheur et chaque expert à la fin des conditions, le chercheur récolte des renseignements sur les difficultés rencontrées et sur l'opinion de l'expert quant à la pertinence des informations fournies. Dans les livrets des conditions 2 et 3, les experts doivent laisser une trace de leur utilisation des informations fournies. Notons que les résultats feront mention de ces traces sans toutefois y consacrer une section particulière. Cette rencontre permet également de récupérer tous les documents relatifs à la condition. L'impression générale des experts assure un contrôle a posteriori sur la clarté des différentes consignes. Chaque expert peut demander de l'aide à tout moment au cours des trois semaines qu'il a pour réaliser son travail. Une période de deux à trois semaines sépare les différentes conditions.

Dans la condition 1, les experts ont simplement accès aux énoncés des items. À la fin de cette condition, chaque expert a fabriqué une matrice individuelle pour chacun des deux tests, soit un total de huit matrices pour l'ensemble des experts. Fabriquer une matrice signifie : (a) rédiger une

liste d'au plus dix attributs cognitifs et (b) décider si un lien existe ou pas entre chaque item et ces attributs.

En préparation à la condition 2, une analyse factorielle est réalisée sur les données des tests A et B afin d'identifier le nombre de dimensions présentes dans chaque test et quels items se regroupent dans ces dimensions. Les analyses factorielles utilisant les corrélations tétrachoriques et le logiciel MPlus (Muthén, et Muthén, 2004) fournissent des solutions comprenant quatre facteurs pour le test A et six facteurs pour le test B (voir annexe F). À partir des regroupements d'items en facteurs, les experts doivent fabriquer huit nouvelles matrices comportant un maximum de dix attributs. Afin de les aider, l'ordre des items est modifié dans le livret de la condition 2 pour que les items d'un même facteur soient regroupés. Il y a 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 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. Elles les informent qu'il est possible qu'ils ne trouvent rien de commun.

Pour préparer la troisième condition, 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-biserial (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 (voir un exemple dans l'annexe G).

Le livret de la condition 3 présente chaque item, ainsi que ses paramètres de difficulté et de discrimination et son portrait fourni par Testgraph. Le livret contient quelques commentaires et suggestions pour guider les experts dans leur interprétation de ces informations. Le contenu du livret

est passé en revue lors de la première rencontre de la condition 3. L'annexe G illustre par un exemple la présentation d'un item dans le livret de la condition 3. Les experts doivent refaire l'exercice une dernière fois et produire les huit dernières matrices.

Les experts fabriquent les matrices dans chaque condition sans avoir accès à leurs travaux précédents afin de comparer la pertinence de chacune des deux sources d'information séparément. Lorsque les experts ont travaillé dans les trois conditions, un total de vingt quatre matrices individuelles est disponible. Par la suite, une synthèse de ces matrices est réalisée pour chaque test et chaque condition (voir figure 4).

2.3.1.4 Les trois phases de construction des matrices synthèses

Le processus d'établissement de standards et de seuils de réussite en éducation fournit un point de départ à la réflexion sur le processus d'élaboration des matrices Q. Une différence entre les deux provient de l'impact de l'ajout ou du retrait d'un attribut sur la structure cognitive globale. Néanmoins, quatre des cinq composantes structurelles du processus d'établissement des standards telles que présentées par Reckase (2000) peuvent être transférées au processus d'élaboration et de validation des différentes matrices Q. Le tableau 4 présente la similarité entre les deux procédures.

Il est inutile de discuter ici de la composante 2. La composante 3 concerne le recrutement des experts présenté au point 2.1. La composante 4 consiste à adapter la méthode MACB (voir 1.3.2). La composante 5 concerne le moyen de remettre les résultats. Elle implique à la fois le format de remise des matrices individuelles inclus dans les livrets des conditions et le processus qui aboutit à la création des matrices synthèses.

Tableau 4.

La similarité entre les processus d'établissement de standards et d'élaboration de la matrice Q

Processus	Composante 1	Composante 2	Composante 3	Composante 4	Composante 5
Établissements de standards	Une autorité pour élaborer une politique	Un domaine évalué	Une sélection de personnes pour porter un jugement sur les niveaux de performance attendus	Une méthode de collecte de ces jugements et des standards	Un moyen de rapporter les résultats
Élaboration d'une matrice Q	Sans objet ici, l'application des modèles implique de recourir au processus d'élaboration des matrices	Un domaine évalué	Une sélection de personnes pour élaborer les attributs cognitifs et les relier aux items	Une méthode de collecte de ces jugements	Un moyen de rapporter les résultats

Dans la version originale du MACB (Vanderwood et al., 1993), les experts doivent définir une liste d'indicateurs. Ils doivent ensuite les classer selon leur importance de manière à obtenir un consensus entre les membres de l'équipe. Cette méthode implique un processus en trois phases. Dans les paragraphes qui suivent, chacune des trois phases de la méthode originale est rapidement décrite. Elles sont mises en parallèle avec les trois phases de la procédure employée pour élaborer les matrices Q puisque la méthode est modifiée.

La phase 1 correspond à l'élaboration des matrices individuelles. Dans la méthode originale (Vanderwood et al., 1993), la phase 1 consiste en une élaboration individuelle des énoncés des indicateurs par les membres du groupe. Ces énoncés seront utilisés dans la phase 2. La première phase de la procédure MACB appliquée à la création des matrices correspond à l'élaboration individuelle des vingt quatre matrices selon le plan illustré par la figure 4. L'ensemble des attributs

identifiés et verbalisés par les experts seront utilisés dans la phase 2. Les liens établis serviront par la suite à faire la synthèse des matrices.

La phase 2 vise à obtenir un consensus. Dans la méthode MACB (Vanderwood et al., 1993), un document contenant tous les indicateurs individuels de la phase 1 est fourni aux membres du groupe. L'objectif consiste à les classer selon leur importance. Les experts doivent s'assurer d'une compréhension commune de chacun des énoncés, quitte à éventuellement effectuer quelques changements si tous les membres du groupe sont d'accord. Il est important de ne pas passer trop de temps sur la formulation, mais de concentrer l'attention sur la compréhension. Commence ensuite le processus de classement des indicateurs individuels. Chaque participant donne individuellement un score entre 0 et 100 à chacun des énoncés, la valeur 100 étant liée à un énoncé de haute importance. Plusieurs énoncés peuvent obtenir le score 100. Les membres du groupe commencent par attribuer les valeurs 100 afin d'avoir un point de référence pour évaluer les autres énoncés.

L'accès à une feuille de calcul sur un ordinateur (sur Excel par exemple) permet à un opérateur d'entrer les scores au fur et à mesure que les membres du groupe ont terminé cette évaluation. Ainsi l'ordinateur calcule rapidement les étendues des scores et les moyennes. Ces statistiques fournissent la base à la discussion qui suit. Le groupe étudie les énoncés pour lesquels les scores montrent une grande variance. Les membres avec un score plus haut ou plus bas que les autres sont amenés à justifier ou expliquer leur choix. Les énoncés pour lesquels les experts sont d'accord ne nécessitent pas de discussion. Pendant le débat, chaque participant peut, s'il le souhaite, changer son score à la lumière des explications apportées par les autres. L'opérateur intègre chaque changement et l'ordinateur calcule instantanément les nouvelles valeurs des paramètres de tendance centrale et de dispersion. La discussion continue si possible jusqu'à obtention d'un consensus.

Cette phase 2 est modifiée lorsqu'il s'agit d'élaborer les matrices Q puisque le but consiste à faire la synthèse des attributs individuels des experts. Relativement à la liste comprenant tous les

attributs élaborés pour chaque test et chaque condition, les experts regroupent les intitulés verbalisés différemment, mais dont le contenu est similaire. Ils définissent tout d'abord les catégories qui serviront au classement et doivent limiter leur nombre à six. Les consignes autorisaient les experts à utiliser jusqu'à dix attributs par matrice afin de leur laisser une certaine marge de manœuvre. Ce nombre est ramené à six afin de simplifier les matrices étant donné la complexité du modèle qui sera utilisé. Les experts doivent classer les attributs dans les catégories au lieu de les classer par importance. Si certains des attributs initiaux n'entrent dans aucune catégorie, ils sont éliminés.

Une fois les six catégories définies, les experts attribuent un score, entre 0 et 100, à chaque attribut de la liste relativement à chacune des six catégories ou à une catégorie intitulée « autre ». Cette dernière est ajoutée afin que chaque attribut puisse toujours avoir un score total de 100.

Cinq fichiers Excel sont utilisés lors de la rencontre, un fichier par expert et un fichier synthèse qui leur est relié. Chaque expert entre directement ses scores dans son propre fichier. Les ordinateurs sont liés en réseau et le fichier synthèse se met à jour au fur et à mesure que les scores sont inscrits dans les fichiers individuels. Cette façon de faire évite le recours à un opérateur, permet de gagner du temps et d'intégrer instantanément les changements lors des discussions qui suivent.

Le fichier synthèse calcule les moyennes et les étendues des scores dans chaque catégorie. Il permet également de s'assurer que les experts n'ont oublié aucun score et que le total (incluant la catégorie « autre ») pour chaque attribut vaut bien 100. Le chercheur fait une sauvegarde de la première version et des sauvegardes régulières des fichiers afin de garder des traces de l'évolution des scores au fil de la discussion. Il prend également en note les attributs sujets à discussion et le contenu de la discussion. Les attributs pour lesquels les experts ne sont pas d'accord font l'objet des débats, si possible jusqu'à obtention d'un consensus. Chacun peut apporter à tout moment des changements à ses scores en fonction des discussions.

La phase 3 vise l'élaboration des matrices synthèses. Dans la méthode MACB (Vanderwood et al., 1993), après la session de travail de la phase 2, la sélection des indicateurs se fait sur la base des données recueillies et des argumentations dans les cas limites. Même si le choix des indicateurs retenus peut se baser sur différentes combinaisons de critères statistiques, le *National Center on Educational Outcome* suggère de s'en tenir à une approche simple. On peut, par exemple, retenir un énoncé dont la valeur moyenne dépasse 90 et l'éliminer lorsqu'elle est inférieure à 50. Les pourcentages de personnes ayant accordé un score de plus de 75 ou un score de moins de 50 sont également utiles à la prise de décision.

Dans le cas présent, la phase 3 consiste à classer les attributs dans les six catégories identifiées à la phase 2. Les traces des discussions de la phase 2 complètent les calculs de la moyenne et des pourcentages d'experts ayant accordé un score supérieur à 75 ou inférieur à 50. Le chercheur utilise ces informations pour classer les attributs dans les catégories et trancher dans les cas litigieux. Un attribut problématique peut éventuellement être éliminé. La figure 5 résume les trois phases de l'application de la méthode MACB à l'obtention d'un consensus entre les experts ainsi que les tâches à accomplir pour classer les attributs dans six catégories.

La dernière partie de la construction des matrices synthèses n'a 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 peut être convertie en remplaçant chaque attribut par le nom de la catégorie à laquelle il appartient. La transformation aboutit à des matrices individuelles de dimension 20 items par 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. Dans une matrice, lorsque plusieurs intitulés originaux appartiennent à la même catégorie, il suffit d'une seule valeur 1 pour attribuer la valeur 1 à la catégorie.

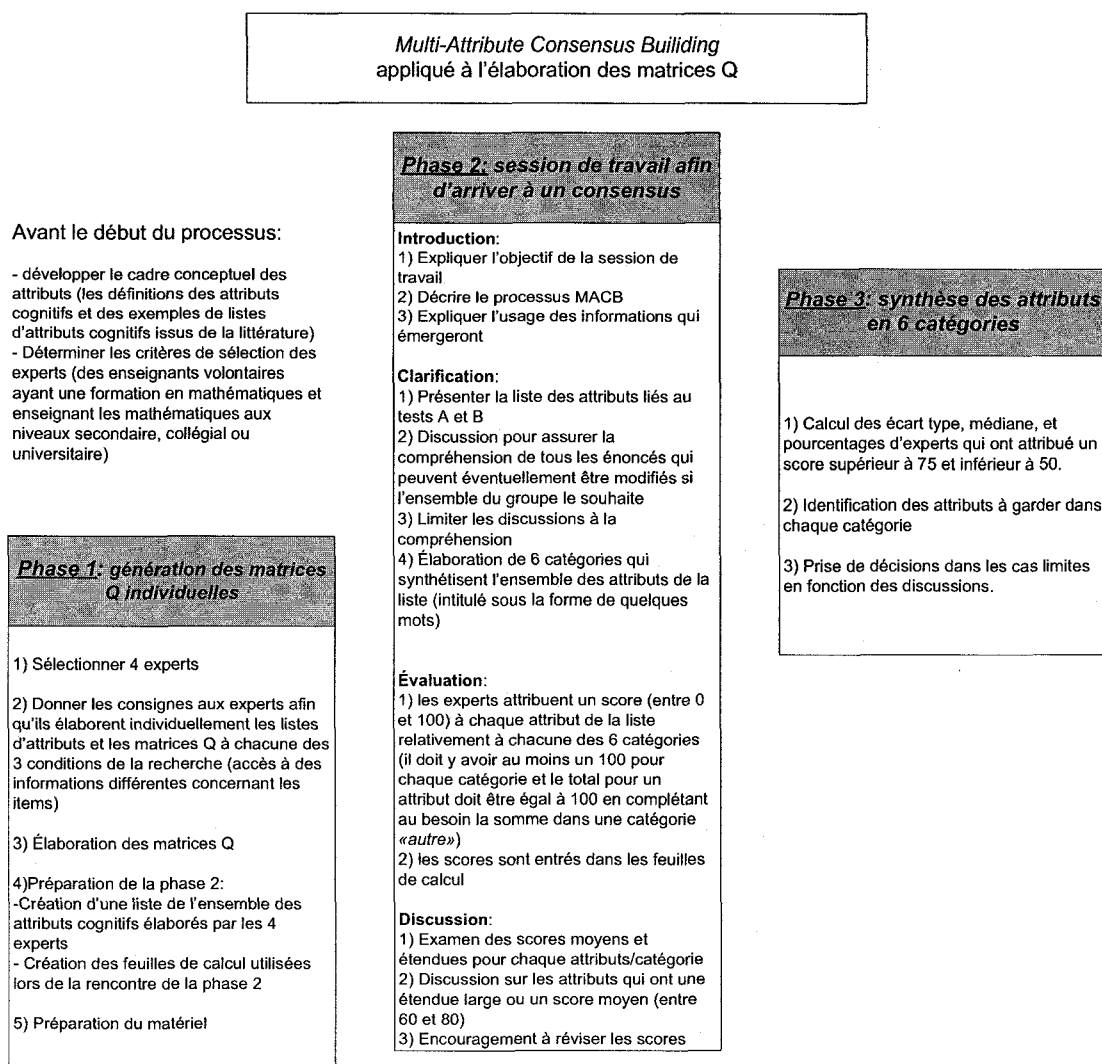


Figure 5. Les trois phases et les tâches à accomplir lors de l'application de la version adaptée de la méthode MACB

Sous cette forme, la concordance des jugements des experts pour chacun des items peut faire l'objet d'une vérification. Dans un premier temps, le calcul du pourcentage des six catégories utilisé par chaque expert permet de s'assurer de la congruence item objectif. De plus, pour un item, la situation peut être décrite comme celle de multiples juges qui accordent un score de 0 ou de 1 à cet item relativement à chacun des six nouveaux attributs (les catégories de la synthèse). Le score est de 1 lorsque le juge estime qu'un lien existe entre l'attribut et l'item et le score est de 0 sinon. Le calcul du coefficient de contingence permet de juger si les experts ont placé les valeurs 1 de façon

homogène ou non. Ce coefficient mesure l'intensité de la liaison entre les valeurs 0 et 1 de deux matrices. Un test khi carré permet de tester l'hypothèse d'un degré d'association nul entre les deux matrices. Ce calcul offre la possibilité de déceler si les quatre experts sont en accord deux à deux ou si l'un d'entre eux a un comportement marginal.

La matrice synthèse a la même forme que les matrices 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. Il faut s'assurer que la matrice ainsi obtenue ne soit pas artificielle et qu'elle soit représentative de l'ensemble du travail des experts. Le calcul du coefficient de contingence peut inclure les valeurs correspondant à la matrice synthèse pour chaque item. Il est alors possible de vérifier si la matrice synthèse est en accord avec la matrice individuelle de chaque expert. Si tel est le cas, elle sera considérée comme représentative du travail de chacun d'eux.

Même si la méthode MACB doit être adaptée, la phase 1 fournit les matrices individuelles de chaque expert sous chaque condition. La phase 2 aboutit à la réalisation de la synthèse des listes d'attributs en intégrant la notion de consensus entre les experts. Cette synthèse des attributs offre la possibilité de fabriquer une matrice représentative des différents experts pour chacune des conditions dans la phase 3. Vingt quatre matrices individuelles et six matrices synthèses sont ainsi élaborées (voir figure 4). Rappelons que les vingt quatre matrices individuelles prennent d'abord la forme originale (24 matrices individuelles originales) fournie par chaque expert puis sont modifiées suite au consensus sur le classement des attributs (24 matrices individuelles modifiées).

2.3.2 Le traitement des matrices Q

Le modèle RUM modélise les données des tests A et B avec les matrices individuelles originales et modifiées et les matrices synthèses. L'objectif consiste à obtenir des indicateurs de la validité et de la fidélité de chacune des matrices et à les comparer. Il s'agit de vérifier si les diverses conditions expérimentales qui président à leur création ont un impact (positif ou négatif) sur la

qualité de la structure cognitive des matrices. Le modèle et la méthode d'estimation du logiciel Arpeggio font l'objet d'une description dans le but de définir ce que représente chaque paramètre et comment s'assurer de sa robustesse. La perspective de comparer les paramètres implique de vérifier la convergence des routines d'estimation et l'ajustement aux données, mais aussi d'appliquer le modèle selon une procédure identique pour chaque matrice. Celle-ci est décrite.

2.3.2.1 Le modèle RUM

Le modèle de fusion (Dibello et al., 1995) a été reparamétrisé par Hartz en 2002 et baptisé RUM. Il 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 jouent un rôle dans la performance des sujets. Il estime des paramètres liés aux sujets et des paramètres liés aux items et à leurs liens avec les attributs dans la matrice Q (voir figure 6).

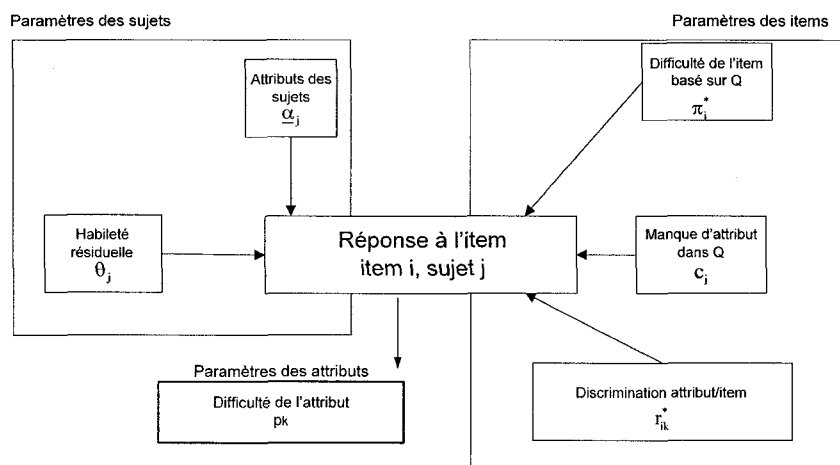


Figure 6. Les paramètres estimés dans le modèle RUM (Hartz, 2002)

Comme le montre la figure 6, un vecteur $\underline{\alpha}_i$ caractérise chaque sujet. Il représente sa maîtrise ou non de chaque attribut présent dans la matrice Q . Le paramètre θ_j correspond à son habileté

résiduelle liée à d'autres attributs cognitifs que ceux de la matrice Q. La valeur de θ_j est la projection de l'ensemble des habiletés résiduelles sur une échelle unidimensionnelle (Roussos et al., 2007).

Les attributs sont caractérisés par leur difficulté notée p_k qui correspond à probabilité de maîtriser l'attribut. Cette probabilité est estimée globalement pour l'ensemble des sujets relativement à chaque attribut. Une probabilité de maîtrise de chaque attribut est également estimée pour chacun des sujets.

Trois paramètres (π , r et c) caractérisent les items. Le paramètre π_i^* correspond à la probabilité d'appliquer convenablement tous les attributs requis par l'item i selon Q, sachant qu'ils sont maîtrisés. Il coïncide avec 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 donc en lien avec la fidélité de la matrice Q. Le paramètre r_{ik}^* représente la pénalité liée au fait de ne pas maîtriser un attribut k relié à l'item dans la matrice. Ce paramètre indique donc si la maîtrise ou non de cet attribut prédit la réussite ou l'échec à l'item. Il est un indicateur de la validité de la matrice. Le paramètre c_i est un deuxième indice de validité qui permet de juger si la matrice Q ne contient pas tous les attributs pertinents (*completeness*). Les équations du modèle font l'objet de l'annexe H.

Il existe un seul paramètre π_i^* et un seul paramètre c pour chaque item du test. Par contre, à chaque lien établi dans la matrice Q, c'est-à-dire à chaque valeur 1, correspond une valeur estimée de r_{ik}^* . La figure 5 illustre la correspondance entre la forme de la matrice Q et les paramètres estimés par le modèle RUM.

$$Q = \begin{matrix} & \begin{matrix} \text{attributs} \\ 1 & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} & \begin{matrix} \text{item 1} \\ \text{item 2} \\ \text{item 3} \end{matrix} \end{matrix}$$



Items	Paramètre π	Paramètre r (att. 1)	Paramètre r (att. 2)	Paramètre r (att. 3)	Paramètre r (att. 4)	Paramètre r (att.5)	Paramètre c
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 5. Un exemple de lien entre la forme de la matrice Q et les paramètres estimés

Un paramètre π_i^* est estimé pour chaque item du test. Sa valeur varie entre 0 et 1 puisqu'il correspond à une probabilité. Une valeur de π_i^* 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). Ce paramètre offre la possibilité de porter un regard sur la force de la structure cognitive du test associé à sa matrice Q. La comparaison des valeurs des paramètres π_i^* pour chaque item obtenues à l'aide des différentes matrices permettra, si des variations sont observées, de porter un jugement sur la fidélité relative des matrices, par item ou globalement.

Les paramètres r_{ik}^* sont des probabilités et varient entre 0 et 1. Comme le montre l'équation (1) de l'annexe H, un sujet qui ne maîtrise pas un l'attribut k_0 requis par l'item i a une probabilité de répondre correctement à l'item i proportionnelle à $r_{ik_0}^*$. Ce paramètre 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 plus la valeur de $r_{ik_0}^*$ est proche de la valeur zéro) plus la probabilité de répondre correctement à l'item est petite. Autrement dit plus la valeur de r_{ik}^* est proche de zéro, plus l'item est discriminant pour l'attribut k .

Une matrice valide devrait établir des liens entre des attributs et les items pour lesquels la maîtrise ou 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. Cet indicateur présente l'avantage d'être facile à comparer d'une matrice à l'autre. Lorsque la moyenne des r_{ik}^* , pour un item i donné, est inférieure à 0,5 (Templin, 2004), la structure cognitive est dite élevée pour l'item i . La moyenne des r_{ik}^* peut également être calculée pour chaque attribut. Les comparaisons des valeurs de r de diverses matrices et des moyennes des valeurs de r pour un item, ou encore pour un attribut, devraient permettre de juger si certaines matrices ont une structure cognitive plus ou moins forte. Le paramètre r^* joue un rôle essentiel dans le processus d'évaluation de la capacité diagnostique d'un test associé à une matrice Q et de la validité de la matrice Q .

Le paramètre c_i quantifie la force de la dépendance aux attributs non présents dans Q . Il varie entre 0 et 3 (Roussos et al., 2007; Templin, 2004). Comme il est inversement relié à cette force, plus la 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. Tout comme pour le paramètre π_i^* , un paramètre c est estimé pour chaque item. Il est donc possible de comparer les valeurs estimées pour chaque item à l'aide de matrices différentes. Les valeurs les plus grandes démontrent dans ce cas que la matrice est plus valide. Elle peut ne l'être que pour certains items.

2.3.2.2 La méthode d'estimation du RUM dans Arpeggio

La version reparamétrisée du modèle de fusion a été développée dans un cadre bayésien et peut être appliquée aux données à l'aide du logiciel Arpeggio. Le paramètre θ_j est établi comme ayant a priori une distribution normale. Les paramètres dichotomiques $\underline{\alpha}_i$ sont modélisés comme étant des variables aléatoires de Bernoulli avec une probabilité de succès notée p_k correspondant à la probabilité de maîtrise de l'attribut k dans la population. Les p_k représentent donc la difficulté des attributs. Ainsi les préalables des $\underline{\alpha}_i$ sont constitués des p_k et des corrélations tétrachoriques entre chaque paire d'attributs. Enfin les paramètres π_i^* , r_{ik}^* et c_i sont définis a priori comme ayant des distributions Beta pour la souplesse qu'elles laissent aux formes des distributions (Roussos et al., 2007).

Hartz (2002) a fait le choix de la méthode d'estimation des chaînes Monte Carlo de Markov (MCMC) en raison de sa grande flexibilité et de la possibilité qu'elle donne de faire des inférences dans les modèles complexes. En effet, une approche de type EM (espérance et maximisation) est plus facile à programmer, mais plus difficile à adapter à ces nouveaux modèles complexes (Roussos et al., 2007). La courte description de la procédure d'estimation MCMC qui suit est largement inspirée de ce que Roussos et ses collègues ont écrit en 2007. La méthode consiste à générer des chaînes de valeurs simulées pour estimer l'ensemble des paramètres. Les périodes du départ ignorées (*burn-in*) correspondent à la partie de la chaîne qui n'est intentionnellement pas utilisée pour l'estimation des distributions a posteriori des paramètres. D'un point de vue théorique, une chaîne avec un nombre suffisant de périodes de départ non utilisées pour les estimations (appelées les phases *burn-in* de la chaîne) devrait aboutir à une bonne approximation des distributions Bayésiennes attendues a posteriori.

De plus l'utilisation de la procédure Metropolis-Hastings avec un échantillonnage de Gibb permet de simplifier les distributions utilisées et de rendre la procédure programmable et opérationnelle. Le logiciel Arpeggio offre à son utilisateur la possibilité d'utiliser des chaînes et des périodes ignorées de la longueur appropriée. Il donne accès aux distributions a posteriori des valeurs des paramètres et de leurs écarts-types. Si l'expérience permet de définir ce qu'est une longueur appropriée, Roussos et ses collègues suggèrent de travailler avec des chaînes aussi longues que 50 000 et des périodes ignorées (*burn-in*) de 40 000.

Le choix initial consistera à suivre les recommandations de Roussos et de ses collègues. Utiliser des chaînes aussi longues présente le désavantage d'augmenter les temps d'analyse que le nombre important de matrices à traiter dans cette recherche vient multiplier. On peut toutefois supposer que si une chaîne converge rapidement, le fait de l'allonger ne modifiera pas les résultats. De plus, la convergence de chaque routine d'estimation sera vérifiée et il sera toujours possible de procéder à des essais avec différentes longueurs de chaînes et de périodes ignorées afin de décider quels choix sont les meilleurs.

2.3.2.3 La convergence et l'ajustement du modèle aux données

D'un point de vue théorique, une chaîne suffisamment longue devrait fournir les distributions attendues. Roussos et ses collègues (2007) suggèrent néanmoins de commencer par s'assurer de la convergence des chaînes vers une solution, que l'on suppose être celle attendue, avant d'interpréter les paramètres obtenus. Cette convergence est parfois difficile à obtenir compte tenu de la complexité du modèle RUM et de l'ajustement du modèle aux données. Pour la vérifier, quatre alternatives existent dont les trois premières sont les plus utiles: (a) les graphiques représentant les différentes valeurs au long des chaînes, (b) les distributions estimées a posteriori, (c) les autocorrélations entre les valeurs estimées par la chaîne et (d) le coefficient R de Gelman et Rubin

(Gelman, Carlin, Stern, et Rubin, 1995). Jang (2005) détaille le processus permettant de s'assurer de la convergence des chaînes et de la fidélité des paramètres obtenus. Elle fournit quelques balises pour procéder à ces vérifications dans la pratique. Le processus implique d'utiliser plusieurs longueurs de chaînes, plusieurs longueurs d'étapes ignorées (*burn-in*), de procéder à la création et à l'interprétation de graphiques pour chaque paramètre estimé.

De nombreux essais devront donc être faits avant de déterminer les paramètres suffisamment robustes pour être considérés comme des indicateurs de la validité ou de la fidélité des matrices. Ces essais et les vérifications de la convergence, notamment sous forme de la création et de l'examen de graphiques, aideront à définir la longueur des chaînes à utiliser pour assurer une convergence adéquate. Le fait de comparer des paramètres issus d'analyses différentes implique bien sûr de procéder de façon rigoureusement identique pour chacune des analyses sous peine d'obtenir des paramètres influencés par des choix différents. La méthode de vérification est essentiellement graphique et les balises d'interprétation sont assez floues impliquant une certaine subjectivité dans le processus.

2.3.2.4 La modélisation RUM appliquée aux données et matrices de la recherche

Pour procéder à des comparaisons de paramètres obtenus à partir d'analyses différentes, la procédure doit être appliquée de façon rigoureusement identique. Dans tous les cas, la chaîne retenue a une longueur de 50000 avec une période ignorée (*burn-in*) de 40000 selon les conseils de Roussos et de ses collègues (2007). La période ignorée correspond à la longueur du début de la chaîne que l'on souhaite négliger lors de l'estimation des valeurs des paramètres a posteriori. Notons que des essais préalables avec des chaînes moins longues ont dans certains cas montré une convergence moins bonne. Ceci explique le choix de garder la longueur 50000 pour l'ensemble des analyses de cette recherche.

Dans un premier temps, deux chaînes sont utilisées simultanément (longueur 50000 et période ignorée (*burn-in*) de 40000). Il s'agit de s'assurer de leur convergence vers les mêmes valeurs des paramètres p_k , π , r et c . Le logiciel Arpeggio estime le coefficient R de Gelman et Rubin (Gelman, Carlin, Stern, et Rubin, 1995) pour chaque paramètre. Une valeur inférieure à 1,2 montre que les deux chaînes convergent sensiblement vers les mêmes valeurs (Hartz, 2002; Jang, 2005; Roussos et al., 2007).

Par la suite, on utilise une seule chaîne de longueur 50000 et de période ignorée (*burn-in*) 40000. Les graphiques permettent de vérifier la convergence des paramètres selon la procédure expliquée au point 2.3.2.3. Le logiciel Arpeggio produit deux fichiers de sortie dans lesquels les valeurs estimées au fil des chaînes sont sauvegardées. Le premier fichier (*itemtime*) concerne les paramètres d'item. Le second fichier (*pktime*) concerne les paramètres de difficulté des attributs p_k . Ils donnent la possibilité de représenter graphiquement les chaînes, la fonction de densité des valeurs et les autocorrélations entre les valeurs successives. Le logiciel R (2006) fournit les trois graphiques pour juger la convergence des chaînes. Une telle méthode de vérification peut sembler arbitraire et subjective, mais aucun autre outil n'est disponible. Le fait de croiser les résultats issus de trois graphiques différents donne une certaine fidélité au processus et constitue une étape importante.

Il faut également vérifier l'ajustement des données au modèle à l'aide du calcul de trois indices. Le premier indice correspond à la moyenne des écarts entre la difficulté observée et la difficulté estimée pour les vingt items. Le deuxième indice est la racine carrée de l'erreur moyenne entre les difficultés observées et estimées au carré. Le troisième indice représente la moyenne des valeurs absolues des écarts entre les difficultés observées et estimées. Une telle façon de faire s'inspire de Jang (2005).

Selon Roussos et ses collègues (2007), la convergence des paramètres c est parfois difficile à obtenir. Dans un tel cas, ils suggèrent d'en fixer la valeur. Le logiciel Arpeggio donne la possibilité

d'imposer à c d'être égal à 10, ce qui a pour effet de forcer le modèle à associer la variance des données aux attributs identifiés dans Q plutôt qu'à d'autres attributs.

Une dernière procédure importante consiste à identifier les valeurs du paramètre r supérieures à 0,9. Une telle valeur de r implique que l'attribut ne joue qu'un rôle minime dans la réussite de l'item. Ces liens peuvent néanmoins avoir une influence plus ou moins grande sur l'échelle commune de probabilité du modèle. Arpeggio procure un indice (*drop-worst*) afin de juger si la suppression du lien dans la matrice a un impact important. En pratique, une valeur de cet indice inférieure à 0,03 implique un changement négligeable alors qu'une valeur supérieure à 0,1 implique que supprimer ce paramètre change la fonction de réponse à l'item (Hartz, 2002; Jang, 2005). Les valeurs de r supérieure à 0,9 dont l'indice (*drop-worst*) est inférieur à 0,03 conduisent à supprimer les liens correspondants dans la matrice pour simplifier le modèle.

Retirer ces liens des matrices n'implique pas forcément qu'ils sont de mauvaise qualité. Il s'agit plutôt de faire le choix d'utiliser un modèle avec moins de paramètres pour gagner en précision et faciliter l'interprétation. Par exemple, un attribut peut être relié à un item, mais les sujets peuvent ne pas en avoir besoin pour répondre à la question. Ils peuvent aussi ne l'utiliser que dans une forme élémentaire. L'élimination des paramètres non informatifs est un aspect intéressant du modèle RUM parce qu'il concentre la puissance statistique sur ce qui permet de faire un diagnostic de qualité (Roussos et al., 2007).

Toutes les modélisations des données de cette recherche suivent la procédure définie a priori et décrite dans le tableau 5.

Tableau 5.

La procédure de modélisation avec Arpeggio appliquée à toutes les matrices synthèses

	Caractéristiques	Intérêt et résultats
Analyse 1	Nombre de chaînes : 2 Longueur des chaînes : 50000 Longueur ignorée (<i>burn-in</i>) : 40000	Estimation des valeurs du R de Gelman et Rubin pour chaque paramètre
Analyse 2	Nombre de chaînes : 1 Longueur des chaînes : 50000 Longueur ignorée (<i>burn-in</i>) : 40000	Vérification graphique de la convergence Estimation des paramètres Identification des paramètres r supérieurs à 0,9
Analyse 3 (Si l'estimation de c pose des problèmes)	Nombre de chaînes : 1 Longueur des chaînes : 50000 Longueur ignorée (<i>burn-in</i>) : 40000 Les valeurs de c sont fixées à 10	Vérification graphique de la convergence Estimation des paramètres et comparaison avec l'analyse 2 Identification des paramètres r supérieurs à 0,9
Analyse 4 (et 5 au besoin)	Nombre de chaînes : 1 Longueur des chaînes : 50000 Longueur ignorée (<i>burn-in</i>) : 40000 (Les valeurs de c sont estimées ou fixées à 10) Modification pas à pas de la matrice Q : les liens entre les attributs et les items pour lesquels la valeur de r est supérieure à 0,9 dans l'analyse 2 ou 3 (et 4) sont supprimés	Vérification graphique de la convergence Estimation des paramètres et comparaison avec les analyses 2 et 3 (et 4) Identification des paramètres r supérieurs à 0,9 (si existants)

3 Chapitre III : résultats

Les résultats présentent d'abord la description des matrices individuelles originales élaborées par les experts dans la phase 1. Par la suite, ils exposent les phases 2 et 3 de la procédure MACB. Les matrices individuelles originales sont modifiées sur la base du classement consensuel obtenu entre les experts, puis elles sont synthétisées. Les résultats décrivent les matrices modifiées et les matrices synthèses, la concordance entre les matrices individuelles modifiées ainsi que la représentativité des matrices synthèses.

Par la suite, Arpeggio permet de modéliser les données des deux tests avec toutes les matrices appropriées. Les résultats illustrent le processus itératif de simplification du modèle de même que celui de la vérification de la convergence et de la robustesse des paramètres estimés. Cette procédure amène à identifier les paramètres estimés par Arpeggio qui sont suffisamment robustes pour être comparés.

Les résultats robustes issus de la modélisation RUM sont organisés en fonction du type de paramètres. Les résultats présentent et comparent les paramètres (p_k) de difficulté des attributs. Viennent ensuite les paramètres π des items qui sont des témoins de la fidélité de la matrice. Enfin, les paramètres r puis c sont présentés et comparés. Ils sont quant à eux associés à la notion de validité de la matrice.

3.1 L'élaboration et la description des matrices

3.1.1 Les matrices individuelles originales (phase 1 de la procédure MACB)

Dans la phase 1, les experts ont élaboré individuellement vingt quatre matrices selon le plan décrit par la figure 4. Lors de la rencontre de remise des résultats 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. L'expert 2 a souligné que le plus difficile avait consisté à comprendre la notion d'attribut cognitif. Les experts 1, 2 et 3 ont travaillé avec des listes de plus de dix attributs qu'ils ont regroupés par la suite.

Dans le livret de la deuxième condition, les experts devaient consigner s'ils avaient ou non trouvé un lien entre les questions des divers regroupements factoriels. La compilation des consignations des quatre experts fait l'objet du tableau 6. Ce tableau met en évidence d'une part que tous les experts n'ont pas toujours trouvé des similitudes entre les items d'un même regroupement et d'autre part une vision différente de la part des quatre experts. En particulier les experts 1 et 3 ont trouvé des liens entre les items de sept regroupements factoriels sur dix alors que l'expert 2 n'a vu des liens entre les questions que dans un seul de ces regroupements.

Dans le livret de la troisième condition, les experts devaient consigner s'ils avaient ou non réussi à tenir compte des informations fournies pour chaque item. La compilation des résultats des experts 2, 3 et 4 fait l'objet du tableau 7. Aucun résultat ne concerne l'expert 1 car il n'a pas complété le formulaire.

Le tableau 7 met en évidence un comportement marginal de l'expert 2 qui n'a pas réussi à prendre en compte les informations fournies pour huit des items du test A et 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. Ainsi, les informations fournies dans la condition 3 ont paru utiles ou très utiles à deux des experts.

Tableau 6.

L'opinion des experts en regard des regroupements de la condition 2

Regroupements factoriels		Identification des experts qui ont vu un lien entre les questions de ce regroupement	Identification des experts qui ont plus ou moins vu un lien entre toutes ou certaines des questions de ce regroupement	Identification des experts qui n'ont pas vu de lien entre les questions de ce regroupement
Test A	Q12 et Q19	1, 2, 3, 4		
	Q7, Q9 et Q10	1,3	2, 4	
	Q2, Q8, Q11, Q15 et Q18	1	3, 4	2
	Q14, Q17 et Q20	3	1	2, 4
Test B	Q1, Q5 et Q7		3, 4	1,2
	Q2, Q9, Q10, Q11 et Q16	4	1, 2, 3	
	Q18 et Q20	1, 3	2	4
	Q8, Q12, Q16 et Q19	1, 3, 4	2	
	Q4 et Q15	1, 3	2, 4	
	Q3 et Q6	1, 3, 4	2	

Tableau 7.

L'opinion des experts en regard de l'utilité des informations de la condition 3

	Nombre d'items du test A pour lesquels les informations ont été :			Nombre d'items du test B pour lesquels les informations ont été :		
	Prises en compte	+ ou – prises en compte	Non prises en compte	Prises en compte	+ ou – prises en compte	Non prises en compte
Expert 1	-	-	-	-	-	-
Expert 2	12	0	8	9	1	10
Expert 3	12	8	0	12	8	0
Expert 4	19	1	0	20	0	0

Au total, cent six libellés différents sont attachés aux tests A et B pour les trois conditions. Ils sont intégralement présentés dans l'annexe D. La comparaison des intitulés proposés par les experts aux catégories présentes dans le tableau de classification des deux tests (voir tableau 2) met en évidence la richesse de l'approche cognitive pour établir un diagnostic. Par exemple, on peut « appliquer des règles ou des algorithmes » ou « faire des opérations algébriques » aussi bien dans une question qui porte sur les fonctions élémentaires que sur le calcul différentiel. Ces deux attributs peuvent ou non s'appliquer à des items communs. Une information diagnostique sur un attribut tel que « décoder le langage mathématiques » ne peut être obtenue à partir des sous scores des items d'une ou plusieurs des catégories définies dans le tableau de classification des test A et B.

La ressemblance des libellés est un résultat important qui facilite le processus de synthèse. En l'absence d'attributs communs aux matrices, seuls les nombres d'attributs et les nombres de liens des matrices individuelles originales sont comparables entre les conditions. Les valeurs de ces deux variables font l'objet du tableau 8. Le nombre d'attributs maximum a été fixé à dix dans les consignes. En pratique, le travail des experts a abouti à des matrices qui en contiennent de huit à douze, mais dix neuf des vingt quatre matrices ont exactement dix attributs. Seul l'expert 2 a dépassé le nombre maximum d'attributs dans la condition 3.

Un test de Friedman (voir annexe A, tableau A2) montre que le test n'influence pas statistiquement le nombre de liens établis dans les matrices ($\chi^2=0,840$ et $p=0,359$). Le fait que le test B est plus difficile que le test A n'a pas joué de rôle statistiquement significatif dans les nombres de liens établis par les experts.

Par contre, un test de signification dans une analyse de variance (voir annexe A, tableau A3) incluant à la fois l'effet de l'expert et l'effet de la condition montre que l'expert joue un rôle dans le nombre moyen de liens ($F(3, 12)=19,918$ et $p=0,000$). La condition a également influencé les nombres moyens de liens ($F(2, 12)=4,665$ et $p=0,032$). Il est le plus grand à la condition 3 et le plus

petit à condition 2. Toutefois l'interaction entre les experts et les conditions n'influence pas statistiquement le nombre moyen de liens ($F(6, 12)=1,958$ et $p=0,152$). Dans cette analyse les données des deux tests ont été regroupées.

Tableau 8.

Le nombre total de liens établis par chaque expert dans chacune des 24 matrices individuelles originales et le nombre d'attributs correspondant

		Expert 1	Expert 2	Expert 3	Expert 4	Nombre moyen de liens
Test A	Condition 1	52(10)	96(10)	45(8)	56(10)	62
	Condition 2	54(10)	75(10)	57(10)	57(10)	61
	Condition 3	48(10)	102(12)	63(10)	69(10)	71
	Total	154	273	165	182	
Test B	Condition 1	52(10)	93(10)	58(10)	83(10)	72
	Condition 2	57(10)	68(9)	50(9)	57(10)	58
	Condition 3	51(10)	85(12)	74(10)	83(10)	
	Total	160	246	182	223	

Note. Le nombre d'attributs est donné entre parenthèses à côté du nombre de liens.

L'examen du tableau 8 montre que l'expert 2 a, dans tous les cas, établi beaucoup plus de liens que ses collègues pour l'ensemble des conditions et les deux tests. C'est d'ailleurs lui qui a proposé douze attributs dans les matrices de la condition 3. L'expert 1 a globalement établi le moins de liens.

En ce qui concerne la comparaison entre les conditions d'élaboration, le tableau 8 et la figure 6 mettent en évidence un nombre de liens plus homogène entre les experts pour les deux tests à la condition 2. Un test d'homogénéité des variances montre que les variances des nombres de liens pour

l'ensemble des deux tests sont statistiquement différentes (la statistique de Levene est égale à 4,953 avec une probabilité de 0,017).

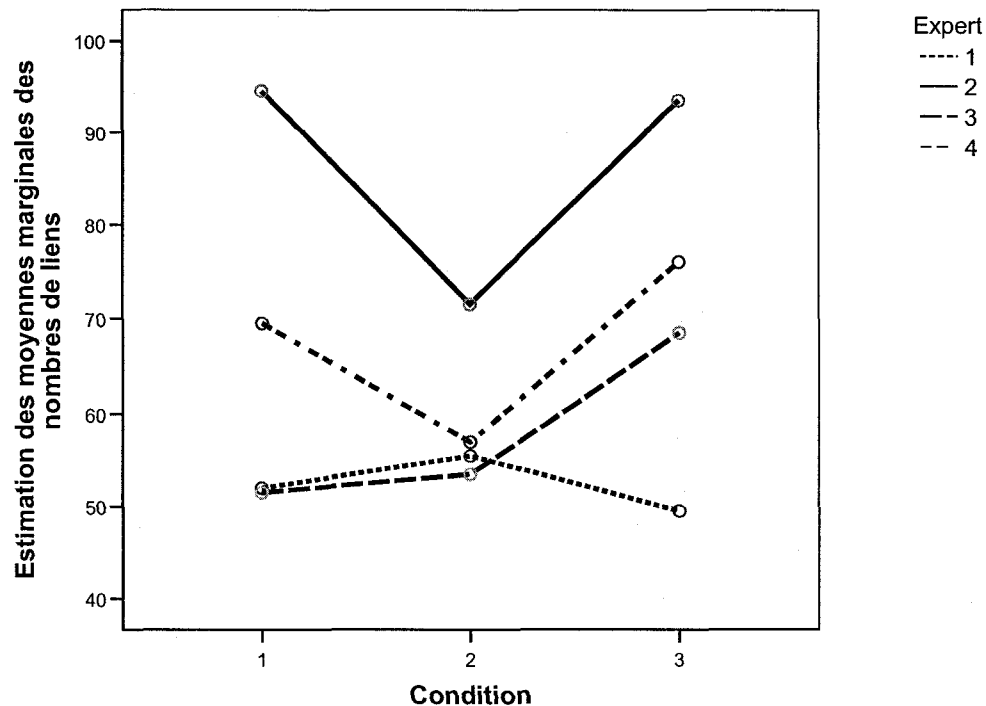


Figure 6. Les nombres moyens de liens dans les matrices individuelles originales en fonction de l'expert et de la condition pour l'ensemble des deux tests.

