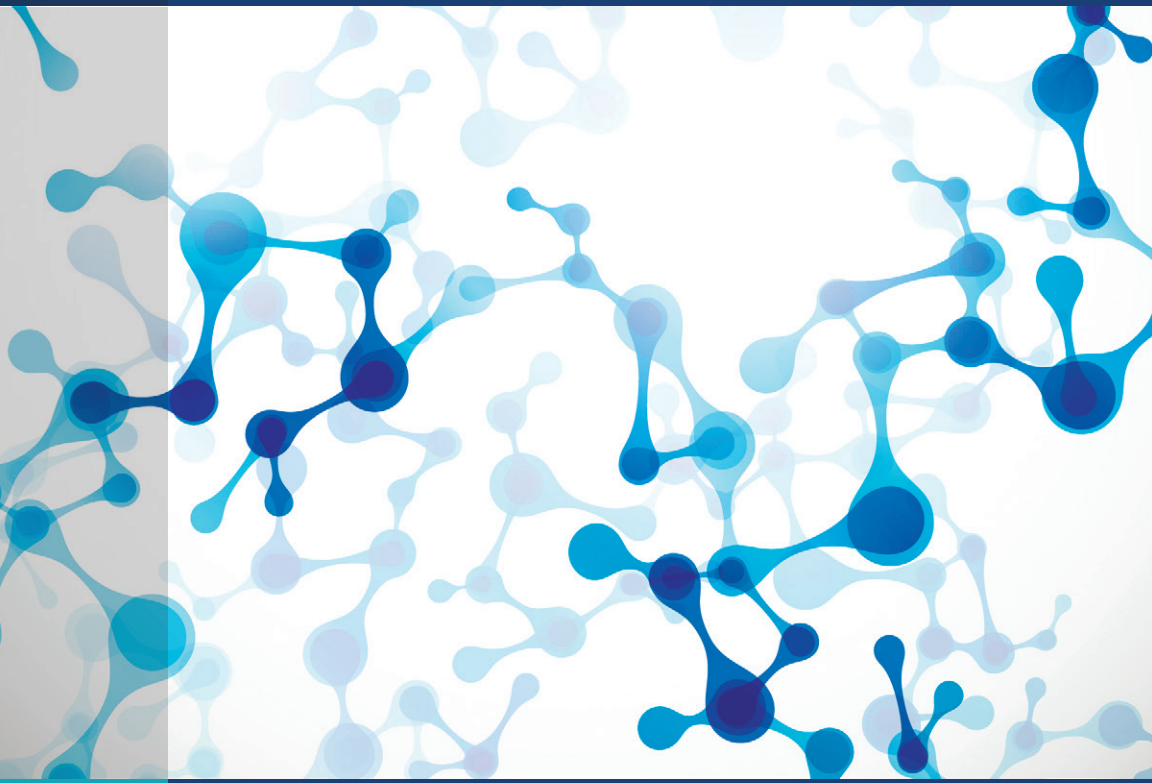
MESURE ET EVALUATION

Sous la direction de
ERIC DIONNE et ISABELLE RAÎCHE

MESURE ET ÉVALUATION DES COMPÉTENCES EN ÉDUCATION MÉDICALE

Regards actuels et prospectifs



Presses de l'Université du Québec

COLLECTION
MESURE ET EVALUATION

Sous la direction de Gilles Raîche,
Université du Québec à Montréal

La collection **Mesure et évaluation** soutient la diffusion de recherches et de travaux fondamentaux, ainsi que de matériel didactique pour les niveaux collégial et universitaire, dans le domaine de la mesure et de l'évaluation en éducation et, plus largement, en sciences humaines.

Les nouveaux enjeux sociétaux et les besoins émergents des milieux de pratique demandent aux intervenants d'être informés des avancées récentes afin de les soutenir dans leur travail. **Mesure et évaluation** offre aussi aux chercheurs un moyen de partager les résultats de leurs travaux avec ces intervenants tout en faisant progresser la recherche, que ce soit en matière de mesure et d'évaluation des apprentissages, de programmes ou encore de méthodologie de recherche.

Les textes publiés sont soumis à un processus d'arbitrage avec le soutien d'évaluateurs externes. La collection **Mesure et évaluation** souscrit à l'adaptation canadienne-française, par la *Revue des sciences de l'éducation*, des règles de publication de l'American Psychological Association.