La figure 6 illustre le fait que l'expert 2 a un comportement assez différent des trois autres experts. Une analyse de variance sans cet expert permet de montrer (voir annexe A, tableau A4) que les trois autres experts ont une influence sur le nombre de liens établis ($F(2, 9)=4,966$ et $p=0,035$). Par contre, lorsque l'expert 2 est enlevé, la condition n'est plus statistiquement significative ($F(2, 9)=1,988$ et $p=0,193$). Même si le nombre de liens évolue d'une condition à l'autre pour les experts, aucune tendance claire ne peut être décelée. La condition 2 semble amener les experts à établir un nombre de liens plus semblable.

L'ensemble des matrices individuelles originales sont utilisées pour modéliser les données des tests A et B. Les résultats sont présentés au point 3.2.1.1.

3.1.2 Les matrices individuelles modifiées

3.1.2.1 Élaboration : les phases 2 et 3 de la procédure MACB

Les experts travaillent séparément lors de la première phase d'élaboration des matrices. Dans la phase 2, ils sont conviés à une rencontre dont l'objectif est de regrouper les attributs individuels dans six catégories qu'ils doivent préalablement définir. Le nombre total d'attributs relatifs au test A, obtenus à la phase 1, est de quatre vingt dix neuf. Ils sont regroupés et présentés en ordre alphabétique afin de mélanger les attributs des trois conditions et des quatre experts. Les sept attributs liés au test B dont les intitulés n'apparaissent pas tels quels dans le test A complètent la liste qui contient donc cent six attributs.

Ces cent six attributs doivent être classés. L'ordre alphabétique offre l'avantage de regrouper des intitulés dans certains cas assez ressemblants (voir annexe D). Les intitulés rassemblés donnent ensuite le détail du contenu de chacune des catégories et constitue une sorte de mode d'emploi pour les relier aux items.

La rencontre de la phase 2 a duré une journée. Elle a permis de définir les catégories recherchées. Afin de se conformer à la consigne et de limiter le nombre de catégorie à six, les experts ont décidé d'éliminer dès le départ les attributs au contenu non purement mathématique. Par exemple, ils ont écarté les attributs liés au fait d'utiliser les leurres pour répondre à la question (format de la question) ou encore « se faire confiance ». Les experts ont également jugé le concept de raisonnement logique comme trop peu précis et on rejeté les attributs y faisant référence.

Les discussions ont rapidement abouti aux définitions de quatre catégories clairement définies (1, 2, 4, 6) alors que deux autres catégories ont fait l'objet de plus de discussions (3 et 5). Sur ces deux dernières catégories, les débats ont d'ailleurs repris lors de la synthèse après que chaque expert

eut mis ses scores. La définition a été affinée au long du débat pour bien faire la différence entre ces deux catégories (3 et 5) et les scores de tous les énoncés en lien avec l'une ou l'autre ont été repris un par un.

Finalement, les six catégories obtenues sont les suivantes :

- 1- Décoder : tout ce qui a trait au fait de lire la question correctement (symboles mathématiques, mots clés)
- 2- Appliquer une technique : calcul algébrique, arithmétique, application de règles ou d'algorithmes
- 3- Faire des liens entre des notions
- 4- Utiliser une représentation graphique (graphique, tableau, dessin ou autre pour aider à trouver la réponse)
- 5- Utiliser des propriétés, définitions, théorèmes, éléments de base et faire des liens entre ces éléments
- 6- Organiser la solution : plusieurs étapes, stratégie

Les dimensions contenues dans le tableau de spécification des deux tests sont de l'ordre du contenu des items : fonctions élémentaires ou calcul différentiel; notion de base, type de fonction ou technique. La mise en parallèle de ces dimensions avec les catégories identifiées par les experts montre que, pour eux, le diagnostic doit se faire sur l'utilisation plutôt que sur le contenu, sur les liens que fait le sujet ou sur les actions qu'il mène. De telles informations ne peuvent donc s'obtenir en atomisant la matière et en découpant le test mais bien en créant des liens entre les items aux différents niveaux identifiés par les catégories élaborées par les experts. Le tableau de spécification découpe les tests selon un contenu, les matrices Q structurent des liens entre les items via des construits latents.

Les discussions entre les experts ont permis d'obtenir un consensus parfait pour cent un des cent six attributs, soit plus de 95%. Dans la plupart des cas de désaccord initial, les explications

fournies par l'expert se souvenant être l'auteur du libellé concerné et du contexte d'application ou du type d'item ont permis de mettre les experts d'accord. Le score moyen de 100/100 a été obtenu pour ces cent un attributs relativement à une seule catégorie. Le fichier Excel n'a pas dû être complété à la phase 3 afin d'ajouter d'autres calculs à ceux de la moyenne et de l'étendue.

Seuls cinq attributs n'ont pas été classés clairement à cause d'intitulés insuffisamment précis (par exemple, « créer des figures ou des équations pour résoudre des problèmes » à cause de la double référence à des figures et à des équations, voir les autres intitulés en annexe E). Comme ni les discussions des experts, ni un retour aux items et aux liens établis avec ces attributs n'ont permis une clarification, ces attributs ont été supprimés d'un commun accord. Il se trouve que la rédaction de quatre de ces cinq attributs revient à l'expert 2. Le résultat de la phase 3 prend la forme de six catégories identifiées par un court intitulé et comportant la liste des attributs qui y ont été classés. Ces listes font l'objet de l'annexe E. Le tableau 9 présente le nombre d'intitulés associés à chaque catégorie.

Tableau 9.

Les regroupements des attributs suite au MACB

Nombre d'attributs individuels classés (total 101)	Catégories
9	Décoder : tout ce qui a trait au fait de lire la question correctement (symboles mathématiques, mots clés)
37	Appliquer une technique : calcul algébrique, arithmétique, application de règles ou d'algorithmes
8	Faire des liens entre des notions (niveau cognitif supérieur)
9	Utiliser une visualisation graphique (graphique, tableau, dessin ou autre pour aider à trouver la réponse)
18	Utiliser des propriétés, définitions, théorèmes, éléments de base et faire des liens entre ces éléments
8	Organiser la solution : plusieurs étapes, stratégie
12 (éliminés)	Autre

3.1.2.2 Description des matrices individuelles modifiées

La mise à jour des matrices individuelles originales de chaque expert se fait en remplaçant les intitulés individuels par la catégorie correspondante puis en attribuant la valeur 1 à la catégorie dès qu'un lien existe entre un de ses intitulés et l'item. Par exemple, si dans une même matrice un item est relié à la fois à « choisir une stratégie » et à « résoudre des problèmes qui nécessitent plusieurs étapes », une seule valeur 1 est placée pour cet item relativement à l'attribut 6 (organisation) dans la matrice modifiée. Si un autre item n'est relié qu'à « choisir une stratégie », une valeur 1 est également placée pour cet item relativement à l'attribut 6 (organisation) dans la matrice modifiée.

Le nombre d'intitulés individuels associé à chaque attribut consensuel lors de la transformation des matrices varie entre 0 et 7. Un test de Friedman (voir annexe A, tableau A5) montre que le test n'influence pas statistiquement le nombre d'intitulés individuels reliés aux attributs consensuels ($\chi^2=1,000$ et $p=0,317$).

Sur les données des deux tests regroupées, une analyse de variance montre que l'expert ($F(3, 72)=17,333$ et $p=0,000$), l'attribut consensuel ($F(5, 72)=489,900$ et $p=0,000$) et la condition ($F(2, 72)=7,125$ et $p=0,002$), de même que toutes les interactions entre ces trois variables ont une influence statistiquement significative sur le nombre moyen d'intitulés individuels reliés aux attributs consensuels (voir annexe A, tableau A6). L'effet expérimental associé aux tests de signification de l'analyse de variance attire l'attention sur le fait que c'est la condition qui a l'impact le moins important ($\eta^2 = 0,165$).

Par contre l'effet expérimental associé à l'attribut ($\eta^2 = 0,971$) montre que les catégories ont été représentées par des nombres d'intitulés individuels très différents. La figure 7 met en évidence la grande représentation de la dimension technique des items (attribut 2) dans les trois conditions. La condition 3 est caractérisée par un nombre d'intitulés individuels reliés à l'attribut 2 qui augmente au

détriment des attributs 5 et 6. Les informations sur les leurres ont amené les experts à faire preuve de plus de précision face à l'aspect technique du processus de réponse.

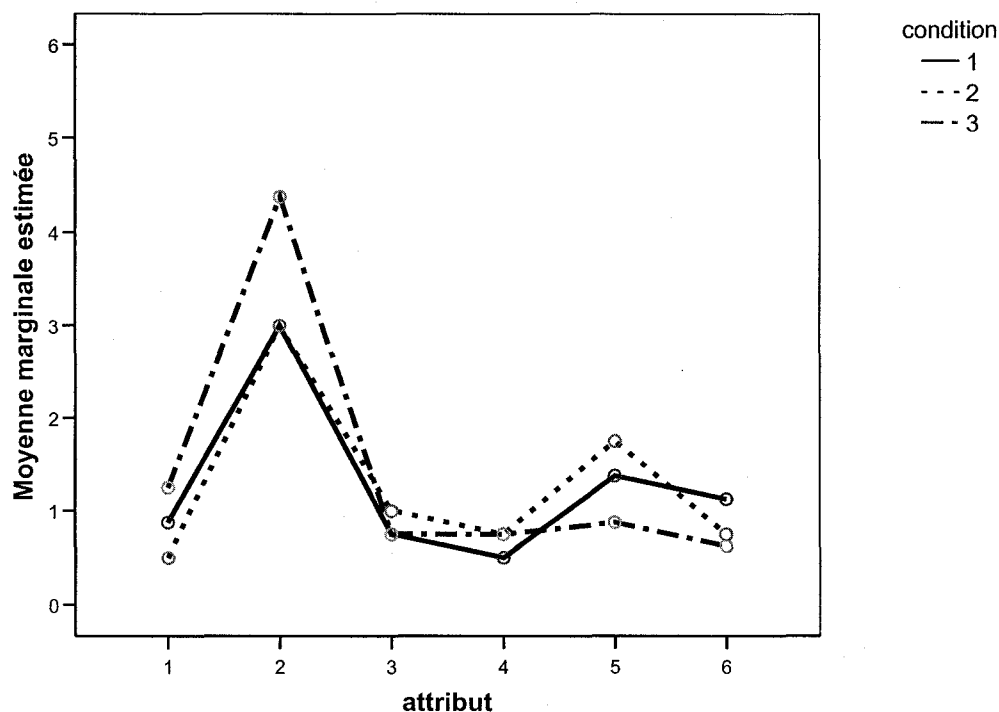


Figure 7. Le nombre moyen d'intitulés individuels relativement aux catégories de la synthèse pour les trois conditions

L'effet significatif de l'expert est associé à un effet expérimental de l'ordre de 0,42. La figure 8 décrit le nombre d'intitulés individuels en lien avec chaque attribut consensuel pour chaque expert dans chaque condition et chaque test. Elle met en évidence les différences de comportements des experts et la similarité des deux tests pour chacun d'eux. Les nombres d'intitulés individuels de l'expert 1 varient peu d'une condition à l'autre. Il est le seul à ne rien avoir relié à l'attribut 1 (décodage) à la condition 3. Il propose plus d'intitulés en lien avec l'aspect technique à la condition 3. L'expert 2 n'utilise que trois attributs aux conditions 1 et 2, mais cinq à la condition 3 (il faut

rappeler que cet expert est l'auteur de quatre des cinq intitulés qui ont été éliminés pour manque de clarté dans la phase 2). L'expert 3 utilise le même nombre d'attributs à chaque condition, mais ceux-ci varient. L'importance grandissante accordée à l'attribut 2 (technique) qui a été de plus en plus détaillé au fil des conditions au détriment des attributs 3 (liens notions) et 5 (propriétés, définitions, et liens) à la condition 3 caractérise l'expert 3. Les sept intitulés de l'expert 3 en lien avec l'attribut 2 influencent le nombre moyen de la condition 3. L'expert 4 a utilisé cinq des six attributs aux deux premières conditions, mais seulement quatre à la dernière condition. Pour cet expert, les attributs 3 et 5 ont été plus abordés aux conditions 2 et 3 qu'à la première. L'aspect technique est également plus représenté à la condition 3 (test B). L'expert 4 est seul à utiliser la catégorie 1 (décodage) à la condition 2, ce qui fera disparaître cette catégorie des matrices synthèses de la condition 2.

Les vingt quatre matrices individuelles modifiées ont la même dimension (20 par 6) et contiennent les mêmes attributs, présentés dans le même ordre. Quand un attribut n'est pas utilisé, la ligne qui lui correspond ne contient que des 0. Chaque matrice individuelle modifiée est caractérisée par : (a) le nombre de liens entre les attributs et les items et (b) le nombre d'attributs qui sont réellement reliés à des items (i.e. pour lesquels les valeurs de la matrice ne sont pas toutes nulles). Les valeurs de ces deux variables font l'objet du tableau 10.

Un test de Friedman (voir annexe A, tableau A7) montre que les nombres de liens établis dans les matrices des deux tests ne sont pas statistiquement différents ($\chi^2=0,800$ et $p=0,371$). Les données des deux tests sont regroupées et une analyse de variance (voir annexe A, tableau A8) inclut l'effet de l'expert et l'effet de la condition. Un test de signification montre que l'expert joue un rôle dans le nombre moyen de liens ($F(3, 12)=5,811$ et $p=0,011$). La condition n'influence pas statistiquement les nombres moyens de liens ($F(2, 12)=1,744$ et $p=0,216$), pas plus que l'interaction entre les experts et les conditions ($F(6, 12)=2,501$ et $p=0,083$).

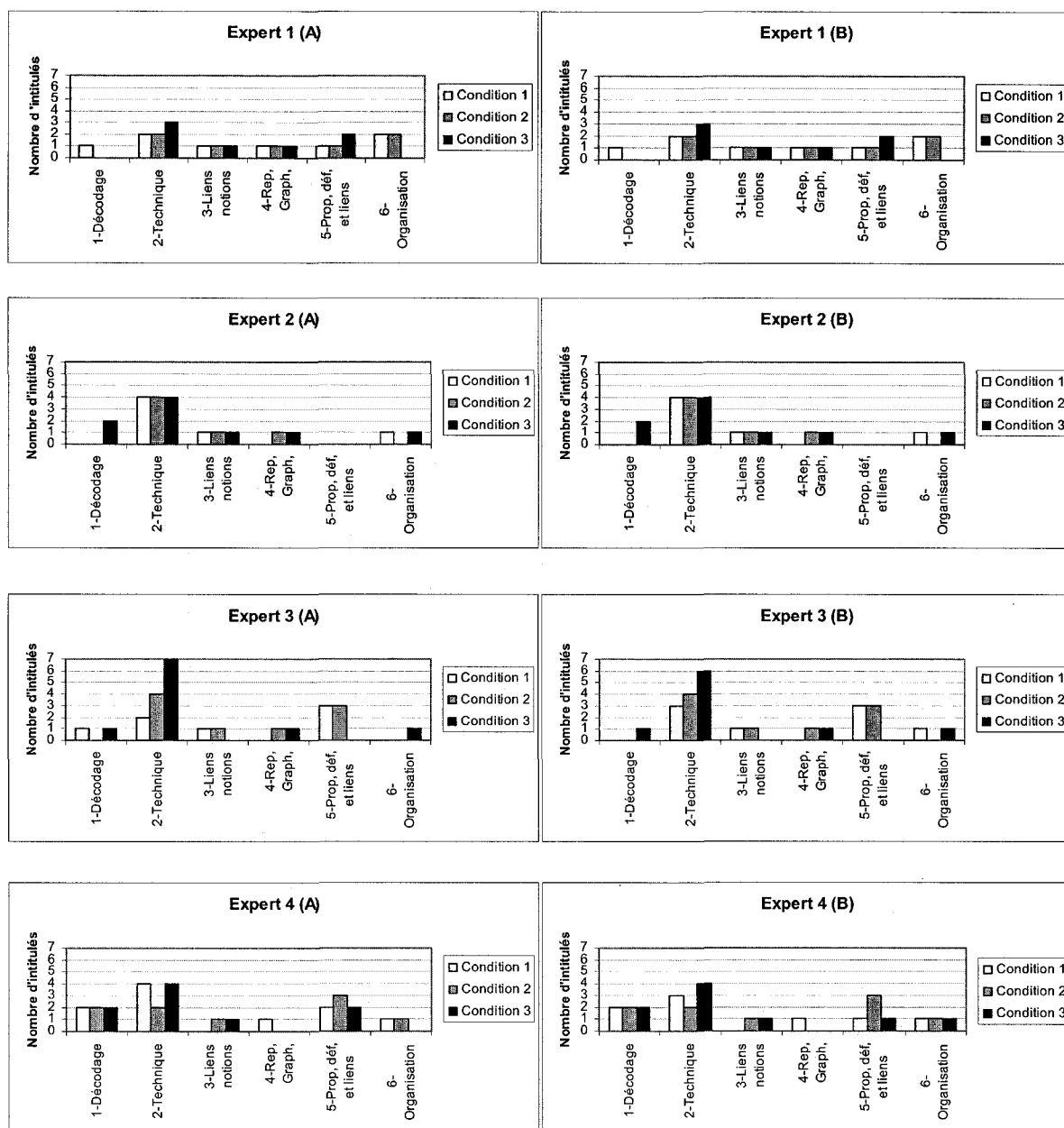


Figure 8. Le nombre d'intitulés dans les matrices individuelles des experts relativement aux catégories de la synthèse

Tableau 10.

Le nombre total de liens établis par chaque expert dans chacune des 24 matrices individuelles modifiées et le nombre d'attributs correspondant

		Expert 1	Expert 2	Expert 3	Expert 4
Test A	Condition 1	48(6)	40(3)	37(4)	44(5)
	Condition 2	38(5)	29(3)	41(4)	47(5)
	Condition 3	34(4)	54(5)	40(4)	46(4)
	Total	120	123	118	137
Test B	Condition 1	42(6)	33(3)	37(4)	61(5)
	Condition 2	43(5)	34(3)	41(4)	46(5)
	Condition 3	37(4)	45(5)	45(4)	57(5)
	Total	122	112	123	164

Note. Le nombre d'attributs est donné entre parenthèses à côté du nombre de liens

Contrairement aux matrices individuelles originales, un test d'homogénéité des variances montre que les variances des nombres de liens des matrices individuelles modifiées pour l'ensemble des deux tests ne sont plus statistiquement différentes d'une condition à l'autre (la statistique de Levene est égale à 0,226 avec une probabilité de 0,800). Elle ne le sont pas non plus d'un expert à l'autre (la statistique de Levene est égale à 0,288 avec une probabilité de 0,061).

Les nombres de liens associés à chaque attribut ne pouvaient pas être comparés dans le cas des matrices individuelles originales à cause des intitulés différents. Dans le cas des matrices individuelles modifiées, les six attributs sont identiques dans les vingt quatre matrices. Il est donc possible de réaliser une analyse de variance (en tenant compte du fait que les experts constituent une mesure répétée) afin de savoir si le nombre de liens avec chacun des attributs consensuels a subi

l'influence des conditions. Les résultats font l'objet du tableau A9 de l'annexe A. La condition a joué un rôle statistiquement significatif dans le nombre de liens établis par les experts pour l'attribut 1 (décodage) ($F(2, 3)=95,559$ et $p=0,002$) et pour l'attribut 5 (propriétés, définitions et liens) ($F(2, 3)=24,765$ et $p=0,014$).

D'après le tableau 10, les experts ont utilisé entre trois (50%) et six (100%) des six attributs consensuels dans leurs matrices (voir également la figure 8). Un test de signification dans une analyse de variance (voir annexe A, tableau A10) montre que l'expert influence statistiquement le nombre d'attributs utilisés ($F(3, 12)=59,667$ et $p=0,000$) de même que l'interaction entre l'expert et la condition ($F(6, 12)=37,667$, $p=0,000$). L'expert 2 a utilisé seulement la moitié des attributs dans ses matrices des conditions 1 et 2. Ce sont les experts 1 et 4 qui utilisent le plus les catégories synthèses. La condition ne joue pas un rôle statistiquement significatif ($F(2, 12)=3,000$ et $p=0,088$).

Le calcul du coefficient de contingence permet de juger si deux experts ont placé leurs valeurs 1 de façon homogène ou non. Il donne la possibilité de déceler si les quatre experts sont en accord deux à deux dans leur manière de positionner les valeurs 1 de leurs matrices d'une même condition pour un même test. Pour chaque test et chaque condition, six comparaisons peuvent être faites. Il y a donc trente six coefficients de contingence à calculer. Un test khi carré permet de tester l'hypothèse nulle disant que le degré d'association entre les deux matrices est nul. Les résultats sont présentés dans le tableau 11.

Onze des trente six coefficients de contingence (c) ne sont pas statistiquement significatifs, trois (2 pour le test A et 1 pour le test B) sont reliés à la condition 1, quatre sont reliés à chacune des conditions 2 (2 pour le test A et 2 pour le test B) et 3 (1 pour le test A et 3 pour le test B). Cela signifie que l'on ne rejette pas l'hypothèse nulle disant que le degré d'association est nul pour ces onze comparaisons et que les experts sont en désaccord sur les positions des valeurs 1 dans leurs matrices. L'expert 2 est le plus impliqué dans les désaccords probablement parce qu'il est seul à

n'utiliser que trois des attributs à plusieurs reprises à la suite de l'élimination de quatre de ses intitulés dans la phase 2 du MACB. Ses huit désaccords interviennent pour les deux tests et toutes les conditions. L'expert 4 est impliqué dans six désaccords pour les deux tests et les trois conditions. L'expert 1 est cinq fois en désaccord avec les autres, particulièrement avec les trois autres experts à la condition 3 pour le test B. C'est le seul cas pour lequel un expert a un comportement aussi marginal. L'expert 3 est impliqué dans trois désaccords pour les deux tests et les trois conditions.

La procédure prévoit qu'au moins deux experts doivent attribuer un 1 pour que la valeur 1 soit utilisée dans la synthèse en regard d'un item et d'un attribut. Il est utile de visualiser le nombre d'experts qui ont utilisé chacun des six attributs dans chaque condition. C'est l'objet de la figure 9. Elle montre que les quatre experts ont utilisé l'attribut 2 (technique) à toutes les conditions et que les autres attributs sont utilisés par deux, trois ou quatre experts sauf pour l'attribut 1 à la condition 2. Imposer qu'au moins deux experts aient établi un lien entre un item et un attribut est donc réaliste. L'expert 4 est le seul à avoir utilisé l'attribut 1 (décodage) à la condition 2 pour les deux tests. Ceci implique que les matrices synthèses de la condition 2 n'ont que des valeurs 0 pour cet attribut. Cela revient à dire que les items n'utilisent pas l'attribut 1 (décodage) dans les deux matrices synthèses de la condition 2.

Les matrices individuelles modifiées sont utilisées pour modéliser les données des tests A et B. Les résultats sont présentés au point 3.2.1.2. Elles servent également à élaborer les matrices synthèses.

Tableau 11.

Les 36 coefficients de contingence pour les matrices individuelles modifiées

Test A		Experts							
		1		2		3		4	
		Coefficient de contingence C	Probabilité p	Coefficient de contingence C	Probabilité p	Coefficient de contingence C	Probabilité p	Coefficient de contingence C	Probabilité p
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.

* $p < 0,05$.

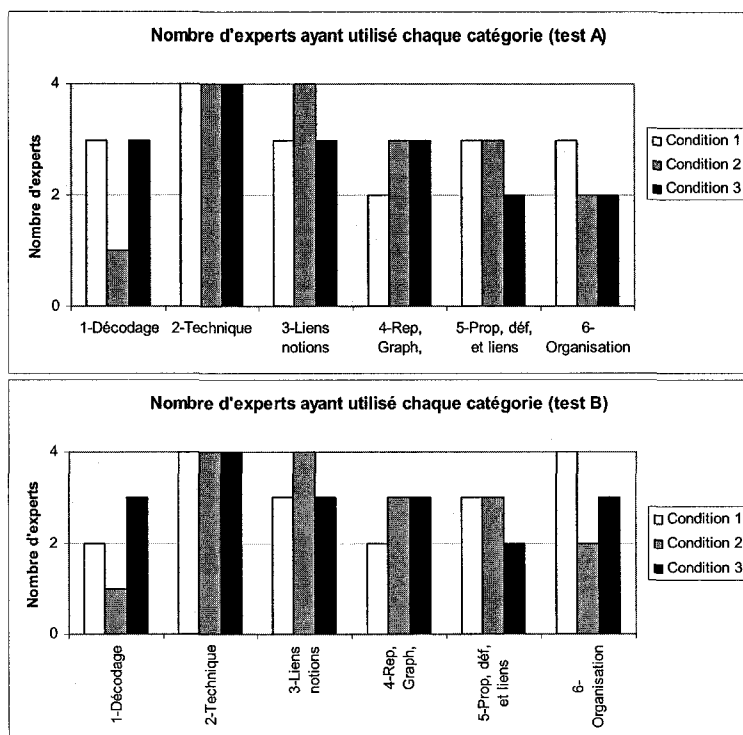


Figure 9. Le comportement des experts relativement aux 6 catégories synthèses

3.1.3 Les matrices synthèses

On obtient les six matrices synthèses en assignant la valeur 1 lorsque deux experts ou plus ont estimé qu'un lien existe entre l'item et l'attribut et la valeur 0 sinon. Tout comme pour les matrices individuelles modifiées, les catégories de la phase 2 du MACB deviennent les attributs des matrices synthèses. Elles portent les appellations de QA1, QA2, QA3 pour le test A et QB1, QB2 et QB3 pour le test B (voir figure 4).

Les six matrices synthèses ont la même dimension (20 items par 6 attributs) et contiennent les mêmes attributs, présentés dans le même ordre. L'attribut 1 n'est pas utilisé dans les matrices de la condition 2, la ligne qui lui correspond ne contient que des 0. Chaque matrice synthèse est caractérisée par : (a) le nombre de liens entre les attributs et les items et (b) le nombre d'attributs qui

sont réellement reliés à des items (i.e. pour lesquels les valeurs de la matrice ne sont pas toutes nulles) (voir tableau 12).

Tableau 12.

Le nombre total de liens établis dans chacune des 6 matrices synthèses et le nombre d'attributs correspondant

		Nombre total de liens (<i>nombre d'attributs</i>)
Test A	Condition 1	52(6)
	Condition 2	46(5)
	Condition 3	51(6)
	Total	149
Test B	Condition 1	47(6)
	Condition 2	48(5)
	Condition 3	56(6)
	Total	151

Un test des rangs de Friedman (voir annexe A, tableau A11) montre que le test n'influence pas statistiquement le nombre total de liens ($\chi^2=0,074$ et $p=0,785$). Un test de signification dans une analyse de variance avec les données des deux tests (voir annexe A, tableau A12) montre que la condition n'influence pas statistiquement le nombre total de liens des matrices synthèses ($F(2, 3)=2,389$ et $p=0,240$).

La figure 10 décrit la répartition des liens pour chaque attribut dans les six matrices synthèses.

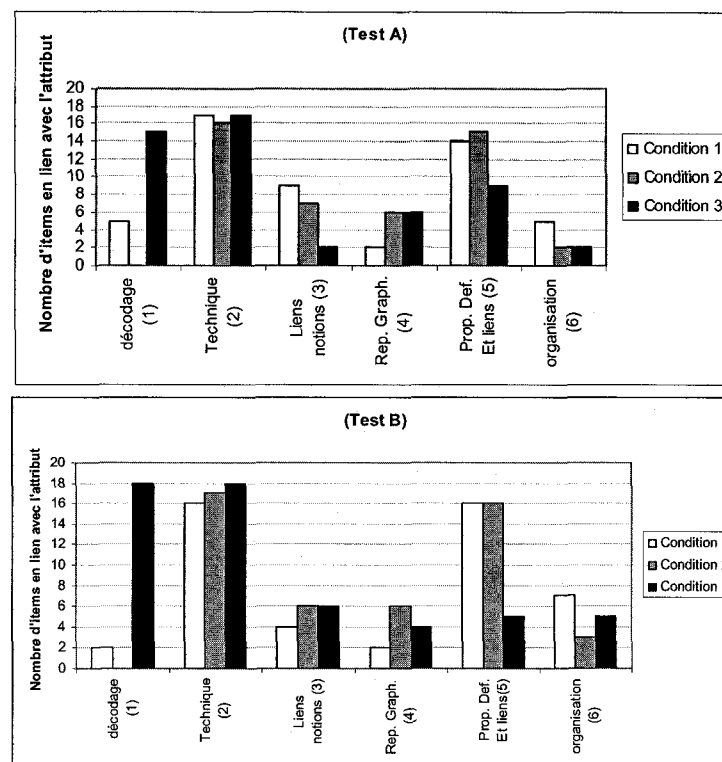


Figure 10. Les nombres de liens pour chaque attribut dans les matrices synthèses au fil des 3 conditions

Un test de Friedman (voir annexe A, tableau A13) montre que le nombre de liens par attribut n'est pas statistiquement influencé par le test ($\chi^2=0,039$ et $p=0,843$). Un test de signification dans une analyse de variance avec les données des deux tests (voir annexe A, tableau A14) montre que la condition n'influence pas statistiquement le nombre de liens pour chaque attribut dans les matrices synthèse ($F(2, 18)=1,265$ et $p=0,306$). Ces nombres de liens sont statistiquement différents d'un attribut à l'autre ($F(5, 18)=56,871$ et $p=0,000$) et l'interaction entre la condition et l'attribut joue un rôle statistiquement significatif dans le nombre de liens ($F(10, 18)=15,041$ et $p=0,000$). Ceci signifie que le nombre de liens est influencé par les conditions pour un ou plusieurs attributs.

En effet, une analyse de variance pour chacun des attributs pris séparément (voir annexe A, tableau A15) vise à mettre en évidence les effets simples pour chaque attribut. Elle permet de déterminer que la condition joue un rôle dans le nombre de liens des matrices synthèses pour

l'attribut 1 (décodage) ($F(2, 3)=50,389$ et $p=0,005$), l'attribut 4 (représentation graphique) ($F(2, 3)=13,000$ et $p=0,033$), l'attribut 5 (propriétés, définitions et liens) ($F(2, 3)=13,000$ et $p=0,033$) et l'attribut 6 ($F(2,3)=2,786$ et $p=0,207$) lorsque le taux d'erreur global est utilisé.

La figure 10 montre que l'attribut 1 (décodage) n'est (comme annoncé au point 3.1.2.2) pas utilisé à la condition 2 et qu'il est beaucoup plus utilisé à la condition 3 pour les deux tests. La fréquence d'utilisation de l'attribut 2 (technique) est semblable d'une condition à l'autre. Celle de l'attribut 3 (liens notions) diminue au fil des conditions pour le test A alors qu'elle augmente un peu pour le test B. L'attribut 4 (représentation graphique) est beaucoup plus utilisé aux conditions 2 et 3 qu'à la condition 1. L'utilisation de l'attribut 5 (propriétés, définitions et liens) tend à beaucoup diminuer à la condition 3. L'attribut 6 (organisation) est plus utilisé à la condition 1.

Avant d'appliquer le modèle RUM aux données des tests A et B avec les matrices synthèses correspondantes, il faut vérifier leur représentativité du travail individuel des experts. Le coefficient de contingence est calculé entre chaque matrice individuelle modifiée et la synthèse correspondante, aboutissant à douze comparaisons pour chaque test. Un test khi carré permet de tester l'hypothèse nulle disant que le degré d'association entre les deux matrices concernées est nul. Les résultats sont présentés dans le tableau 13.

Les résultats montrent que la matrice synthèse est en accord avec chacun des experts, pour chaque test et chaque condition. Attribuer la valeur 1 à la synthèse dès que deux experts ont établi un lien semble fournir une représentation adéquate des choix de chacun des quatre experts et ce pour les trois conditions.

Les six matrices synthèses peuvent être considérées comme représentatives du travail individuel des experts. Elles sont utilisées pour modéliser les données des tests A et B. Les résultats sont présentés dans la partie 3.2.1.3.

Tableau 13.

Les 12 coefficients de contingence entre les matrices synthèses et les matrices individuelles modifiées

		Experts							
		1		2		3		4	
		Coefficient de contingence c	Probabilité p	Coefficient de contingence c	Probabilité p	Coefficient de contingence c	Probabilité p	Coefficient de contingence c	Probabilité p
Test A									
Condition 1	Synthèse A1	0,623*	0,000	0,356*	0,000	0,503*	0,000	0,437*	0,000
Condition 2	Synthèse A2	0,562*	0,000	0,458*	0,000	0,627*	0,000	0,388*	0,000
Condition 3	Synthèse A3	0,452*	0,000	0,430*	0,000	0,519*	0,000	0,597*	0,000
Test B									
Condition 1	Synthèse B1	0,628*	0,000	0,390*	0,000	0,604*	0,000	0,459*	0,000
Condition 2	Synthèse B2	0,645*	0,000	0,424*	0,000	0,512*	0,000	0,403*	0,000
Condition 3	Synthèse B3	0,269*	0,002	0,587*	0,000	0,460*	0,000	0,582*	0,000

*p<0,05.

3.2 Le traitement des matrices

Le modèle RUM est systématiquement appliqué à l'aide du logiciel Arpeggio à toutes les matrices individuelles originales, à toutes les matrices individuelles modifiées sur la base du consensus et à toutes les matrices synthèses. Une procédure rigoureusement identique s'applique à chacune des modélisations afin de comparer les paramètres estimés dans les trois conditions lorsque ceux-ci sont suffisamment robustes.

3.2.1 L'application du modèle RUM aux matrices individuelles originales

La procédure exposée dans le tableau 4 guide la modélisation des données des deux tests avec les matrices individuelles originales. La vérification graphique de la convergence telle que décrite dans la méthodologie au point 2.3.2.3 montre des problèmes importants de convergence des paramètres estimés (Dans le souci de ne pas répéter l'exposé de la procédure de vérification de la convergence plusieurs fois au fil du texte, celle-ci sera décrite en détail pour les matrices synthèses au point 3.2.3.1).

Dans le cas des matrices individuelles originales, les paramètres estimés sont donc des indicateurs peu robustes de la qualité des matrices, leurs comparaisons deviennent sans intérêt. Le nombre trop important d'attributs explique peut-être le manque de convergence puisque le modèle est d'autant plus compliqué que le nombre d'attributs est grand.

3.2.2 L'application du modèle RUM aux matrices individuelles modifiées

Les matrices individuelles originales des quatre experts ont été modifiées à l'aide des attributs consensuels (phase 2 et 3 du MACB). Les matrices individuelles modifiées ont été décrites au point 3.1.2.2. Elles servent à modéliser les données des deux tests en suivant la procédure du tableau 4. Plusieurs des paramètres estimés sont à nouveau jugés peu robustes à cause des problèmes de convergence observés. L'hypothèse d'un nombre d'attributs trop important lors de la modélisation avec les matrices individuelles originales ne semble pas suffisante pour expliquer les problèmes de convergence puisque ceux-ci persistent avec des matrices réduites à six attributs. Les paramètres estimés ne sont pas comparés.

3.2.3 L'application du modèle RUM aux matrices synthèses et les simplifications réalisées sur ces matrices

3.2.3.1 L'application du modèle RUM aux matrices synthèses

Les données du test A sont modélisées à l'aide du modèle RUM en utilisant les matrices synthèses QA1, QA2 et QA3. Il en va de même pour les données du test B avec QB1, QB2 et QB3. Chaque première analyse utilise deux chaînes de longueur 50000 et de période ignorée de 40000 afin de calculer les valeurs des paramètres R de Gelman et Rubin (Gelman et al., 1995). À chaque paramètre estimé correspond une valeur de R.

Toutes les valeurs de R liées aux paramètres π , r et c ainsi qu'à la difficulté p_k des attributs pour le test A sont inférieures à 1,2. Cela signifie que les deux chaînes convergent vers les mêmes valeurs pour tous ces paramètres. Pour le test B, toutes les valeurs de R estimées pour la condition 1 sont inférieures à 1,2. Trois des valeurs de R pour la condition 2 (avec QB2) et cinq des valeurs de R pour la condition 3 (avec QB3) sont supérieures à 1,2 suggérant un problème de convergence. Selon Roussos et ses collègues, la triple procédure graphique de vérification de la convergence décrite dans le chapitre 2 est toutefois plus utile pour vérifier la convergence.

Cette procédure graphique est appliquée de façon systématique à toutes les analyses qui se déroulent comme prévu par le tableau 4. La deuxième analyse utilise une seule chaîne de longueur 50000 et de période ignorée 40000. Les graphiques produits représentent les valeurs estimées au fil de la chaîne, les distributions des estimations des paramètres (ou fonctions de densité) et les autocorrélations entre les valeurs estimées au long de la chaîne. Les figures 11, 12 et 13 donnent trois exemples de graphiques qui servent à vérifier la convergence. La figure 11 correspond au paramètre de difficulté de l'attribut 5 (paramètre p_{k5}) avec la matrice QA1. La figure 12 correspond au paramètre π de l'item 9 (paramètre π_9) avec la matrice QA1. La figure 13 correspond au paramètre c

de l'item 19 (paramètre c_{19}) avec la matrice QA1. Ces trois paramètres sont représentatifs des différents cas rencontrés.

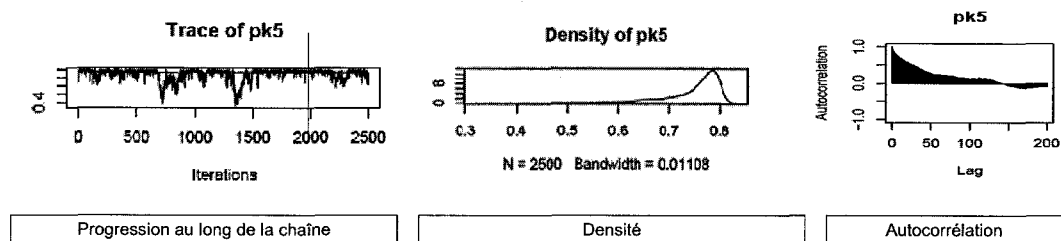


Figure 11. Les trois graphiques correspondant à la difficulté de l'attribut 5 (valeur pk_5) avec la matrice QA1

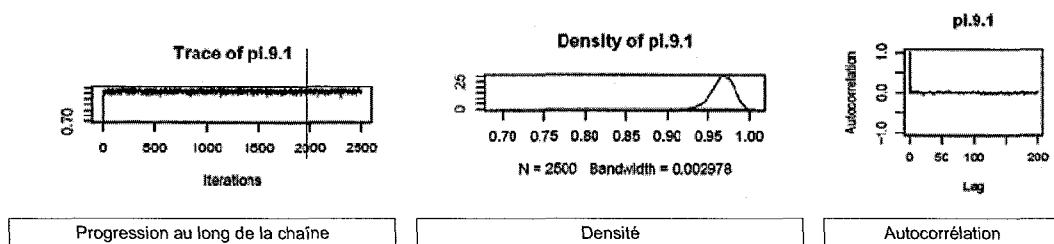


Figure 12. Les trois graphiques correspondant au paramètre π de l'item 9 (valeur π_9) avec la matrice QA1

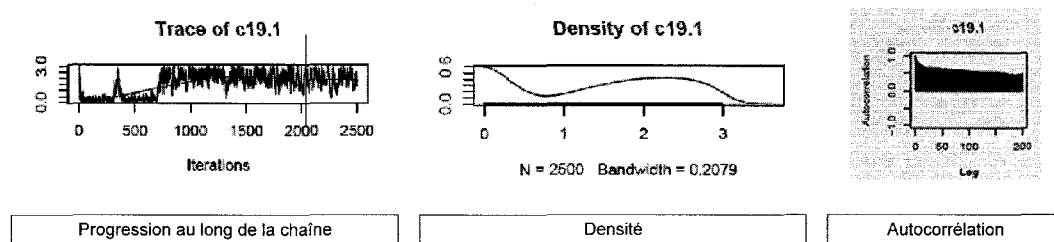


Figure 13. Les trois graphiques correspondant au paramètre c de l'item 19 (valeur c_9) avec la matrice QA1

Le paramètre p_k de l'attribut 5 (figure 11) montre une assez bonne convergence puisque le graphique illustrant la progression le long de la chaîne présente des valeurs assez stables au delà de

la période ignorée (*burn-in*). Les graphiques de ce type ont été nombreux. Notons que les estimations sont sauvegardées toutes les vingt valeurs entre 0 et 50 000 itérations. Ceci explique l'axe des abscisses gradué de 0 à 2500 (2500 correspond à 50000 divisé par 20). La période ignorée (*burn-in*) est indiquée par un trait vertical à la valeur 2000 (qui correspond à 40000 divisé par 20). En outre, le graphique de densité présente un mode unique assez précis entre les valeurs 0,7 et 0,8. Enfin, le graphique des autocorrélations décroît rapidement et la valeur est inférieure à 0,3 à *lag* 100 et inférieure à 0,2 à *lag* 200. Selon Hartz (2002) et Jang (2005), ces valeurs assurent une bonne convergence.

La convergence du paramètre π_9 (figure 12) est excellente. La progression le long de la chaîne présente des variations très faibles, le mode du graphique de densité est unique et précis et le graphique présentant les autocorrélations montre une décroissance vers 0 très rapide et sans fluctuation. Le processus de vérification systématique de la convergence a amené à observer d'assez nombreux graphiques de ce type.

La convergence liée au paramètre c_{19} (figure 13) est mauvaise. Les valeurs estimées fluctuent beaucoup et le graphique de densité présente deux modes distincts. Le manque de convergence est confirmé par le graphique des autocorrélations qui ne décroît pas rapidement et dont les valeurs restent supérieures à 0,2 à *lag* 300. Certains paramètres ont présenté des graphiques de ce type, en particulier les paramètres c .

Le fait de décider si la convergence est acceptable ou non en se basant sur la forme des graphiques a un côté arbitraire et subjectif heureusement compensé par le triple regard fourni par les trois graphiques. La figure 13 illustre le problème de convergence de plusieurs des paramètres c . Dans un tel cas, Roussos et ses collègues (2007) recommandent de fixer les valeurs de c . C'est le choix qui a été fait et systématiquement appliqué à toutes les analyses. Les conséquences liées à la suppression de c sont discutées au point 3.2.3.2.

Une chaîne de longueur 50000 avec une période ignorée (*burn-in*) de 40000 est ensuite utilisée dans laquelle les paramètres c de chaque item sont fixés. Les résultats de cette analyse servent à identifier les liens entre les attributs et les items caractérisés par une valeur de r supérieure à 0,9. Il faut s'assurer que les valeurs de l'indice (*drop-worst*) sont inférieures à 0,03 avant de simplifier la matrice Q en supprimant les valeurs 1 correspondantes. À l'issue de l'analyse utilisant cette matrice simplifiée, on réitère la procédure si nécessaire.

Le tableau 14 présente le nombre de valeurs enlevées dans chaque matrice synthèse à la fin du processus. Il contient le nombre de liens avant simplification, le nombre et le pourcentage de liens qui ont été supprimés, les attributs concernés et le nombre de liens des matrices simplifiées. Notons que tous les indices (*drop-worst*) relatifs aux changements opérés sur les matrices sont inférieurs à 0,03. Ils impliquent une transformation négligeable du modèle global.

Tableau 14.

La simplification des matrices synthèses

		Nombre total de liens dans les matrices synthèses	Nombre de liens supprimés et %	Attributs concernés	Nombre total de liens dans les matrices synthèses simplifiées
Test A	Condition 1	52	9 (17%)	1,2,3,3,3,4,4,5,5	43
	Condition 2	46	6 (13%)	4,4,5,5,5,5	40
	Condition 3	51	6 (12%)	1,1,1,3,5,5	45
Test B	Condition 1	47	5 (11%)	2,2,5,5,6	42
	Condition 2	48	6 (12%)	2,2,2,2,4,5	42
	Condition 3	56	5 (9%)	2,2,2,2,2	51

Une analyse de variance (voir annexe A, tableau A16) montre que le nombre moyen de liens supprimés n'est pas statistiquement différent d'un test à l'autre ($F(1, 4)=2,500$ et $p=0,189$) et un test

de Friedman montre que la condition (voir tableau A17 de l'annexe A) n'influence pas statistiquement le nombre de liens qui sont supprimés ($\chi^2=1,000$ et $p=0,317$).

Le tableau 14 permet d'attirer l'attention sur deux différences entre les deux tests. D'abord, c'est à la condition 1 et pour le test A que le plus de liens ont été jugés inutiles. Le fait que cette matrice a été élaborée en premier par les experts fournit peut-être une explication à cette particularité. Les cinq autres nombres de suppressions sont assez semblables (5 et 6).

De plus, les attributs concernés par les simplifications sont variables pour le test A, même si l'attribut 5 (propriétés, définitions et liens) apparaît à plusieurs reprises, alors que la quasi-totalité des liens enlevés aux conditions 2 et 3 et deux des cinq liens enlevés à la condition 1 concernent l'attribut 2 (technique) pour le test B. Il semble donc que les experts aient utilisé abusivement l'attribut 2 pour le test B, sans que tous les liens ne jouent un rôle important dans le modèle. Cette tendance s'est de plus accentuée au fil des conditions. On peut se demander si la plus grande difficulté du test B peut expliquer ce comportement des experts. Les matrices QA1, QA2, QA3, QB1, QB2 et QB3 simplifiées sont renommées QA1s, QA2s, QA3s, QB1s, QB2s et QB3s.

La convergence est vérifiée graphiquement pour chaque analyse utilisant les matrices synthèses simplifiées. Elle est jugée acceptable pour l'ensemble des paramètres des analyses impliquant le test A aux trois conditions et le test B à la condition 1. Les graphiques générés pour les analyses impliquant le test B aux conditions 2 et 3 sont légèrement moins satisfaisants comme pouvaient le laisser prévoir les quelques valeurs de R supérieures à 1,2 pour ces deux cas. Néanmoins, la convergence est considérée comme acceptable car les graphiques sont loin d'être aussi problématiques que ceux de la figure 13.

L'ajustement du modèle aux données est examiné du point de vue des items. Les indices d'ajustement font l'objet du tableau 15. Les valeurs présentées montrent que les difficultés des items

estimées par le modèle sont très semblables aux difficultés des items observées. L'ajustement du modèle aux données est donc considéré comme satisfaisant.

Tableau 15.

Les indices d'ajustement des modèles retenus pour les trois conditions

	Moyenne des écarts entre les difficultés observées et estimées	Racine carrée de l'erreur moyenne au carré	Moyenne de l'écart absolu entre la difficulté estimé et la difficulté observée
Test A			
Condition 1	-0,0224	0,0268	0,0224
Condition 2	-0,0158	0,0214	0,0163
Condition 3	-0,0121	0,0188	0,0134
Test B			
Condition 1	-0,0115	0,0198	0,0130
Condition 2	-0,0057	0,0109	0,0071
Condition 3	-0,0138	0,0225	0,0154

3.2.3.2 Les conséquences des choix

Avant de décrire les matrices synthèses simplifiées, un regard sur les conséquences des choix qui ont été faits peut être porté. Le premier choix concerne la longueur des chaînes utilisées. Les problèmes de convergence rencontrés avec les matrices individuelles ont amené à ne pas utiliser des chaînes plus courtes que celles conseillées par Roussos et ses collègues (2007). Étant donné que les mêmes longueurs ont été respectées pour l'ensemble des analyses et que les paramètres sont jugés robustes dans le cas des matrices synthèses, ce choix ne devrait pas avoir d'influence sur les résultats.

Un deuxième choix a consisté à supprimer des matrices tous les liens pour lesquels la valeur du paramètre r était supérieure à 0,9. Ces simplifications devraient avoir un impact négligeable sur le modèle selon les valeurs des paramètres (*drop-worst*).

Le dernier choix concerne le fait de fixer la valeur de c dans toutes les analyses. Une telle façon de faire force le modèle à associer la variance des données aux attributs identifiés dans Q plutôt qu'à d'autres attributs. Afin de voir quelles sont les conséquences de ce choix, les paramètres p_k , π et r estimés dans le premier modèle contenant deux chaînes et le modèle retenu sont systématiquement comparés. Ces comparaisons aboutissent aux considérations suivantes.

Le fait de fixer la valeur de c implique une diminution des paramètres p_k . La probabilité de maîtriser les attributs présents dans la matrice diminue lorsque l'influence de l'habileté résiduelle est supprimée du modèle. Les paramètres π ont également tendance à diminuer. Le fait de fixer la valeur du paramètre c a pour effet de diminuer la probabilité qu'un sujet qui maîtrise tous les attributs pertinents pour un item le réussisse. Les variations observées sont inférieures à 0,175 et toutes celles qui excèdent 0,1 sont des diminutions. Par comparaison, Jang n'a observé que des diminutions qui peuvent aller jusqu'à 0,2. Plus la valeur de π est grande, plus la matrice est fidèle. La fidélité a donc tendance à diminuer lorsqu'on n'estime pas le paramètre c dans le modèle. Étant donné que toutes les analyses sont faites de la même manière, ceci aura un impact négligeable sur les comparaisons des paramètres π des différentes conditions.

En ce qui concerne les paramètres r , rappelons que plus la valeur est petite, plus forte est la structure cognitive. Les lignes des matrices pour tous les items reliés à un seul attribut (sauf l'item A10 avec QA3 et l'item B14 avec QB2) sont plus valides lorsqu'on fixe c puisque le paramètre r correspondant diminue. Pour ces items la suppression du paramètre c a pour effet de renforcer la structure cognitive.

Pour les items reliés à plus d'un attribut, certaines valeurs de r augmentent et d'autres diminuent sans qu'il soit possible de dessiner une tendance. Toutefois, à l'exception des analyses avec QA1s, le nombre de valeurs qui diminuent est toujours supérieur au nombre de valeurs qui augmentent comme en témoigne le tableau 16. Étant donné que la suppression du paramètre c oblige les attributs présents dans la matrice à expliquer les variations, il semble logique que les paramètres r montrent dans plusieurs cas une augmentation de la puissance diagnostique du modèle.

Tableau 16.

Les différences entre l'analyse initiale et l'analyse retenue pour l'ensemble des matrices

	QA1	QA2	QA3	QB1	QB2	QB3
Nombre de paramètres r qui diminuent	19 (44%)	21 (52,5%)	24 (53%)	28 (67%)	28 (67%)	32 (63%)
Nombre de paramètres r qui augmentent	24 (56%)	19 (47,5%)	21 (47%)	14 (33%)	14 (33%)	19 (37%)
Total	43	40	45	42	42	51

Une précision doit être ajoutée en ce qui concerne le test B associé à la matrice de la condition 3. Dans ce cas particulier, le fait de fixer la valeur de c a eu pour effet de modifier les liens associés à une valeur de r supérieure à 0,9. Ceux qui ont été enlevés à la suite du processus pas à pas sont différents de ceux qui auraient été identifiés dans la toute première analyse utilisant deux chaînes. Pour les items 8, 10, 12 et 19 du test B avec la matrice de la troisième condition, les paramètres r des attributs 1 et 2 connaissent des variations dont l'amplitude est comprise entre 0,5 et 0,72. Ce phénomène n'apparaît pas dans les cinq autres matrices. Il se peut qu'il soit une conséquence d'un manque de convergence des paramètres. Il conviendra donc d'être très prudent dans les conclusions qui pourraient découler des paramètres liés à la matrice QB3.

Le fait de fixer la valeur du paramètre c et de simplifier les matrices a globalement eu pour effet une diminution des paramètres. Ceci montre l'importance d'utiliser une procédure

rigoureusement identique pour chacun des tests et chacune des conditions puisque l'objectif consiste à faire des comparaisons entre les paramètres des différentes conditions pour les deux tests.

Les modélisations RUM utilisant les six matrices synthèses simplifiées fournissent les paramètres d'items estimés qui seront comparés afin de définir si certaines conditions fournissent des matrices plus valides et plus fidèles. Ces matrices synthèses simplifiées sont tout d'abord décrites. Les paramètres issus de la modélisation RUM sont ensuite présentés puis comparés.

3.2.3.3 La description des matrices synthèses simplifiées

Le tableau 17 décrit les six matrices synthèses simplifiées. Celui-ci récapitule diverses informations concernant QA1s, QA2s et QA3s pour le test A et QB1s, QB2s et QB3s pour le test B. Il précise les nombres de changements entre les conditions.

Les résultats des analyses des nombres de liens sont semblables pour les matrices synthèses et les matrices synthèses simplifiées. Un test de Friedman (voir annexe A, tableau A18) montre que le nombre de liens par attribut n'est pas statistiquement influencé par le test ($\chi^2=0,343$ et $p=0,558$). Un test de signification dans une analyse de variance avec les données des deux tests (voir annexe A, tableau A19) montre que la condition n'influence pas statistiquement le nombre de liens pour chaque attribut dans les matrices synthèse simplifiées ($F(2, 18)=1,140$ et $p=0,342$). Ces nombres de liens sont statistiquement différents d'un attribut à l'autre ($F(5, 18)=30,670$ et $p=0,000$). L'interaction entre la condition et l'attribut joue un rôle statistiquement significatif dans le nombre de liens ($F(10, 18)=8,709$ et $p=0,000$). Ceci signifie que la condition influence le nombre de liens associés à un ou plusieurs attributs.

D'ailleurs, une analyse de variance pour chacun des attributs pris séparément (voir annexe A, tableau A20) permet de déterminer que la condition joue un rôle dans le nombre de liens des matrices synthèses simplifiées pour l'attribut 1 ($F(2, 3)=18,900$ et $p=0,020$), l'attribut 4 ($F(2,$

3)=6,333 et $p=0,084$), l'attribut 5 ($F(2, 3)=8,167$ et $p=0,061$) et l'attribut 6 ($F(2, 3)=2,545$ et $p=0,226$). Ces résultats sont similaires à ceux obtenus avant la simplification des matrices synthèses

Tableau 17.

La description des six matrices synthèses simplifiées

	Item non utilisé	Attribut non utilisé	Nombre moyen d'att. par item	Nombre total de liens / nombre possible de liens	Nombre de changements entre les conditions			
Test A								
Condition 1	Item 11	Attribut 4	2,25	43 /120		Condition 1	Condition 2	Condition 3
Condition 2	-	Attribut 1	2	40 /120	Condition 1	0	-	-
Condition 3	-	-	2,15	45 /120	Condition 2	26 (22%)	0	-
					Condition 3	37 (31%)	33(27,5%)	0
Test B								
Condition 1	Item 5	-	2,21	42 /120		Condition 1	Condition 2	Condition 3
Condition 2	Item 5	Attribut 1	2,21	42 /120	Condition 1	0	-	-
Condition 3	Item 5	-	2,68	51 /120	Condition 2	26 (22%)	0	-
					Condition 3	41 (34%)	39 (32,5%)	0

Pour le test A, le nombre de liens entre les attributs et les items varie entre 40 et 45. Les simplifications aboutissent à une matrice de la condition 1 de laquelle l'item 11 et l'attribut 4 ont été supprimés. Toutes les valeurs des paramètres r de l'item 11 étaient supérieures à 0,9 et l'attribut 4 n'était relié qu'à l'item 11. Le fait que cet item est le plus facile du test A avec un taux de bonnes réponses de 92% aurait pu expliquer que sa réussite ne donne que peu d'information sur la maîtrise ou non des attributs qui lui étaient reliés si le même phénomène s'observait aux deux autres conditions, ce qui n'est pas le cas.

Pour le test B, le nombre de liens varie entre 42 et 51. L'item 5 a été supprimé des analyses à chacune des trois conditions. Les valeurs des paramètres r de cet item aux conditions 1 et 3 étaient supérieures à 0,9, le choix a été fait de le supprimer également de la condition 2. Cet item est facile avec un pourcentage de bonnes réponses de 83%. Rappelons que l'attribut 1 n'était relié à aucun item dans les matrices synthèses de la condition 2 pour les deux tests. Un test de Friedman (voir annexe A, tableau A21) montre que le nombre de liens par item est comparable pour les trois conditions pour le test A ($\chi^2=1,583$ et $p=0,453$) et pour le test B ($\chi^2=4,378$ et $p=0,112$).

Chaque matrice contient cent vingt valeurs (0 ou 1). Le tableau 17 présente les nombres de changements opérés entre les différentes conditions pour les deux tests. C'est entre les conditions 1 et 2 qu'on observe le moins de changements pour les deux tests. C'est entre les conditions 1 et 3 que les nombres de changements sont les plus grands pour les deux tests. La figure 14 détaille les nombres de changements pour chaque item du test A et du test B, entre les conditions 1 et 2, 1 et 3 et 2 et 3. Les items des deux tests sont ordonnés de manière à présenter les nombres de changements en ordre croissant.

La mise en parallèle de la figure 14 et du regroupement factoriel fourni aux experts à la condition 2 aboutit à la création du tableau 18. Celui-ci présente le nombre de changements entre les conditions 1 et 2 et entre les conditions 1 et 3 en tenant compte de l'appartenance à un regroupement factoriel. Le tableau met d'abord en évidence le fait que le nombre moyen de liens varie toujours plus entre les conditions 1 et 3 qu'entre les conditions 1 et 2 pour chacun des regroupements factoriels, de même que pour les items isolés. En outre, il y a plus de changements pour le test B que pour le test A. Ceci peut suggérer que les experts ont plus tenu compte des informations fournies pour le test B que pour le test A, ou encore que la structure du test B se prêtait mieux à l'utilisation de ces informations.

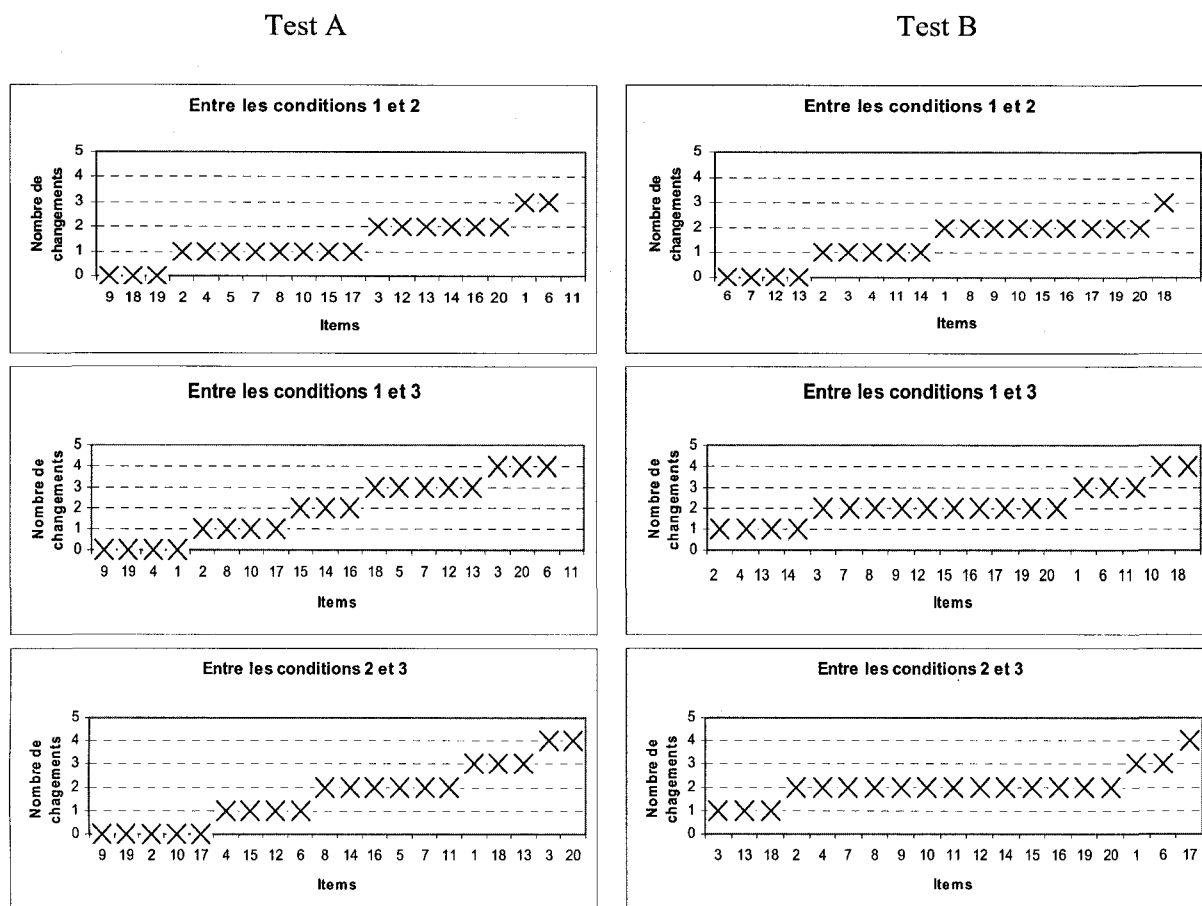


Figure 14. Les distributions par ordre croissant des changements dans les matrices relativement aux items du test A et du test B

On sait déjà que les deux tests diffèrent en difficulté et ont un nombre différent d'items dans chaque catégorie du tableau de spécification. On peut se demander si la comparaison de la structure factorielle de chaque test à son tableau de spécification donne des résultats semblables ou non. Le tableau 19 illustre le fait que la moitié des regroupements factoriels du test A (facteurs 1 et 3) et les deux tiers de ceux du test B (facteurs 1, 2, 4 et 6) sont composés à 60% ou plus d'items d'une même case du tableau de spécification. Ainsi, le regroupement factoriel a un lien plus étroit avec le tableau de spécification pour le test B que pour test A.

D'après l'opinion des experts quant à leur exploitation des regroupements factoriels (voir tableau 6), tous ont vu des relations entre tous ou certains des items du facteur 1 pour le test A et des

facteurs 2, 4 et 6 pour le test B. En ce qui concerne le facteur 3 du test A, seul l'expert 2 n'a pas identifié ce qui rapprochait les items. Les experts 1 et 2 n'ont pas trouvé ce qui regroupait les items du facteur 1 dans le cas du test B. On peut en déduire une tendance des experts à mieux exploiter l'information fournie par le regroupement factoriel lorsque les items sont clairement associés à un contenu commun identifié dans le tableau de spécification.

Tableau 18.

La mise en parallèle des changements et de la structure factorielle pour les tests A et B

	Nombre de changements entre les conditions 1 et 2	Moyenne	Nombre de changements entre les conditions 1 et 3	Moyenne
Test A				
Facteur 1 : 12, 19	$2+0=2$	1	$3+0=3$	1,5
Facteur 2 : 7, 9, 10	$1+0+1=2$	0,7	$3+0+1=4$	1,3
Facteur 3 : 2, 8, 11, 15, 18	$1+1+0^*+1+0=3$	0,6	$1+1+0^*+2+3=7$	1,4
Facteur 4 : 14, 17, 20	$2+1+2=5$	1,7	$2+1+4=7$	2,3
Items n'appartenant pas à un facteur: 1, 3, 4, 5, 6, 13, 16	$3+2+1+1+3+2+2=14$	2	$0+4+0+3+4+3+6=20$	2,9
Test B				
Facteur 1 : 1, 5, 7	$2+0^*+0=2$	0,7	$3+0^*+2=5$	1,7
Facteur 2 : 2, 9, 10, 11, 16	$1+2+2+1+2=8$	1,6	$1+2+4+3+2=12$	2,4
Facteur 3 : 18, 20	$2+3=5$	2,5	$4+2=6$	3
Facteur 4 : 8, 12, 16, 19	$2+0+2+2=6$	1,5	$2+2+2+2=8$	2
Facteur 5 : 4, 15	$1+2=3$	1,5	$1+2=3$	1,5
Facteur 6 : 3, 6	$1+0=1$	0,5	$2+3=5$	2,5
Items n'appartenant pas à un facteur: 13, 14, 17	$0+1+2=3$	1	$1+1+2=4$	1,3

Note. Les valeurs 0 marquées d'une * correspondent à des items pour lesquels la comparaison ne peut pas être faite (l'item 11 du test A n'est pas utilisé à la condition 1 et l'item 5 du test B est éliminé).

Tableau 19.

La mise en parallèle de la structure factorielle des tests A et B et du tableau de spécification

	% d'items Fonctions élémentaires Notions	Fonctions	% d'items Calcul différentiel 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%

Il reste à vérifier si les attributs reliés aux items d'un même facteur mettent en évidence ce qui rapproche les items. Ceci se fait en comptant le nombre d'items de chaque facteur qui ont un lien avec chaque attribut dans la matrice synthèse simplifiée de la condition 2. Le tableau 20 présente les nombres d'items obtenus à la condition 2, mais également aux conditions 1 et 3 afin de rendre possible les comparaisons. En ce qui concerne l'attribut 1, le tableau 20 illustre bien sa plus grande utilisation à la condition 3 pour tous les facteurs des deux tests.

Les regroupements factoriels des items du test A semblent essentiellement partager les attributs 2 (technique) et 5 (propriétés, définitions et liens). Le facteur 1 et le facteur 4 sont reliés majoritairement à l'attribut 2, alors que les facteurs 1 et 3 le sont aux attributs 2 et 5 ensemble. On

retrouve la dimension technique dans tous les regroupements factoriels du test B sauf dans le premier. Le facteur 2 se distingue par des items reliés à tous les attributs. Les facteurs 3 et 4 sont caractérisés par les attributs 2, 5 et 6. Le facteur 1 se concentre sur l'attribut 5 et les facteurs 5 et 6 sur les attributs 2 et 5.

L'attribut 1, de même que l'attribut 4 ne correspondent pas à un changement important du nombre d'items qui leur sont reliés entre la condition 1 et la condition 2. Les facteurs 1, 2 et 3 du test A et les facteurs 1 et 6 du test B voient leur nombre d'items reliés aux attributs rester relativement stable. Il semble donc que les regroupements factoriels des items qui les composent ne fournissent pas un éclairage très différent sur ces items.

En ce qui concerne le facteur 4 du test A, seule la condition 2 fournit des liens avec l'attribut 3 (liens notions) pour deux des trois items en même temps que disparaissent deux liens avec l'attribut 5 (propriétés, définition et liens). La condition 2 amène les matrices synthèses à faire disparaître tous les liens entre les items du facteur 2 du test B et l'attribut 6 (organisation). Pour finir, des quatre items du facteur 4 du test B, deux perdent leur lien avec l'attribut 2 (technique) alors que tous les items sont reliés à l'attribut 5. Les regroupements factoriels semblent donc avoir changé la perception des experts relativement aux items du facteur 4 du test A et du facteur 4 du test B, les experts ont dit avoir utilisé les informations fournies pour le facteur 4 du test B, mais pas pour le facteur 4 du test A (voir tableau 6). Notons que le facteur 4 du test A contient des items de différentes catégories du tableau de spécification, mais que 75% des items du facteur 4 du test B appartiennent à une même catégorie du tableau de spécification du test B. Il faudra étudier la qualité des liens identifiés à la condition 2 pour ces items à l'aide des paramètres r estimés par le RUM.

Indépendamment des regroupements factoriels, la figure 14 met en évidence une attitude des experts qui change selon les items au fil des conditions. Les experts ont relié certains items aux mêmes attributs dans les trois conditions (par exemple, les items 9 et 19 du test A) alors qu'ils ont

fait des choix très différents pour d'autres (par exemple, l'item 20 du test A). Cette constatation amène à répartir les items dans quatre groupes afin de chercher à savoir si l'influence des conditions varie selon que les liens évoluent ou non d'une condition à l'autre.

Tableau 20.

Nombre d'items en lien avec chaque attribut pour chaque facteur et chaque condition

Condition	Attribut 1			Attribut 2			Attribut 3			Attribut 4			Attribut 5			Attribut 6		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Test A																		
Facteur 1 : 12, 19	0	0	1	2	2	2	1	0	0	0	0	0	1	2	2	0	0	0
Facteur 2 : 7, 9, 10	0	0	1	3	3	3	1	0	0	0	0	0	1	1	1	2	1	0
Facteur 3 : 2, 8, 11, 15, 18	0	0	2	4	4	4	1	2	0	0	1	1	3	4	3	1	0	1
Facteur 4 : 14, 17, 20	0	0	2	2	3	2	0	2	0	0	0	1	3	1	1	0	0	0
Test B																		
Facteur 1 : 1, 5, 7	0	0	2	1	1	1	0	1	0	0	1	1	2	2	0	0	0	0
Facteur 2 : 2, 9, 10, 11, 16	1	0	5	2	3	4	2	3	3	2	2	2	4	4	1	3	0	1
Facteur 3 : 18, 20	0	0	2	1	2	2	0	1	1	0	1	1	2	1	2	1	2	2
Facteur 4 : 8, 12, 16, 19	0	0	4	4	2	2	0	0	0	0	0	0	1	4	0	1	0	0
Facteur 5 : 4, 15	0	0	2	2	1	1	1	0	1	0	0	0	1	2	1	0	0	0
Facteur 6 : 3, 6	0	0	2	2	2	2	0	0	0	0	0	0	2	1	0	1	1	0

Note. Lorsque le nombre d'items présente une différence de 2 items ou plus entre la condition 1 et la condition 2, les deux valeurs sont encadrées.

Le premier groupe contient les items pour lesquels on observe une stabilité de la part des experts dans leur façon de les relier aux attributs. Les items du groupe 1 présentent un changement ou moins entre les conditions 1 et 2 et entre les conditions 1 et 3. Le groupe 2 contient les items qui présentent au moins deux changements entre la condition 1 et la condition 2 seulement alors que le

groupe 3 contient ceux pour lesquels deux changements ou plus se produisent entre les conditions 1 et 3 uniquement. Dans le groupe 4 se trouvent les items pour lesquels au moins deux liens ont changé entre les conditions 1 et 2 et entre les conditions 1 et 3. Le tableau 21 présente les items appartenant à chacun des quatre groupes pour les deux tests. Les items en caractères gras sont ceux pour lesquels une différence de deux liens ou plus existe entre les conditions 2 et 3. Le tableau 21 illustre le fait que l'item 1 du test A est seul dans le groupe 2 et que plusieurs items qui n'ont pas changé entre les conditions 1 et 2 changent à la condition 3 (items, 5, 7, 15 et 18 du test A et 3, 6, 7, 11 et 12 du test B). Le fait que plus de changements existent entre les matrices du test B que du test A se retrouve ici.

Tableau 21.

Les items des tests A et B classés selon l'attitude des experts

	Groupe 1 (items stables)	Groupe 2 (changent entre 1 et 2)	Groupe 3 (changent entre 1 et 3)	Groupe 4 (items variables)
Items A	2 – 4 – 8 – 9 – 10 – 17 – 19	1	5 – 7 – 15 – 18	3 – 6 – 12 – 13 – 14 – 16 – 20
Items B	2 – 4 – 13 – 14	aucun	3 – 6 – 7 – 11 – 12	1 – 8 – 9 – 10 – 15 – 16 – 17 – 18 – 19 – 20

Note. L'item 11 du test A et l'item 5 du test B ne sont pas classés puisqu'ils ne sont pas utilisés à toutes les conditions.

Les items en caractère gras varient entre les conditions 2 et 3.

Les tests A et B comportent chacun dix items sur les fonctions élémentaires et dix items sur le calcul différentiel, mais le test B est plus difficile que le test A. Un bref aperçu du contenu des items est présenté dans le tableau 22. Les items y sont présentés selon les groupes du tableau 21 et identifiés selon que leur contenu est relié aux fonctions élémentaires (*) ou au calcul différentiel (**). Le tableau inclut la difficulté des items, chaque groupe contient des items faciles, de difficultés moyennes et difficiles. Toutefois, les trois items les plus difficiles du test A (A6, A12 et A13) et l'item le plus difficile du test B (B10) sont dans le groupe 4

Tableau 22.

Le contenu des items de chacun des 4 groupes

	Test A		Test B
Groupe 1			
A2* (0,63)	Calcul algébrique	B2** (0,38)	Calculer la pente d'une courbe en un point
A4* (0,82)	Calcul algébrique	B4** (0,38)	Calculer une limite simple (sans indétermination)
A8* (0,40)	Résoudre une équation contenant une valeur absolue	B13* (0,19)	Calcul algébrique avec racine carrée et valeur absolue
A9** (0,86)	Donner la dérivée de $\ln(x)$	B14* (0,79)	Calcul algébrique
A10* (0,60)	Calcul algébrique		
A17* (0,35)	Calcul algébrique		
A19* (0,42)	Calcul arithmétique (logarithmes)		
Groupe 2			
A1* (0,68)	Composée de fonction		
Groupe 3			
A5** (0,57)	Calculer la dérivée en un point	B3* (0,32)	Calcul arithmétique
A7** (0,35)	Calculer une limite (forme indéterminée)	B6** (0,27)	Calculer une limite (forme indéterminée)
A15* (0,49)	Résoudre une équation trigonométrique	B7* (0,55)	Calcul algébrique (logarithmes)
A18** (0,74)	Calculer la pente d'une courbe en un point	B11* (0,81)	Trouver le domaine de définition d'une fonction
		B12** (0,78)	Calculer la dérivée d'une fonction composée

Groupe 4

A3** (0,66)	La notion de fonction (continuité)	B1* (0,51)	La notion de fonction
A6* (0,29)	Résoudre une inéquation	B8** (0,84)	Calculer la dérivée d'une fonction composée
A12* (0,21)	Calcul algébrique avec racine carrée et valeur absolue	B9** (0,36)	La notion de fonction
A13** (0,22)	Trouver l'équation de la droite tangente	B10* (0,15)	La notion de fonction (périodicité)
A14** (0,84)	Calculer la dérivée d'une fonction composée	B15** (0,95)	Donner la dérivée de la fonction racine carrée
A16** (0,62)	La notion de fonction (dérivabilité)	B16* (0,44)	Calcul algébrique
A20** (0,83)	La notion de fonction (minimum)	B17** (0,45)	La notion de fonction
		B18** (0,39)	La notion de fonction (minimum)
		B19** (0,75)	Calculer la dérivée d'une fonction composée
		B20* (0,53)	Résoudre une inéquation

Note. La difficulté des items apparaît entre parenthèse à coté du numéro de chaque item. Les items dont le contenu est un calcul algébrique ou arithmétique apparaissent en caractères gras. Les items reliés aux fonctions élémentaires sont identifiées par * et ceux liés au calcul différentiel sont identifiées par **.

Le tableau 22 met en évidence un comportement différent des experts pour plusieurs items comparables appartenant aux deux tests. Tous les items du test A liés à un calcul algébrique ou arithmétique appartiennent au groupe 1 sauf l'item A12 qui est dans le groupe 4. L'énoncé de l'item A12 demande de calculer $\sqrt{a^2 - b^2}$ lorsque $a = 5 \cos x$ et $b = 3 \cos x$. L'énoncé de l'item B13 demande de calculer $\sqrt{a^2 + b^2}$ lorsque $a = 3 \sin x$ et $b = 4 \sin x$. Malgré leurs énoncés similaires et leurs difficultés comparables (la difficulté de A12 vaut 0,21 et celle de B13 vaut 0,19), l'item B13 appartient au groupe 1.

L'item A18 demande de calculer la pente en $x = 1/2$ de la courbe d'équation $y = \sin \pi x$ et l'item B2 demande la pente en $x = 1$ de la courbe d'équation $y = \sin \pi x$. Malgré leur ressemblance, l'item A18 appartient au groupe 3 et l'item B2 appartient au groupe 1. Les items A9 et B15 demandent de donner la dérivée d'une fonction de référence ($f(x) = \ln x$ et $f(x) = \sqrt{x}$) et sont très faciles (avec une difficulté de 0,86 et 0,95). Malgré leur similarité, ils se trouvent dans le groupe 1 et le groupe 4. On peut se demander si ces variations dans le comportement des experts sont voulues ou non. Par exemple, il se pourrait que les experts aient identifié des erreurs commises dans les conditions précédentes et aient tenté de les corriger par la suite. Rien ne permet toutefois de valider une telle supposition.

Pour les autres items, les experts font preuve de plus de cohérence. Les trois items en lien avec le calcul de la dérivée d'une fonction composée sont dans les groupes 3 et 4. Les deux items pour lesquels une limite est à calculer en levant l'indétermination sont dans le groupe 3. Tous les items en lien avec la notion de fonction font partie du groupe 4, de même que les deux items demandant de résoudre une inéquation. On remarque qu'il y a plus d'item du test A dans le groupe 1 et plus d'item du test B dans le groupe 4. Ceci a peut-être un lien avec le fait que le test B est plus difficile que le test A.

Les deux tests contiennent dix items liés aux fonctions élémentaires et dix items liés au calcul différentiel identifiés par une étoile (fonction élémentaires) ou deux étoiles (calcul différentiel) dans le tableau 22. La figure 15 présente le nombre d'items liés aux fonctions élémentaires ou au calcul différentiel de chacun des quatre groupes. Le groupe 1 du test A est essentiellement formé d'items liés aux fonctions élémentaires, alors que les groupes 3 et 4 sont plutôt formés d'items reliés au calcul différentiel. En ce qui concerne le test B, le groupe 4 contient plus d'items liés au calcul différentiel, mais les items du groupe 1 sont partagés équitablement entre les deux contenus. Le test A est plus facile, les items liés aux fonctions élémentaires restent plus stables alors que ce sont ceux

reliés au calcul différentiel sont plus influencés par les conditions. Il se peut que la difficulté plus grande du test B ait amené les experts à changer d'avis au fil des conditions autant pour les items liés aux fonctions élémentaires que pour ceux liés au calcul différentiel.

Les matrices des trois conditions présentent des différences dans leurs liens entre les items et les attributs. Les experts ont un comportement qui varie selon les items, et selon le test. Le test A est plus facile que le test B les experts tendent à être plus influencés par les conditions pour les items reliés au calcul différentiel. Le test A comporte plus d'items restés stables que le test B, ils sont plutôt reliés aux fonctions élémentaires.

Les comparaisons des indicateurs de qualité devront tenir compte de ces groupes d'items et de leur contenu tout en gardant à l'esprit la différence qui existe entre les deux tests.

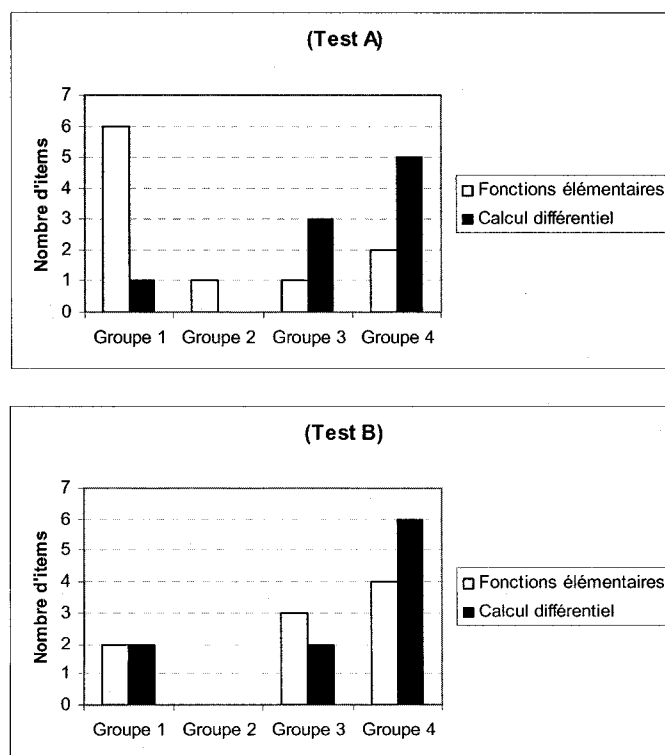


Figure 15. Le nombre d'items dans les quatre groupe en fonction du contenu, pour les tests A et B

3.3 Les comparaisons des paramètres

Cette recherche vise à définir si la validité et la fidélité d'une matrice subissent l'influence des conditions de son élaboration. Le modèle RUM a été choisi car il fournit trois paramètres d'items afin de porter un jugement sur la validité et la fidélité de la matrice utilisée. Ces paramètres ont été estimés pour les matrices individuelles originales, les matrices individuelles modifiées et les matrices synthèses simplifiées. Étant donné que seuls les paramètres obtenus à l'aide des matrices synthèses simplifiées ont été jugés robustes, ils sont les seuls à être comparés.

Le paramètre π correspond à la difficulté de l'item relativement aux attributs qui lui sont reliés. La valeur attendue de ce paramètre est 1 et une valeur supérieure à 0,6 signifie que les sujets ont une bonne chance de répondre correctement à l'item s'ils maîtrisent ces attributs. Autrement dit, si les liens établis dans la matrice définissent ce que les sujets utilisent réellement dans leur résolution de la question, la valeur sera élevée. Une telle valeur témoigne d'une utilisation régulière des attributs liés à cet item et donc de la fidélité de la matrice pour cet item.

À chaque valeur 1 présente dans la matrice Q coïncide un paramètre r . Il existe donc un paramètre r estimé pour chaque couple item-attribut identifié par Q. Ce paramètre correspond à la pénalité due au fait de ne pas maîtriser l'attribut. Il varie entre 0 et 1 et plus sa valeur est petite plus l'item est discriminant pour l'attribut. Si l'item n'est pas assez discriminant pour l'attribut, la validité du lien est faible. Une moyenne des valeurs de r inférieure à 0,5 pour un item indique que la structure cognitive est élevée et témoigne de la validité de la matrice pour cet item.

Le paramètre c est un indicateur de la validité de la matrice Q puisqu'il quantifie la force de dépendance du processus de réponse à l'item à des attributs non présents dans la matrice. Toutefois, le choix a été fait de fixer les valeurs de c . Aucun regard ne sera porté sur la validité de Q à l'aide de

ce paramètre. La difficulté (p_k) des attributs est d'abord présentée. Viennent ensuite les résultats concernant les paramètres π et r . Ils sont présentés et comparés d'une condition à l'autre.

3.3.1 La difficulté p_k des attributs : résultats et comparaisons

La difficulté des attributs (p_k) est le premier paramètre qui est comparé d'une condition à l'autre pour les matrices synthèses simplifiées. Elle correspond à la probabilité de maîtriser l'attribut. Arpeggio estime la difficulté globale de chaque attribut pour l'ensemble des sujets ainsi que la probabilité de maîtrise de chacun des sujets. Les comparaisons ne permettent pas de juger de la validité et de la fidélité des matrices. Il s'agit simplement de savoir si la difficulté des attributs varie d'un test à l'autre ou d'une condition à l'autre. Un test de signification de Friedman (voir annexe A, tableau A22) montre que les valeurs de p_k ne sont pas statistiquement différentes pour les deux tests ($\chi^2=0,498$ et $p=0,480$). Même si le test B est plus difficile que le test A, il n'y a pas de différence entre la difficulté des attributs estimées pour les deux tests. Ceci suggère peut-être qu'aucun attribut ne prend convenablement en compte la plus grande difficulté des items du test B.

Un test de signification dans une analyse de variance (voir annexe A, tableau A23) incluant les données des deux tests montre que la condition ($F(1, 1)=4,908$ et $p=0,270$) ne joue aucun rôle dans la valeur moyenne de p_k (attribut est considérée comme une mesure répétée).

La figure 16 représente la difficulté de chaque attribut pour les trois conditions et pour les deux tests. Étant donné que le test B est plus difficile que le test A, on pourrait s'attendre à ce que certains des attributs soient plus difficiles dans le cas du test B. Sur les seize paires de valeurs présentées, sept valeurs indiquent une plus grande difficulté des attributs pour le test B : trois d'entre elles concernent la condition 1 et trois concernent la condition 2, une seule des valeurs plus petites s'observe à la condition 3. À la condition 1, la difficulté des items du test B semble prise en compte essentiellement par l'attribut 5, alors que les attributs 2, 3, 4 et 6 pour la condition 2 et les attributs 2

et 4 pour la condition 3 jouent ce rôle. La condition 2 attribuerait donc la différence de difficulté entre les deux tests à plus d'attributs.

Le tableau 23 présente les valeurs des difficultés des attributs classés du plus difficile au plus facile pour chaque test et chaque condition.

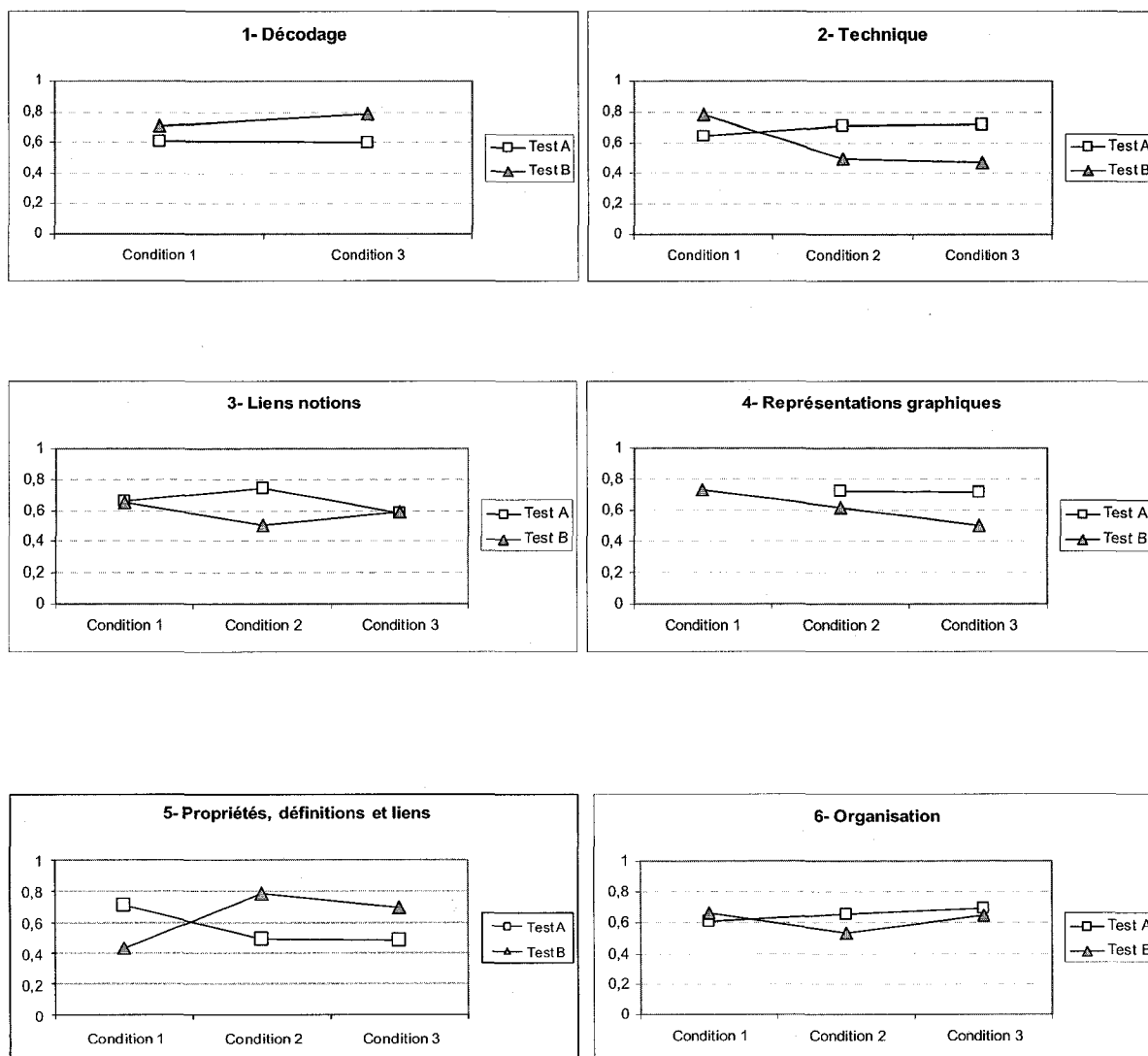


Figure 16. La difficulté des attributs pour les trois conditions et les deux tests

Tableau 23.

La difficulté des attributs

Condition 1		Condition 2		Condition 3	
Attribut	Difficulté	Attribut	Difficulté	Attribut	Difficulté
Test A					
1-Décodage	0,61	5-Prop. déf. et liens	0,50	5-Prop. déf. et liens	0,49
6-Organisation	0,61	6-Organisation	0,65	3-Liens notions	0,58
2-Technique	0,65	2-Technique	0,71	1-Décodage	0,60
3-Liens notions	0,66	4-Rep. graphique	0,72	6-Organisation	0,70
5-Prop. déf. et liens	0,71	3-Liens notions	0,75	4-Rep. graphique	0,71
				2-Technique	0,72
Difficulté moyenne	0,65		0,67		0,63
Étendue	0,10		0,25		0,23
Test B					
5-Prop. déf. et liens	0,44	2-Technique	0,50	2-Technique	0,47
3-Liens notions	0,66	3-Liens notions	0,50	4- Rep. graphique	0,50
6-Organisation	0,67	6-Organisation	0,54	3-Liens notions	0,59
1-Décodage	0,71	4- Rep. graphique	0,61	6-Organisation	0,65
4- Rep. graphique	0,73	5-Prop. déf. et liens	0,79	5-5-Prop. déf. et liens	0,70
2-Technique	0,79			1-Décodage	0,79
Difficulté moyenne	0,66		0,59		0,61
Étendue	0,35		0,29		0,31

Note. Les valeurs des difficultés inférieures ou égales à 0,5 apparaissent en caractères gras.

Dans l'hypothèse que des attributs plus difficiles à maîtriser pourraient expliquer la difficulté de certains items, on peut se demander si les items les plus difficiles de chaque test sont ou non reliés

aux attributs les plus difficiles. Pour ce faire, les items de difficulté inférieure à 0,5 sont identifiés. Selon ce critère, huit des vingt items du test A et dix des dix neuf items du test B sont considérés comme difficiles (Rappelons que l'item 5 du test B a été retiré des analyses). Les attributs de difficulté inférieure à 0,5 sont identifiés dans le tableau 23 par une valeur de p_k en caractère gras. Le tableau 24 présente la proportion d'items difficiles et d'items faciles reliés à chacun des attributs identifié comme difficile. Une telle façon de présenter les résultats vise à explorer la cohérence de chaque matrice. Le calcul du coefficient de discrimination permet d'identifier si les attributs difficiles font la différence entre les items faciles et difficiles. Comme aucun attribut n'est identifié comme difficile pour le test A à la première condition, celui-ci ne fait pas partie du tableau 24.

Tableau 24.

Les nombres d'items faciles et difficiles reliés aux attributs les plus difficiles

Attribut difficile	Condition 1			Attribut difficile	Condition 2			Attribut difficile	Condition 3		
	Nbre d'items difficiles	Nbre d'items faciles	Coeff. de discrimination		Nbre d'items difficiles	Nbre d'items faciles	Coeff. de discrimination		Nbre d'items difficiles	Nbre d'items faciles	Coeff. de discrimination
Test A											
-	-	-	-	5 (Prop. déf. et liens)	6/8	4/12	0,42	5 (Prop. déf. et liens)	6/8	1/12	0,67
Test B											
5 (Prop. déf. et liens)	9/10	4/9	0,46	2 (Technique)	8/10	5/9	0,24	2 (Technique)	9/10	4/9	0,46
				3 (Liens notions)	5/10	1/9	0,39	4 (Rep. Graphique)	3/10	1/9	0,19
				2 et/ou 3	10/10	6/9	0,33	2 et/ou 4	10/10	5/9	0,44

Note. La valeur du coefficient de discrimination est une approximation obtenue en faisant la différence entre la proportion d'items difficiles et la proportion d'item faciles. Les valeurs en caractères gras indiquent une excellente discrimination (supérieure à 0,40).

Pour le test A, six des huit items difficiles sont reliés à l'attribut 5 identifié comme le plus difficile dans les conditions 2 et 3. Dans tous les cas, il y a plus d'items difficiles que de faciles qui

sont reliés aux attributs difficiles. Les deux valeurs du coefficient de discrimination sont supérieures à 0,40 et montrent une excellente discrimination de l'attribut 5, identifié comme le plus difficile, pour les items. On peut en déduire une certaine cohérence pour les conditions 2 et 3 et le test A.