MESURE ET ÉVALUATION DES COMPÉTENCES EN ÉDUCATION MÉDICALE

DANS LA MÊME COLLECTION

INTRODUCTION À LA MODÉLISATION D'ÉQUATIONS STRUCTURELLES
AMOS dans la recherche en gestion
Lili Zheng, Michel Plaisent, Cataldo Zuccaro et Prosper Bernard
ISBN-978-2-7605-4738-4, 118 pages

CONSTRUIRE DES GRILLES D'ÉVALUATION DESCRIPTIVES AU COLLÉGIAL
Guide d'élaboration et exemples de grille
France Côté
ISBN-978-2-7605-4101-6, 192 pages

DES MÉCANISMES POUR ASSURER LA VALIDITÉ DE L'INTERPRÉTATION
DE LA MESURE EN ÉDUCATION, **Volume 1 – LA MESURE**
Sous la direction de Gilles Raîche, Karine Paquette-Côté et David Magis
Avec la collaboration de Diane Leduc et d'Hélène Meunier
ISBN-978-2-7605-2685-3, 148 pages

DES MÉCANISMES POUR ASSURER LA VALIDITÉ DE L'INTERPRÉTATION
DE LA MESURE EN ÉDUCATION, **Volume 2 – L'ÉVALUATION**
Sous la direction de Gilles Raîche, Karine Paquette-Côté et David Magis
Avec la collaboration de Diane Leduc et d'Hélène Meunier
ISBN-978-2-7605-2687-7, 178 pages

DES MÉCANISMES POUR ASSURER LA VALIDITÉ DE L'INTERPRÉTATION
DE LA MESURE EN ÉDUCATION, **Volume 3 – ASPECTS PRATIQUES**
Sous la direction de Gilles Raîche, Pascal Ndinga et Hélène Meunier
ISBN-978-2-7605-3593-0, 172 pages

Membre de L'ASSOCIATION NATIONALE DES ÉDITEURS DE LIVRES

Presses de l'Université du Québec
Le Delta I, 2875, boulevard Laurier, bureau 450, Québec (Québec) G1V 2M2
Téléphone : 418 657-4399 *Télécopieur :* 418 657-2096
Courriel : puq@puq.ca *Internet :* www.puq.ca

Sous la direction de
ERIC DIONNE et ISABELLE RAÎCHE

MESURE ET ÉVALUATION DES COMPÉTENCES EN ÉDUCATION MÉDICALE

Regards actuels et prospectifs

Presses de l'Université du Québec

Financé par le gouvernement du Canada
Funded by the Government of Canada
Canada

Conseil des arts du Canada
Canada Council for the Arts

SODEC Québec

Révision
Gislaine Barrette

Correction d'épreuves
Mélissa Guay

Conception graphique
Richard Hodgson

Image de couverture
iStock

Mise en page et adaptation numérique
Studio C1C4

ISBN 978-2-7605-4427-7
ISBN PDF 978-2-7605-4428-4
ISBN EPUB 978-2-7605-4429-1

Dépôt légal : 3e trimestre 2017
› Bibliothèque et Archives nationales du Québec
› Bibliothèque et Archives Canada

À ma fille Marianne (9 ans) qui a déjà compris
les enjeux de la mesure et de l'évaluation… !
ED

À ma famille pour leur soutien
inconditionnel…
IR

REMERCIEMENTS

Nous désirons remercier le Centre d'appui pédagogique en santé pour la francophonie (CAPSAF) de la Faculté de médecine de l'Université d'Ottawa pour le soutien financier ayant permis la publication de cet ouvrage.

TABLE DES MATIÈRES

REMERCIEMENTS IX

LISTE DES FIGURES XVII

LISTE DES TABLEAUX XIX

LISTE DES SIGLES XXIII

INTRODUCTION
Regards actuels et prospectifs sur la mesure et l'évaluation des compétences en éducation médicale 1
Eric Dionne et Isabelle Raîche

Contenu de l'ouvrage 5

Bibliographie 7

PARTIE 1
LA MESURE DES COMPÉTENCES 9

CHAPITRE 1
Démonstration d'une méthodologie mettant à profit les modèles de Rasch : l'exemple d'une échelle de mesure de l'offre active de services de santé en français 11
Julie Grondin, Eric Dionne, Jacinthe Savard et Lynn Casimiro