En ce qui concerne le test B, l'attribut 5 discrimine bien les items à la condition 1. À la condition 2, les deux attributs les plus difficiles sont les attributs 2 et 3. L'attribut 2 discrimine peu les items mais l'attribut 3 les discrimine convenablement. Par contre, à la condition 3, l'attribut 2 est difficile et discrimine bien les items alors que l'attribut 4 est difficile mais discrimine mal les items. La cohérence est donc problématique pour l'attribut 2 à la condition 2 et l'attribut 4 à la condition 3 pour le test B.

En plus d'estimer la difficulté globale de chacun des attributs, Arpeggio estime leur difficulté pour chacun des sujets. La difficulté correspond ici à la probabilité de maîtrise de l'attribut par chacun des sujets. Les figures 17 et 18 présentent les distributions des difficultés de chaque attribut en fonction du score obtenu par les sujets, et ce pour chaque condition. La figure 17 concerne le test A et la figure 18 le test B. L'axe des abscisses de chaque graphique représente le score total obtenu au test et l'axe des ordonnées représente la difficulté de l'attribut pour le sujet. Comme l'attribut 1 n'apparaît pas à la condition 2 pour les deux tests et l'attribut 4 n'apparaît pas à la condition 1 pour le test A, on ne dispose d'aucun graphique dans ces trois cas.

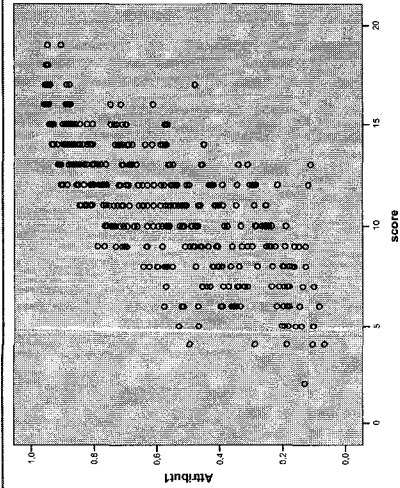
Condition 3

Condition 2

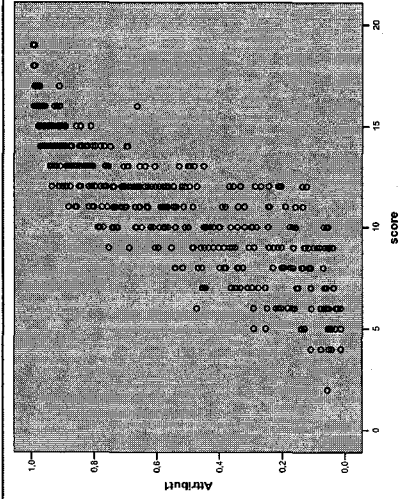
Condition 1

Attribut

1

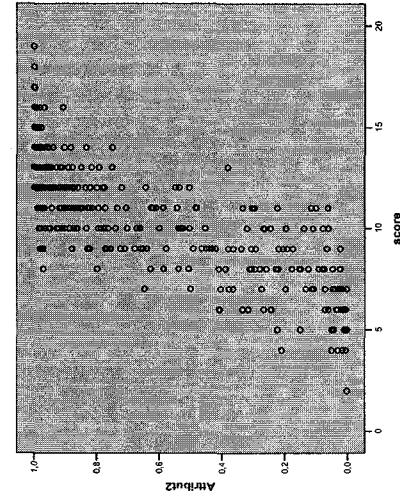
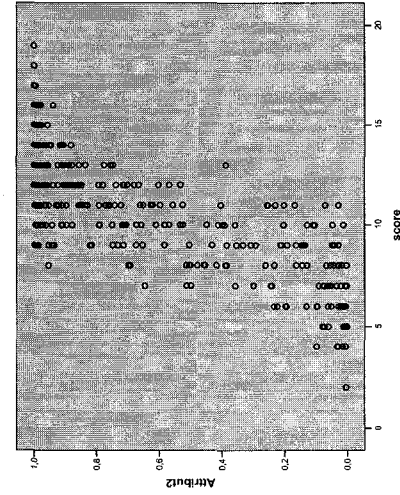
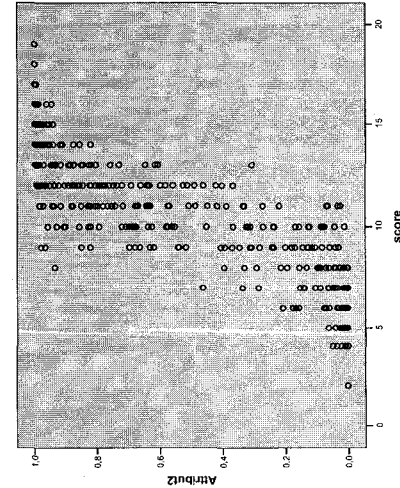


S/O



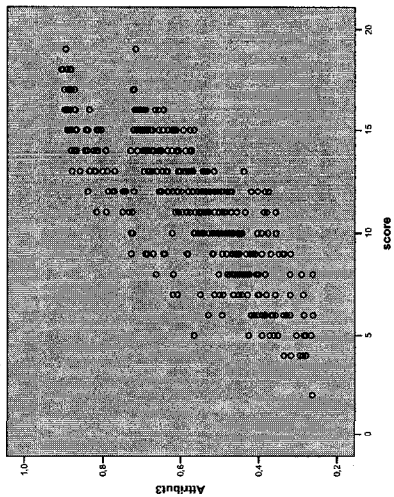
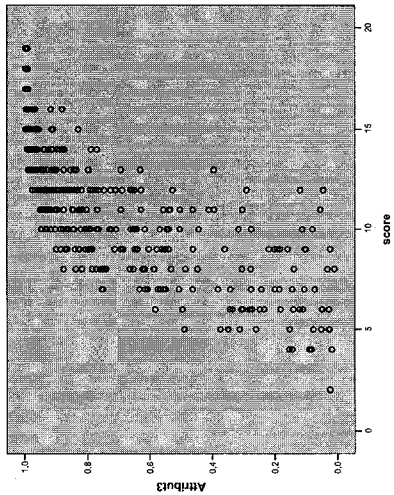
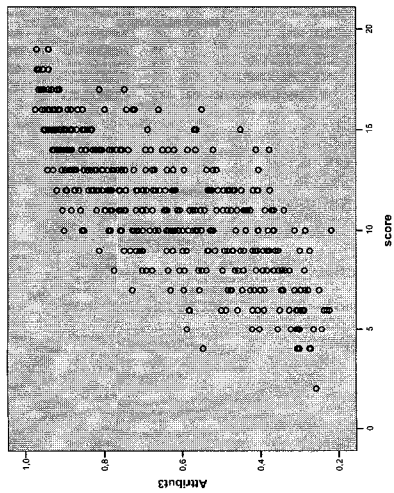
Attribut

2



Attribut

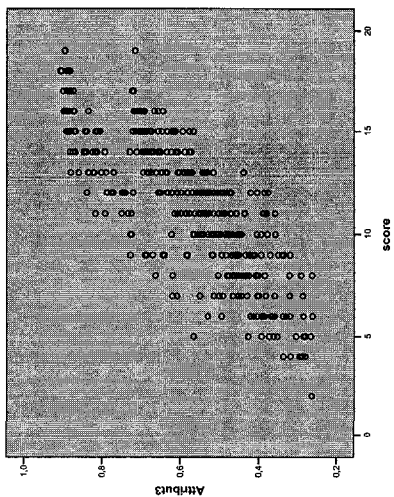
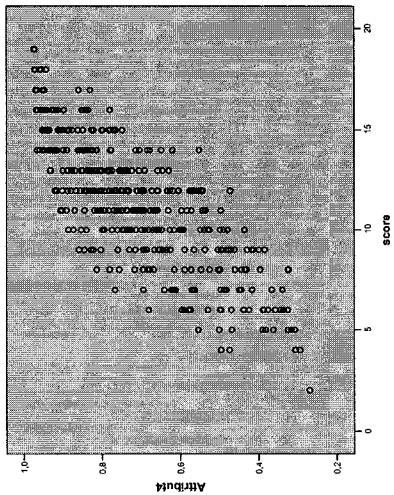
3



Attribut

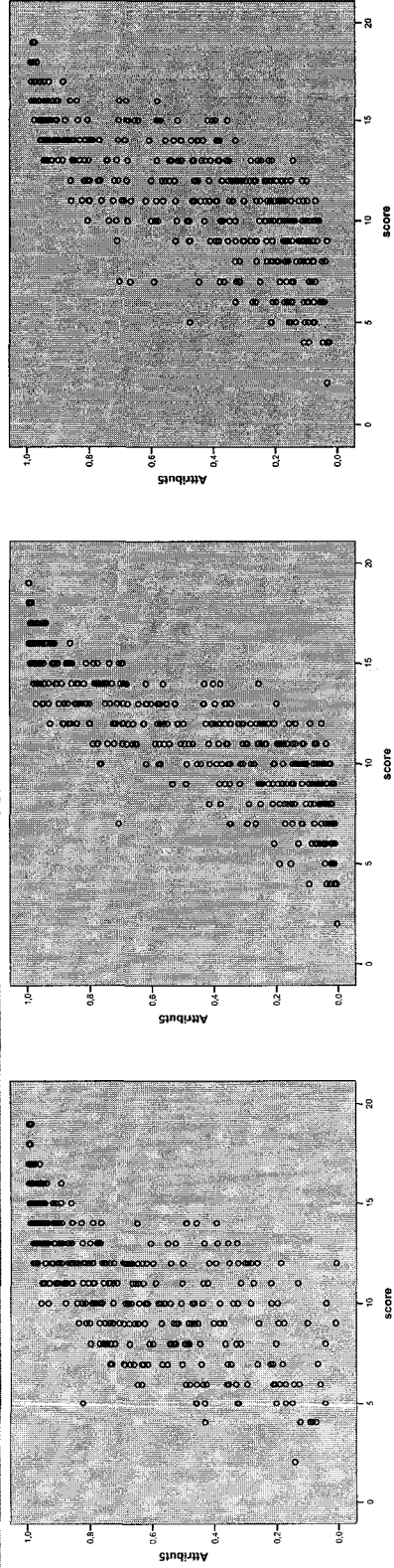
4

S/O



Atribut

5



Atribut

6

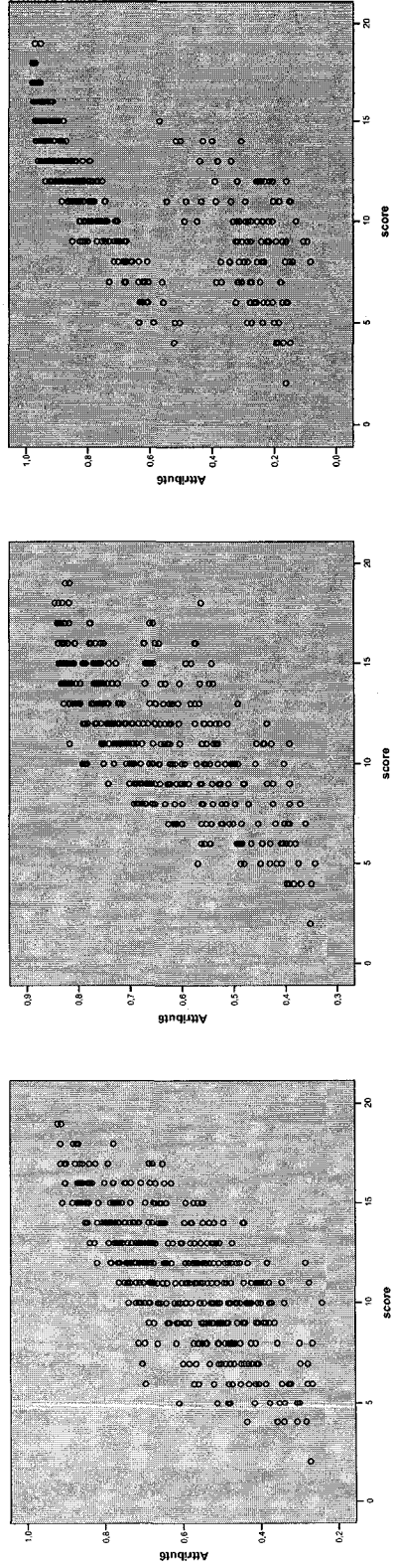


Figure 17. Les distributions des difficultés des attributs pour chaque condition en fonction du score total obtenu au test

La figure 17 donne la possibilité de faire des comparaisons entre les conditions pour le test A. Le nuage de points lié à l'attribut 1 est beaucoup plus homogène à la condition 3 qu'à la condition 1, mais n'est pas disponible à la condition 2. Pour l'attribut 2, les trois nuages ont des formes analogues. Dans les cas de l'attribut 3, de l'attribut 4 et de l'attribut 6, le nuage de points de la condition 3 se sépare en deux parties ce qui suggère pour chacun de ces attributs deux contenus distincts qui auraient probablement dû être distingués lors de l'élaboration de la matrice. Pour l'attribut 5, le nuage de points de la condition 2 est le plus homogène.

La condition 3 se caractérise par plusieurs nuages de points en deux parties. Notons qu'un retour aux intitulés individuels des experts composant ces regroupements ne fournit aucune piste d'explication aux deux parties des nuages des attributs 3, 4 et 6 de la condition 3. La condition 2 présente les portraits les plus cohérents pour le test A, car les nuages de points de la condition 1 sont généralement un peu moins homogènes que ceux de la condition 2.

La figure 18 donne la possibilité de faire des comparaisons entre les conditions pour le test B. Les nuages représentant les attributs 1 et 6 sont plus homogènes à la condition 1. Toutefois, le nuage de la condition 2 pour l'attribut 1 n'est pas disponible. Les nuages des attributs 4 et 5 se séparent à la condition 3. Pour les attributs 2 et 3, les nuages de la condition 2 sont plus homogènes. Pour l'attribut 4, les nuages des conditions 1 et 2 sont semblables et pour l'attribut 5 celui de la condition 1 est plus homogène.

En considérant la figure 18 dans son ensemble, la condition 2 fournit un portrait plus cohérent que la condition 1 ou équivalent pour trois attributs (2, 3 et 4). Tout comme pour le test A, la condition 3 pose problème pour plusieurs attributs. Le manque d'homogénéité des nuages de points et la séparation plus ou moins claire en plusieurs ensembles de points suggèrent un manque de cohérence dans l'utilisation des attributs pour cette condition.

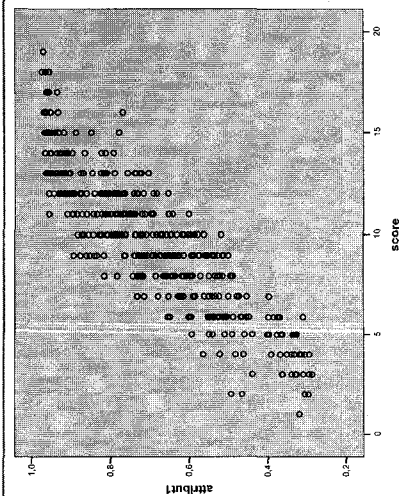
(B)Étape 1

Étape 2

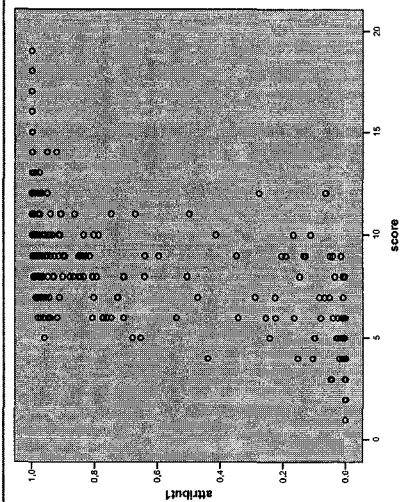
Étape 3

Attribut

1

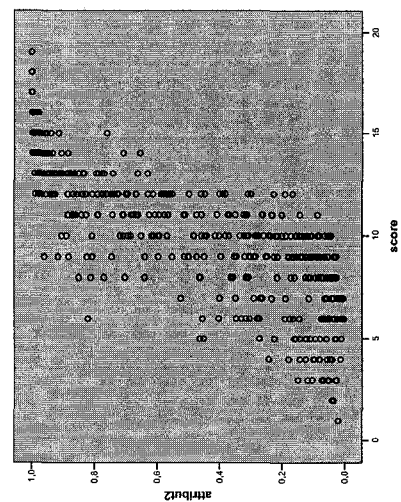
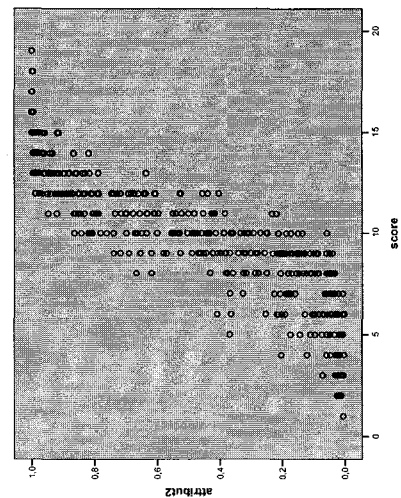
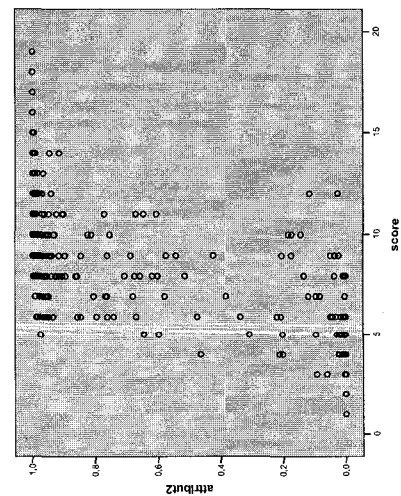


S/O



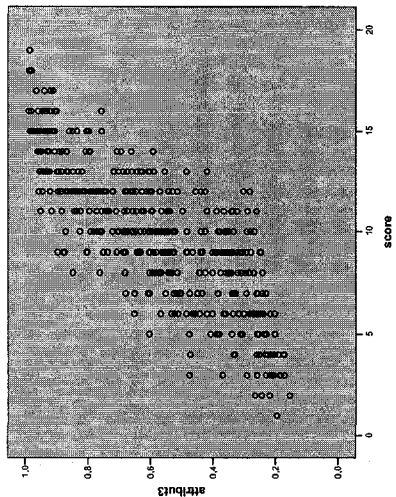
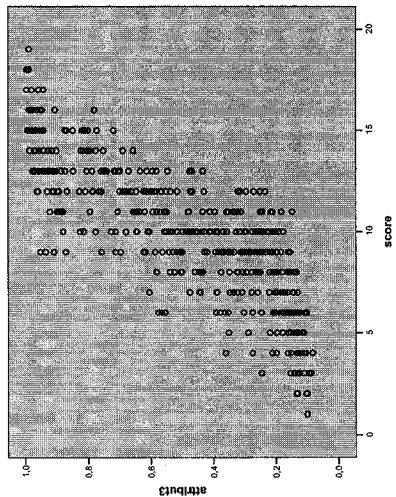
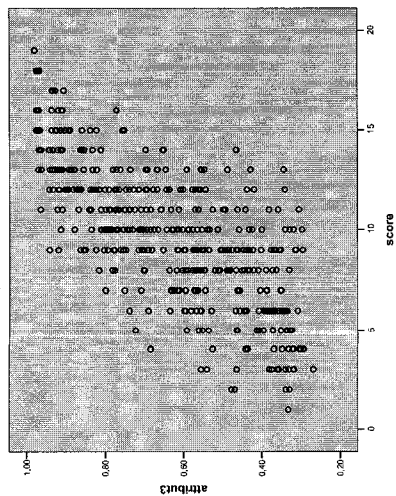
Attribut

2



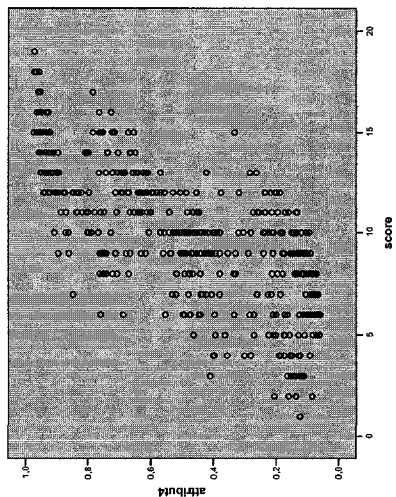
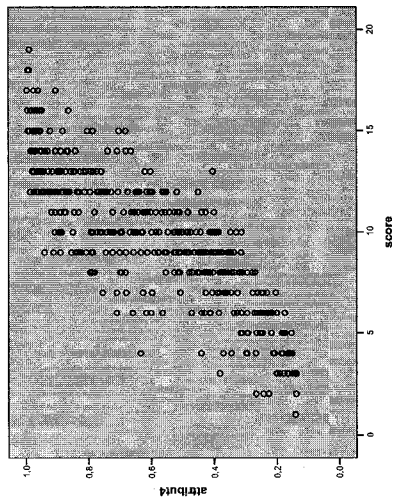
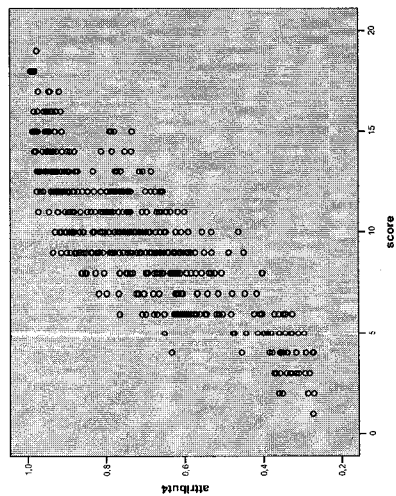
Attribut

3



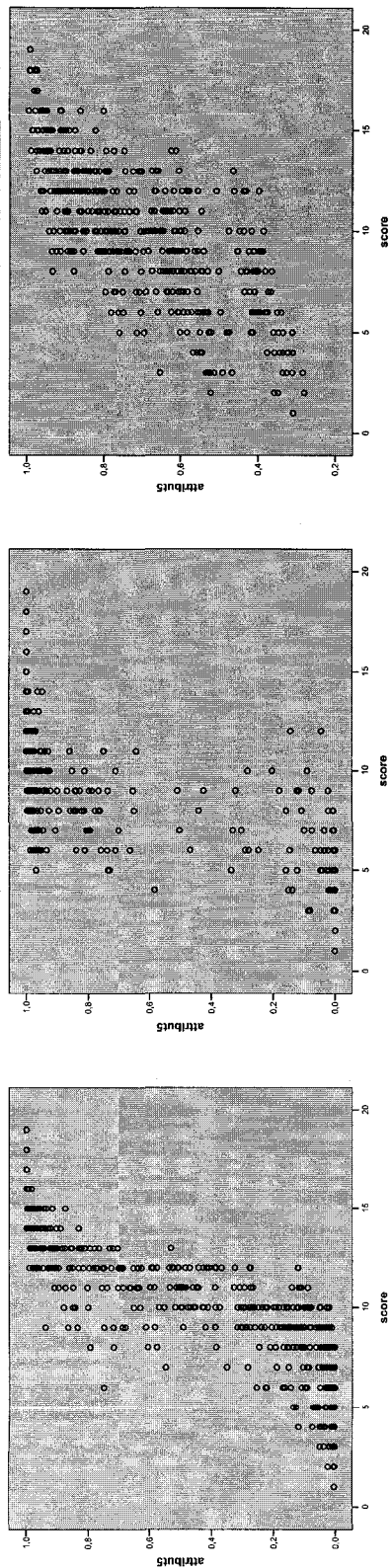
Attribut

4



Attribut

5



Attribut

6

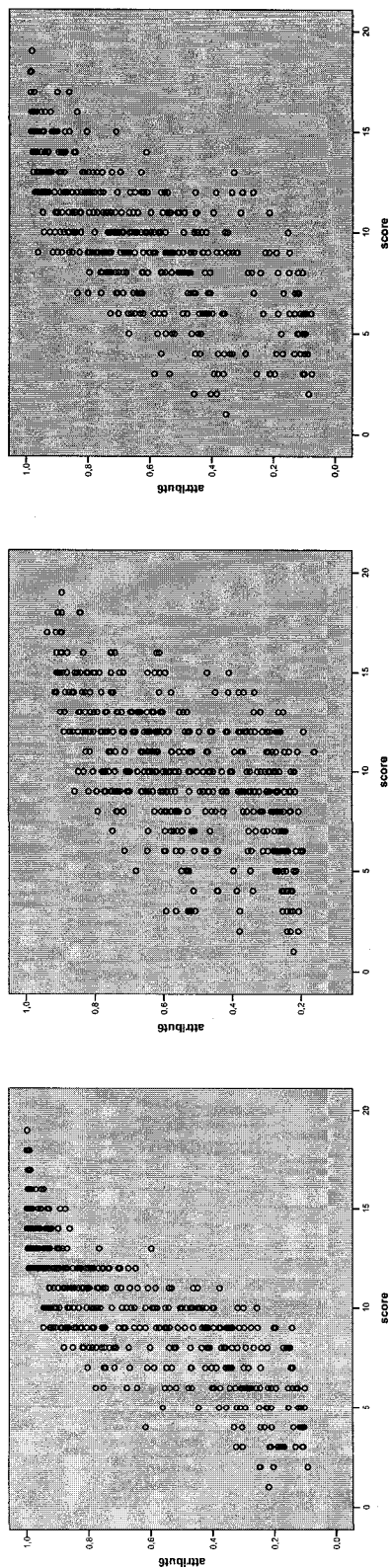


Figure 18. Les distributions des difficultés des attributs pour chaque condition et chaque attribut en fonction du score total obtenu au test

B

Les premières comparaisons concernent les paramètres de difficulté des attributs liés aux matrices synthèses simplifiées. Elles montrent que dans les deux tests, les items difficiles sont le plus souvent reliés aux attributs difficiles assurant ainsi une certaine cohérence aux matrices. En outre, le manque d'homogénéité des distributions des difficultés en fonction des scores obtenus au test par les sujets montre que la condition 3 pose le problème de la définition de certains des attributs. En effet, les nuages de points de la condition 3 se scindent parfois en deux parties plus ou moins distinctement. L'explication peut provenir de deux construits différents dans la définition des attributs concernés. Une explication basée sur un fonctionnement différentiel entre deux groupes de sujets est moins probable puisque le problème ne se pose pas aux conditions 1 et 2.

3.3.2 Les paramètres d'items π : résultats et comparaisons

Les paramètres d'items du modèle RUM peuvent être comparés dans le but de porter un regard sur la validité et la fidélité des matrices. Selon Roussos et ses collègues (2007), le paramètre π est le plus facile à interpréter. Arpeggio estime un paramètre π pour chacun des vingt items de chaque test. Ce paramètre correspond à la probabilité que les sujets appliquent correctement les attributs reliés à l'item lorsqu'ils les maîtrisent et donne une indication sur la fidélité de Q. Il varie entre 0 et 1 et sa valeur attendue devrait être proche de 1. Roussos et ses collègues suggèrent qu'une valeur de π inférieure à 0,6 indique un item dont la difficulté est sous-estimée par les attributs. Les paramètres présentés proviennent des analyses utilisant les matrices synthèses simplifiées.

Un test de Friedman (voir annexe A, tableau 24) montre que les valeurs de π ne sont pas différentes pour les deux tests ($\chi^2=0,004$ et $p=0,951$). Un test de Friedman (voir annexe A, tableau A25) montre que la condition n'influence pas la valeur des paramètres π ($\chi^2=4,500$ et $p=0,105$). Par contre, de manière logique, l'item joue un rôle statistiquement significatif dans la valeur du

paramètre π (voir annexe A, tableau A26) pour le test A ($F(19, 39)=149,848$ et $p=0,000$) et pour le test B ($F(18, 54)=138,337$ et $p=0,000$).

La difficulté joue un rôle dans la valeur des paramètres π . Un test de signification dans une analyse de variance (voir annexe A, tableau A27) montre que la difficulté des items joue un rôle dans les valeurs de π pour le test A ($F(18, 40)=70,458$ et $p=0,000$) de même que pour le test B ($F(17, 39)=43,315$ et $p=0,000$). De façon générale, la valeur de π est d'autant plus grande que l'item est facile.

Par exemple, le paramètre π de l'item A12 (calculer $\sqrt{a^2 - b^2}$ lorsque $a = 5 \cos x$ et $b = 3 \cos x$) vaut 0,28 dans la condition 1. Cet item est difficile avec un taux de réussite de 21%. Dans la condition 1, cet item a des liens avec les attributs 2 (technique) et 3 (liens notions). La petite valeur du paramètre π signifie que la probabilité que les sujets qui maîtrisent ces deux attributs les appliquent correctement à ces items est faible. Ces attributs ne prennent donc pas convenablement en compte cet item difficile. Au contraire, le paramètre π de l'item 9 vaut 0,86 dans la condition 1. Cet item demande de donner la dérivée de la fonction $\ln x$ et est facile avec un taux de réussite de 86%. Dans la condition 1, la matrice synthèse le relie aux attributs 2 (technique) et 5 (propriétés, définitions et liens). La valeur élevée du paramètre π signifie que les sujets qui maîtrisent ces deux attributs ont de bonnes chances de les utiliser convenablement pour répondre à cette question. Ces deux attributs représentent donc bien la question.

Les items ont été classés en fonction du comportement des experts dans quatre groupes définis dans le tableau 21. Le groupe 1 contient les items pour lesquels le jugement des experts est stable entre les trois conditions et le groupe 4 ceux pour lesquels les experts ont constamment changé d'opinion. Un test statistique dans une analyse de variance (voir annexe A, tableau A28) montre que l'appartenance à un groupe n'influence pas la valeur de π pour le test A ($F(3, 53)=1,742$ et $p=0,170$) et pour le test B ($F(2, 54)=1,949$ et $p=0,152$). Notons que la valeur moyenne de π est la plus petite

pour le groupe 4 dans le cas du test A et pour le groupe 1 pour le test B montrant une différence entre les deux tests.

La figure 19 présente les valeurs de π pour chaque item et chaque condition. Le test A et le test B font l'objet de deux graphiques séparés. Les items sont classés du plus difficile au plus facile. La figure 19 illustre le fait que les items pour lesquels la valeur du paramètre π est inférieure à 0,6 sont toujours des items difficiles (par exemple l'item A12). Dans le cas du test A, tous les items dont le pourcentage de bonnes réponses est inférieur ou égal à 0,4 ont une valeur de π inférieure à 0,6 ou très proche de 0,6. Dans le cas du test B, seuls certains des items difficiles, dont les deux plus difficiles, ont une telle valeur de π . Pour les sept items les plus difficiles du test A, la condition 3 donne des paramètres π plus grands pour les items 12, 13, 19. Pour l'item 8 un gain est obtenu de façon similaire aux conditions 2 et 3 relativement à la condition 1. Pour les autres items difficiles, π varie assez peu d'une condition à l'autre. En ce qui concerne le test B, les items 10 et 13 ont les plus petites valeurs de π . C'est la condition 3 qui leur fournit la valeur la plus grande de même que pour les items 9 et 16. Les conditions 1 et 3 fournissent des valeurs plus grandes que la condition 2 pour l'item 3. La condition 2 est favorable à l'item 6 au sens d'une valeur de π plus grande. Les conditions 2 et 3 sont favorables aux items 4 et 18.

Pour l'ensemble des items, la figure 20 permet de visualiser si la condition 2 ou la condition 3 aboutit ou non à des valeurs de π plus grandes que la condition 1 ainsi que l'amplitude des différences observées,

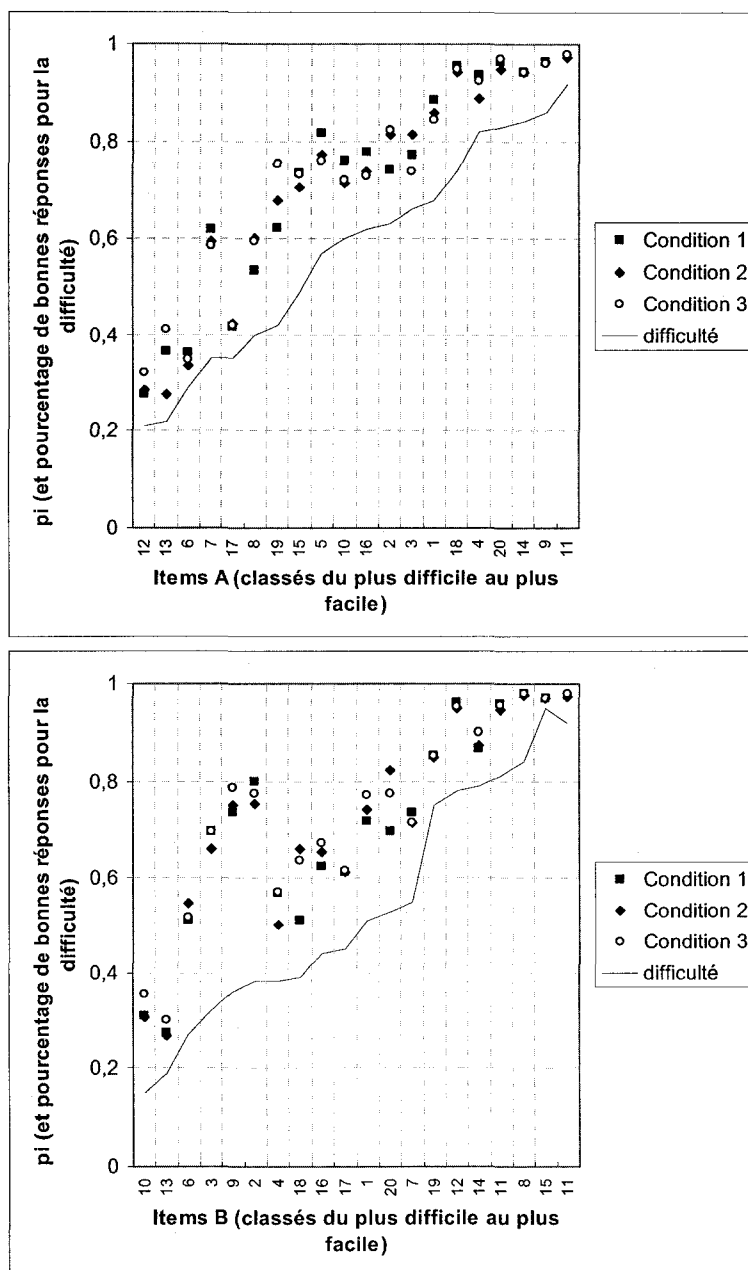


Figure 19. Les valeurs de π présentées pour les items des tests A et B classés en ordre de difficulté décroissante, pour les 3 conditions.

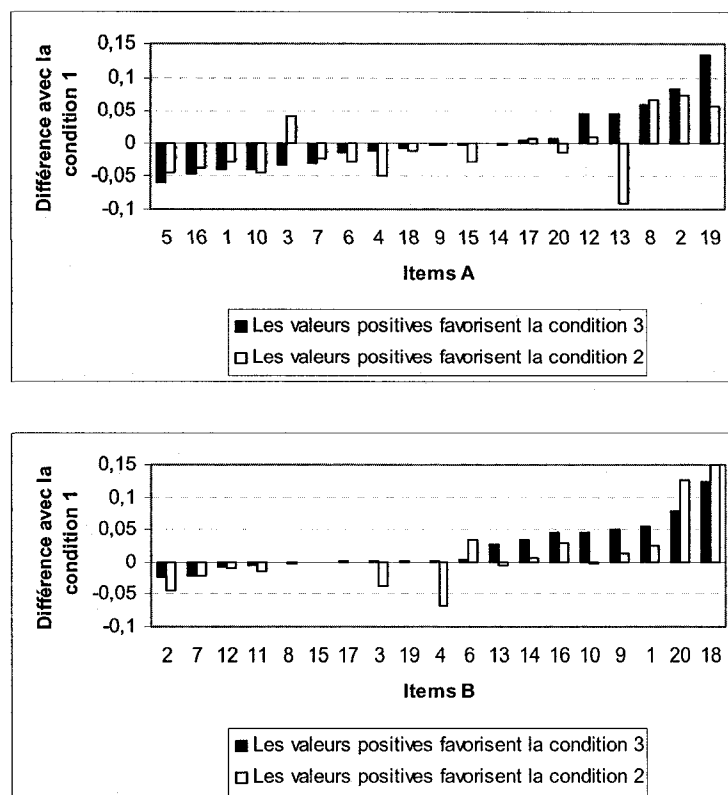


Figure 20. La différence entre les valeurs de π obtenues aux conditions 2 et 3 et celles obtenues à la condition 1. Les items sont classés pour des différences entre les conditions 1 et 3 de plus en plus grandes.

La figure 20 met en évidence le fait que les différences entre les valeurs de π entre la condition 1 et les conditions 2 et 3 ne dépassent jamais une amplitude de 0,15. Seuls les paramètres π de l'item 19 du test A et des items 18 et 20 du test B varient de plus de 0,1. Pour le test A, il y a à peu près autant d'items pour lesquels le paramètre π est plus grand à la condition 1 qu'à la condition 3 que le contraire. C'est également le cas pour les conditions 1 et 2. Pour le test B, il y a plus d'items pour lesquels la condition 3 fournit des valeurs de π plus grandes. La condition 2 donne des paramètres π beaucoup plus grands pour les items 18 et 20. Quant aux autres items, il y a à peu près autant d'items pour lesquels les valeurs sont plus grandes à la condition 1 qu'à la condition 2.

Dans le cadre de la préparation des renseignements additionnels fournis aux experts à la condition 2 (voir 2.3.1.3), les données des deux tests ont fait l'objet d'analyses factorielles qui ont abouti à classer treize des vingt items du test A et dix sept des vingt items du test B dans quatre facteurs pour le test A et six pour le test B. Ces regroupements factoriels constituaient l'information fournie aux experts à la condition 2 (voir annexe F, tableau F2). La figure 21 met en évidence le fait que la moyenne des valeurs de π des items n'appartenant à aucun regroupement factoriel est inférieure à celle des items appartenant à un regroupement factoriel pour les deux tests. Un test de signification dans une analyse de variance (voir annexe A, tableau 29) montre que seule l'appartenance à un regroupement factoriel influence statistiquement la valeur de π ($F(1, 104)=4,010$ et $p=0,048$). Ainsi les valeurs de π correspondant aux items qui partagent une dimension avec d'autres items sont statistiquement supérieures à celles des autres items pour les trois conditions. Néanmoins, le fait d'informer les experts de l'appartenance des items à un regroupement factoriel dans la condition 2 n'influence pas les valeurs de π .

En résumé, les items les plus difficiles ont des valeurs de π souvent inférieures à 0,6. Globalement, les conditions améliorent peu ce problème, même si les conditions 2 ou 3 voient la valeur π de certains items augmenter. Le fait que certains items aient une dimension commune avec d'autres items semble fournir des valeurs de π plus grandes.

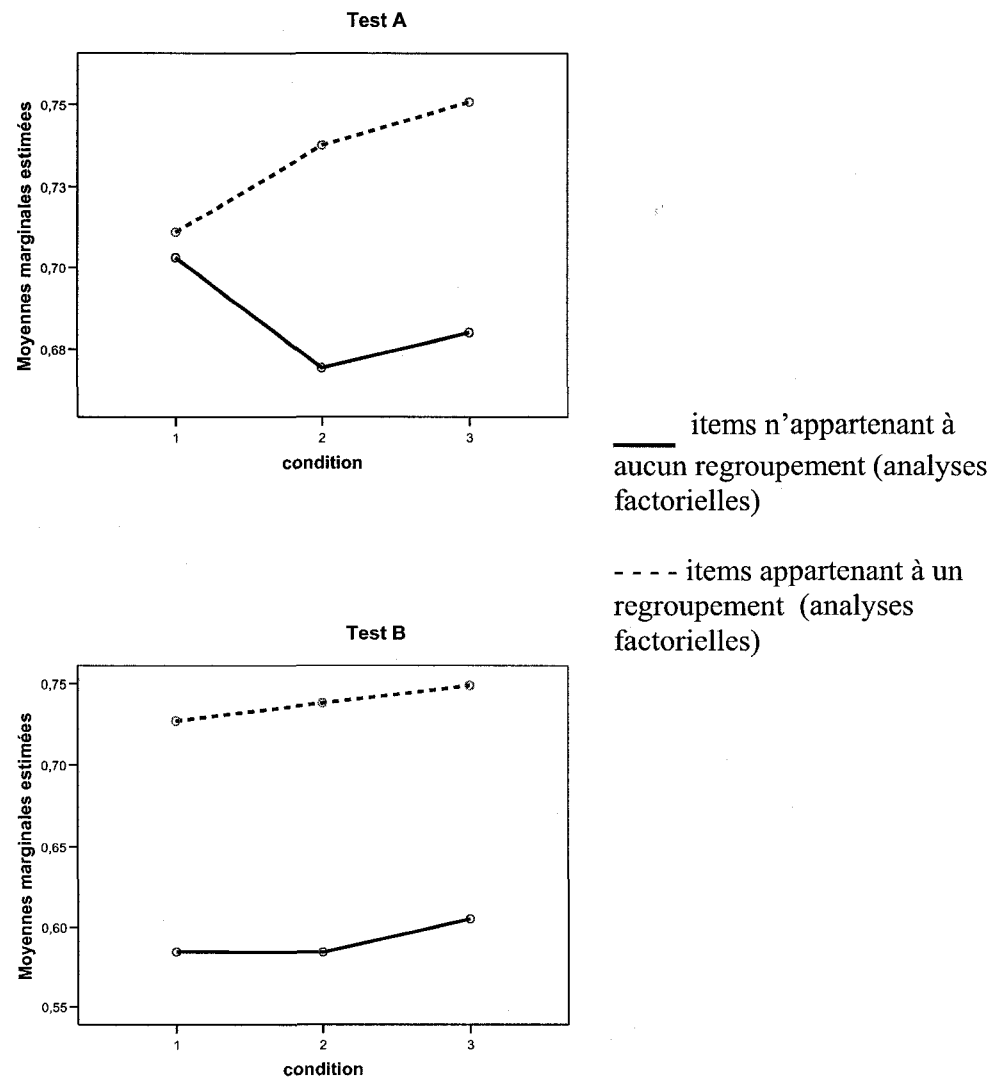


Figure 21. Les valeurs moyennes de π pour les items appartenant ou non à un regroupement selon les analyses factorielles pour les deux tests

3.3.3 Les paramètres d'items r : résultats et comparaisons

Il reste à comparer les paramètres r estimés par les modèles pour chacune des trois matrices synthèses simplifiées. Ces paramètres indiquent si les items discriminent convenablement les attributs ou pas. Chaque valeur de r donne une information sur la validité d'un lien établi entre un item et un attribut dans la matrice Q.

Les paramètres r sont inversement proportionnels à la force de l'évidence que l'item fournit sur la maîtrise de l'attribut. Par exemple, la valeur de $r = 0,12$ obtenue à la condition 1 par le lien entre l'item 7 du test A et l'attribut 2 (technique) indique que cet item est un très bon discriminant pour cet attribut. Autrement dit, le sujet qui trouve la bonne réponse à la question 7 (Trouver la limite quand x tend vers 1 de $\frac{\ln x}{x^3 - 1}$) a de bonnes chances de maîtriser l'attribut 2 (technique). Par contre la valeur de r pour le lien entre cet item et l'attribut 5 (propriétés, définitions et liens) est de 0,84. Ainsi le fait de réussir cet item ne donne pas d'informations claires sur la maîtrise de l'attribut 5 (propriétés, définitions et liens).

Tout d'abord, un test de signification dans une analyse de variance (voir annexe A, tableau A30) montre que la valeur moyenne des valeurs de r des trois conditions est différente pour les deux tests ($F(1, 261)=11,941$ et $p=0,001$). La moyenne des valeurs de r du test A est supérieure à celle du test B. Un test de Friedman (voir annexe A, tableau A31) montre que la condition n'influence pas les valeurs de r pour le test A ($\chi^2=0,278$ et $p=0,870$) et pour le test B ($\chi^2=0,318$ et $p=0,853$). La condition n'influence pas la moyenne des valeurs de r de chaque groupe d'item tels que défini dans le tableau 21 pas plus que la moyenne des items reliés à chacun des deux contenus (fonctions élémentaires et calcul différentiel) (voir annexe A, tableau A32).

En ce qui concerne les regroupements d'items identifiés pour la condition 2 par les analyses factorielles, un test de signification dans une analyse de variance (voir annexe A tableau, A33)

montre que le fait de partager une dimension avec d'autres items influence de façon significative la moyenne des valeurs de r pour le test A ($F(2, 126)=5,713$ et $p=0,018$) et pour le test B ($F(1, 133)=8,764$ et $p=0,004$). La figure 22 permet de visualiser les valeurs moyennes de r pour chaque condition et chaque test selon que les items appartiennent ou non à un regroupement issu des analyses factorielles. Cette figure met en évidence le fait que les valeurs de r sont plus petites pour les items qui partagent un facteur avec d'autres que pour les items isolés, aussi bien pour le test A que pour le test B.

Rappelons que le regroupement des items 14, 17 et 20 dans le facteur 4 du test A a amené les experts à identifier l'attribut 3 comme commun à deux items sur les trois. Les valeurs de r correspondantes sont 0,73 et 0,69 pour les items 14 et 20 du test A. Le regroupement des items 8, 12, 16 et 20 du test B dans le facteur 4 a amené les experts à identifier l'attribut 5 comme commun aux quatre items. Les valeurs de r correspondantes sont de 0,26, 0,27, 0,48 et 0,37. Même si les valeurs de r obtenues pour les items du test B démontrent des liens appropriés, notons que la moyenne de r pour ces items varie peu d'une condition à l'autre.

Le tableau 25 présente les valeurs moyennes de r pour chaque test à chacune des trois conditions de même que globalement. En ce qui concerne le test A, la moyenne des valeurs de r est influencée (voir annexe A, tableau A34) par l'item ($F(19, 59)=5,555$ et $p=0,000$) et par la combinaison item-attribut ($F(44, 59)=5,669$ et $p=0,000$). Dans le cas du test B, seul l'item joue un rôle dans la valeur moyenne de r ($F(18, 62)=7,051$ et $p=0,000$). De plus le groupe d'item (selon le comportement des experts, voir tableau 21) influence la valeur de r pour le test A ($F(3, 118)=4,173$ et $p=0,008$) et ne l'influence pas pour le test B ($F(2, 132)=0,410$ et $p=0,665$) (voir annexe A, tableau A35).

Des analyses de variances sont faites pour chaque attribut pour le test A d'une part et pour le test B d'autre part (voir annexe A, tableaux A36 et A37). Elles montrent que la condition n'influence

pas la moyenne de r de façon statistiquement significative pour chacun des attributs. Des analyses de variance séparées pour chaque item (voir annexe A, tableaux A38 et A39) montrent que seuls deux items ont une valeur moyenne de r qui est influencée par les conditions. Ces items sont l'item 17 du test A et l'item 9 du test B. Ces deux items ont un contenu différent, mais tous deux sont difficiles. Rien ne semble toutefois distinguer ces deux items des autres.

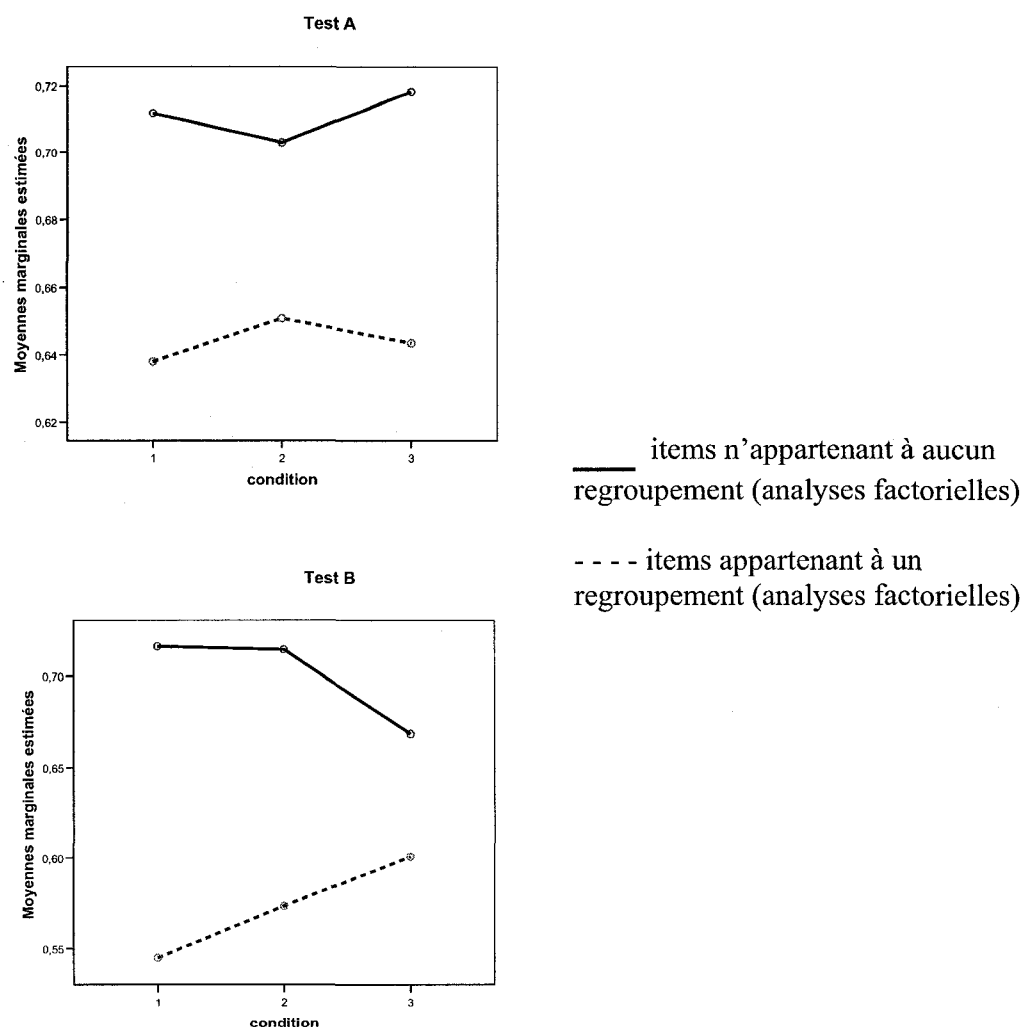


Figure 22. Les valeurs moyennes de r pour les items appartenant ou non à un regroupement selon les analyses factorielles pour les deux tests

Tableau 25.

Les valeurs moyennes de r pour chaque test, chaque condition et globalement

	Test A	Test B
Condition 1	0,67	0,58
Condition 2	0,67	0,60
Condition 3	0,67	0,61
Globalement	0,67	0,60

Les résultats qui précèdent montrent que quel que soit le regard porté, les valeurs des paramètres r ne sont pas statistiquement influencées par les conditions. Ainsi l'hypothèse d'une influence des conditions sur la validité des matrices n'est pas démontrée par l'ensemble des tests présentés. Il reste à examiner les valeurs de r afin de juger si la structure cognitive des trois conditions est bonne ou non et si on observe des différences entre elles. Les valeurs de r sont comprises entre 0 et 1, les plus petites valeurs démontrant les liens les plus forts entre l'item et l'attribut dans la matrice Q et les valeurs les plus grandes indiquant des liens moins appropriés. Ainsi, une structure cognitive forte devrait présenter une majorité de valeurs inférieures à 0,5 ou encore mieux à 0,3. La figure 23 présente les pourcentages de valeurs de r inférieures à 0,3, comprises entre 0,3 et 0,5, puis entre 0,5 et 0,7 et enfin supérieures à 0,7 pour chacune des conditions et les deux tests dans le cas des trois matrices synthèses simplifiées.

Comme on pouvait s'y attendre le portrait des deux tests est assez différent et le pourcentage de valeurs inférieures à 0,5 plus important pour B explique un r moyen plus petit pour ce test que pour le test A. De plus, il y a peu de différences entre les trois conditions pour chaque test pris séparément. En ce qui concerne le test A, le pourcentage de valeurs inférieures à 0,5 est très faible, et

ce pour l'ensemble des trois conditions. Environ la moitié des paramètres r sont supérieurs à 0,7 et près de 90% sont supérieurs à 0,5 pour les trois conditions. Même s'il y a une très légère augmentation du nombre de paramètres inférieurs à 0,5 à la condition 3, aucune condition n'améliore la structure cognitive de la matrice reliée au test A car le nombre de valeurs de r suggérant une forte structure reste très petit. Ainsi, la structure cognitive des matrices synthèses simplifiées n'est pas de très bonne qualité pour A. Il est possible que le fait que sept items du test A n'appartiennent à aucun regroupement factoriel (voir annexe F, tableau F2) explique en partie la différence entre la structure des deux tests puisque ceci influence la valeur de r .

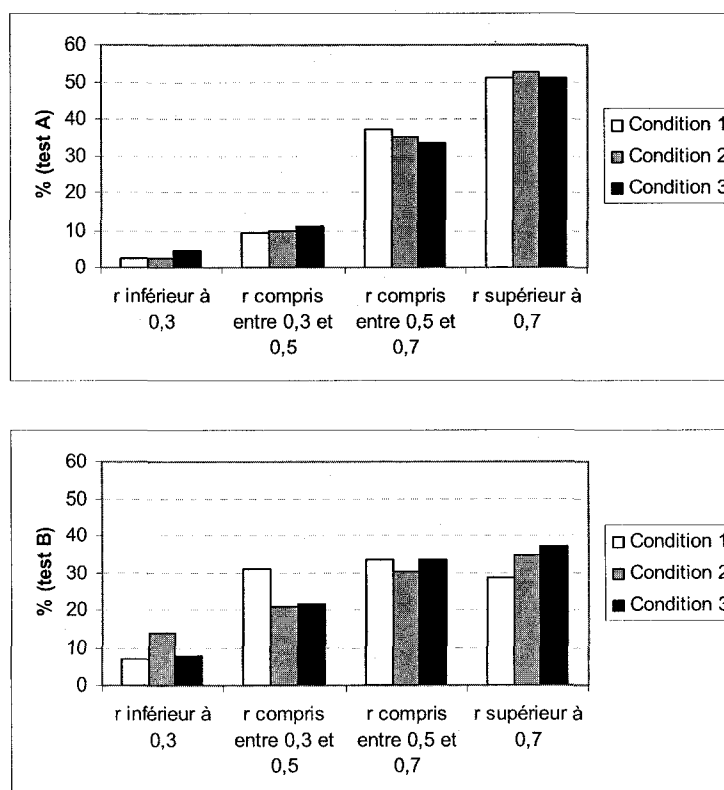


Figure 23. Les pourcentages de valeurs r dans 4 intervalles pour les tests A et B et les 3 conditions

En ce qui concerne le test B, plus d'un tiers des valeurs de r sont inférieures à 0,5 à chaque condition et en contrepartie, il y a moins de valeurs supérieures à 0,7. Ainsi, comme attendu, la

structure des matrices est plus forte pour le test B que pour le test A. Il est à noter que seulement trois items du test B n'appartiennent pas à un regroupement factoriel (voir annexe F, tableau F2). Il y a également plus de variations entre les conditions pour le test B que pour le test A. La condition 2 fournit le plus de valeurs de r inférieures à 0,3 et le moins de valeurs comprises entre 0,5 et 0,7. C'est la condition 1 qui offre le plus de valeurs de r comprises entre 0,3 et 0,5 et le moins de valeurs supérieures à 0,7. La condition 3 comporte plus de valeurs supérieures à 0,5 que les deux autres conditions. Si les trois conditions sont à peu près équivalentes pour le test A, il semble que la condition 2 fournisse la matrice avec la meilleure structure cognitive puisqu'elle a le plus de valeurs de r inférieures à 0,5.

Étant donné qu'une valeur de r est estimée pour chaque lien entre un item et un attribut et que les liens ne sont pas les mêmes d'une condition à l'autre, il est difficile de comparer individuellement les valeurs des paramètres. Une approche consiste à calculer puis comparer les moyennes des valeurs de r pour chaque attribut dans chaque matrice. Les résultats sont présentés dans la figure 24.

L'attribut 1 n'apparaît pas à condition 2. La variation entre les conditions 1 et 3 est plus importante pour le test A que pour le test B et les attributs 3 et 6 varient le plus entre les conditions. Pour ces deux attributs, on observe la valeur la plus petite à la condition 3. Néanmoins ceci ne se retrouve pas dans le test B pour lequel au contraire les moyennes liées aux attributs 3 et 6 et à la condition 3 sont supérieures aux deux autres conditions. Pour les autres attributs, et ce dans les deux tests, les variations des moyennes sont assez faibles comme on pouvait s'y attendre à la suite des analyses de variance. Globalement, les valeurs moyennes de r sont inférieures pour le test B.

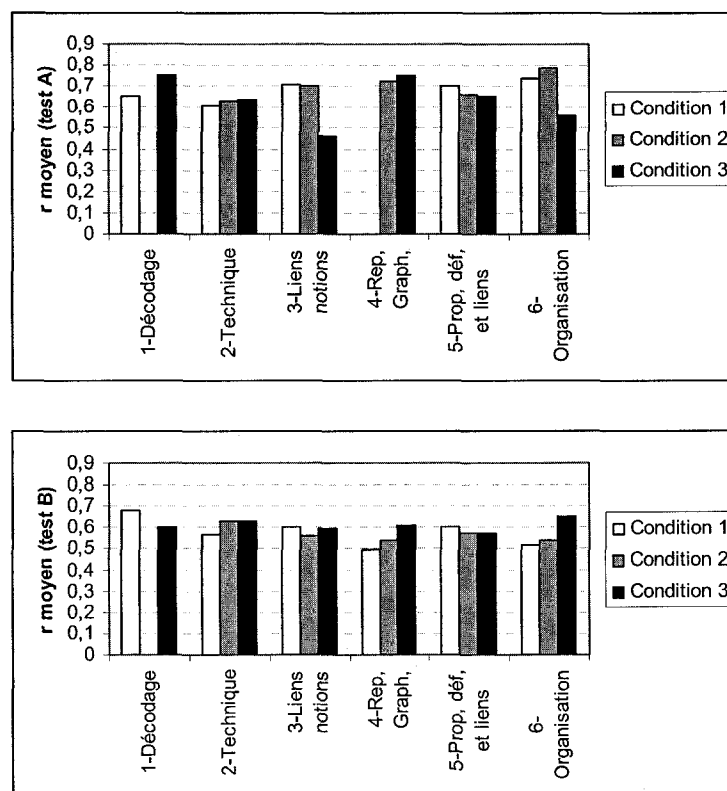


Figure 24. Les valeurs moyennes des paramètres r relativement aux 6 attributs pour les trois conditions et les deux tests

Les moyennes peuvent également être calculées par item plutôt que par attribut. La figure 25 présente le nombre d'items ayant un r moyen inférieur à 0,3, compris entre 0,3 et 0,5, compris entre 0,5 et 0,7 et supérieur à 0,7. Les tests A et B font l'objet de deux graphiques séparés. La figure 25 met à nouveau en évidence la meilleure structure des matrices pour le test B puisqu'entre cinq et huit items correspondent à une valeur de r inférieure à 0,5 alors que seulement un ou deux items ont de telles valeurs pour le test A. De plus, entre sept et neuf items du test A ont des valeurs moyennes de r supérieures à 0,7 contre seulement cinq items du test B. Pour le test A, les conditions 1 et 2 sont relativement équivalentes (rappelons que l'item 11 ne fait pas partie de la première condition), douze items ont un r moyen inférieur à 0,7 pour chacune d'elles. En ce qui concerne le test B, la condition 2 est clairement celle qui fournit le plus de r moyens inférieurs à 0,5 avec huit items contre six à la

condition 1 et cinq à la condition 3. C'est aussi bien entendu à la condition 2 que le moins d'items ont des valeurs moyennes de r supérieures à 0,5.

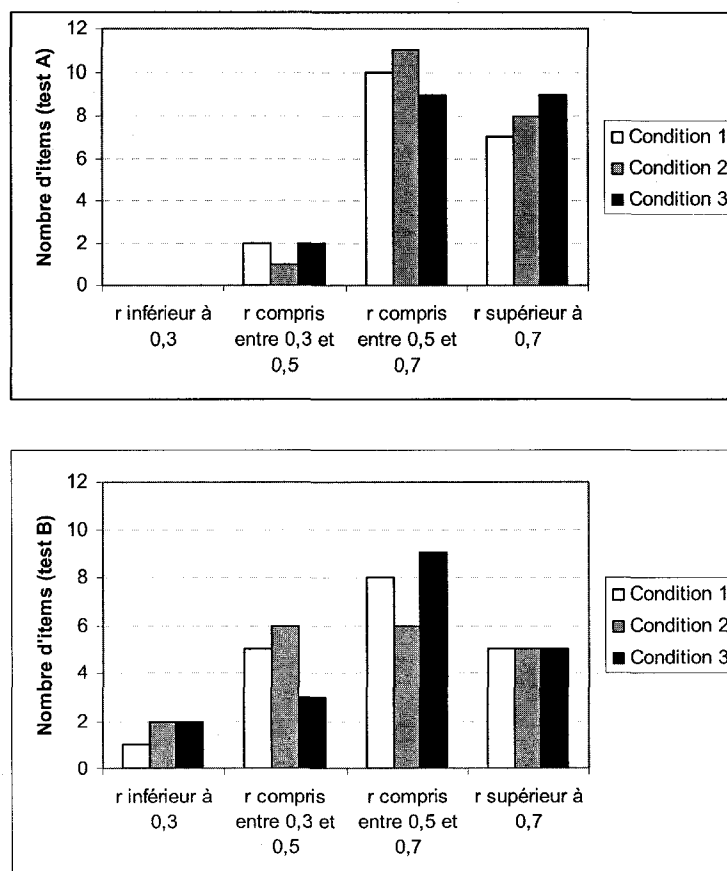


Figure 25. Le nombre d'items relativement à leur r moyen pour chaque condition et chaque test

Note. Le test A comporte 19 items à la condition 1 et le test B en comporte 19 aux trois conditions.

Globalement et du point de vue des valeurs moyennes de r pour les items, si les conditions 1 et 2 sont assez équivalentes pour le test A, la condition 2 se distingue en donnant la matrice avec la plus forte structure cognitive dans le cas du test B. La condition 3 ne semble améliorer la structure des matrices ni pour le test A ni pour le test B.

La figure 26 présente maintenant les valeurs moyennes de r lorsque les items sont regroupés selon que leur contenu est lié aux fonctions élémentaires ou au calcul différentiel.

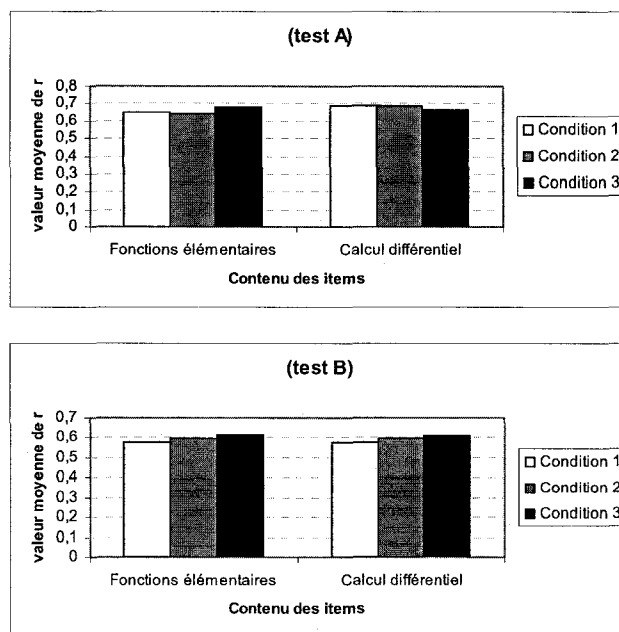


Figure 26. Les valeurs moyennes de r lorsque les items sont regroupés selon leur contenu pour chaque condition et chaque test

Vu sous cet angle, il y a à nouveau peu de variations d'une condition à l'autre et le test B a la meilleure structure. Pour le test A, excepté pour la condition 3, les items reliés aux fonctions élémentaires présentent des valeurs moyennes de r très légèrement inférieures à celles des items reliés au calcul différentiel. Pour le test B, les deux contenus sont équivalents en termes de valeurs moyennes de r . Si les conditions 1 et 2 sont semblables pour le test A, la condition 1 donne les valeurs de r les plus petites pour le test B.

Lors de l'examen des différentes matrices, les items ont été séparés en quatre groupes selon que les experts ont ou non beaucoup évolué dans leur façon de les caractériser d'une condition à l'autre (voir tableau 21). La figure 27 utilise ces groupes pour présenter les valeurs moyennes de r pour les trois conditions et les deux tests.

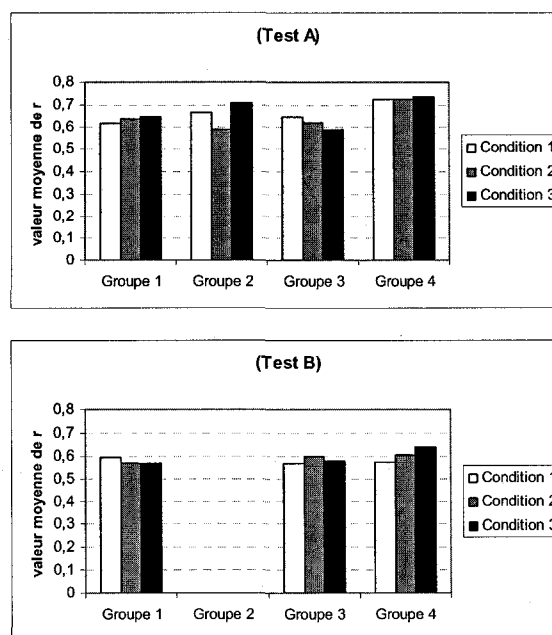


Figure 27. Les valeurs moyennes de r lorsque les items sont regroupés selon leur appartenance à un groupe pour chaque condition et chaque test

Les variations sont encore très faibles ici. Rappelons que le groupe 2 ne contient qu'un seul item du test A et aucun du test B. Seul l'item 1 du test A a changé seulement entre les conditions 1 et 2. Pour cet item, la condition 2 fournit la plus petite valeur de r . Sinon, la condition 1 est le plus souvent associée à la plus petite valeur moyenne de r (4 fois sur 7).

Chaque valeur de r peut être examinée sans faire de moyennes selon divers regroupements. L'annexe I fournit le détail des valeurs des paramètres r (voir tableaux I1 et I2). Elle contient une figure par attribut afin de visualiser l'évolution au fil des trois conditions pour les deux tests (voir figures I1, I2, I3, I4, I5 et I6). Chaque figure fait l'objet de commentaires à l'intérieur de l'annexe I. L'interprétation des graphiques de l'annexe I est synthétisée dans le tableau 26.

Le tableau 26 présente, pour chaque attribut, chaque condition et chaque test, le nombre de valeurs de r plus petites dès qu'au moins une comparaison peut être faite pour un item. Étant donné que certaines de ces différences sont petites et d'autres importantes, le choix est fait de ne prendre en

compte que les différences entre la plus grande et la plus petite valeur supérieures à 0,1. Le fait de comparer les valeurs des paramètres r correspondant à des liens identiques dans plusieurs matrices aide à porter un jugement sur la structure cognitive globale de chaque matrice. On peut supposer qu'un lien qui a été établi à chaque condition a des chances d'être pertinent, mais que ce sont les autres liens qui améliorent ou non la structure globale.

Même si la littérature suggère que des valeurs de r inférieures à 0,5 démontrent une structure forte et des valeurs inférieures à 0,3 une structure excellente, le tableau 26 présente le nombre total de valeurs de r inférieures à 0,4 et 0,6 afin que plus de comparaisons soient possibles. Ces nombres de valeurs sont présentés pour chaque attribut, chaque condition et chaque test. Enfin, comme certains liens n'existent que dans une seule condition, le tableau 25 dénombre ceux auxquels correspond une valeur de r inférieure à 0,4 ou à 0,6 pour chaque test et chaque condition. Les informations contenues dans le tableau 26 permettent d'identifier si une condition est ou non favorable à chacun des six attributs.

L'attribut 1 concerne tout ce qui a trait au fait de lire la question correctement, qu'il s'agisse de comprendre les mots clés, le langage ou encore les symboles mathématiques. La particularité de cet attribut tient au fait que les consignes semblent avoir influencé les experts au point de les amener à utiliser de façon systématique cet attribut pour la grande majorité des items dans la condition 3. Celle-ci fournit huit liens avec une valeur de r inférieure à 0,6 pour le test B.

L'attribut 2 concerne la dimension technique de résolution des items comme le calcul algébrique ou arithmétique ou l'application de règles ou d'algorithmes. La condition 1 a été plus propice pour établir des liens de qualité avec l'attribut 2 puisque les comparaisons sont plus souvent en faveur de la condition 1, que sous cette condition on trouve le plus de valeurs inférieures à 0,4 et à 0,6 et qu'elle seule identifie deux liens pour le test B pour lesquels la valeur de r est inférieure à 0,4.

Tableau 26.

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	
Attribut 1	A	B	A	B	A	B	A	B	A	B
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.

L'attribut 3 consiste à faire des liens entre des notions et l'attribut 4 concerne tout ce qui a trait à une représentation graphique pour trouver la réponse à la question. La condition 2 semble légèrement plus favorable à ces deux attributs lorsqu'on regarde les comparaisons pour les deux tests. Elle seule fournit des valeurs de r inférieures à 0,4 pour le test B dans le cas de l'attribut 3 et en fournit le plus dans le cas de l'attribut 4 pour le test B.

L'attribut 5 vise l'utilisation des propriétés, des définitions, des théorèmes ainsi que les liens que l'on doit faire entre ces éléments de base pour répondre à la question. Cet attribut est celui, avec l'attribut 3, qui a le plus posé de problème dans sa définition lors de la phase 2 du MACB et occasionné des discussions. C'est également la condition 2 qui fournit le plus de valeurs de r inférieures à 0,4 et à 0,6 pour les deux tests relativement à l'attribut 5. Les comparaisons l'avantagent pour le test A et elle fournit des liens avec des valeurs de r inférieures à 0,6 pour les deux tests.

L'attribut 6 concerne l'organisation de la réponse comme le fait de procéder en plusieurs étapes ou encore celui d'utiliser une stratégie. La condition 1 fournit le plus de petites valeurs pour le test B et est la seule à en identifier certaines dans le cas de l'attribut 6.

En résumé, les moyennes des valeurs de r montrent que la structure du test B est meilleure que celle du test A. Ces moyennes ne sont pas influencées par la condition aussi bien globalement que calculées par attribut, par item (sauf pour deux items), par contenu, par groupe d'items (selon le comportement des experts, voir tableau 21) ou par regroupement (selon les analyses factorielles de la condition 2, voir annexe F, tableau F2) et n'ont pas permis de dessiner de tendance en ce qui concerne une amélioration pour l'une ou l'autre des conditions. Les pourcentages de valeurs de r inférieures à 0,3, comprises entre 0,3 et 0,5, comprises entre 0,5 et 0,7 et supérieures à 0,7 ne permettent pas d'identifier une condition plus favorable au test A, mais la condition 2 favorise le test B car elle fournit le plus de petites valeurs.

Même si les comparaisons des valeurs de r sont complexes, elles aboutissent à des considérations propres à chaque attribut. L'attribut 1 (décodage) tire profit de la condition 3. Les valeurs de r reliées aux attributs 2 (technique) et 6 (organisation) ne tirent pas d'avantage des informations fournies aux conditions 2 et 3. Les valeurs de r des attributs 3 (liens notions), 4 (représentation graphique) et 5 (propriétés, définitions et liens) ont tendance à être plus petites à la condition 2.

3.4 Une illustration du diagnostic des sujets

La revue de la littérature a montré que les avantages de ces modèles sur les épreuves de maîtrise viennent de la grande variété des attributs cognitifs utilisables et du fait de pouvoir dépasser une vision atomisée du contenu du test. Plutôt que de présenter un exemple fictif, elle annonçait l'utilisation des données de la recherche pour montrer concrètement l'avantage de cette approche. Les paragraphes qui suivent ne constituent donc pas un résultat en soi, mais une démonstration concrète de l'intérêt des modèles cognitifs sous la forme du diagnostic cognitif à l'aide des données de la recherche comme annoncé au point 1.2.1.

La figure 28 présente le diagnostic de huit sujets pour le test B modélisé avec la matrice synthèse simplifiée de la condition 2. Dans chaque graphique, les deux sujets présentés ont le même score. Les sujets 151 et 23 ont un score total de 6 sur 20, les sujets 111 et 35 en ont un de 10 sur 20, les sujets 105 et 52 en ont un de 12 sur 20 et les sujets 378 et 223 ont un score total de 19 sur 20.

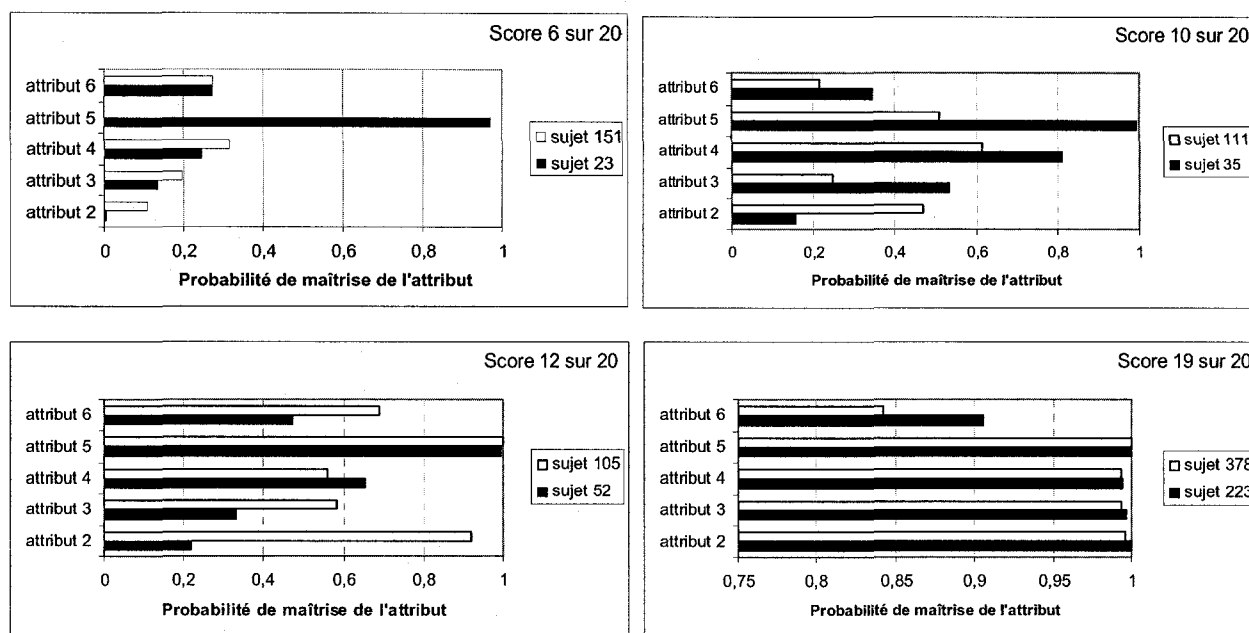


Figure 28. Le diagnostic de huit sujets relativement aux attributs, en fonction de leur score total pour le test B et la matrice synthèse simplifiée de la condition 2

La figure 28 met en évidence le fait que deux sujets ayant répondu convenablement au même nombre de questions peuvent avoir des forces et faiblesses très différentes, et ce aussi bien pour les scores faibles, moyens ou élevés. En effet, les barres de longueurs différentes représentant la probabilité de maîtrise de chaque attribut mettent en évidence les différences entre les deux sujets illustrés. Par exemple, les sujets 23 et 151 n'ont convenablement répondu qu'à six des vingt items, mais le sujet 23 a une très haute probabilité de maîtriser l'attribut 5 alors que le sujet 151 a une probabilité nulle relativement à cet attribut. Les sujets 111 et 35 ont dix bonnes réponses sur vingt questions, mais ils présentent des probabilités de maîtrise très différentes pour tous les attributs. Les sujets 105 et 52 ont un score de 12 sur 20 et une très bonne probabilité de maîtriser l'attribut 5, mais ils diffèrent dans leur maîtrise des autres attributs, particulièrement en ce qui concerne l'attribut 2. Enfin, les sujets 378 et 223 n'ont pas correctement répondu à une seule question à cause d'un petit problème

d'organisation de leur solution (attribut 6), mais le sujet 378 a une probabilité de maîtriser cet attribut légèrement plus petite que le sujet 223.

La visualisation du portrait de ces quelques sujets met en évidence la pertinence de la modélisation cognitive de leurs réponses dans une perspective diagnostique annoncée par la littérature (par exemple par Roussos et al., 2007 ou par Stout, 2004). Les informations fournies ici seraient utiles autant à l'étudiant qu'à son professeur afin de remédier aux problèmes identifiés. La réflexion qui suit se base sur un examen plus approfondi des items réussis par les sujets 23 et 151 et de leurs liens avec l'attribut 5 tels que présentés par le tableau 27.

Le tableau 27 inclut les quinze items qui sont reliés à l'attribut 5 dans la matrice QB2s. Étant donné que l'item 5 a été retiré de cette matrice et que les deux sujets ont bien répondu à cette question, il reste cinq bonnes réponses à étudier pour les deux sujets examinés. Les cinq bonnes réponses du sujet 23 sont reliées à l'attribut 5 tandis que trois bonnes réponses du sujet 151 le sont. Vu sous cet angle, il ne semble pas qu'il y ait une grande différence entre les deux sujets. Un examen des valeurs du paramètre r associé à chacun des liens de la matrice montre que les valeurs sont inférieures à 0,5 pour les items 8, 12, 13, 16 et 19 suggérant un fort lien entre l'attribut 5 et ces items. La différence entre les deux sujets provient du fait que le sujet 23 a bien répondu aux questions 8, 12, 16 et 19 alors que le sujet 151 n'a répondu à aucune de ces questions correctement.

Pour illustrer la spécificité des modèles cognitifs relativement à profil de maîtrise basé sur un découpage des items relié au tableau de spécification, le tableau 27 précise la classification de chaque item dans le tableau de spécification du test B. Le sujet 23 a des scores partiels de 4/5 en calcul différentiel-technique, de 1/5 en calcul différentiel-notion et de 1/5 en fonctions élémentaires-fonction. Ces autres scores partiels sont de 0/5. Le sujet 151 a obtenu 2/5 en fonctions élémentaires-notion, en fonctions élémentaires-fonction et en calcul différentiel-notion. Ces autres scores partiels valent 0/5. Cette approche montre une différence entre les deux sujets puisque le sujet 23 a plus de

difficulté en ce qui a trait aux fonctions élémentaires alors que le sujet 151 a des difficultés dans les deux domaines. Toutefois, elle cible le contenu, mais pas les habiletés qui font défaut à ces deux sujets. L'approche cognitive montre que les sujets 23 et 151 ont des difficultés semblables sauf en ce qui concerne l'attribut 5 (propriétés, définitions et liens) qui ne pose pas de problèmes au sujet 23, mais qui n'est pas du tout maîtrisé par le sujet 151. La plus grosse difficulté du sujet 23 concerne l'attribut 2 (technique).

Tableau 27.

Les items réussis par les sujets 23 et 151 ainsi que les liens entre les items et l'attribut 5

Items	Classification dans le tableau de spécification	Sujet 23	Sujet 151	Attribut 5
Item 1	Fonc. Él. notion		X	X
Item 2	Cal. Diff. notion			X
Item 3	Fonc. Él. fonction		X	
Item 4	Cal. Diff. technique		X	X
Item 5	Fonct. Él. fonction	X	X	
Item 6	Cal. Diff. notion			X
Item 7	Fonc. Él. fonction			X
Item 8	Cal. Diff. technique	X		X*
Item 9	Cal. Diff. notion			
Item 10	Fonc. Él. notion			X
Item 11	Fonc. Él. notion			X
Item 12	Cal. Diff. technique	X		X*
Item 13	Fonc. Él. fonction			X*
Item 14	Fonc. Él. Notion		X	
Item 15	Cal. Diff. technique	X	X	X
Item 16	Fonc. Él. notion	X		X*
Item 17	Cal. Diff. notion			X
Item 18	Cal. Diff. notion			X
Item 19	Cal. Diff. Technique	X		X*
Item 20	Fonc. Él. fonction			

Note. Pour les sujets, les X indiquent les bonnes réponses. Pour l'attribut 5, les X indiquent un lien entre l'attribut 5 et l'item correspondant. Les * identifient les liens pour lesquels les valeurs du paramètre r sont inférieures à 0,5, elles correspondent aux X en caractères gras.

4 Chapitre IV : interprétation, discussion et conclusion

4.1 L'interprétation et la discussion

Le cadre conceptuel de cette recherche identifie quatre sources potentielles d'influence sur le processus d'élaboration de la matrice Q des modèles cognitifs dans une approche inductive. Trois d'entre elles ont été fixées pour ne laisser varier que la quatrième qui correspond aux conditions d'élaboration. Des trois conditions expérimentales mises en place, la première ne donnait que les énoncés des items aux experts, la seconde fournissait les items regroupés selon le résultat d'analyses factorielles, la troisième condition procurait aux experts les valeurs des coefficients de difficulté et de discrimination des items et un graphique montrant le fonctionnement des leurres.

Cette étude a modélisé les données de deux tests à choix multiples en mathématiques comportant vingt items. La méthode de travail adoptée a consisté en un travail individuel de chaque expert suivi par une séance de travail collectif visant à obtenir une liste d'attributs cognitifs qui fasse consensus. La tâche individuelle de chaque expert comportait : (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.

La méthode MACB a été identifiée dans la littérature pour faire travailler les experts de façon individuelle dans la phase 1, puis pour obtenir un consensus de leur part sur le classement des attributs dans six catégories à la phase 2, la phase 3 devant faire la synthèse des résultats obtenus et aboutir à la création de matrices synthèses. Les modifications à la procédure originale du MACB ont été documentées et appliquées dans la recherche.

Les modifications n'ont pas soulevé de problèmes particuliers, les experts ont facilement compris ce qui était attendu et la phase 2 a duré une journée, comme prévu. La méthode MACB modifiée a abouti à un consensus relativement à une liste d'attributs cognitifs et la ressemblance des énoncés individuels a facilité le processus. Cette liste consensuelle a permis de modifier les matrices individuelles des experts puis de les regrouper en une matrice synthèse par test et par condition.

Cette étude vise à déterminer si le fait de placer les experts dans l'une des trois conditions expérimentales a une influence sur les matrices individuelles et les matrices synthèses qu'ils produisent dans chacune des conditions. L'étude comparative de la forme des matrices obtenues dans les trois conditions fournit une réponse à la première question de la recherche qui se pose ainsi :

1. Comment les conditions expérimentales influencent-elles les matrices Q de chaque expert et les matrices synthèses?

L'examen de la concordance entre les matrices et la représentativité des matrices synthèses des trois conditions permet de répondre à la deuxième question qui est :

2. Comment les conditions expérimentales influencent-elles la concordance entre les matrices individuelles et la représentativité des matrices synthèses ?

Les comparaisons des valeurs des paramètres de validité et de fidélité estimés par le modèle RUM pour chacune des matrices permettent de répondre à la troisième question de recherche dont l'intitulé est le suivant :

3. Comment les conditions expérimentales influencent-elles la validité et la fidélité des matrices Q?

La première question vise à identifier comment les conditions ont influencé les différentes matrices. La recherche a modélisé quatre sortes de matrices. Les experts ont élaboré des matrices individuelles originales qui ont été modifiées suite au consensus. Ces dernières ont été regroupées en matrices synthèses que l'application du modèle RUM a amené à simplifier. Au terme de la procédure d'élaboration des matrices, les matrices individuelles originales (24) de chaque expert pour chaque

condition et chaque test, ces mêmes matrices modifiées (24) sur la base du classement des attributs en six catégories et les matrices synthèses pour l'ensemble des quatre experts (6) qui ont ensuite été simplifiées (6) constituent le matériel analysé pour répondre à la première question.

Les matrices individuelles originales se caractérisent par le nombre de liens qu'elles contiennent. Ces nombres sont semblables pour les deux tests, mais influencés par les conditions. Les nombres de liens les plus petits et les plus homogènes s'observent à la condition 2, ils sont très hétérogènes dans les conditions 1 et 3. Les attributs individuels ont fait l'objet d'un regroupement consensuel. Les nombres d'intitulés individuels associés à chaque attribut consensuel sont semblables pour les deux tests, mais ils varient en fonction de l'expert, de l'attribut visé et de la condition.

Le processus de regroupement des intitulés a amené à éliminer cinq intitulés individuels dont quatre appartenaient à un même expert (expert 2). En outre, l'examen des informations laissées par cet expert sur sa perception des différentes consignes a montré qu'il s'est distingué de ses collègues. En effet, contrairement aux autres experts, il a peu tenu compte des regroupements factoriels de la condition 2 et des informations de la condition 3. Cette constatation amène à se questionner sur le choix des experts qui a un impact pratique sur les résultats. L'expert 2 était le plus diplômé, mais son expérience d'enseignement concernait peu les niveaux secondaire et collégial. En outre, il n'a passé que six de ses vingt années d'enseignement à enseigner les mathématiques. En comparaison, deux des experts avaient une expérience d'enseignement uniquement au secondaire alors que le troisième enseigne à l'école Polytechnique et connaît donc parfaitement la population visée par le test. En conséquence, il semble qu'une connaissance approfondie de la population des étudiants qui passent le test soit un critère de sélection essentiel lors du recrutement des experts.

Dans les trois conditions, l'aspect technique des questions (attribut 2) fait l'objet du plus de représentations. Dans la condition 3, les experts ont fait preuve d'une plus grande précision face à

cet aspect des questions 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. Le consensus a amené à modifier les matrices individuelles originales.

Les matrices individuelles modifiées contiennent les mêmes six attributs. Les experts en ont utilisé entre 50% et 100% dans leurs matrices individuelles sans être statistiquement influencés par les conditions. Leur expertise leur a permis de déterminer dans la condition 1, sans autre aide que les énoncés des questions, une liste d'attributs qui n'a pas augmenté aux conditions 2 et 3. Au contraire, un ou même plusieurs attributs ont disparu au fil des conditions suivantes. La mise en place des conditions 2 et 3 n'a donc pas aidé les experts à identifier de nouveaux attributs, elle les a seulement amenés à parfois en modifier l'intitulé et à leur relier des items parfois différents.

Le nombre total de liens établis dans les matrices individuelles modifiées est semblable pour les deux tests et n'a pas varié selon les conditions. La forme des matrices individuelles modifiées a permis la comparaison des nombres de liens associés à chacun des six attributs séparément. Seul l'attribut 1 est relié à des nombres d'items très différents aux trois conditions.

Le profil particulier de l'attribut 1 qui passe d'une utilisation faible aux deux premières conditions à une utilisation quasi systématique à la troisième amène à se questionner. L'exemple fourni aux experts dans le livret de la condition 3 faisait référence à une difficulté de reconnaissance de l'écriture pour expliquer la différence entre un leurre attrayant et la bonne réponse. Cet exemple avait pour but d'illustrer le type de réflexion attendu, mais il semble avoir influencé les experts au point de les voir utiliser cet attribut pour presque tous les items et de biaiser les résultats pour cet attribut.

Le regroupement des matrices individuelles modifiées fournit une matrice synthèse relativement à chacun des deux tests pour chaque condition. Les conditions n'ont pas influencé leur nombre total de liens, mais les nombres de liens reliés à l'attribut 1 (décodage), à l'attribut 4

(représentation graphique) et à l'attribut 5 (propriétés, définitions et liens) varient en fonction des conditions. L'attribut 1 (décodage) n'apparaît pas dans les matrices de la condition 2 alors que la presque totalité des items des deux tests lui sont reliés à la condition 3. Les conditions 2 et 3 mènent à une certaine augmentation de l'utilisation de l'attribut 4 et la condition 3 à une diminution de celle de l'attribut 5.

Lors de la modélisation des données avec Arpeggio et les matrices synthèses, la procédure aboutit à identifier les liens qui peuvent être supprimés pour alléger le modèle sans en changer la structure et à simplifier les matrices synthèses. Les conditions n'influencent pas le nombre de suppressions, mais les liens retirés sont essentiellement d'ordre technique pour le test B. Le nombre total de liens dans ces matrices simplifiées ne subit pas l'influence des conditions sauf en ce qui concerne l'attribut 1 (décodage).

L'examen des changements opérés dans les matrices synthèses simplifiées entre les trois conditions montre une variation 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 conditions ont eu tendance à plus influencer les experts pour les items reliés au calcul différentiel. Il n'est pas possible de savoir si c'est la plus grande facilité du test A qui 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 si le fait que les experts ont toujours travaillé sur le test A avant le test B joue un rôle.

En outre, 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 spécification que celui du test A. En particulier, le facteur 4 du test B contient quatre items dont le contenu porte à 75% sur des techniques en calcul différentiel selon le tableau de spécification. La condition 2 a identifié l'attribut 5 (propriétés, définitions et liens) comme étant partagé par les quatre items et fournit des liens très

valides au sens du paramètre r du modèle RUM. Il semble que lorsqu'une relation très claire existe entre le contenu des items selon le tableau de spécification et le résultat des analyses factorielles, le regroupement des items aide les experts à trouver ce que partagent ces items.

Ce résultat va dans le même sens que ce que présentent Leighton et ses collègues (2007) dans une récente étude sur l'emploi de méthodes exploratoires et confirmatoires pour identifier les dimensions cognitives dans une évaluation à grande échelle en science. En effet, leur étude suggère que des tableaux de spécification des tests soigneusement élaborés avec des catégories mutuellement exclusives aident les chercheurs à identifier les processus cognitifs utilisés par les sujets. Dans le cas d'une approche inductive utilisant des items préexistants, il est donc important de prendre le temps de valider ou d'améliorer les tableaux de spécification des items avant de procéder à l'élaboration des matrices Q.

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. L'ensemble de ces résultats suggère que le test B a une meilleure structure que le test A qui a peut-être amené les experts à mieux tirer parti des informations fournies à la condition 2.

La deuxième question interroge sur l'impact des conditions sur la concordance entre les matrices individuelles et sur la représentativité des matrices synthèses. Les matrices individuelles modifiées sur la base du consensus ont une forme semblable qui a permis le calcul d'un coefficient de contingence. Sur les trente six valeurs calculées, onze ont montré un manque de concordance entre deux matrices. Étant donné que quatre intitulés individuels de l'expert 2 ont été abandonnés suite au consensus, ses matrices individuelles modifiées ne contiennent que trois des six attributs.

Ceci explique que les matrices de cet expert ont montré un manque 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. Aucun problème de concordance n'est observé pour le test A. Les matrices individuelles modifiées sont toujours concordantes pour la condition 1, la concordance pose le plus de problèmes à la condition 3 pour le test B. De plus, dans ce dernier cas, l'expert 1 n'est systématiquement pas d'accord avec les trois autres experts selon le coefficient de contingence.