1. Le contexte 12

2. Le cadre théorique 13

2.1. Les modèles de la famille de Rasch 13

2.2. La démarche de modélisation des scores de Tennant et Conaghan 14
2.3. Les modèles choisis 15
2.4. La qualité de l'ajustement des données au modèle 17
2.5. Les échelles de réponses 19
2.6. Les postulats d'unidimensionnalité et d'indépendance locale 22
2.7. Le fonctionnement différentiel d'items (FDI) 25
2.8. Les indices de fidélité 26
3. La méthodologie 27
3.1. Les sujets 27
3.2. L'instrument 28
3.3. Le logiciel 28
4. Les résultats et la discussion 28
4.1. Les modèles utilisés 28
4.2. La qualité de l'ajustement 29
4.3. Les catégories de réponses 31
4.4. Les indices sur les postulats d'unidimensionnalité et d'indépendance locale 39
4.5. La détection du FDI 42
4.6. La qualité de l'échelle de mesure 44
4.7. La fidélité 46
Conclusion 47
Bibliographie 49

CHAPITRE 2
L'analyse psychométrique d'outils d'évaluation en pédagogie des sciences de la santé : une comparaison des conclusions selon les approches classique et de Rasch 53
Jean-Sébastien Renaud

1. Le contexte 54
2. Le cadre conceptuel 57
2.1. La théorie classique des tests 57
2.2. L'approche de Rasch 58
3. La méthodologie 59
3.1. L'échantillon 59
3.2. L'instrument 60
3.3. Les analyses 61
4. Les résultats 62
4.1. L'approche de la théorie classique des tests 62
4.2. L'approche de Rasch 64
Conclusion 69
Bibliographie 72

CHAPITRE 3
Exploration des scores à un test de concordance de script sous la loupe de la modélisation de Rasch 77
Eric Dionne, Julie Grondin et Marie-Eve Latreille

1. La problématique. 78
2. Le modèle de Rasch 82
3. La méthodologie 84
   - 3.1. Les participants. 85
   - 3.2. La transformation des scores 85
   - 3.3. Les méthodes de détermination des scores 85
   - 3.4. Le déroulement des analyses 86
4. Les résultats et la discussion 87
   - 4.1. Les modèles étudiés 87
   - 4.2. Les échelles de mesure 88
   - 4.3. Les statistiques d'ajustement pour les items et les répondants 96
   - 4.4. Les indices de fidélité 99
   - 4.5. L'indépendance locale 101
   - 4.6. La dimensionnalité. 102
   - 4.7. La comparaison des modèles 105

   Conclusion. 106

   Bibliographie 107

PARTIE 2
L'ÉVALUATION DES COMPÉTENCES 111

CHAPITRE 4
Les avancées technologiques, les enjeux et les défis de la notation automatisée en éducation dans le domaine de la santé. 113
Maxim Morin, André-Philippe Boulais et André F. De Champlain

1. Les définitions 115
2. Les repères historiques. 116
3. La conception des engins 117
   - 3.1. L'identification de manifestations observables 119
   - 3.2. L'établissement des critères de notation 120
4. Des exemples de stratégies de notation. 122
   - 4.1. L'évaluation de la prise en charge de patients à l'aide de simulation 122
   - 4.2. L'évaluation de la prise de notes cliniques 125
   - 4.3. L'évaluation de la prise de décisions cliniques 127
   - 4.4. L'évaluation des habiletés techniques de chirurgiens 129
   - 4.5. La synthèse 130

5. Les enjeux et défis 131
5.1. L'opérationnalisation du construit. 132
5.2. Le glissement lors de la correction humaine 133
5.3. L'accès à des outils d'élaboration des critères 133
5.4. La robustesse des techniques de TALN. 134
Conclusion. 134
Bibliographie 137

CHAPITRE 5
Une approche pragmatique de validation en éducation médicale : l'application du modèle de Kane à un outil d'évaluation du raisonnement clinique. 143
Thomas Pennaforte et Nathalie Loye

1. Le prétexte à l'utilisation du modèle de Kane : la validation d'un outil d'évaluation du raisonnement clinique. 147
1.1. L'assise théorique de notre exemple 147
1.2. Le problème à résoudre 147
1.3. L'idée de développement 149
2. Le modèle de Kane. 150
2.1. Les principes généraux 150
2.2. La définition des termes 151
2.3. Le principe des inférences 154
3. L'application du modèle de Kane pour soutenir le processus de validation de l'outil d'évaluation du raisonnement clinique : une démarche pragmatique 158
3.1. Première étape : la définition du trait. 158
3.2. Deuxième étape : la définition de l'argument de validité . . . 158
3.3. Troisième étape : la collecte d'éléments de preuve. 159
Conclusion. 169
Bibliographie 170