La discussion peut se poursuivre ici quant à l'importance de la sélection des experts. L'expert 2 a montré un comportement marginal dans l'élaboration de ses matrices individuelles, ce qui aboutit logiquement à un problème d'accord entre cet expert et les trois autres. Les résultats montrent que l'expert 1 est également en désaccord avec ses collègues à la condition 3 pour le test B. Cet expert enseigne au niveau supérieur à l'école Polytechnique. Il est donc confronté à la population d'étudiants visée par le test après qu'ils aient passé le test et non avant comme les experts qui enseignent au secondaire. Cet expert se distingue par son utilisation des informations de la condition 3. On peut se demander si sa connaissance différente des étudiants l'a amené à mieux tenir compte des informations fournies à la condition 3 ou pas. Il n'est pas possible de répondre à cette question étant donné que les matrices individuelles n'ont pas fourni de paramètres d'items assez robustes pour être comparés. De plus, la perception de l'expert 3 ne peut pas aider à porter un jugement puisqu'il n'a pas rempli le questionnaire de la condition 3. Quoi qu'il en soit, il ressort clairement que le choix des experts joue un rôle essentiel dans le processus. Il est probable qu'un nombre d'experts plus grand est souhaitable lorsqu'on doit élaborer une matrice afin que leurs différences individuelles n'influencent pas trop les matrices synthèses élaborées.

En ce qui concerne les matrices synthèses, le calcul des coefficients de contingence entre chaque matrice individuelle et la synthèse correspondante permet de juger leur représentativité du panel d'experts. Ces calculs montrent que la matrice synthèse est représentative de chaque matrice individuelle dans les trois conditions. Le fait de fournir des informations sur la difficulté et le pouvoir de discrimination des items ainsi que sur le fonctionnement des leurres semble amener les experts à un moins bon accord entre eux, mais la représentativité des matrices synthèses n'en souffre pas. Le fait que les coefficients de contingence aient permis d'identifier le comportement marginal de certains experts, mais montrent que toutes les matrices sont représentatives du panel d'experts fournit une certaine validation à la procédure de calcul choisie. En outre la représentativité des matrices est un bon indicateur de la pertinence de la méthode MABC pour établir la synthèse des matrices individuelles.

La troisième et dernière question vise à définir de quelle manière les conditions affectent la validité et la fidélité des matrices. Le modèle RUM a été choisi parce qu'il permet d'estimer trois paramètres liés à la validité et à la fidélité de la matrice utilisée. Lors de la modélisation des données avec les matrices individuelles originales et modifiées, plusieurs problèmes de convergence ont empêché l'obtention de paramètres suffisamment robustes pour être comparés d'une condition à l'autre. Aucune explication satisfaisante n'a pu être trouvée même si Hartz (2002), Jang (2004) et Roussos et al. (2007) signalent des problèmes fréquents de convergence.

Les matrices synthèses ont fourni des paramètres très robustes pour le test A aux trois conditions et pour le test B à la condition 1. La robustesse des paramètres estimés aux conditions 2 et 3 pour le test B a été jugée acceptable malgré la remise en cause de celle de la condition 3 à plusieurs reprises dans les analyses. L'ensemble des paramètres d'ajustements étaient convenables pour les deux tests et les trois conditions. En outre, les analyses ont permis de vérifier que le fait de fixer la valeur du paramètre c augmente la robustesse des estimations et améliore la structure cognitive

comme annoncé dans la littérature par 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 qui ont fait l'objet des comparaisons proviennent des matrices synthèses simplifiées. Il n'existe pas de lien direct entre les paramètres de difficulté (p_k) et la validité ou la fidélité des matrices. Il est surprenant de constater qu'aucune différence de difficulté des attributs ne découle de la différence de difficulté entre les deux tests. Toutefois, la condition 2 fournit quatre attributs plus difficiles pour B que pour A. Cette condition semble donc de ce point de vue mieux prendre en compte la différence entre les deux tests.

Le résultat le plus intéressant vient des représentations graphiques des valeurs de la difficulté des attributs 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 deux parties. Un manque de convergence peut fournir une explication à ce phénomène dans le cas du test B, mais on observe également ce problème dans le cas du test A, pour lequel la convergence est bonne. Cette information suggère un manque de validité des attributs à la condition 3 puisque le contenu n'est pas clairement défini. Les nuages de points des conditions 1 et 2 sont homogènes. Stout rapporte l'existence de ce problème dans une conférence présentée en 2004. La solution qu'il propose consiste à chercher à diviser les attributs concernés. Comme ce problème n'a pas été observé aux deux autres conditions, les informations de la condition 3 semblent avoir abouti à des attributs au contenu moins homogène.

Le paramètre π est un indicateur de la fidélité de la matrice. Les valeurs moyennes de ce paramètre sont semblables pour les deux tests. La valeur des paramètres π dépend de la difficulté des items puisque les items les plus difficiles manquent de fidélité au sens d'une valeur de π trop petite. Les conditions n'améliorent pas ce problème pas plus que l'appartenance à l'un des groupes d'items créés selon le comportement des experts (voir tableau 21). Ceci signifie que même lorsque les

experts ont opéré beaucoup de changements avec des liens qui varient beaucoup, il n'en résulte pas une meilleure prise en considération de la difficulté des items difficiles. En particulier, le fait d'informer les experts sur la difficulté des items ne les a pas amenés à mieux identifier les attributs en lien avec les items difficiles.

Dans une perspective diagnostique, l'information cognitive est importante pour l'ensemble des items. Même si les tests diagnostiques comportent traditionnellement des items moins difficiles que les tests de sélection, des attributs pertinents reliés aux items difficiles sont essentiels pour que la structure de la matrice soit forte. Selon le 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 en plus grand nombre. Le fait que les attributs prennent mal en considération la difficulté relative des deux tests pourrait également fournir une explication aux valeurs de π trop petites pour les items difficiles.

Le paramètre r est un indicateur de la validité de la matrice. Les valeurs estimées montrent que la structure cognitive du test B est meilleure que celle du test A. Les conditions n'influencent pas les valeurs de r sauf en ce qui concerne l'attribut 1 (décodage). Les conditions 1 et 2 sont plus propices à une matrice plus valide selon les paramètres r . En particulier, la condition 2 offre la matrice avec la structure cognitive la plus forte pour le test B. De plus, elle 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.

La comparaison de la structure du test selon le tableau de spécification et selon les regroupements factoriels suggère que les experts ont plus de facilité à exploiter les regroupements d'items lorsque ceux-ci sont associés à des contenus semblables. Les différences observées entre les

deux tests trouvent peut-être là une explication plausible puisque les facteurs ont mieux été mis en relation avec le tableau de spécification et ont mieux été pris en compte par les experts pour B que pour A. En outre, les analyses ont montré que l'attribut 5 était présent dans la plupart des facteurs pour les deux tests et que les liens avec l'attribut 3 subissaient l'influence de la condition 2. On peut supposer que la condition 2 a aidé les experts à mieux cibler ces deux attributs. L'attribut 6 concerne l'organisation pour trouver la réponse. Les informations additionnelles fournies aux conditions 2 et 3 ont minimisé l'intérêt des experts face à cet aspect du processus de réponse qui se sont concentrés sur le contenu plutôt que sur la démarche.

Les matrices élaborées à la condition 3 sont les moins valides au sens du paramètre r . 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 et les paramètres de validité r sont statistiquement différents selon que les items sont ou non associés à d'autres par les analyses factorielles. Les items qui partagent une dimension avec d'autres 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). En ce qui concerne le paramètre π , il est possible que la validité des items les plus difficiles soit moins aisée à obtenir parce qu'ils partagent moins avec les autres items. En y réfléchissant, 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 va dans le sens de la nécessité d'élaborer un tableau de spécification dont chaque catégorie vise un contenu précis et comporte plus d'un item.

Dans la condition 1, les experts n'avaient accès qu'aux énoncés des items des deux tests. Les listes d'attributs cognitifs qu'ils ont élaborées dans cette première condition ont été les plus complètes et aucun problème de concordance n'a été observé entre leurs matrices individuelles. L'aspect technique des items est bien cerné dans cette condition et ne l'est pas au détriment des autres aspects. C'est également dans la condition 1 que les experts ont le mieux pris en compte l'organisation dans le processus de réponse aux questions. 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 synthèses simplifiées.

Dans la condition 2, les experts devaient considérer les items tels que regroupés par une analyse factorielle. Cette information supplémentaire les a amenés à établir un nombre de liens très semblables d'un expert à l'autre et globalement plus petit qu'aux deux autres conditions. À une exception près, la concordance entre les matrices individuelles n'a pas été problématique. Les experts ont presque tous négligé l'attribut 1 concernant le décodage des questions. Par contre, 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 (p_k) 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 présents dans les matrices.

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 sa 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 (attribut 1) et biaisé les résultats, (c) les matrices de l'expert 1 n'étaient pas en concordance avec celles des trois autres experts pour le test B, (d) des anomalies ont été observées dans les modélisations successives pour le test B et la convergence était

moins bonne , (e) les valeurs des paramètres p_k 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.

4.2 Les limites

Une première limite à cette recherche émane du fait que les conditions ont été considérées comme indépendantes et traitées comme telles dans les résultats. La méthodologie prévoyait de comparer la condition 2 et la condition 3 à la condition 1 afin de juger de la pertinence de chaque source d'information séparément. Les conditions 2 et 3 auraient pu faire l'objet de plus de comparaisons, ce qui n'a pas été fait étant donné l'identification de nombreux problèmes reliés à la condition 3.

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 (autant que possible) de ne pas tenir compte de leur travail précédent à chaque nouvelle condition.

Les experts ont 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.

Toutefois, il se peut que le plan expérimental adopté ait amené 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 galvauder 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 pour laquelle ils étaient déstabilisés ou lassés. Toutefois, rien ne permet de valider ou d'invalider cette hypothèse.

Les choix des experts et de leur nombre ont été faits dans une approche essentiellement pragmatique. Il est 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 obtenues. En ce qui concerne le choix de ces experts, les résultats ont 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 suite au consensus. Le fait d'inclure les matrices de cet expert dans le processus de synthèse des matrices constituait peut-être une mauvaise décision qui a biaisé les résultats. En outre l'expert 1 s'est distingué à la condition 3. Il se peut qu'une meilleure connaissance des étudiants de niveau universitaire plutôt que secondaire ou collégial de la part des experts 1 et 2 ait eu un impact sur les résultats.

L'ensemble de la recherche repose sur une succession de choix théoriques et pratiques. En ce qui concerne les choix théoriques, la condition 2 se base sur les résultats d'analyses factorielles exploratoires puis confirmatoires. Étant donné que plusieurs modèles s'ajustaient aux données, le modèle retenu n'est pas forcément le meilleur, ce qui a pu influencer les résultats de cette condition.

Les experts auraient peut-être eu plus de facilité à interpréter d'autres regroupements. À la vue de l'importance des liens entre la structure factorielle et le tableau de spécification, il apparaît a posteriori qu'une comparaison des facteurs possibles au tableau de spécification du test aurait pu aider à retenir les facteurs, ce qui n'a pas été fait.

Le modèle RUM a montré sa pertinence par la richesse des comparaisons qu'il a rendues possibles. Toutefois, la complexité de ce modèle n'a pas permis de comparer les matrices individuelles à cause des problèmes de convergence, ce qui était pourtant la visée originale de la recherche. Ces problèmes de convergence ont amené à élaborer des matrices synthèses selon une procédure qui a probablement influencé les résultats. En outre, même si la modélisation RUM à l'aide d'Arpeggio a suivi rigoureusement les étapes fournies par la littérature, ce modèle est trop récent pour que l'ensemble des implications dues au choix de la longueur des chaînes utilisées ou des liens à supprimer soient parfaitement connues et contrôlées. Il se peut que le choix d'un modèle plus simple comme le modèle DINA ait abouti à des estimations plus robustes et conduit à des résultats différents.

En ce qui concerne les choix pratiques, les tests utilisés dans cette recherche constituent en eux-mêmes une limite puisqu'ils ont imposé une démarche complètement a posteriori. En outre les tableaux de spécification n'ont pas suffisamment été exploités, ni pour soutenir les décisions dans les analyses factorielles, ni pour aider les experts. L'identification des attributs dans un tel contexte a rendu la tâche d'autant plus difficile.

Les items portant sur les fonctions élémentaires et le calcul différentiel ont été retenus plus ou moins arbitrairement. D'autres items auraient peut-être conduit à d'autres résultats, ces choix limitent une éventuelle généralisation des résultats. L'idée originale de la recherche visait à comparer chaque expert d'une condition à l'autre. La nécessité de fabriquer une matrice synthèse a amené à un regroupement des attributs individuels et à en éliminer certains qui auraient été gardés avec un plus

grand nombre de catégories autorisées par les consignes. La méthode appliquée a démontré sa pertinence, mais elle aurait pu être différente. Le chercheur a donc biaisé les résultats par l'intermédiaire des nombreuses décisions prises : nombre d'attributs, nombre de catégories, méthode de travail, choix des items, etc.

Un biais a été introduit dans la condition 3 sous la forme de l'exemple qui illustrait les consignes. En effet, cet exemple suggérait que la différence entre le leurre le plus choisi et la bonne réponse tenait peut-être à un problème de compréhension du langage mathématique. Ceci a influencé les experts jusqu'à aboutir à une utilisation massive de l'attribut 1 (décodage).

Les résultats tendent à montrer que les conditions dans lesquelles les experts ont accès aux items regroupés par une analyse factorielle fournissent une matrice Q avec une structure cognitive un peu plus forte pour les deux tests qui font l'objet de cette étude. Toutefois, ces résultats sont issus de l'analyse de deux cas particuliers et ne peuvent en aucun cas être généralisés sans que d'autres recherches soient menées. Pourtant, même si les indications de meilleure qualité des matrices de la condition 2 sont assez ténues, 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 montre, elle aussi, la pertinence de cette approche. Il semble en particulier que les items qui partagent une dimension commune avec d'autres items dans les tests soient définis de manière plus valide et plus fidèle que les autres items, indépendamment du fait d'en informer ou non les experts. En outre, un soin particulier doit être porté à l'élaboration du tableau de spécification afin que sa structure soutienne efficacement le travail des experts dans leur utilisation des résultats des analyses factorielles. D'autres études devront chercher à vérifier ces résultats.

Les comparaisons ne portent que sur les modélisations impliquant les matrices synthèses. Il est dommage de n'avoir pas pu utiliser les matrices individuelles des experts pour vérifier la compatibilité des résultats concernant les synthèses et les individus.

4.3 Les recommandations

Les modifications apportées à la méthode MACB originale ont permis d'utiliser cette méthode afin d'établir une liste d'attributs cognitifs qui fasse consensus auprès des experts. Au delà de conseiller d'utiliser cette méthode pour obtenir la liste des attributs, il apparaît qu'elle aurait pu servir à placer les liens dans les matrices synthèses. En fournissant les matrices individuelles modifiées à l'issue de la phase 2 du MACB aux experts, ils pourraient tous ensemble reprendre les liens et décider de ceux qui sont finalement pertinents en cherchant à obtenir un consensus.

Une telle façon de faire n'impliquerait plus de vérifier la représentativité de la matrice synthèse telle que créée dans cette recherche puisque le consensus interviendrait non seulement au niveau des listes d'attributs, mais également au niveau des liens qu'elle contient. De telles matrices synthèses seraient encore plus représentatives du panel d'experts que celles utilisées dans cette recherche. Le seul inconvénient à cette façon de faire est le temps de travail supplémentaire demandé aux experts.

Les interprétations des résultats montrent que la condition 3 est peu favorable à l'obtention de matrices de qualité. Il ne faut pas oublier que le biais introduit dans les consignes doit nuancer cette recommandation, mais il semble logique de déconseiller de fournir la difficulté des items, leur pouvoir discriminant et 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 se peut également que ce type d'informations n'aide pas les experts à élaborer les attributs requis par chaque item. Les experts n'ont peut-être pas réussi à identifier des attributs caractérisant les items les plus difficiles parce que les attributs sont toujours

utilisés à la fois par des items faciles et difficiles. De même, les experts n'ont pas forcément vu le lien entre le pouvoir de discrimination ou l'utilisation des leurres et les attributs reliés à l'item.

Le fait de réaliser des analyses factorielles sur les données et de regrouper les items qui partagent une même dimension semble améliorer la structure cognitive des matrices puisque les comparaisons ont souvent été à l'avantage de la condition 2. Le fait de n'avoir utilisé que deux tests constitue une limite importante à cette recherche, d'autant que certains des résultats ne se retrouvent pas de façon uniforme pour les deux tests. Le fait que le test A est plus facile que le test B, le fait qu'il a toujours été traité en premier par les experts et le fait que le tableau de spécification valide mieux la structure factorielle pour le test B que pour le test A ont certainement joué un rôle dans les différences.

Une recommandation importante lorsqu'on élabore une matrice Q concerne donc le tableau de spécification du test. Il faut d'une part avoir accès à un tableau de spécification du test soigneusement élaboré et d'autre part en tenir compte pour interpréter les regroupements factoriels des items. Plus le lien entre le tableau de spécification et les facteurs est étroit et mieux les experts sont en mesure d'identifier les attributs que partagent les items. Ainsi, l'élaboration d'un tableau de spécification du test aussi précis que possible doit précéder la fabrication de la matrice Q dans une approche inductive.

Cette recherche met en évidence l'importance des consignes données aux experts. Le fait de leur avoir fourni un exemple pour les aider à comprendre ce qui était attendu dans la condition 3 a biaisé les résultats liés à l'attribut 1. Cette façon de procéder n'est donc pas à recommander pour expliquer leur tâche aux experts ou est à revoir avec d'autres types d'exemples.

4.4 La conclusion

Les modèles cognitifs fournissent des renseignements individuels sur les forces et faiblesses des sujets à partir des seules réponses des sujets sans aucune trace de leur démarche. Ces modèles informent les enseignants et les étudiants sur les difficultés rencontrées et en conséquence sur les moyens d'y remédier. Roussos et ses collègues (2007) affirment que le développement actuel du modèle RUM permet d'envisager de mettre en place des études pilotes pour utiliser ce modèle sur le terrain afin d'exploiter des évaluations à grande échelle à des fins de diagnostic et d'évaluation formative ou pour élaborer des épreuves diagnostiques visant à soutenir les élèves dans leurs apprentissages. Une image valant mille mots, la figure 28 témoigne de la valeur d'un diagnostic sous forme de la probabilité pour un sujet de maîtriser un ensemble d'attributs indispensables à la résolution d'items en mathématiques. Il est facile d'en imaginer l'intérêt dans n'importe quel autre domaine.

C'est la matrice Q et les attributs cognitifs qu'elle contient et relie aux items qui donnent leur caractère spécifique à ces modèles et les différencient des autres approches diagnostiques. Un défi essentiel à relever consiste à les élaborer pour qu'elles garantissent la qualité du diagnostic posé. Cette étude montre que le défi se situe également dans la robustesse des paramètres estimés vu les nombreux problèmes de convergence rencontrés.

Cette recherche exploratoire apporte une contribution multiple à l'avancée des connaissances concernant ces modèles. Tout d'abord, la méthode pour faire travailler les experts a démontré sa pertinence par le fait que les matrices synthèses obtenues étaient représentatives de l'ensemble des experts et qu'elles ont permis d'obtenir des estimateurs robustes des paramètres du modèle. De plus,

les résultats attirent l'attention sur la rédaction des consignes à fournir aux experts qui peuvent très facilement avoir une influence non attendue sur les matrices produites.

Cette recherche a démontré que les items plus difficiles ont tendance à être moins bien représentés par les attributs cognitifs élaborés par les experts. Dans la perspective de fournir le meilleur diagnostic possible, ces items doivent pourtant être reliés à des attributs pertinents. Il semble donc qu'il faille porter une attention particulière à ces items lorsqu'on élabore la matrice.

Enfin, des trois conditions expérimentales de cette recherche, la condition 3 n'a pas aidé les experts dans leur tâche, soit parce que les informations étaient trop nombreuses, soit parce qu'elles n'étaient pas pertinentes. En outre, l'analyse des leurres des questions à choix multiple semble faire perdre de vue la question dans son ensemble. Les experts élaborent des attributs adéquats lorsqu'ils n'ont que l'énoncé de la question dans les mains. Toutefois, le fait de regrouper les items sur la base des résultats d'une analyse factorielle semble constituer un apport d'information utile. En outre, les items qui partagent une dimension avec d'autres items correspondent à des liens plus valides et plus fidèles dans les matrices indépendamment du fait d'en informer les experts. Lorsque les experts ont accès à des regroupements factoriels d'items, ils semblent trouver plus facilement ce que partagent les items lorsque l'analyse factorielle et le tableau de spécification du test présentent des informations congruentes.

Cette étude a démontré qu'il est possible d'appliquer un modèle cognitif à un test élaboré par des enseignants pour leur propre usage et d'obtenir des paramètres robustes. Toutefois, il est curieux de constater que seules les matrices synthèses ont fourni de tels paramètres au contraire des matrices individuelles de chaque expert. Les résultats obtenus ne sont pas généralisables puisqu'ils correspondent à l'étude de deux cas particuliers de tests, mais ils ouvrent la voie à des recherches visant à expliquer les problèmes de convergence des estimations et en particulier dans quelle mesure ceux-ci sont influencés par la forme des matrices.

De nombreuses recherches devront encore se pencher sur les modèles cognitifs dans la perspective d'améliorer la création des matrices et de rendre leur utilisation plus souple. Il faudra également se pencher sur les problèmes de convergence des routines d'estimation afin de comprendre ce qui les influence.

Références

- Bechger, T.M., Verstralen, H.H.F.M., & Verhelst, N.D. (2000). *Equivalent logistic test models*. RetD report 2000-4, Arnhem: Cito.
- Birenbaum, M., Tatsuoka, C., & Yamada, T (2004). Diagnostic assessment in TIMSS-R : Between-countries and within country comparisons of eighth graders' mathematics performance. *Studies in Educational Evaluation*, 30, 151-173.
- Boaler, J. (2002). Exploring the nature of mathematical activity: Using theory, research and "working hypotheses" to broaden conceptions of mathematics knowing. *Educational Studies in Mathematics*, 51, 3-21.
- Brown, R.S. (2000). *Using latent class analysis to set academic performance standards*. Paper presented at the annual meeting of the American Educational Research Association, New-Orleans, LA.
- Buck, G., & 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), 119-157.
- Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple choice test of second language reading comprehension. *Language testing*, 47(3), 423-466.
- Buckendahl, C.W., Smith, R.W., Impara, J.C., & Plake, S. (2000). *A comparison of the Angoff and Bookmark standard setting methods*. Paper presented at the Annual meeting of the Mid-Western Educational Research association, Chicago, October, 2000.
- Buckendahl, C.W., Smith, R.W., Impara, J.C., & Plake, B.S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational measurement*, 39(3), 253-263.

- Chipman, S.F., Nichols, P.D., & Brennan, R.L. (1995). Introduction. Dans P.D. Nichols, S.F. Chipman et R.L. Brennan (Ed.), *Cognitively diagnostic assessment* (pp. 1-18). Hillsdale, NJ: Erlbaum.
- 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* (pp. 305-326). Hillsdale, NJ: Erlbaum.
- Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. New York: HBJ College Publishers.
- De la Torre, J., & Douglas, J.A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69(3), 333-353.
- DiBello, L.V. , Stout, W.F., & Roussos, L.A (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. Dans P.D. Nichols, S.F. Chipman et R.L. Brennan (Ed.), *Cognitively diagnostic assessment* (pp. 361-389). Hillsdale, NJ: Erlbaum.
- DiBello, L., Roussos, L. A., Stout, W. (2006). *Special topics section chapter: Review of cognitively diagnostic assessment and a summary of psychometric models*. Draft.
- Dimitrov, D.M. (1994). *Hierarchical latent trait approach in test analysis*. Paper presented at the Annual Meeting of the Midwestern Educational Research Association, Chicago, October, 1994.
- Dimitrov, D.M., & Raykov, T. (2003). Validation of cognitive structures: a structural equation modeling approach. *Multivariate Behavioral Research*, 38(1), 1-23.
- Embretson, S.E. (1985a). Introduction to the problem of test design. Dans S.E. Embretson (Eds), *Test design: Developments in psychology and psychometrics*. 3-17. New York: Academic Press.

- Embretson, S.E. (1985b). Multicomponent latent trait models for test design. Dans S.E. Embretson (Eds), *Test design: Developments in psychology and psychometrics* (pp. 279-294). New York: Academic Press.
- Embretson, S.E. (1993). Psychometric models for learning and cognitive processes. Dans N. Fredericksen, R.J. Mislevy et I.I. Bejar (Eds), *Test theory for a new generation of tests* (pp. 125-150). Hillsdale, NJ: Erlbaum.
- Embretson, S.E. (1999). Cognitive psychology applied to testing. Dans F.T. Durso, R.S. Nickerson, R.W. Schvaneveldt, S.T. Dumais, D.S. Lindsay et M.T.H. Chi (Eds), *Handbook of applied cognition* (pp 629-658). New York: Willey.
- Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4, 272-299.
- Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 37, 359-374.
- Garofalo, J., & Lester, F.K.J. (1985). Metacognition, cognitive monitoring, and mathematical performance. *Journal for Research in Mathematics Education*, 16(3), 163-176.
- Gierl, M.J., Leighton, J.P., & Hunka, S.M. (2000). Exploring the logic of Tatsuoaka's Rule-Space model for test development and analysis. *Educational Measurement : Issues and Practices*, 19(3), 34-44.
- Gierl, M.J. (1997). *An investigation of the cognitive foundation underlying the rule-space model*. Thèse de doctorat, Proquest Digital Dissertation.

- Gitomer, D.H., & Rock, D. (1993). Addressing process variables in test analysis. Dans N. Fredericksen, R.J. Mislevy et I.I. Bejar (Eds), *Test theory for a new generation of tests* (pp. 125-150). Hillsdale, NJ: Erlbaum.
- Hair, J. F. Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate analysis, fifth edition*. Upper Saddle River, NJ: Prentice Hall.
- Hakstian, R., & Whalen, T. E. (1976). A K-sample significance test for independent alpha coefficients. *Psychometrika*, 41(2), 219-231.
- 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.
- Hu, L., & Bentler, P., M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, 4, 1-55.
- 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.
- Jaeger, R.M. (1989). Certification of student competence. In: R.L. Linn (ED.). *Educational measurement (3rd edition)*. New York: American Council on Education and Macmillan.
- Jaeger, R.M. (1996). Setting standards for complex performances: An iterative, judgmental policy capturing strategy. *Educational Measurement: Issues and Practice*, 14(4), 16-20.
- Jiang, H. (1996). *Applications of computational statistics in cognitive diagnosis and IRT modeling*. Thèse de doctorat non publiée. University of Illinois, Urbana-Champaign.
- 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.

- Junker, B.W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258-272.
- Kane, M.T. (2001). So much remain the same: Conception and status of validation in setting standards. Dans G.J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 53-88) Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Kilpatrick, J. (2001). Understanding mathematical literacy: The contribution of research. *Educational Studies in Mathematics*, 47(1), 101-116.
- Laveault, D., & Grégoire, J. (1997). *Introduction aux théories des tests en sciences humaines*. De Boeck Université, Paris, Bruxelles.
- Legendre, R. (2005). *Dictionnaire actuel de l'éducation*, 3^{ème} édition. Guérin, Montréal.
- Leighton, J.P., Gierl, M.J., & Hunka, S.M. (2002). *The attribute hierarchy model for cognitive assessment*. Paper presented at the annual meeting of the National Council for Measurement in Education, Nouvelle Orléans, April, 2002.
- Leighton, J. P., Goikert, R. J., & 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), 141-189.
- Levy, R., & Mislevy, R.J. (2004). *Specifying and refining a measurement model for a simulation-based assessment*. National center for research on evaluation: rapport de recherche CSE Report 619.
- Lewis, D. R., Erickson, R. N., Johnson, D. R., & Bruininks, P. H. (1991). *Using multi-attribute utility evaluation: Technique in special education*. Unpublished manuscript, University of Minnesota, Institute on community integration, Minneapolis.

- Lipkins, R.H., Jones, J.P., & Halkitis, P.N (1996). *On the trial of the deviant rater : influence of item effects and method on consistency*. Paper presented at the annual meeting of the American Educational Research Association in New York, 8-12, New-York. (avril)
- Loye, N. (2002). *Une étude de l'accord et de la fidélité inter-juges*. Mémoire de maîtrise non publié.
- Lunz, M.E. (2000). Setting standards on performance examinations. *Objective measurement, Theory into Practice*, 5, 181-199.
- Lunz, M.E., Stahl, J.A., & Wright, B.D. (1994). Inter-judge reliability and decision reproducibility. *Educational and psychological measurement*, 54, 913-925.
- Marsh, H. W., Ellis, L. A., Parada, R. H., & Richards, G. (2005). A short version of the self description questionnaire II : Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment*, 17(1), 81-102.
- McGinty, D., & Neel, J.H. (1996). *Judgmental; standard setting using a cognitive component model*. Paper presented at 1996 the annual meeting of the national council on measurement in education, New York.
- Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, 50, 741-749.
- Milewski, G.B., & 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 , 2002.
- Mislevy, R.J. (1993). Foundations of a new test theory. Dans N. Frederiksen, R.J. Mislevy, et I. Bejar (Eds), *Test theory for a new generation of tests*. Hilldale, New Jersey: Lawrence Erlbaum Associates, Publishers.
- Mislevy, R.J. (1995). Probability-based inference in cognitive diagnosis. Dans P.D. Nichols, S.F. Chipman et R.L. Brennan (Ed.), *Cognitively diagnostic assessment* (pp. 43-71). Hillsdale, NJ: Erlbaum.

- Mislevy, R.J. (2004). *A brief introduction to evidence-centered design*. National center for research on evaluation: rapport de recherche CSE Report 632.
- Mislevy, R.J., Almond, R.G., Yan, D., & Steinberg, L.S. (2000). *Bayes Nets in educational assessment: where do the numbers come from?* National center for research on evaluation: rapport de recherche CSE Report 518.
- Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3-62.
- Murphy, K.R., et Davidshofer, C.O. (1988). *Psychological testing*. Englewood Cliffs, NJ: Prentice Hall.
- Muthèn, B., & Muthèn, L. (2004). *Mplus user's guide*.
- Nichols, P.D. (1994). A framework for developing cognitively diagnostic assessments. *Review of educational research*, 64(4), 575-603.
- Nichols, P.D., Chipman, S.F., & Brennan R.L. (Ed.) (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.
- Noss, R., Healy, L. & Hoyle, C. (1997). The construction of mathematical meanings: Connecting the visual and the symbolic. *Educational Studies in Mathematics*, 33, 203-233.
- R (2006). The R project for statistical computing [computer software and manual]. Obtenu à l'adresse <http://www.r-project.org/>
- Raymond, M.R., & Reid, J.B. (2001). Who made thee a judge? Selecting and training participants for setting standards . Dans G.J. Cizek (Ed). *Setting performance standards: concepts, Methods, and perspectives* (pp. 159-173). Mahwah, NJ: Erlbaum.
- Ramsey, J. O. (1996). A geometrical approach to item response theory. *Behaviormetrika*, 23, 3-17.
- Ramsey, J. O. (2000). Testgraph [computer software and manual]. Obtenu à l'adresse <ftp://ego.psych.mcgill.ca/pub/ramsay/testgraf/>

- Ramsey, J. O., & Silverman, B. W. (1997). *Functional data analysis*. New York: Springer.
- Reckase, M. D. (2000). A survey and evaluation of recently developed procedures for setting standards on educational tests. Dans: M.L. Bourque, S. Byrd, (Ed.). *Student performance standards on the national assessment of educational progress: Affirmation and improvements. A study initiated to examine a decade of achievement level setting on NAEP*, (pp.43-69). National assessment governing board, Washington, DC.
- Reckase, M.D. (2001). Innovative methods for helping standards-setting participants to perform their task: the role of feedback regarding consistency, accuracy, and impact. Dans G.J. Cizek (Ed). *Setting performance standards: concepts, Methods, and perspectives* (pp. 159-173). Mahwah, NJ: Erlbaum.
- Roussos, L., DiBello, L., Stout, W., Hartz, S., Henson, R. A., 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 (pp. 275-318). New York, NY: Cambridge University Press.
- 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, 95-125.
- Rupp, A.A., & Templin, J. (2006). *The effect of Q-matrix misspecification on parameter estimates and misclassification rates in the DINA model*. Document non publié.
- Rupp, A.A., & Watermann, R. (2005). *Curricular validity and cognitive item demands as explanatory mechanisms for differential item functioning in the German TOSCA study*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada, April , 2005.

Sadesky, G. (2003). *Cluster analysis and its application in standard setting*. Consulté en ligne le 08 août 2005 à l'adresse :

<http://www.education.ualberta.ca/educ/psych/crame/files/SadeskyCSSE2003.pdf>

Sadesky, G.S., & Gushta, M.M. (2004). *Standard setting using the attribute hierarchy model*.

Consulté en ligne le 08 août 2005 à l'adresse :

<http://www.education.ualberta.ca/educ/psych/crame/files/AHM.pdf>

Scallon, G. (1988). *L'évaluation formative des apprentissages, 1 et 2*. Les presses de l'Université Laval, Québec.

Scallon, G. (2000). *L'évaluation formative*. ERPI, Québec.

Sijtsma, K., & Verweij, A.C. (1999). Knowledge of solution strategies and IRT modeling of items for transitive reasoning. *Applied Psychological Measurement*, 23(1), 55-68.

Sireci, S.G., Robin, F.R., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education*, 12(3), 301-323.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika*, 67(4), 485-518.

Stout, W. (2004). *The Reparameterized unified model skills diagnostic system*. Conference presented at the Spearman diagnostics conference, Philadelphia, PA.

Tate, M. W., & Brown, S. M. (1970). Note on the Cochran Q test. *Journal of the American Statistical Association*, 65(329), 155-160.

Tatsuoka, K.K. (1983). Rule-space: an approach for dealing with misconceptions based on item response theory. *Journal of educational measurement*, 20, 345-354.

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* (pp. 305-326). Hillsdale, NJ: Erlbaum.

- Tatsuoka, K.K. (2005). *Rule space method: Cognitively diagnostic tool*. Paper presented at the National Council for Measurement in Education Training Program, Montreal, April, 2005.
- Tatsuoka, K.K. , Birenbaum, M., Lewis, C., & Sheehan, K.M. (1993). Proficiency scaling based on conditional probability functions for attributes. *Educational Testing Service*. Princeton, NJ.
- Tatsuoka, K.K., & Tatsuoka, M.M. (1997). Computerizing cognitive diagnostic adaptive testing : Effect on remedial instruction as empirical validation. *Journal of educational Measurement*, 34(1), p3-20.
- Templin, J. L. (2004). *Generalized linear mixed proficiency models for cognitive diagnosis*. University of Illinois, Urbana-Champaign.
- Thissen, D., Steinberg, L., & Fitzpatrick, A.R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement*, 26(2), 161-176.
- Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of applied psychology*, 4, 25-29.
- Vanderwood, M., Ysseldyke, J., & Thurlow, M. (1993). *Consensus building: A process for selecting educational outcomes and indicators*. National center on educational outcomes, report 2, University of Minnesota.
- Wang, W. (1998). Rasch analysis of distractors in multiple-choice items. *Journal of Outcome Measurement*, 2(1), 43-65.
- Wilson, M., & Case, H. (2000). An examination of variation in rater severity over time: A study in rater drift. *Objective Measurement, Theory into Practice*, 5, 113-133.
- Wolfe, E. W., & Kao, C.-W., (1996). *Expert/novice differences in the focus and procedures used by essay scorers*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Yan, D., Almond, R., & Mislevy, R. (2003). *Empirical comparisons of cognitive diagnostic models*.

ETS, NJ : Princeton. Article non publié.

Yu, C.-Y., & Muthén, B. (2002). *Evaluation of model fit indices for latent variable models with categorical and continuous outcomes*. Technical report.

Zieky, M.J. (2001). So much has changed: How the setting of cutscores has evolved since the 1980s.

Dans G.J. Cizek (Ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives* (pp. 29-51) Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

ANNEXE A

Résultats des analyses de variance et des tests non paramétriques. Tous les effets sont fixes.

Tableau A1.

Influence du test sur la moyenne des tests A et B

Source	ddl.	Moyenne au carré	F	p	η^2
Test	1	100,875	8,918*	0,003	0,847
erreur	902	(11,311)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A2

Influence du test sur le nombre moyen de liens dans les matrices individuelles originales (test non paramétrique de Friedman)

Source	ddl.	Moyenne au carré	χ^2	p
Test	1	57,042	0,840	0,359
résidu	11	(68,860)		

*p<0,05

Tableau A3

Influences de l'expert et de la condition sur le nombre moyen de liens dans les matrices individuelles originales

Source	df	Moyenne au carré	F	p.	η^2
Condition	2	316,667	4,665*	0,032	0,670
Expert	3	1351,931	19,918*	0,000	1,000
Condition x Expert	6	132,889	1,958	0,152	0,490
erreur	12	(67,875)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A4

Influence de la condition et de l'expert sur le nombre moyen de liens dans les matrices individuelles originales (sans l'expert 2)

Source	ddl.	Moyenne au carré	F	p.	η^2
Condition	2	141,556	1,988	0,193	0,306
Expert	2	353,722	4,966*	0,035	0,525
Condition * Expert	4	117,889	1,655	0,243	0,424
erreur	9	(71,222)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A5

Influence du test sur le nombre moyen d'intitulés individuels reliés aux catégories du MACB (test non paramétrique de Friedman)

Source	ddl.	Moyenne au carré	χ^2	p.
Test	1	0,042	1,000	0,317
résidu	11	(0,042)		

*p<0,05

Tableau A6

Influence de la condition, de l'expert et de l'attribut sur le nombre moyen d'intitulés individuels reliés aux catégories du MACB

Source	ddl.	Moyenne au carré	F	p	η^2
condition	2	,396	7,125*	0,002	0,165
expert	3	,963	17,333*	0,000	0,419
Attribut (catégorie)	5	27,217	489,900*	0,000	0,971
condition * expert	6	,359	6,458*	0,000	0,350
condition * attribut	10	1,638	29,475*	0,000	0,804
expert * attribut	15	2,824	50,833*	0,000	0,914
condition * expert * attribut	30	1,378	24,808*	0,000	0,912
erreur	72	(0,056)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A7

Influence du test sur le nombre moyen de liens dans les matrices individuelles modifiées (test de Friedman)

Source	ddl.	Moyenne au carré	χ^2	p
Test	1	22,042	0,800	0,371
résidu	12	(28,042)		

*p<0,05

Tableau A8

Influence de la condition et de l'expert sur le nombre moyen de liens dans les matrices individuelles modifiées

Source	df	Moyenne au carré	F	p	η^2
Condition	2	48,042	1,744	0,216	0,225
Expert	3	160,042	5,811*	0,011	0,592
Condition * Expert	6	68,875	2,501	0,083	0,556
erreur	12	(27,542)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A9

Influence de la condition sur le nombre moyen de liens dans les matrices individuelles modifiées pour chaque attribut séparément (expert est une mesure répétée)

Source	df	Moyenne au carré	F	p	η^2
Attribut 1					
Condition	2	135,375	95,559*	0,002	0,895
erreur	3	(1,417)			
Attribut 2					
Condition	2	12,125	7,658	0,066	0,836
erreur	3	(1,583)			
Attribut 3					
Condition	2	3,125	0,833	0,515	0,357
erreur	3	(3,750)			
Attribut 4					
Condition	2	5,792	3,089	0,187	0,673
erreur	3	(1,875)			
Attribut 5					
Condition	2	70,167	24,765*	0,014	0,943
erreur	3	(2,833)			
Attribut 6					
Condition	2	7,167	1,132	0,430	0,430
erreur	3	(6,333)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

* $p < 0,05$

Tableau A10

Influence de l'expert et de la condition sur le nombre moyen d'attributs consensuels utilisés dans matrices individuelles modifiées

Source	df	Moyenne au carré	F	p	η^2
expert	3	2,486	59,667*	0,000	0,937
condition	2	,125	3,000	0,088	0,333
expert * condition	6	1,569	37,667*	0,000	0,950
erreur	12	(0,042)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

* $p < 0,05$

Tableau A11

Influence du test sur le nombre moyen de liens (total) dans les matrices synthèses (test non paramétrique de Friedman)

Source	df	Moyenne au carré	χ^2	p
Test	1	0,667	0,074	0,785
résidu	2	(13,167)		

*p<0,05

Tableau A12

Influence de la condition sur le nombre moyen de liens (total) dans les matrices synthèses

Source	df	Moyenne au carré	F	p	η^2
Condition	2	21,500	2,389	0,240	0,614
erreur	3	(9,000)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A13

Influence du test sur le nombre moyen de liens (par attribut) dans les matrices synthèses (test de Friedman)

Source	df	Moyenne au carré	χ^2	p
Test	1	0,111	0,039	0,843
résidu	17	(2,993)		

*p<0,05

Tableau A14

Influence de la condition et de l'attribut sur le nombre moyen de liens (par attribut) dans les matrices synthèses

Source	df	Moyenne au carré	F	p	η^2
Condition	2	3,583	1,265	0,306	0,123
Attribut	5	161,133	56,871*	0,000	0,940
Condition * Attribut	10	42,617	15,041*	0,000	0,893
erreur	18	(2,833)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

* $p < 0,05$

Tableau A15

Influence de la condition sur le nombre moyen de liens dans les matrices synthèses pour chaque attribut séparément

Attribut 1	Source	df	Moyenne au carré	F	p	η^2
Attribut 1	Condition	2	151,167	50,389*	0,005	0,971
	erreur	3	(3,000)			
Attribut 2	Condition	2	0,667	1,333	0,385	0,471
	erreur	3	(0,500)			
Attribut 3	Condition	2	4,167	0,595	0,606	0,284
	erreur	3	(7,000)			
Attribut 4	Condition	2	8,667	13,000*	0,033	0,897
	erreur	3	(0,667)			
Attribut 5	Condition	2	45,500	13,000*	0,033	0,897
	erreur	3	(3,500)			
Attribut 6	Condition	2	6,500	2,786	0,207	0,650
	erreur	3	(2,333)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

* $p < 0,05$. Si les valeurs de F sont considérées comme indépendantes, le taux d'erreur de l'ensemble est égal à $(1 - (1 - 0,05)^6) = 0,26$. Les valeurs de F correspondant à une valeur de p inférieure à 0,26 sont en caractères gras.

Tableau A16

Influence de test sur le nombre moyen de liens supprimés dans les matrices synthèses lors de l'application du modèle RUM

Source	df	Moyenne au carré	F	p	η^2
Test	1	4,167	2,500	0,189	0,385
erreur	4	(1,667)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A17

Influence de la condition sur le nombre moyen de liens supprimés dans les matrices synthèses lors de l'application du modèle RUM (test de Friedman)

Source	df	Moyenne au carré	χ^2	p
Condition	1	2,250	1,000	0,317
résidu	1	(2,250)		

*p<0,05

Tableau A18

Influence du test sur le nombre moyen de liens (par attribut) dans les matrices synthèses simplifiées (test de Friedman)

Source	df	Moyenne au carré	χ^2	p
Test	1	1,361	0,343	0,558
erreur	17	(4,126)		

*p<0,05

Tableau A19

Influence de la condition et de l'attribut sur le nombre moyen de liens (par attribut) dans les matrices synthèses

Source	df	Moyenne au carré	F	p	η^2
Condition	2	4,528	1,140	0,342	0,112
Attribut	5	121,828	30,670*	0,000	0,895
Condition * Attribut	10	34,594	8,709*	0,000	0,829
erreur	18	(3,972)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A20

Influence de la condition sur le nombre moyen de liens dans les matrices synthèses simplifiées pour chaque attribut séparément

Source	df	Moyenne au carré	F	p	η^2
Attribut 1					
Condition	2	126,000	18,900*	0,020	0,926
erreur	3	(6,667)			
Attribut 2					
Condition	2	0,167	0,034	0,966	0,402
erreur	3	(4,833)			
Attribut 3					
Condition	2	4,500	0,900	0,494	0,375
erreur	3	(5,000)			
Attribut 4					
Condition	2	9,500	6,333	0,084	0,809
erreur	3	(1,500)			
Attribut 5					
Condition	2	32,667	8,167	0,061	0,845
erreur	3	(4,000)			
Attribut 6					
Condition	2	4,667	2,545	0,226	0,629
erreur	3	(1,833)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05. Si les valeurs de F sont considérées comme indépendantes, le taux d'erreur de l'ensemble est égal à $(1-(1-0,05)^6)=0,26$. Les valeurs de F correspondant à une valeur de p inférieure à 0,26 sont en caractères gras.

Tableau A21

Influence de la condition sur le nombre moyen de liens par item dans les matrices synthèses simplifiées (test de Friedman)

Source	df	Moyenne au carré	χ^2	p
Test A				
Condition	2	0,317	1,583	0,453
résidu	38	(0,404)		
Test B				
Condition	2	1,421	4,378	0,112
résidu	36	(0,606)		

*p<0,05

Tableau A22

Influence du test sur la valeur moyenne de pk (test de Friedman)

Source	df	Moyenne au carré	F	p
Test	1	0,008	0,498	0,480
résidu	15	(0,017)		

*p<0,05

Tableau A23

Influence de la condition sur la valeur moyenne de pk (attribut est une mesure répétée)

Source	df	Moyenne au carré	F	p	η^2
condition	1	0,004	4,908	0,270	0,831
erreur	1	(0,001)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A24

Influence du test sur la valeur moyenne de π dans les matrices synthèses simplifiées (test de Friedman)

Source	df	Moyenne au carré	χ^2	p
test	1	0,000	0,004	0,951
résidu	55	(0,001)		

*p<0,05

Tableau A25

Influence de la condition sur la valeur moyenne de π dans les matrices synthèses simplifiées (test est une mesure répétée)

Source	df	Moyenne au carré	F	p
Condition	2	0,002	4,500	0,105
résidu	74	(0,001)		

*p<0,05

Tableau A26

Influence de l'item sur la valeur moyenne de π avec les matrices synthèses simplifiées

Source	df	Moyenne au carré	F	p	η^2
Test A					
Item	19	0,145	149,848*	0,000	0,986
erreur	39	(0,001)			
Test B					
Item	18	0,123	138,337*	0,000	0,985
erreur	54	(0,001)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A27

Influence de la difficulté des items sur la valeur moyenne de π avec les matrices synthèses

simplifiées

Source	df	Moyenne au carré	F	p	η^2
Test A					
difficulté	18	0,150	70,458*	0,000	0,969
erreur	40	(0,002)			
Test B					
difficulté	17	0,126	43,315*	0,000	0,950
erreur	39	(0,003)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

* $p < 0,05$

Tableau A28

Influence de l'appartenance à un groupe sur la valeur moyenne de π avec les matrices synthèses

simplifiées

Source	df	Moyenne au carré	F	p	η^2
Test A					
groupe	3	0,079	1,742	0,170	0,090
erreur	53	(0,045)			
Test B					
groupe	2	0,076	1,949	0,152	0,067
erreur	54	(0,039)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

* $p < 0,05$

Tableau A29

Influences de la condition et de l'appartenance à un regroupement (AF) sur la valeur moyenne de π avec les matrices synthèses simplifiées

Source	df	Moyenne au carré	F	p	η^2
condition	2	0,002	0,039	0,962	0,001
app. à un regroupement	1	0,186	4,010*	0,048	0,037
test	1	0,038	0,815	0,369	0,008
condition * app.	2	0,003	0,054	0,947	0,001
condition * test	2	0,000	0,006	0,994	0,000
app. * test	1	0,046	0,998	0,320	0,010
condition * app. * test	2	0,002	0,037	0,964	0,001
Erreur	104	(0,046)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A30

Influence du test sur la valeur moyenne des paramètres r avec les matrices synthèses simplifiées

Source	df	Moyenne au carré	F	p	η^2
Test	1	0,362	11,941	0,001	0,044
erreur	261	(0,030)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A31

Influence de la condition sur la valeur moyenne des paramètres r pour les deux tests et pour chaque test séparément avec les matrices synthèses simplifiées (tests de Friedman)

Source	df	Moyenne au carré	χ^2	p
Test A				
condition	2	0,003	0,278	0,870
résidu	78	(0,024)		
Test B				
condition	2	0,005	0,318	0,853
résidu	82	(0,033)		

*p<0,05

Tableau A32

Influence de la condition sur les valeurs moyennes de r pour les items de chacun des deux tests regroupés selon leur groupe ou selon leur contenu (tests de Friedman)

Source	df	Moyenne au carré	χ^2	p
Test A				
Groupe 1				
Condition	2	0,013	1,011	0,603
résidu	20	(0,027)		
Groupe 2				
Condition	-	-	-	-
résidu				
Groupe 3				
Condition	2	0,003	0,181	0,914
résidu	18	(0,042)		
Groupe 4				
Condition	2	0,000	0,009	0,996
résidu	30	(0,007)		
Contenu 1				
Condition	2	0,003	0,210	0,900
résidu	34	(0,026)		
Contenu 2				
Condition	2	0,003	0,262	0,877
erreur	40	(0,024)		
Test B				
Groupe 1				
Condition	2	0,02	0,253	0,881
résidu	14	(0,022)		
Groupe 2				
Condition	-	-	-	-
résidu				
Groupe 3				
Condition	2	0,002	0,127	0,939
résidu	18	(0,041)		
Groupe 4				
Condition	2	0,005	0,298	0,861
résidu	36	(0,035)		
Contenu 1				
Condition	2	0,007	0,750	0,687
résidu	32	(0,020)		
Contenu 2				
Condition	2	0,002	0,137	0,934
résidu	46	(0,035)		

*p<0,05

Tableau A33

Influence de l'appartenance à un regroupement (AF) sur la valeur moyenne des paramètres r avec les matrices synthèses simplifiées

Source	df	Moyenne au carré	F	p	η^2
Test A					
App.	2	0,139	5,713	0,018	0,043
Erreur	126	(0,024)			
Test B					
App.	1	0,292	8,764	0,004	0,062
Erreur	133	(0,033)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

* $p < 0,05$

Tableau A34

Influence de l'item et de l'attribut sur la valeur moyenne des paramètres r avec les matrices synthèses simplifiées

Source	df	Moyenne au carré	F	p	η^2
Test A					
item	19	0,036	5,555*	0,000	0,641
attribut	5	0,015	2,353	0,052	0,166
Item*attribut	44	0,037	5,669*	0,000	0,809
erreur	59	(0,007)			
Test B					
item	18	0,144	7,051*	0,000	0,672
attribut	5	0,017	0,830	0,533	0,063
Item*attribut	49	0,016	0,783	0,812	0,382
erreur	62	(0,020)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

* $p < 0,05$

Tableau A35

Influence de l'appartenance à un groupe sur la valeur moyenne de r avec les matrices synthèses simplifiées

Source	df	Moyenne au carré	F	p	η^2
Test A					
groupe	3	0,095	4,173*	0,008	0,096
erreur	118	(0,023)			
Test B					
groupe	2	0,025	0,410	0,665	0,006
erreur	132	(0,036)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A36

Influence des conditions sur la valeur moyenne de r pour chaque attribut avec les matrices synthèses simplifiées pour le test A

Source	df	Moyenne au carré	F	p	η^2
Attribut 1					
Condition	1	0,027	3,267	0,092	0,189
erreur	14	(0,008)			
Attribut 2					
Condition	2	0,004	0,107	0,899	0,005
erreur	46	(0,037)			
Attribut 3					
Condition	2	0,028	2,328	0,144	0,297
erreur	11	(0,012)			
Attribut 4					
Condition	1	0,002	0,256	0,626	0,031
erreur	8	(0,009)			
Attribut 5					
Condition	2	0,014	0,523	0,599	0,037
erreur	27	(0,026)			
Attribut 6					
Condition	2	0,031	2,270	0,185	0,431
erreur	6	(0,014)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A37

Influence des conditions sur la valeur moyenne de r pour chaque attribut avec les matrices

synthèses simplifiées pour le test B

Source	df	Moyenne au carré	F	p	η^2
Attribut 1					
Condition	1	0,011	0,256	0,619	0,014
erreur	18	(0,042)			
Attribut 2					
Condition	2	0,017	0,374	0,690	0,020
erreur	37	(0,045)			
Attribut 3					
Condition	2	0,002	0,066	0,937	0,010
erreur	13	(0,037)			
Attribut 4					
Condition	2	0,010	0,211	0,814	0,050
erreur	8	(0,046)			
Attribut 5					
Condition	2	0,002	0,074	0,929	0,005
erreur	31	(0,033)			
Attribut 6					
Condition	2	0,027	0,989	0,403	0,152
erreur	11	(0,027)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

* $p < 0,05$

Tableau A38

Influence des conditions sur la valeur moyenne de r pour chaque item avec les matrices synthèses simplifiées pour le test A

	Source	df	Moyenne au carré	F	p	η^2
Item 1						
	Condition	2	0,005	0,344	0,744	0,256
	erreur	2	(0,014)			
Item 2						
	Condition	2	0,005	1,652	0,377	0,623
	erreur	2	(0,003)			
Item 3						
	Condition	2	0,004	1,363	0,337	0,353
	erreur	5	(0,003)			
Item 4						
	Condition	2	0,003	4,254	0,190	0,810
	erreur	2	(0,001)			
Item 5						
	Condition	2	0,004	0,136	0,876	0,064
	erreur	4	(0,026)			
Item 6						
	Condition	2	0,002	0,676	0,550	0,213
	erreur	5	(0,003)			
Item 7						
	Condition	2	0,009	0,066	0,937	0,019
	erreur	7	(0,130)			
Item 8						
	Condition	2	0,001	0,022	0,979	0,009
	erreur	5	(0,038)			
Item 9						
	Condition	-	-	-	-	-
	erreur					
Item 10						
	Condition	2	0,040	0,728	0,638	0,593
	erreur	1	(0,055)			
Item 11						
	Condition	-	-	-	-	-
	erreur					
Item 12						
	Condition	2	0,001	0,164	0,854	0,076
	erreur	4	(0,007)			
Item 13						
	Condition	2	0,006	0,933	0,432	0,189
	erreur	8	(0,007)			
Item 14						
	Condition	2	0,001	0,675	0,573	0,310
	erreur	3	(0,002)			
Item 15						

Item 16	Condition erreur	2	0,003	0,202	0,824	0,075
		5	(0,014)			
Item 17	Condition erreur	2	0,003	0,308	0,756	0,170
		3	(0,010)			
Item 18	Condition erreur	2	0,049	34,719*	0,028	0,972
		2	(0,001)			
Item 19	Condition erreur	2	0,020	1,136	0,392	0,312
		5	(0,018)			
Item 20	Condition erreur	2	0,004	0,081	0,924	0,051
		3	(0,052)			
	Condition erreur	2	0,000	0,057	0,946	0,036
		3	(0,007)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

Tableau A39

Influence des conditions sur la valeur moyenne de r pour chaque item avec les matrices synthèses simplifiées pour le test B

Source	df	Moyenne au carré	F	p	η^2
Item 1					
Condition	2	0,017	0,603	0,602	0,287
erreur	3	(0,029)			
Item 2					
Condition	2	0,004	0,309	0,744	0,081
erreur	7	(0,013)			
Item 3					
Condition	2	0,005	0,577	0,634	0,366
erreur	2	(0,008)			
Item 4					
Condition	2	0,005	0,305	0,748	0,092
erreur	6	(0,015)			
Item 5					
Condition	-	-	-	-	-
erreur					
Item 6					
Condition	2	0,025	1,206	0,374	0,325
erreur	5	(0,021)			
Item 7					
Condition	2	1,61E-005	0,011	0,989	0,007
erreur	0	-			
Item 8					

	Condition erreur	-	-	-	-	-
Item 9	Condition erreur	2	0,052	6,761*	0,023	0,659
		7	(0,008)			
Item 10	Condition erreur	2	0,039	0,934	0,432	0,189
		8	(0,042)			
Item 11	Condition erreur	2	0,002	1,273	0,440	0,560
		2	(0,002)			
Item 12	Condition erreur	-	-	-	-	-
Item 13	Condition erreur	2	0,000	0,015	0,985	0,007
		4	(0,019)			
Item 14	Condition erreur	2	0,015	3,112	0,372	0,862
		1	(0,005)			
Item 15	Condition erreur	2	-	-	-	-
		5				
Item 16	Condition erreur	2	0,003	0,223	0,812	0,129
		3	(0,015)			
Item 17	Condition erreur	2	0,001	0,184	0,835	0,039
		9	(0,003)			
Item 18	Condition erreur	2	0,005	0,342	0,718	0,064
		10	(0,014)			
Item 19	Condition erreur	-	-	-	-	-
Item 20	Condition erreur	2	0,020	0,800	0,500	0,242
		5	(0,025)			

Note. Les valeurs entre parenthèses représentent les erreurs moyennes au carré.

*p<0,05

ANNEXE B

Test de signification des valeurs alpha de deux tests indépendants

Hakstian et Whalen (1976) fournissent une méthode pour réaliser un test de signification sur les valeurs du coefficient alpha. Dans le cas de deux échantillons indépendants, la statistique du test est notée M. Elle suit approximativement une distribution de khi carré avec un degré de liberté. Elle permet de tester l'hypothèse nulle selon laquelle les valeurs alpha des deux tests sont semblables.

M est obtenu à l'aide de la formule suivante :
$$M = \frac{J-1}{18J} \left\{ \sum_{k=1}^2 B_k - \frac{\left[\sum_{k=1}^2 B_k (1-r_{\alpha k})^{-1/3} \right]^2}{\sum_{k=1}^2 B_k (1-r_{\alpha k})^{-2/3}} \right\}$$
 dans laquelle

$B_k = (9n_k - 11)^2 / (n_k - 1)$. De plus, J est le nombre d'items (J=20); n est la taille de l'échantillon

($n_1=426$ et $n_2=478$); $r_{\alpha k}$ est la valeur du coefficient alpha estimé ($r_{\alpha 1}=0,659$ et $r_{\alpha 2}=0,709$).

La statistique M du test est égale à 2,67. Elle est inférieure à la valeur critique du khi carré à un degré de liberté qui vaut 3,84. On ne peut donc pas rejeter l'hypothèse nulle.

ANNEXE C

Définitions et exemples d'attributs tirés de la littérature

L'annexe A présente les définitions et exemples d'attributs cognitifs issus de la littérature tels que fournis aux experts lors des trois conditions de l'élaboration individuelle des matrices (phase 1).

✧ Définition 1 : (traduction libre)

Gierl (1997)

Définition d'un attribut : Un attribut est la description de procédures, d'habiletés, de stratégies et de connaissances qu'un sujet doit posséder pour résoudre un test. Les attributs ne sont pas observables directement, ils doivent être inférés à partir des réponses des sujets.

✧ Définition 2 : (traduction libre)

Buck et al. (1998)

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.

Voici maintenant quelques listes d'attributs utilisées dans des recherches :
Ces listes vous sont fournies afin de vous aider à comprendre ce que peut être un attribut. Vous pouvez vous en inspirer afin de fabriquer vos propres listes en lien avec les questions sur lesquelles vous devez travailler.

✧ Liste 1 : (traduction libre)

Milewski (2002)

1. utiliser des concepts et opérations de base pour résoudre des problèmes arithmétiques
2. comprendre la géométrie et les coordonnées en géométrie
3. comprendre les propriétés des nombres et leurs relations
4. utiliser les probabilités, les statistiques de base, les représentations graphiques
5. créer des figures ou des équations pour résoudre des problèmes
6. appliquer des règles en algèbre ou en géométrie
7. relier les thèmes entre eux
8. considérer plusieurs solutions pour résoudre un problème
9. organiser et gérer les informations pour résoudre des problèmes nécessitant plusieurs conditions
10. reconnaître des formes équivalentes
11. raisonner logiquement
12. chercher la solution par essais et erreurs

13. résoudre des problèmes qui semblent peu familiers
14. reconnaître les mots clés
15. utiliser les choix de réponse pour répondre à la question
16. voir qu'un problème ne fournit pas les informations nécessaires pour trouver une seule solution

✧ **Liste 2 : (traduction libre)**

Rupp (2005)

1. se souvenir des définitions et des théorèmes
2. comprendre les relations entre les termes mathématiques
3. faire des opérations arithmétiques
4. faire des opérations non arithmétiques
5. interpréter des graphiques statistiques
6. comprendre un énoncé long ou compliqué
7. visualiser le problème et ses solutions possibles
8. faire une suite d'opérations compliquées pour résoudre un problème
9. établir des liens entre des équations et des expressions
10. regarder un problème complexe sous des angles divers
11. faire des opérations métamathématiques

✧ **Liste 3 : (traduction libre)**

(Hartz, 2002)

1. symboles complexes (exposants, fractions, racines ...) / manipulation numériques
2. géométrie et trigonométrie avec créativité (applications non standard des règles, combinaison originale de plusieurs règles, éventuellement manipulation algébrique préalable)
3. application simple des règles de bases de la géométrie
4. manipulation algébrique en plusieurs conditions
5. problème en mots à traduire en langage mathématique pour la résolution
6. factorisation algébrique ou numérique
7. calculs (du genre à être facilité par l'utilisation de la calculatrice)

❏ Liste 4 : (traduction libre)

Tatsuoka (1996)

1. arithmétique : nombres et opérations (nombres pairs et impairs, nombres premiers, facteurs, fractions, ordre, pourcentages, carré de nombre positifs, racine carrée de petits nombres positifs comme 4, 9, 16..., moyenne de nombres positifs).
2. algèbre élémentaire (travailler avec des variables (addition et soustraction seulement), équations linéaires. Expressions algébriques de degré 1, valeurs absolues, nombres irrationnels comme $\sqrt{3}$)
3. algèbre intermédiaire : travailler avec des équations quadratiques factorisables, expressions algébriques complexes, fonctions, ensembles, probabilités simples, variables avec exposants, statistiques de base (mode, médiane)
4. géométrie et géométrie analytique : travailler avec le périmètre, l'aire et le volume de triangles, cercles, rectangles et autres objets géométriques. En géométrie analytique, travailler avec les points, droites et leurs expressions dans un système de coordonnées.
5. traduire des problèmes *en mots en langage mathématique*
6. intuition mathématique : peut reconnaître un problème mal posé et le corriger. Lorsque plusieurs stratégies sont possibles, être capable de choisir la plus simple ou la plus rapide. Des opérations simples avec des nombres ne sont pas incluses dans cet attribut.
7. application de connaissances : peut se souvenir et interpréter le savoir concernant les concepts de base et les propriétés en arithmétique, algèbre élémentaire, géométrie et géométrie analytique. Peut faire des calculs en arithmétique : (fractions, puissances), en géométrie, avec des nombres relatifs, des valeurs absolues, peut calculer des médianes ou des modes.
8. application de règles et d'algorithmes : peut résoudre des équations et des systèmes d'équations, dériver des expressions algébriques, calculer des expressions algébriques, faire des substitutions de nombres ou de variables dans des expressions algébriques.
9. raisonnement et pensée logique : peut faire un raisonnement de cause à effet, porter un jugement sur le sens des nombres, identifier les conditions nécessaires et suffisantes et les utiliser adéquatement pour résoudre des problèmes, définir si une solution est générale ou un cas particulier. Habileté à raisonner dans l'espace.
10. pensée analytique et structure de problème : peut appliquer des processus mentaux de haut niveau pour résoudre des problèmes. Peut décomposer un problème en plusieurs éléments et les organiser afin de résoudre le problème.
11. travailler avec des figures, tableaux et graphiques : peut comprendre des tableaux, des figures et des graphiques. Dans un problème, peut faire le lien entre l'énoncé et les tableaux, figures ou graphiques. Peut fabriquer des figures pour représenter le problème et faciliter sa résolution.
12. habileté à passer des examens : peut tirer profit de la forme de la question et trouver la solution sans résoudre le problème de la manière attendue. Par exemple, peut utiliser les choix de réponse pour trouver la solution sans faire le raisonnement attendu.
13. degré de complexité : peut résoudre des problèmes qui nécessitent plusieurs conditions, qu'elles soient implicites ou explicites. Peut définir, ordonner et prioriser les conditions par lesquelles passer pour résoudre le problème et résoudre le problème condition par condition.
14. lecture : peut comprendre des questions contenant une négation, une comparaison, les termes *au moins*, *au plus*, *devrait* ou la notion d'augmenter ou de diminuer.
15. attention : peut se souvenir de ce qui est demandé en détail, identifier les contraintes, les pièges, peut lire et comprendre un énoncé long et complexe.
16. peut traduire l'énoncé en mots en énoncé mathématiques dans le cas où les variables et les relations sont explicitées dans le texte et sont simples.

ANNEXE D

Les 106 énoncés d'attributs

L'annexe D présente la liste des 106 attributs utilisée lors de la recherche d'un consensus entre les experts pour le classement des attributs en six catégories (phase 2).

Énoncés des attributs (par ordre alphabétique)	
1	Application de règles et d'algorithmes (peut résoudre des équations, dériver des expressions, ...)
2	Application de règles et/ou d'algorithmes (formule quadratique, uv, u/v de dérivées) incluant la substitution des nombres variables dans des expressions
3	Appliquer des connaissances: définitions et propriétés des fonctions élémentaires, équations de droites, ...
4	Appliquer des connaissances: définitions, propriétés de fonctions élémentaires, relations entre fonctions
5	Appliquer des règles en algèbre ou en géométrie
6	Appliquer des règles et des algorithmes de résolution et de calcul
7	Appliquer des règles ou des algorithmes: résoudre, dériver, substituer des nombres ou des variables
8	Appliquer des règles, des algorithmes: produit scalaire, dérivées, intégration, etc.
9	Calculer efficacement et détecter les erreurs éventuelles
10	Calculer sans erreurs et être attentif aux sources d'erreurs éventuelles
11	Calculs à être facilités par l'utilisation de la calculatrice
12	Calculs pouvant être facilité par une calculatrice
13	Choisir une stratégie
14	Comprendre adéquatement les propriétés de valeur abs. de x et ses liens avec racine carrée et x carré
15	comprendre la géométrie et les coordonnées en géométrie
16	Comprendre la notion de limite et de continuité (et discontinuité) d'une fonction, en particulier dans le voisinage de $x=0$ (/par0)
17	Comprendre le concept de fonction (et la composée) et/ou reconnaître / analyser la représentation graphique de la fonction
18	Comprendre le concept de fonction (trigonométrique, valeur absolue, racine carrée, 1er ou 2ème degré) et les termes s'y rattachant (minimum, règle, image, composée)
19	Comprendre le concept de limite, continuité et de dérivée
20	Comprendre les propriétés des nombres et leurs relations

21	Comprendre les relations entre les termes mathématiques
22	Connaître les propriétés particulières des fonctions exponentielles et logarithmiques (limites, dérivées, valeurs usuelles,...) et leurs liens avec d'autres fonctions
23	Connaître les règles de dérivation des fonctions (dérivée d'une somme, produit, composée, etc...)
24	Créer des figures ou des équations pour résoudre des problèmes
25	Décoder correctement l'écriture mathématique
26	Décoder le langage mathématique et la représentation symbolique s'y rattachant
27	Décoder le langage symbolique et/ou comprendre le sens des termes utilisés en maths tels que limites, fonction continue, dérivable, domaine de définition, etc.
28	Effectuer des opérations sur des expressions algébriques ou numériques contenant des fractions
29	Établir des liens entre différentes notions: dérivées, pente, point critique, accroissement, tangente, etc.
30	Établir des liens: racine-carrée-valeur absolue, dérivabilité-continuité, dérivée-pente,...
31	Factorisations algébriques ou numériques
32	Faire des liens entre différentes notions: dérivée, pente, tangente, etc.
33	Faire des opérations arithmétiques
34	Faire des opérations arithmétiques et algébriques: addition, fonction, exposant, racine carrée, factorisation, inéquation, etc.
35	Faire des substitutions de nombre et/ou de variables dans des expressions algébriques
36	Faire le lien entre les propriétés d'une fonction et son graphique
37	Faire une lecture appropriée, faire des comparaisons (p.e. la plus grande) et/ou structurer les étapes de la résolution d'un problème
38	Fonction définition, notation, graphe, domaine
39	Fonctions élémentaires (trigo, exp, Ln, exposants, valeur absolue, etc.): définitions, propriétés et liens entre elles
40	Habileté à manipuler correctement des termes algébriques en plusieurs conditions
41	La mémorisation des dérivées de certaines fonctions (logarithmiques, exponentielles et trigonométriques)
42	Maîtriser les notions se rapportant aux fonctions (dom, min, max, variation, continuité, dérivabilité, graphique, signal, V.A.) et les liens existants entre elles
43	Maîtriser les opérations sur les fonctions (addition, division, composition,...) ainsi que les règles de dérivation

44	Maîtriser les propriétés des fonctions exponentielles et logarithmiques
45	Maîtriser les propriétés particulières de fonctions polynomiales, de la fonction valeur absolue et racine carrée
46	Manipulation de termes algébriques (ou autres ex: logarithmiques) en plusieurs conditions
47	Manipulation des symboles complexes en algèbre élémentaire et intermédiaire (racine, carré, +, -, sup. inf, valeur absolue)
48	Manipulation numériques et algébriques: équations, inéquations (résoudre), droite pour équation
49	Manipuler des expressions algébriques (réécrire, factoriser, simplifier, développer)
50	Manipuler des expressions algébriques (réécrire, réduire, simplifier, ...)
51	Manipuler des expressions algébriques: addition, soustraction, multiplication, exposant, factorisation, équation, inéquation...
52	Manipuler des termes algébriques (incluant travailler avec des équations quadratiques factorisables et des inéquations)
53	Organiser et gérer les informations pour résoudre des problèmes nécessitant plusieurs étapes
54	Organiser la résolution d'un problème nécessitant plusieurs étapes
55	Passer par une étape intermédiaire pour répondre à la question
56	Pouvoir résoudre des équations et /ou inéquations (1er degré, 2ème degré, factorisables ou pas, trigonométriques, valeurs absolues)
57	Procéder par élimination des choix de réponse
58	Raisonnement logiquement
59	Raisonnement logiquement: ne rien oublier, porter (?) fréquemment sur la ou les réponses
60	Reconnaître des formes équivalentes
61	Reconnaître l'écriture symbolique mathématique
62	Reconnaître les mots clés
63	Reconnaître les symboles mathématiques utilisés
64	Reconnaître quand une fonction est continue
65	Relier les thèmes entre eux
66	Repérer les mots clés dans une question et en comprendre la signification
67	Repérer les mots clés et en comprendre la signification

68	Repérer les mots clés, en comprendre la signification, et établir des liens entre eux
69	Représentation graphique d'une situation problème et faire l'analyse pour faciliter la résolution du problème
70	Représenter graphiquement une fonction et pouvoir décrire ses propriétés
71	Résoudre des équations et/ou des inéquations (1er ou 2eme degré, factorisables ou pas), équations trigonométriques
72	Résoudre des problèmes qui nécessitent plusieurs étapes
73	Se faire confiance
74	se souvenir de la dérivée de certaines fonctions (exponentielles, logarithmiques, trigonométriques)
75	Se souvenir de la dérivée de certaines fonctions (trigo et log), des propriétés des exposants, logarithmes et des algorithmes de dérivées (produit et ou quotient)
76	symboles complexes (exposants, fractions, racines,) / manipulations numériques
77	Techniques de base: mettre en facteur, évaluer, dériver
78	Tenir compte de plusieurs éléments
79	Tirer profit de la formulation de la question et des choix de réponse (procéder par élimination ou une stratégie plus rapide)
80	Tracer un ou plusieurs graphiques pour répondre à la question
81	Travailler avec des figures, graphes, tableaux, lire ou construire des figures
82	Travailler avec des figures, peut fabriquer des figures pour représenter le problème et faciliter sa solution
83	Travailler avec des figures, tableaux, graphiques allant jusqu'à fabriquer des figures
84	Travailler avec des fractions (numériques ou non) et liens avec les exposants
85	Travailler avec les fonctions logarithmiques et exponentielles
86	Travailler sur les fonctions linéaires, équations, propriétés, graphiques, paramètres et leurs rôles
87	Travailler sur les fonctions trigonométriques
88	Utiliser certaines techniques: binôme de Newton, règle de l'Hôpital
89	Utiliser correctement les algorithmes (de dérivée), les lois (de l'Hôpital, des exposants), les propriétés (des logarithmes)
90	Utiliser des algorithmes (de dérivation) et/ou des règles (incluant substituer des nombres et/ou des variables dans des expressions algébriques)

91	Utiliser des concepts et opérations de base pour résoudre des problèmes arithmétiques
92	Utiliser des concepts et opérations de base pour résoudre des problèmes arithmétiques
93	Utiliser des graphes (donnés ou à construire)
94	Utiliser des lois et/ou des propriétés et/ou définitions (exposants et logarithmes)
95	Utiliser les choix de réponse: trouver la solution par examen
96	Utiliser les choix de réponses pour trouver la solution
97	Utiliser les choix de réponses pour trouver la solution (procéder par élimination, et/ou raisonnement logique)
98	Utiliser les propriétés ou des définitions pour résoudre un problème
99	Visualiser le problème et ses solutions possibles
100	(B) Peut tirer profit de la forme de la question et trouver la solution sans résoudre le problème de la façon attendue
101	(B) Introduire une étape intermédiaire adéquate pour répondre à la question (changement de variable, représentation, ...)
102	(B) Structurer les étapes de résolution d'un problème nécessitant plusieurs étapes
103	(B) Tracer des graphiques, des tableaux ou schématiser pour répondre à une question
104	(B) Connaître les règles de dérivation et les appliquer efficacement
105	(B) Travailler sur les fonctions élémentaires ($ x $, x^n , racine $n^{\text{ième}}$ de x , trigonométriques, exponentielles et logarithmiques)
106	(B) Déployer un raisonnement en respectant les règles de la logique

ANNEXE E

Les 106 attributs regroupés en six catégories tels qu'obtenus par consensus entre les experts.

Catégorie 1 : (9 attributs)

Décoder : tout ce qui a trait au fait de lire la question correctement (symboles ss, mots clés)

25	Décoder correctement l'écriture mathématique
26	Décoder le langage mathématique et la représentation symbolique s'y rattachant
27	Décoder le langage symbolique et/ou comprendre le sens des termes utilisés en maths tels que limites, fonction continue, dérivable, domaine de définition, etc.
61	Reconnaître l'écriture symbolique mathématique
62	Reconnaître les mots clés
63	Reconnaître les symboles mathématiques utilisés
66	Repérer les mots clés dans une question et en comprendre la signification
67	Repérer les mots clés et en comprendre la signification
68	Repérer les mots clés, en comprendre la signification, et établir des liens entre eux

Catégorie 2 : (37 attributs)

Appliquer une technique : calcul algébrique, arithmétique, application de règles ou d'algorithmes

1	Application de règles et d'algorithmes (peut résoudre des équations, dériver des expressions, ...)
2	Application de règles et/ou d'algorithmes (formule quadratique, uv, u/v de dérivées) incluant la substitution des nombres variables dans des expressions
5	Appliquer des règles en algèbre ou en géométrie
6	Appliquer des règles et des algorithmes de résolution et de calcul
7	Appliquer des règles ou des algorithmes: résoudre, dériver, substituer des nombres ou des variables
8	Appliquer des règles, des algorithmes: produit scalaire, dérivées, intégration, etc.
9	Calculer efficacement et détecter les erreurs éventuelles
10	Calculer sans erreurs et être attentif aux sources d'erreurs éventuelles
11	Calculs à être facilités par l'utilisation de la calculatrice
12	Calculs pouvant être facilité par une calculatrice
23	Connaître les règles de dérivation des fonctions (dérivée d'une somme, produit, composée, etc...)
28	Effectuer des opérations sur des expressions algébriques ou numériques contenant des fractions
31	Factorisations algébriques ou numériques
33	Faire des opérations arithmétiques

34	Faire des opérations arithmétiques et algébriques: addition, fonction, exposant, racine carrée, factorisation, inéquation, etc.
35	Faire des substitutions de nombres et/ou de variables dans des expressions algébriques
40	Habileté à manipuler correctement des termes algébriques en plusieurs conditions
43	Maîtriser les opérations sur les fonctions (addition, division, composition,...) ainsi que les règles de dérivation
46	Manipulation de termes algébriques (ou autres ex: logarithmiques) en plusieurs conditions
47	Manipulation des symboles complexes en algèbre élémentaire et intermédiaire (racine, carré, +, -, sup. inf, valeur absolue)
48	Manipulation numériques et algébriques: équations, inéquations (résoudre), droite pour équation
49	Manipuler des expressions algébriques (réécrire, factoriser, simplifier, développer)
50	Manipuler des expressions algébriques (réécrire, réduire, simplifier, ...)
51	Manipuler des expressions algébriques: addition, soustraction, multiplication, exposant, factorisation, équation, inéquation...
52	Manipuler des termes algébriques (incluant travailler avec des équations quadratiques factorisables et des inéquations)
56	Pouvoir résoudre des équations et /ou inéquations (1er degré, 2ème degré, factorisable ou pas, trigonométriques, valeurs absolues)
71	Résoudre des équations et/ou des inéquations (1er ou 2ème degré, factorisable ou pas), équations trigonométriques
75	Se souvenir de la dérivée de certaines fonctions (trigo et log), des propriétés des exposants, logarithmes et des algorithmes de dérivées (produit et ou quotient)
76	Symboles complexes (exposants, fractions, racines,) / manipulations numériques
77	Techniques de base: mettre en facteur, évaluer, dériver
84	Travailler avec des fractions (numériques ou non) et liens avec les exposants
88	Utiliser certaines techniques: binôme de Newton, règle de l'Hôpital
89	Utiliser correctement les algorithmes (de dérivée), les lois (de l'Hôpital, des exposants), les propriétés (des logarithmes)
90	Utiliser des algorithmes (de dérivation) et/ou des règles (incluant substituer des nombres et/ou des variables dans des expressions algébriques)
91	Utiliser des concepts et opérations de base pour résoudre des problèmes arithmétiques
92	Utiliser des concepts et opérations de base pour résoudre des problèmes arithmétiques
104	(B) Connaître les règles de dérivation et les appliquer efficacement

Catégorie 3 : (8 attributs)

Faire des liens entre des notions

16	Comprendre la notion de limite et de continuité (et discontinuité) d'une fonction, en particulier dans le voisinage de $x=0$ (/par0)
19	Comprendre le concept de limite, continuité et de dérivée

29	Établir des liens entre différentes notions: dérivées, pente, point critique, accroissement, tangente, etc.
30	Établir des liens: racine-carrée-valeur absolue, dérivabilité-continuité, dérivée-pente,...
32	Faire des liens entre différentes notions: dérivée, pente, tangente, etc
36	Faire le lien entre les propriétés d'une fonction et son graphique
42	Maîtriser les notions se rapportant aux fonctions (dom, min, max, variation, continuité, dérivabilité, graphique, signe, V.A.) et les liens existants entre elles
65	Relier les thèmes entre eux

Catégorie 4 : (9 attributs)

Utiliser une visualisation graphique (graphique, tableau, dessin ou autre pour aider à trouver la réponse)

69	Représentation graphique d'une situation problème et faire l'analyse pour faciliter la résolution du problème
70	Représenter graphiquement une fonction et pouvoir décrire ses propriétés
80	Tracer un ou plusieurs graphiques pour répondre à la question
81	Travailler avec des figures, graphes, tableaux, lire ou construire des figures
82	Travailler avec des figures, peut fabriquer des figures pour représenter le problème et faciliter sa solution
83	Travailler avec des figures, tableaux, graphiques allant jusqu'à fabriquer des figures
93	Utiliser des graphes (donnés ou à construire)
99	Visualiser le problème et ses solutions possibles
103	Tracer des graphiques, des tableaux ou schématiser pour répondre à une question

Catégorie 5 : (18 attributs)

Utiliser des propriétés, définitions, théorèmes, éléments de base et faire des liens entre ces éléments

3	Appliquer des connaissances: définitions et propriétés des fonctions élémentaires, équations de droites, ...
4	Appliquer des connaissances: définitions, propriétés de fonctions élémentaires, relations entre fonctions
14	Comprendre adéquatement les propriétés de valeur abs. de x et ses liens avec racine carrée et x carré
17	Comprendre le concept de fonction (et la composée) et/ou reconnaître / analyser la représentation graphique de la fonction
18	Comprendre le concept de fonction (trigonométrie, valeur absolue, racine carrée, 1er ou 2ème degré) et les termes s'y rattachant (minimum, règle, image, composée)
22	Connaître les propriétés particulières des fonctions exponentielles et logarithmiques (limites, dérivées, valeurs usuelles,...) et leurs liens avec d'autres fonctions
38	Fonction définition, notation, graphe, domaine

39	Fonctions élémentaires (trigo, exp, Ln, exposants, valeur absolue, etc.): définitions, propriétés et liens entre elles
41	La mémorisation des dérivées de certaines fonctions (logarithmiques, exponentielles et trigonométriques)
44	Maîtriser les propriétés des fonctions exponentielles et logarithmiques
45	Maîtriser les propriétés particulières de fonctions polynomiales, de la fonction valeur absolue et racine carrée
74	Se souvenir de la dérivée de certaines fonctions (exponentielles, logarithmiques, trigonométriques)
85	Travailler avec les fonctions logarithmiques et exponentielles
86	Travailler sur les fonctions linéaires, équations, propriétés, graphiques, paramètres et leurs rôles
87	Travailler sur les fonctions trigonométriques
94	Utiliser des lois et/ou des propriétés et/ou définitions (exposants et logarithmes)
98	Utiliser les propriétés ou des définitions pour résoudre un problème
105	(B) Travailler sur les fonctions élémentaires ($ x $, x^n , racine nième de x , trigonométriques, exponentielles et logarithmiques)

Catégorie 6 : (8 attributs)

Organiser la solution : plusieurs étapes, stratégie

13	Choisir une stratégie
37	Faire une lecture appropriée, faire des comparaisons (p.e. la plus grande) et/ou structurer les étapes de la résolution d'un problème
53	Organiser et gérer les informations pour résoudre des problèmes nécessitant plusieurs étapes
54	Organiser la résolution d'un problème nécessitant plusieurs étapes
55	Passer par une étape intermédiaire pour répondre à la question
72	Résoudre des problèmes qui nécessitent plusieurs étapes
101	(B) Introduire une étape intermédiaire adéquate pour répondre à la question (changement de variable, représentation, ...)
102	(B) Structurer les étapes de résolution d'un problème nécessitant plusieurs étapes

12 attributs éliminés par consensus

15	Comprendre la géométrie et les coordonnées en géométrie
57	Procéder par élimination des choix de réponse
58	Raisonnement logique
59	Raisonnement logique: ne rien oublier, porter attention fréquemment sur la ou les réponses
73	Se faire confiance
78	Tenir compte de plusieurs éléments

79	Tirer profit de la formulation de la question et des choix de réponse (procéder par élimination ou une stratégie plus rapide)
95	Utiliser les choix de réponse: trouver la solution par examen
96	Utiliser les choix de réponses pour trouver la solution
97	Utiliser les choix de réponses pour trouver la solution (procéder par élimination, et/ou raisonnement logique)
100	(B) Peut tirer profit de la forme de la question et trouver la solution sans résoudre le problème de la façon attendue
106	(B) Déployer un raisonnement en respectant les règles de la logique

5 Attributs supprimés par manque de consensus

21	Comprendre les relations entre les termes mathématiques
60	Reconnaître des formes équivalentes
64	Reconnaître quand une fonction est continue
20	Comprendre les propriétés des nombres et leurs relations
24	Créer des figures ou des équations pour résoudre des problèmes

ANNEXE F

La description des analyses factorielles exploratoires et confirmatoires dont les résultats sont fournis aux experts lors de la condition 2

Les indices d'ajustement RMSEA (*root mean squared error of approximation*) et SRMR (*standardized root mean squared residual*) sont utilisés afin de définir le modèle le plus adéquat dans les analyses factorielles exploratoires. Hu et Bentler (1999) suggèrent de tenir compte de deux coefficients pour s'assurer de l'ajustement des modèles. Selon eux, une valeur du coefficient RMSEA inférieure à 0,05 et une valeur du coefficient SRMR inférieure à 0,06 démontrent un ajustement satisfaisant. Marsh et ses collègues (2005) précisent que l'ajustement est convenable pour une valeur du RMSEA inférieure à 0,08 mais très bon en dessous de 0,05.

Comme le montre le tableau 3, les valeurs des coefficients RMSEA sont toujours largement inférieures à 0,05 et ne sont parfois pas estimées lorsqu'elles deviennent de plus en plus petites, donc inférieure à 0,05. Les valeurs du coefficient SRMR suggèrent que 4 ou 5 facteurs sont présents dans le test A et que 4, 5 ou 6 facteurs sont présents dans le test B. Selon Fabrigar et ses collègues (1999) plusieurs considérations sont importantes lorsque vient le temps de décider du nombre de facteurs à retenir ainsi que de leur composition pour établir le modèle de l'analyse factorielle confirmatoire qui va suivre. Tout d'abord, ils suggèrent d'examiner avec attention les différentes solutions après rotation orthogonale ou oblique afin de définir une solution plausible. Ils suggèrent également de privilégier une rotation oblique (p.e., *promax*) puisqu'elle permet une corrélation entre les facteurs. L'approche utilisée ici est un peu intuitive mais va dans le sens de ce qui est suggéré par Fabrigar et ses collègues.

C'est à l'aide des coefficients de saturation obtenus après une rotation oblique *promax* que les différentes solutions sont étudiées en gardant à l'esprit que les modèles d'analyses factorielles

exploratoires à plus d'un facteur n'ont pas de solution unique et qu'il faut en choisir une (Fabrigar et al., 1999). En outre la théorie n'est pas d'un grand secours lorsque vient le temps de s'assurer de la pertinence des regroupements qui sont faits ici. Il est à noter que dans cette recherche, c'est plutôt le travail théorique que les experts vont faire sur la base des regroupements qui est source d'intérêt. Les modèles d'analyses factorielles confirmatoires sont élaborées à partir des items pour lesquels les coefficients de saturation *promax* sont les plus grands dans les différentes solutions obtenues. Ainsi un premier modèle d'analyse confirmatoire à 4 facteurs est appliqué aux données du test A et un premier modèle à 6 facteurs est appliqué aux données du test B.

Les indices d'ajustement choisis sont alors les statistiques CFI (*comparative fit index*) et TLI (*Tucker-Lewis index*). Hu et Bentler (1999) suggèrent que l'ajustement est satisfaisant dès que ces indices sont tous les deux supérieurs à 0,95 et que les valeurs des coefficients RMSEA sont inférieures à 0,06. À l'aide d'études de simulation, Yu et Muthén (2002) ont montré que ces valeurs s'appliquent également dans le cas de variables catégorielles. Les deux analyses factorielles confirmatoires respectent parfaitement ces valeurs limites et présentent donc un ajustement convenable. De plus les coefficients de saturation obtenus sont tous significatifs au seuil de 0,05. Toutefois ces deux modèles sont ensuite légèrement modifiés sur la base de considérations différentes pour chacun des deux tests. Dans les modèles, chacun des items est associé à un coefficient de détermination R^2 . Ce dernier exprime le pourcentage de la variance des variables qui est expliqué par le modèle. En ce qui concerne le test A, deux items sont associés à un coefficient R^2 inférieur à 10%. Cela signifie donc que ce modèle explique 1,5% de la variance de l'un (item 6) et 4,5% de l'autre (item 17) alors que le modèle explique entre 15 et 65% de la variance des autres items. Cette remarque suggère que ces deux items n'ont peut-être pas leur place dans le modèle. Après essais, le modèle finalement retenu est celui dans lequel l'item 6 est enlevé mais pas l'item 17 car alors l'ajustement est moins bon. La solution finale se trouve dans le tableau 4. En ce qui

concerne le test B, le problème ne se pose pas puisque chacun des items a un coefficient de détermination d'au moins 20%. Mais une autre question concerne l'un des items (item 16) qui présente un coefficient de saturation élevé relativement à deux facteurs dans les analyses exploratoires et qui a été placé, un peu arbitrairement, dans un seul facteur dans l'analyse confirmatoire. Une deuxième analyse confirmatoire est menée dans laquelle cet item est placé dans les deux facteurs suggérés par l'exploration des données. Il est intéressant de noter que l'ajustement de ce modèle est meilleur que celui du précédent. Vu l'usage qui doit être fait des regroupements d'items dans cette recherche, la dernière décision prise ici est de garder ce dernier modèle. Le tableau 3 présente les indices d'ajustement obtenus pour les analyses factorielles exploratoires et confirmatoires retenues pour les tests A et B.

Tableau F1.

Les indices d'ajustement des analyses factorielles exploratoires et confirmatoires pour les tests A et B

facteurs	Test A					Test B				
	EFA RMSEA	SRMR	CFI	CFA TLI	RMSEA	EFA RMSEA	SRMR	CFI	CFA TLI	RMSEA
1	0,022	0,0839				0,034	0,0899			
2	0,011	0,0741				0,009	0,0687			
3	-	0,0640				-	0,0608			
4	-	0,0539	0,955	0,951	0,028	-	0,0545			
5	-	0,0454				-	0,0490			
6	impossible					-	0,0406	0,972	0,974	0,025

Note. Les valeurs en caractères gras indiquent un bon ajustement au sens de Hu et Bentler (1999)

De plus, une vérification de la position de l'item dans les tests d'origine et des distributions des scores pour les items de chaque regroupement est faite afin de s'assurer qu'aucun de ces regroupements n'est dû au fait que les items présentent beaucoup de valeurs manquantes. Les résultats font ensuite l'objet d'un rapport qui est fourni aux experts au début de la condition 2. Le livret de la condition 2 contient le tableau 4, les consignes et les items regroupés selon le tableau 4. Il est à noter que l'item 16 apparaît à deux reprises dans les regroupements du test B et que l'attention

des experts est attirée sur ce fait.

Tableau F2.

Le rapport des analyses factorielles tel que fourni aux experts lors de la condition 2

**Regroupements de questions à considérer
pour établir la liste des attributs et les relier aux questions**

Note : les attributs doivent être élaborés et reliés aux questions en tenant compte de toutes les questions

Test A

Question 12	Question 7	Question 2	Question 14
Question 19	Question 9	Question 8	Question 17
	Question 10	Question 11	Question 20
		Question 15	
		Question 18	

Test B

Question 1	Question 2	Question 18	Question 8	Question 4	Question 3
Question 5	Question 9	Question 20	Question 12	Question 15	Question 6
Question 7	Question 10		Question 16		
	Question 11		Question 19		
	Question 16				

ANNEXE G

Un exemple de la documentation fournie aux experts à la condition 3.

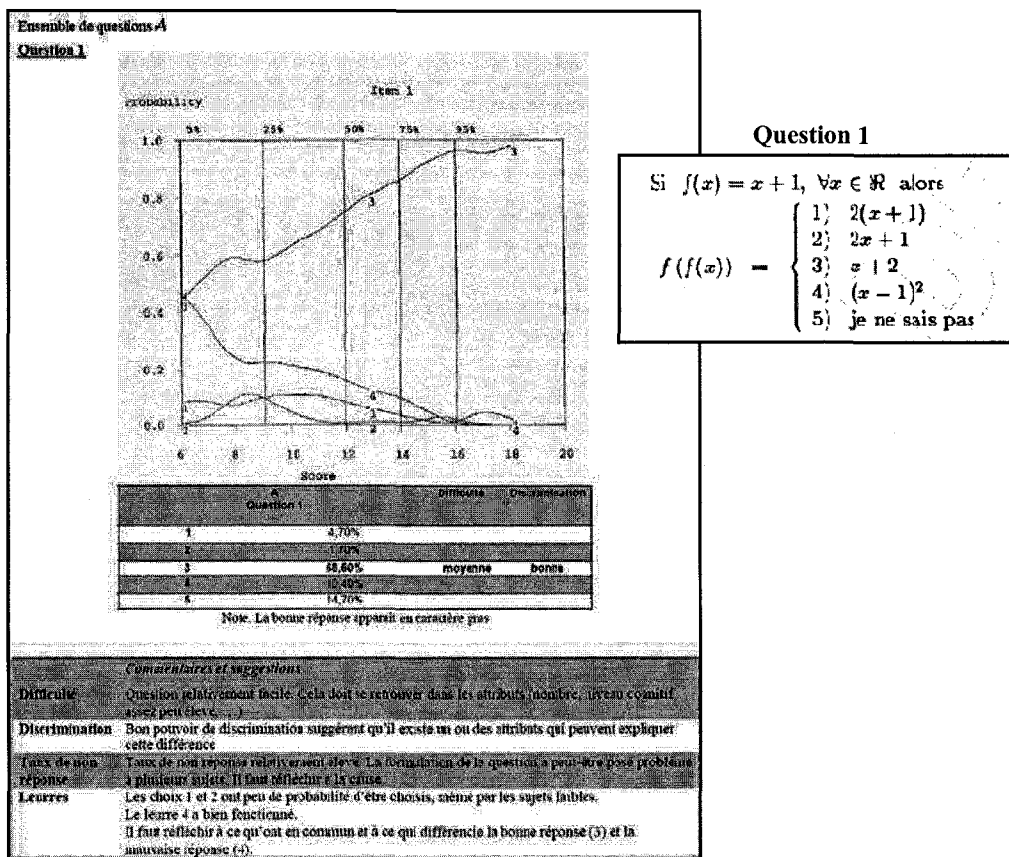


Figure G1: Informations fournies aux experts dans le livret de la condition 3 relativement à l'item 1 du test A

ANNEXE H

Les équations du modèle RUM

Soient les paramètres $\pi_{ik} = P(Y_{ijk} = 1 | \alpha_{jk} = 1)$ et $r_{ik} = P(Y_{ijk} = 1 | \alpha_{jk} = 0)$ dans lesquels Y_{ijk} fait référence à un événement 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 . 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 de la matrice Q . De plus, les applications correctes ou non des différents attributs sont supposées indépendantes pour un couple (α, θ) donné.

$$P(X_{ij} = 1 | \underline{\alpha}_j, \theta_j) = \pi_i^* \prod_{k=1}^K r_{ik}^{(1-\alpha_{jk}) \times q_{ik}} P_{c_i}(\theta_j) \quad (1)$$

dans laquelle :

$$P_{c_i}(\theta_j) = \text{logistic}(-1, 7(\theta_j - c_i))$$

Avec,

- $\pi_i^* = \prod_{k=1}^M \pi_{ik}^{q_{ik}} = P(\text{appliquer convenablement tous les attributs requis par l'item } i \text{ sachant qu'ils sont maîtrisés}).$ Ce paramètre correspond à la difficulté de l'item i relativement aux M attributs reliés à l'item i dans la matrice Q .
- $r_{ik}^* = \frac{r_{ik}}{\pi_{ik}} = \frac{P(Y_{ijk} = 1 | \alpha_{jk} = 0)}{P(Y_{ijk} = 1 | \alpha_{jk} = 1)}$ est la pénalité correspondant au fait de ne pas maîtriser l'attribut k .
- $P_{c_i}(\theta_j) = P(\text{convenablement appliquer les attributs reliés à l'item } i \text{ conditionnellement à } \theta_j \text{ qui est l'habileté résiduelle du sujet } j).$ Ce terme est un modèle de Rasch pour lequel la difficulté est le paramètre $-c_i$.

- c_i est un indice permettant de juger si la matrice Q ne contient pas tous les attributs pertinents (*completeness*).

Tableau I2

Les valeurs de r pour le test B et les 3 conditions

Test B	Cond.	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	Item 17	Item 18	Item 19	Item 20
1-Décodage	1		0,65															0,71			
	2																				
	3	0,75	0,56	0,67	0,62		0,33	0,6	0,27	0,59	0,81	0,8	0,28	0,43		0,88	0,41	0,79	0,79	0,37	0,9
2-Technique	1		0,56	0,68	0,58		0,43	0,63	0,25				0,28	0,38	0,87	0,88	0,53	0,82	0,71	0,37	
	2		0,42	0,51	0,63		0,24	0,65				0,75	0,88	0,55	0,8		0,45	0,78	0,85		0,7
	3		0,34	0,53	0,84		0,23	0,69			0,27	0,76	0,89	0,7			0,48	0,74	0,88		0,82
3-Liens notions	1		0,5		0,48					0,67								0,76			
	2	0,66	0,46							0,33	0,29							0,86	0,78		
	3		0,44		0,57					0,45	0,51							0,8	0,8		
4-Rep. Graph.	1	0,5	0,35	0,56	0,78		0,42	0,65		0,43	0,36	0,71	0,88	0,63				0,8	0,65		0,68
	2	0,61								0,22	0,38							0,7	0,77		
	3	0,39								0,44	0,79							0,82			
5-Prop, déf, et liens	1						0,43			0,55	0,4				0,77		0,24				0,73
	2	0,8	0,61		0,6		0,53	0,65	0,26		0,77	0,83	0,27	0,48		0,88	0,48	0,82	0,81	0,37	
	3			0,68							0,64			0,45				0,52			0,58
6-Organisation	1																				
	2						0,67												0,54		0,41
	3									0,53					0,62			0,83	0,71		0,59

L'attribut 1 fait l'objet de la figure I1. Il concerne tout ce qui a trait au fait de lire la question correctement, qu'il s'agisse de comprendre les mots clés, le langage ou encore les symboles mathématiques. La figure I1 montre que les matrices de la condition 1 relient l'attribut 1 à seulement quatre items du test A et deux du test B alors qu'à la condition 3 de nombreux items lui sont reliés. En particulier, les items pour lesquels les experts ont changé entre la condition 1 et 3 ou plus souvent sont presque tous reliés à cet attribut à la condition 3. Cet attribut n'apparaît pas dans les matrices de la condition 2.

Les quatre items du test A et les deux du test B qui sont présents à la condition 1 se retrouvent tous à la condition 3. Les valeurs de r des deux conditions sont semblables pour les items 1 et 4 du test A et les items 2 et 17 du test B. Par contre la structure de la matrice de la condition 3 fait beaucoup augmenter les valeurs pour les items 13 et 16 du test A. Globalement pour le test A, trois des quatre valeurs de la condition 1 sont inférieures à 0,6 alors que toutes les valeurs de la condition 3 sont supérieures à 0,6. Globalement pour le test B, les deux valeurs de r de la condition 1 sont supérieures à 0,6 alors que huit valeurs de la condition 3 sont inférieures à 0,6. Parmi celles-ci, quatre sont plus petites que 0,4. Pour le test A, aucun des liens identifiés à la seule condition 3 ne présente de valeur de r inférieure à 0,6, mais pour le test B, sept des seize nouvelles valeurs sont inférieures à 0,6 et quatre d'entre elles le sont à 0,4.

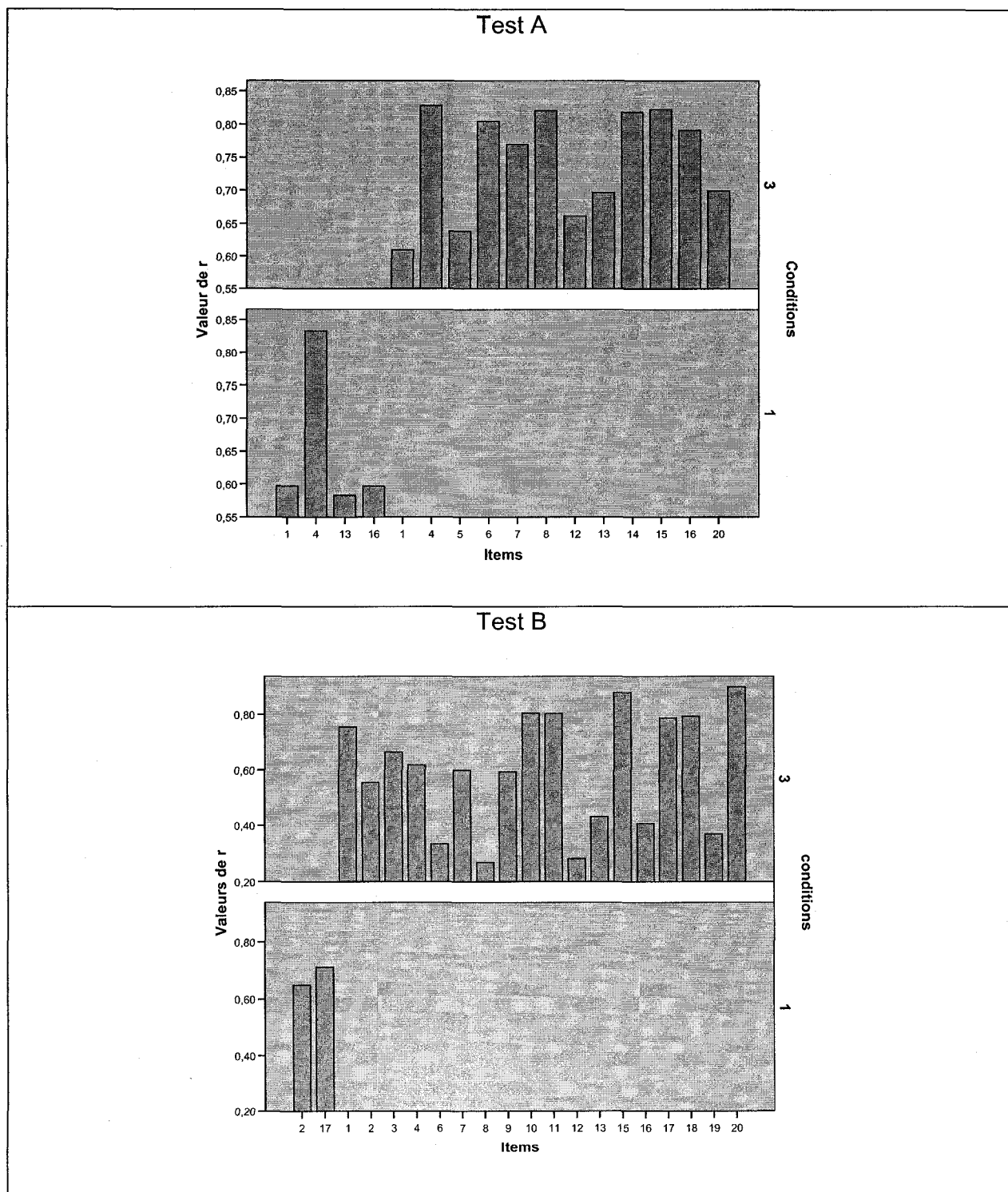


Figure II. Les valeurs r pour l'attribut 1 selon les conditions et les tests A et B

L'attribut 2 fait l'objet de la figure I2. L'attribut 2 concerne la dimension technique de résolution des items comme le calcul algébrique ou arithmétique ou l'application de règles ou d'algorithmes. Les descriptions des matrices ont montré que cet attribut était non seulement relié à la plupart des items des deux tests aux trois conditions, mais que la condition 3 avait amené les experts à représenter cette dimension avec plusieurs intitulés particulièrement détaillés à la condition 3.

Les experts sont peu influencés par les conditions en ce qui a trait à l'aspect purement technique des questions. En effet la figure I2 montre que quatorze items du test A (items 2, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 18 et 19) et dix items du test B (items 2, 3, 4, 6, 7, 12, 13, 16, 17 et 18) sont reliés à l'attribut 2 à chacune des trois conditions. Sur les quatorze comparaisons possibles pour le test A, six présentent une variation de plus de 0,1. Quatre de ces six valeurs de r sont favorables à la condition 1 dans le sens d'une valeur plus petite à cette condition. Sur les dix comparaisons possibles pour le test B, huit présentent une variation de plus de 0,1 dont quatre sont favorables à la condition 1 avec des valeurs beaucoup plus petites pour deux items (items 12 et 13). Les quatre autres donnent l'avantage à part égale aux deux autres conditions.

De façon globale, le nombre de valeurs de r inférieures à 0,6 est supérieur pour les deux tests à la condition 1. En effet, pour le test A sept valeurs sont inférieures à 0,6 à la condition 1 contre six et cinq aux deux conditions suivantes. Pour le test B, huit valeurs sont inférieures à 0,6 contre cinq pour chacune des deux autres conditions. D'autre part, les items 8 et 19 du test B sont reliés à l'attribut 2 à la seule condition 1 avec des valeurs de r inférieures à 0,4. Ainsi, les conditions 2 et 3 perdent de vue ces deux liens de qualité, mais la condition 3 est la seule à identifier un lien avec l'item 10 dont la valeur de r est inférieure à 0,3.

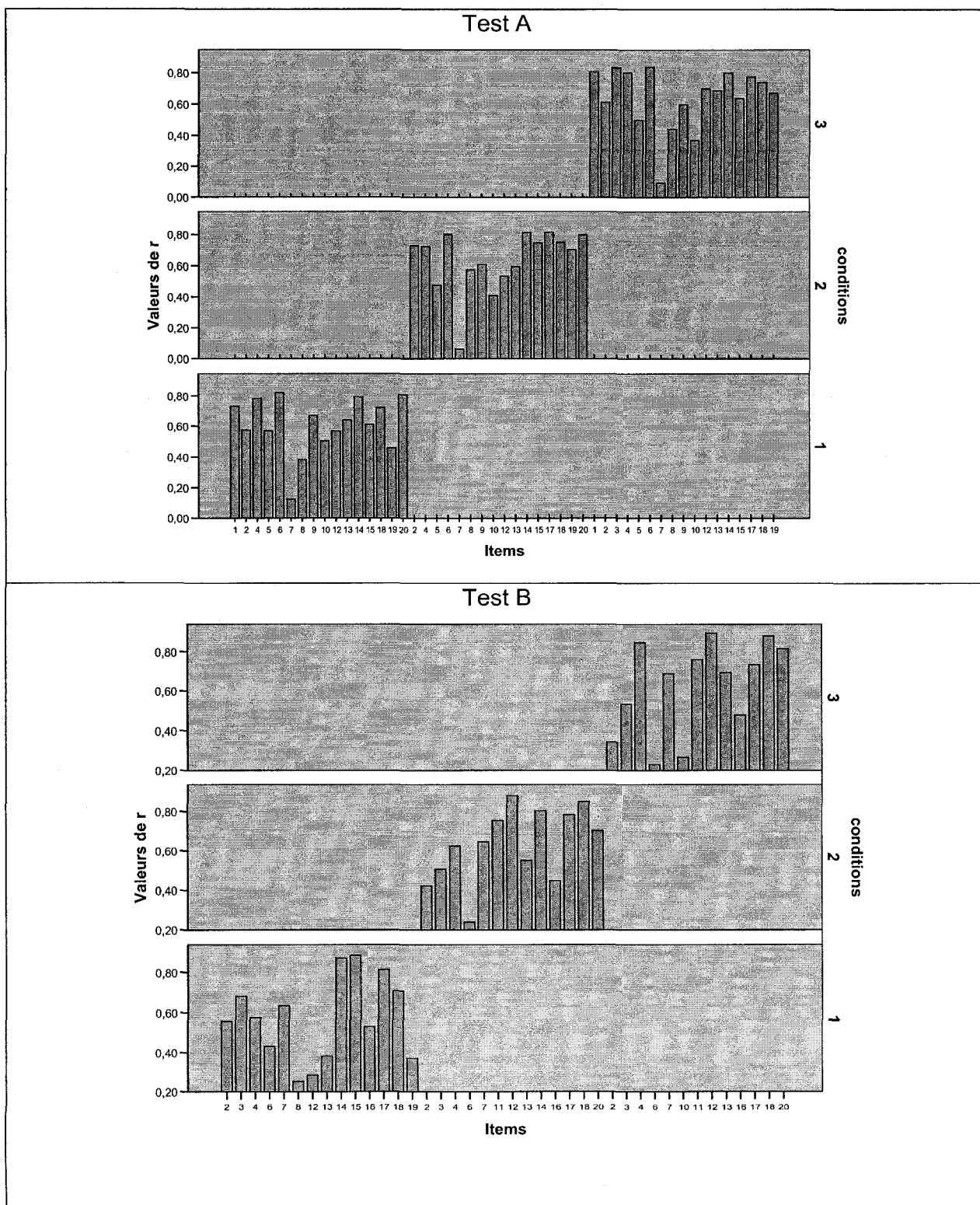


Figure I2. Les valeurs r pour l'attribut 2 selon les conditions et les tests A et B

L'attribut 3 fait l'objet de la figure I3. L'attribut 3 consiste à faire des liens entre des notions. Cet attribut a été l'objet de nombreuses discussions quant à son contenu exact lors de la phase 2 du MACB. De ces discussions ressortait que les experts voyaient cet attribut comme étant d'un assez haut niveau cognitif. En effet, il implique de faire des liens entre des notions qui sont des abstractions contrairement à l'attribut 5 dans lequel les liens sont faits entre des objets concrets tels des définitions ou des notions de base.

Au contraire de l'attribut 2, les conditions ont beaucoup influencé les experts lorsqu'ils ont relié les items à l'attribut 3. En effet, un seul item du test A (item 13) et trois items du test B (items 2, 9 et 17) ont été reliés à cet attribut aux trois conditions. En ce qui concerne le test A, c'est d'ailleurs le seul item qui subsiste à la condition 3 avec une valeur de r inférieure à 0,5 et plus petite qu'aux deux autres conditions. Toutefois, un seul lien existant dans la matrice avec cet attribut, ce résultat est sujet à réserve. La valeur de r pour cet attribut est inférieure à 0,6 à la condition 2 alors qu'elle est supérieure à 0,6 à la condition 1. Dans le cas du test B, la valeur de r reliée à l'item 2 varie assez peu entre les trois conditions. Elle est la plus proche de 0,4 à la condition 3. Pour l'item 9, c'est la condition 2 qui fournit la plus petite valeur inférieure à 0,4. Enfin, l'item 17 a des valeurs de r toujours supérieures à 0,6. La condition 1 est seule à identifier un lien entre l'item 5 et l'attribut 3 avec une valeur de r de l'ordre de 0,5.

De manière plus globale, une seule valeur de r est inférieure à 0,6 pour chacune des trois conditions pour le test A. Par contre, le test B présente deux valeurs inférieures à 0,6 à la condition 1 contre 3 à la condition 2 (dont deux sont plus petites que 0,4) et quatre à la condition 3. La condition 1 est la seule à établir un lien entre l'attribut 3 et l'item r avec une valeur de r inférieure à 0,6.

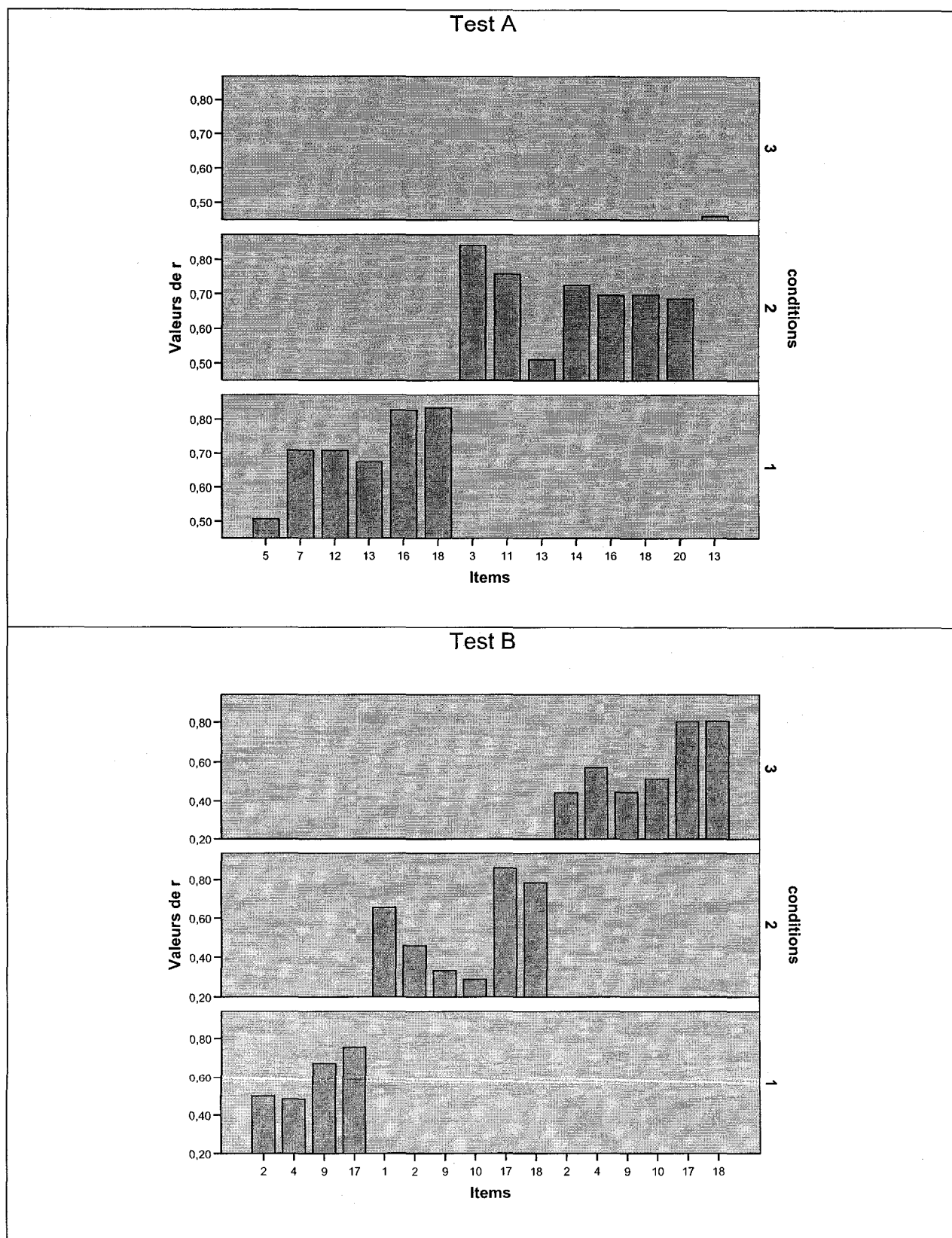


Figure I3. Les valeurs r pour l'attribut 3 selon les conditions et les tests A et B

L'attribut 4 fait l'objet de la figure I4. L'attribut 4 concerne tout ce qui a trait à une représentation graphique pour trouver la réponse à la question. Il n'est pas présent dans la matrice de la condition 1 du test A. Pour ce dernier, trois items (items 3, 6 et 16) ont un lien avec l'attribut 4 aux conditions 2 et 3. Deux de ces trois valeurs sont plus petites à la condition 2, mais avec une différence de moins de 0,1. La seule différence de plus de 0,1 donne l'avantage à la condition 3 pour l'item 3. Pour le test B, deux items (items 9 et 10) ont un lien avec cet attribut aux trois conditions. Pour ces deux items, la plus petite valeur est obtenue à la condition 2. De plus, l'item 1 est relié à l'attribut 4 aux conditions 2 et 3. La condition 3 fournit une valeur plus petite pour cet item. Aucun lien spécifique à une condition n'a de valeur inférieure à 0,6, ni pour A, ni pour B.

Globalement, aucune valeur n'est inférieure à 0,6 pour le test A. En ce qui concerne le test B, deux valeurs sont inférieures à 0,6 à la condition 1, deux valeurs sont inférieures à 0,4 à la condition 2 et deux valeurs sont inférieures à 0,6 à condition 3, dont l'une est plus petite que 0,4.

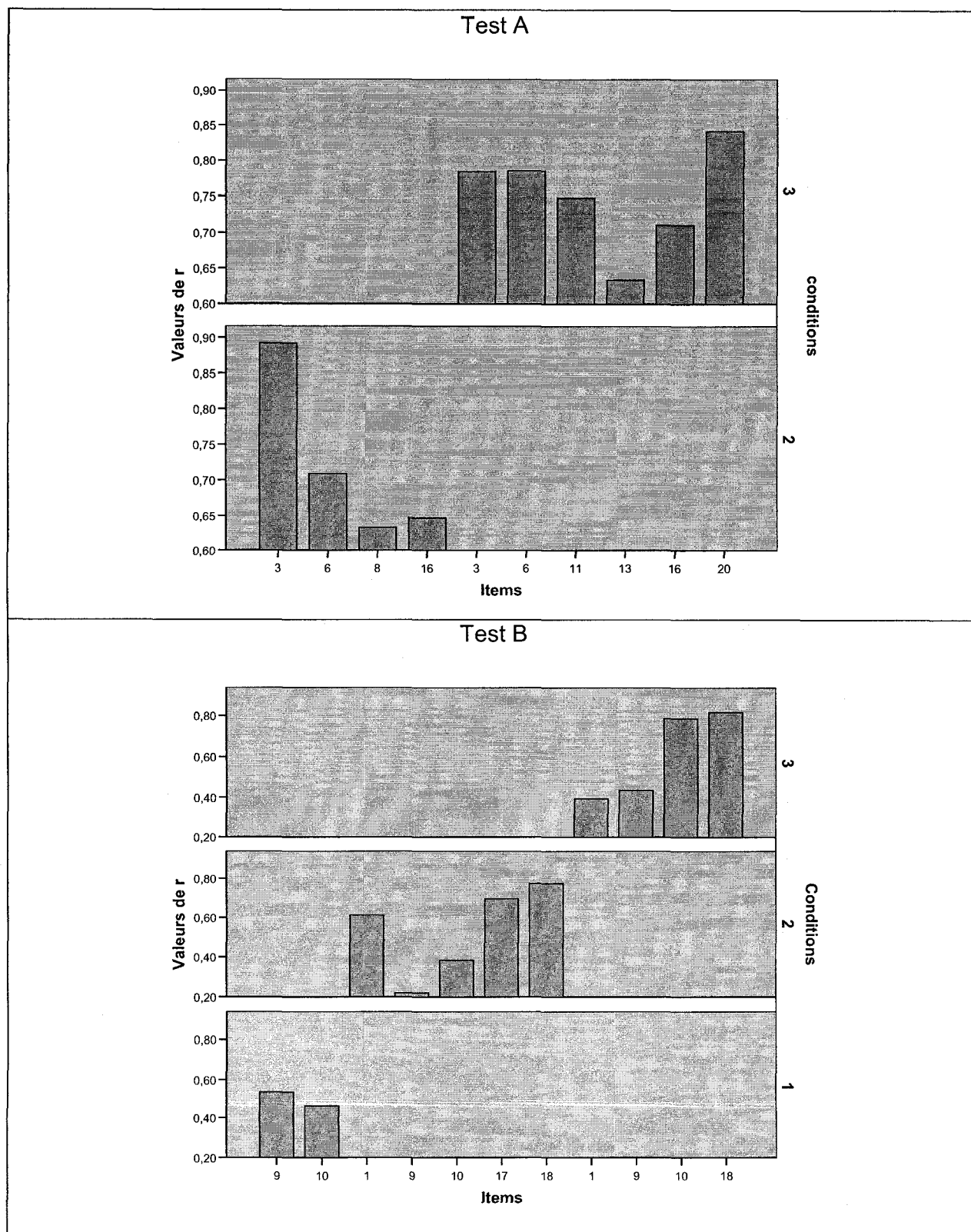


Figure 14. Les valeurs r pour l'attribut 4 selon les conditions et les tests A et B

L'attribut 5 fait l'objet de la figure I5. L'attribut 5 vise l'utilisation des propriétés, des définitions, des théorèmes ainsi que les liens que l'on doit faire entre ces éléments de base pour répondre à la question. Cet attribut est celui, avec l'attribut 3, qui a le plus posé de problème dans sa définition lors de la phase 2 du MACB et occasionné des discussions.

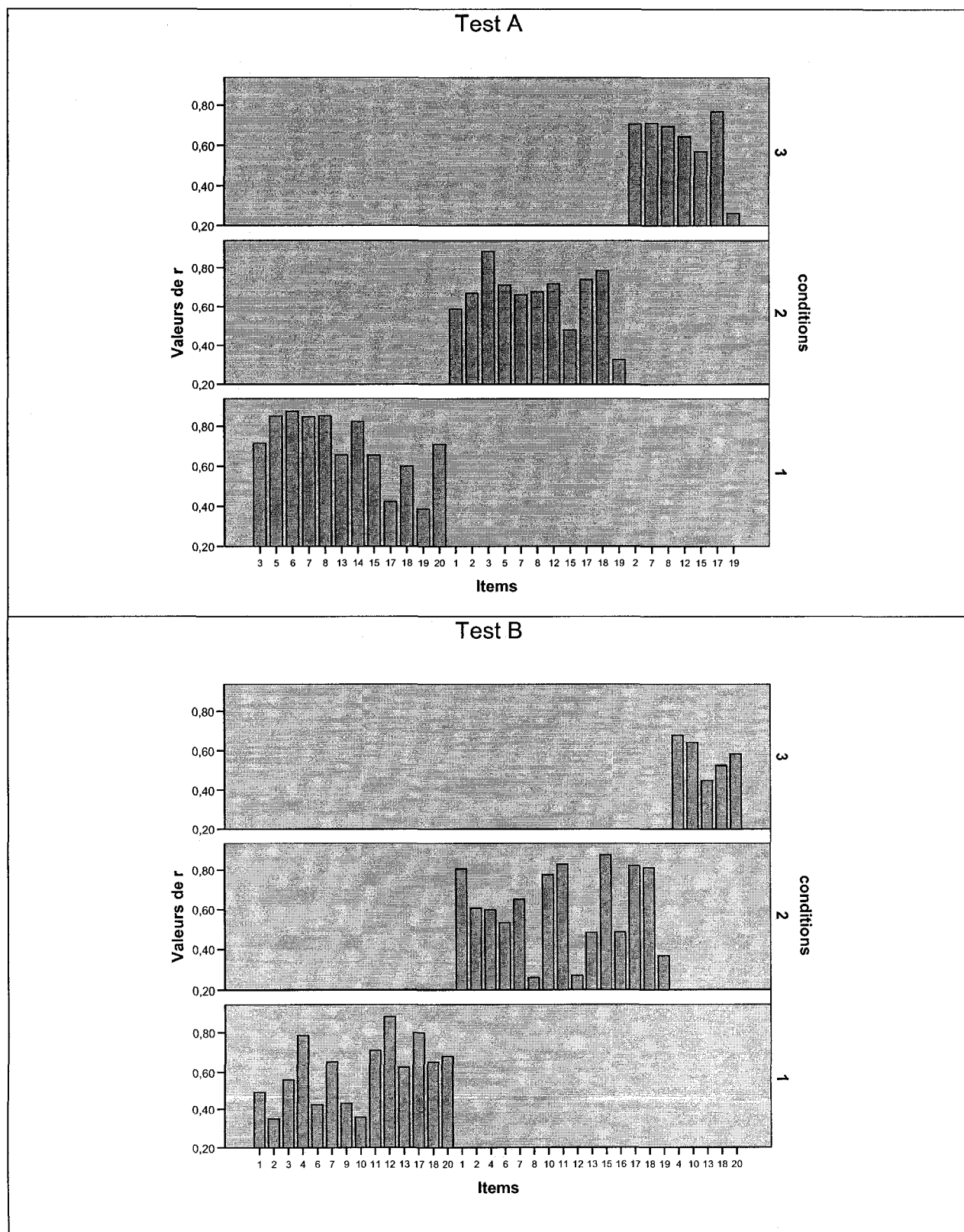
Cinq items du test A (items 7, 8, 15, 17 et 19) et quatre items du test B (items 4, 10, 13 et 18) sont reliés à cet attribut aux trois conditions. Pour les items 7 et 8 du test A, les valeurs de r sont toujours plus grandes que 0,6. Pour l'item 15 du test A, la valeur la plus petite est proche de 0,4 pour la condition 1. Pour l'item 19 du test A, les valeurs sont toujours inférieures à 0,4, mais elle est la plus petite à la condition 3. Enfin pour l'ensemble des huit comparaisons possibles présentant un écart d'au moins 0,1, la condition 1 fournit la plus petite valeur dans trois cas et la condition 2 dans quatre cas. Pour le test B, la condition 2 fournit la plus petite valeur r pour l'item 4, l'item 12 et l'item 13. Cette valeur est d'environ 0,6 pour l'item 4. Pour l'item 10, c'est la condition 1 qui fournit une valeur de r inférieure à 0,4 alors que les deux autres conditions ont des valeurs plus grandes que 0,6, alors que la condition 3 joue ce rôle pour l'item 18. Toutes ces comparaisons correspondent à une variation d'au moins 0,1.

La condition 3 a amené les experts à minimiser le nombre de liens avec l'attribut 5 pour les deux tests. Ainsi trois items du test A (items 3, 5, 18) et sept items du test B (items 1, 2, 6, 7, 11, 12, 17) lui sont reliés aux conditions 1 et 2, mais plus à la condition 3. Sur ces dix items, six ont une valeur inférieure à la condition 1, alors que la condition 2 est plus favorable à deux de ces items, particulièrement pour l'item 12 du test B qui passe d'une valeur supérieure à 0,8 à une valeur inférieure à 0,3. Pour les deux autres items, il y a équivalence entre les valeurs obtenues aux deux conditions. La condition 2 est la seule à établir un lien entre l'attribut 5 et l'item 1 pour une valeur de l'ordre de 0,6. En prenant en considération l'ensemble des comparaisons possibles deux à deux, sur

les neuf présentant une variation d'au moins 0,1, cinq favorisent la condition 1 avec des valeurs de r plus petites.

Pour le test A, la condition 2 est la seule à identifier un lien entre l'attribut 5 et l'item 19 avec une valeur de r inférieure à 0,4. Pour le test B, chaque condition identifie des liens spécifiques avec des valeurs de r inférieures à 0,4 ou à 0,6. La condition 2 est celle qui en fournit le plus avec trois liens inférieurs à 0,6 dont deux sont plus petits que 0,4.

Globalement, trois valeurs sont inférieures à 0,6, dont une l'est à 0,4, pour les conditions 1 et 2 et le test A. La condition 3 n'offre que 2 valeurs inférieures à 0,6 dont une est inférieure à 0,4 pour ce même test. En ce qui concerne le test B, six valeurs sont inférieures à 0,6 dont deux sont inférieures à 0,4 à la condition 1; sept sont inférieures à 0,6 dont 3 sont inférieures à 0,4 pour le test B à la condition 2 et deux valeurs sont inférieures à 0,6 pour la condition 3.



Enfin le dernier et sixième attribut concerne l'organisation de la réponse comme le fait de procéder en plusieurs étapes ou encore celui d'utiliser une stratégie. L'attribut 6 fait l'objet de la figure 30. Les conditions ont beaucoup influencé les experts puisqu'aucun item n'a été relié à cet attribut aux trois conditions pour le test A. Les items 3 et 7 se retrouvent dans les conditions 1 et 2, mais la différence entre les valeurs de r est inférieure à 0,1. Un seul item du test B (item 20) est relié à l'attribut 6 aux trois conditions. Pour cet item, la condition 2 fournit la plus petite valeur qui est de l'ordre de 0,4. Sur les autres comparaisons possibles, trois fournissent un écart de plus de 0,1 entre les valeurs de r et chacune des trois conditions fournit l'une des plus petites valeurs. La condition 1 est seule à identifier deux liens avec l'attribut 6 (items 10 et 16) pour lesquels la valeur de r est inférieure à 0,4.

Globalement, aucune valeur de r n'est inférieure à 0,6 pour le test A. La condition 1 fournit le plus de valeur de r inférieures à 0,6 pour le test B, deux de ces valeurs sont inférieures à 0,4. Aux conditions 2 et 3, deux valeurs sont à chaque fois inférieures à 0,6. L'une des deux est de l'ordre de 0,4 à la condition 2.

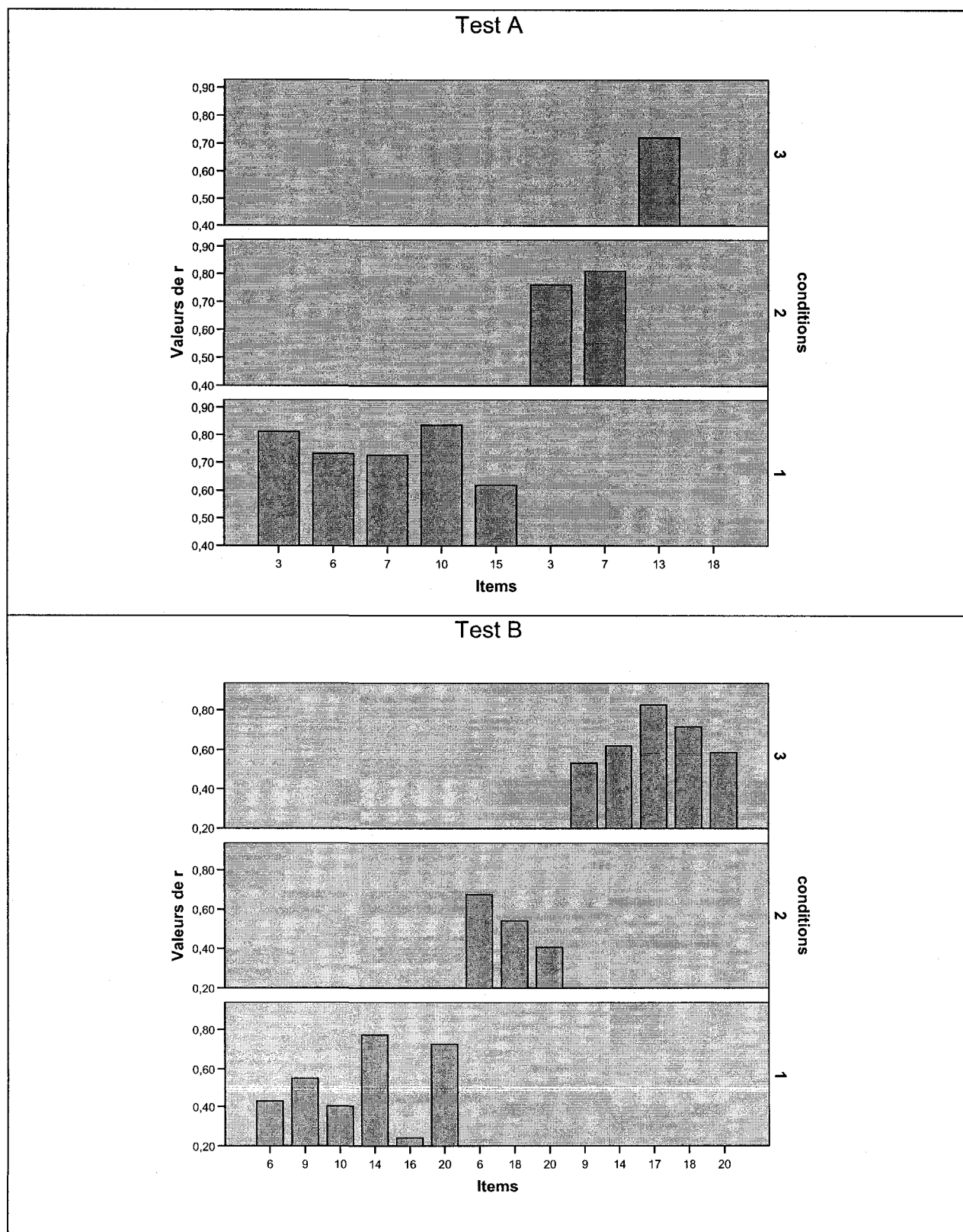


Figure 16. Les valeurs r pour l'attribut 6 selon les conditions et les tests A et B

ANNEXE J

Les attestations d'approbation déontologique



Université d'Ottawa University of Ottawa

Services de interventions de recherche et déontologie Research Ethics and Ethics Services

Le 10 mai 2006

Dany Laveault,
Faculté d'éducation
Université d'Ottawa
145 Jean-Jacques Lussier (469)
INTERNE

Nathalie Loye
633 rue des vignobles
Rosebère, QC J7A 4P9

Objet : Augmenter le pouvoir diagnostique des questions à choix multiple grâce à la modélisation des réponses (Dossier # 02-06-14)

Chers Monsieur Laveault et Madame Loye,

Le Comité d'éthique de la recherche en Sciences Sociales et Humanités de l'Université d'Ottawa (CER en SSH) a examiné votre demande d'approbation et y a accordé une catégorie 1a (approbation). Vous trouverez donc ci-joint le certificat d'approbation déontologique pour votre projet de recherche.

Au cours de votre étude, toute modification au protocole ou aux formulaires ne peut être introduite sans l'approbation préalable écrite du CER. Vous devez aussi aviser le CER dans les plus brefs délais de tout événement ou expérience indésirables vécus par les participants.

Cette attestation d'approbation déontologique est valide jusqu'au 24 mars 2007. Veuillez soumettre un rapport annuel à la Responsable de l'éthique en recherche en mars 2007 pour soit fermer le dossier ou faire demande d'extension. Ce rapport se trouve à l'adresse suivante : http://www.ssd.uottawa.ca/deontologie/application_den_f.asp

Si vous avez des questions, n'hésitez pas à me contacter au poste 1787.

Veillez agréer mes sentiments les meilleurs.

Catherine Paquet
Responsable de l'éthique en recherche
CC : Richard Clément, Président du CER en SSH

550 rue Christopher 550 Christopher Street
Ottawa (Ontario) K1N 6N5 Canada Ottawa, Ontario K1N 6N5 Canada
(613) 562-5841 • Télec. (613) 562-5538
http://www.uottawa.ca/serveurs/du_ssd/recherche/index.html



Université d'Ottawa University of Ottawa
 Services de recherches de recherche et déontologie Research Grants and Ethics Services

**COMITÉ D'ÉTHIQUE DE LA RECHERCHE
 EN SCIENCES SOCIALES ET HUMANITÉS**

ATTESTATION D'APPROBATION DÉONTOLOGIQUE

La présente attestation certifie que le Comité d'éthique de la recherche en Sciences Sociales et Humanités de l'Université d'Ottawa a examiné la demande d'approbation déontologique pour le projet intitulé Augmenter le pouvoir diagnostique des questions à choix multiple grâce à la modélisation des réponses (Dossier # 02-06-14), présenté par Nathalie Loyer et supervisé par Dany Laveault de la Faculté d'éducation. Le Comité d'éthique a déterminé que la demande respectait les principes déontologiques établis par l'Énoncé de politique des trois conseils et par les règles de procédure des Comités d'éthique de l'Université d'Ottawa et a donc accordé une catégorie 1a (approbation) à ce projet.

La présente attestation est valide pour un an à partir de la date indiquée ci-dessous.

Catherine Paquet
 Responsable de l'éthique en recherche
 Pour le Président du CÉR en Sciences Sociales et Humanités
 Richard Clément

24 mars 2006
 Date

330 rue Cumberland 330 Cumberland Street
 Ottawa (Ontario) K1N 6N5 Canada Ottawa (Ontario) K1N 6N5 Canada
 (613) 562-5941 • Téléc. (613) 562-5338
<http://www.uottawa.ca/cer/cer/en/ethics/index.html>



Université d'Ottawa University of Ottawa

Services de autorisations de recherche et déontologie Research Grants and Ethics Services

Le 10 mai 2006

Dany Laveault
Faculté d'éducation
Université d'Ottawa
145 Jean-Jacques Lussier (469)
INTERNE

Nathalie Loyer
633 rue des vignobles
Rosebère, QC J7A 4P9

Objet : Augmenter le pouvoir diagnostique des questions à choix multiple grâce à la modélisation des réponses (Dossier # 02-06-14)

Chers Monsieur Laveault et Madame Loyer,

Le Comité d'éthique de la recherche en Sciences Sociales et Humanités de l'Université d'Ottawa (CER en SSH) a examiné votre demande d'approbation et y a accordé une **catégorie 1a** (approbation). Vous trouverez donc ci-joint le certificat d'approbation déontologique pour votre projet de recherche.

Au cours de votre étude, toute modification au protocole ou aux questionnaires ne peut être introduite sans l'approbation préalable écrite du CER. Vous devez aussi aviser le CER dans les plus brefs délais de tout événement ou expérience indésirables vécus par les participants.

Cette attestation d'approbation déontologique est valide jusqu'au 24 mars 2007. Veuillez soumettre un rapport annuel à la Responsable de l'éthique en recherche en mars 2007 pour soit fermer le dossier ou faire demande d'extension. Ce rapport se trouve à l'adresse suivante : http://www.ssd.uottawa.ca/deontologie/application_chen_f.asp

Si vous avez des questions, n'hésitez pas à me contacter au poste 1787.

Veillez agréer mes sentiments les meilleurs.

Catherine Paquet
Responsable de l'éthique en recherche
CC : Richard Clément, Président du CER en SSH

151, rue Clarkstown 2ND Clarkstown Street
Ottawa (Ontario) K1N 6N5 Canada Ottawa, Ontario K1N 6N5 Canada
(613) 562-5941 • Tél/Fax (613) 562-5938
téléphone et télécopieur en français disponibles



Université d'Ottawa University of Ottawa

Service des universitaires de recherche et d'éducation Research Grants and Ethics Services

COMITÉ D'ÉTHIQUE DE LA RECHERCHE EN SCIENCES SOCIALES ET HUMANITÉS

ATTESTATION D'APPROBATION DÉONTOLOGIQUE

La présente attestation certifie que le Comité d'éthique de la recherche en Sciences Sociales et Humanités de l'Université d'Ottawa a examiné la demande d'approbation déontologique pour le projet intitulé Augmenter le pouvoir diagnostique des questions à choix multiple grâce à la modélisation des réponses (Dossier # 02-06-14), présenté par Nathalie Loye et supervisé par Dany Laveault de la Faculté d'éducation. Le Comité d'éthique a déterminé que la demande respectait les principes déontologiques établis par l'Énoncé de politique des trois conseils et par les règles de procédure des Comités d'éthique de l'Université d'Ottawa et a donc accordé une catégorie 1a (approbation) à ce projet.

La présente attestation est valide pour un an à partir de la date indiquée ci-dessous.

Catherine Poquet
Responsable de l'éthique en recherche
Pour le Président du CÉR en Sciences Sociales et Humanités
Richard Clément

24 mars 2006
Date

150 rue Cumberland • 155 Charles Road, Suite
Ottawa, Ontario K1M 5N5 Canada Ottawa, Ontario K1N 6N5 Canada
(613) 562-1141 • Téléc. Fax (613) 562-5738
<http://www.uottawa.ca/~services/ethic/research/ethic.html>