CHAPITRE 6
L'utilisation de la formation par concordance comme modalité d'évaluation formative pour entraîner à la prise de décision opératoire. 177
Isabelle Raîche et Bernard Charlin

1. La problématique et une revue de la littérature 178
1.1. La problématique : le rôle de la prise de décision dans l'expertise chirurgicale et les limites du mode d'enseignement actuel 179
1.2. L'évaluation formative : quelques fondements théoriques . . . 181
2. La création d'un dispositif d'évaluation formative. 184
2.1. La planification générale 185
2.2. La définition des contenus à inclure dans le dispositif d'évaluation formative 187

2.3. L'élaboration d'un tableau de spécification 189
2.4. La création d'items 191
2.5. La conception et l'assemblage du dispositif d'évaluation formative 193

3. Le projet pilote : le sondage de la réaction des apprenants et des enseignants et les considérations logistiques 195
3.1. Les réactions des apprenants 195
3.2. Les réactions des enseignants 197
3.3. Les considérations logistiques 198

4. La discussion 199
4.1. La synthèse des résultats 199
4.2. Les liens avec la littérature sur l'évaluation formative 201
4.3. Les limites du projet 203
4.4. Les recherches futures 203

Conclusion 204

Annexe A. Étapes du développement du dispositif d'évaluation suivant le modèle de Downing 205

Annexe B. Contenus essentiels pour la prise de décision dans la colectomie droite laparoscopique 206

Annexe C. Résumé du projet pilote sur la validité apparente et la faisabilité de cette formation par concordance 207

Bibliographie 208

CHAPITRE 7
Le rôle de l'évaluation de programme dans le domaine de la santé 213
Maud Mediell et Eric Dionne

1. Les fondements de l'épidémiologie et des pratiques évaluatives en santé publique au Canada 216
1.1. Qu'est-ce que l'épidémiologie ? 216
1.2. Qu'entend-on par interventions complexes en santé ? 218
1.3. Deux tendances : l'épidémiologie évaluative et l'évaluation économique des interventions en santé 219

2. L'évaluation développementale 227
2.1. Une description 227
2.2. Exemple : le programme international de Médecine et les humanités 229
2.3. Les défis de l'évaluation développementale 231

Conclusion 232

Bibliographie 233

CONCLUSION 237
Eric Dionne et Isabelle Raîche

NOTICES BIOGRAPHIQUES 239

LISTE DES FIGURES

FIGURE 1.1
Mesure moyenne des sujets ayant opté pour chacune des catégories de réponses suivant le modèle PC 32

FIGURE 1.2
Mesure des seuils entre chaque catégorie de réponses selon le modèle PC 33

FIGURE 1.3
Courbes de probabilité des catégories de réponses de l'item 1 selon le modèle PC 35

FIGURE 1.4
Sommaire de la structure des catégories de réponses de l'item 1 selon le modèle PC, tel que produit par le logiciel Winsteps 36

FIGURE 1.5
Courbes de probabilité des catégories de réponses de l'item 1 selon le modèle PC-regroup2 38

FIGURE 1.6
Sommaire de la structure des catégories de réponses de l'item 1 selon le modèle PC-regroup2, tel que produit par le logiciel Winsteps 39

FIGURE 1.7
Représentation de Wright suivant le modèle RS. 45

FIGURE 2.1
Adéquation entre la distribution des sujets (histogramme du haut) et des items (histogramme du bas) 68

FIGURE 3.1
Exemple d'une vignette traduite du TCS construit par Latreille (2012) 79

FIGURE 3.2
Position des répondants (e : étudiante ; p : praticienne), du niveau d'expérience et des items sur l'échelle de mesure avec la méthode 1 94

FIGURE 3.3
Position des répondants (e : étudiante ; p : praticienne), du niveau d'expérience et des items sur l'échelle de mesure avec la méthode 2 95

FIGURE 3.4
Distribution des statistiques d'ajustement *infit* et *outfit* basées sur le carré moyen (CM) pour les items et triées en ordre croissant avec la méthode 1 et la méthode 2 97

FIGURE 3.5
Distribution des statistiques d'ajustement *infit* et *outfit* basées sur la valeur standardisée (Z) pour les items et triées en ordre croissant avec la méthode 1 et la méthode 2 98

FIGURE 3.6
Distribution des statistiques d'ajustement pour les répondants *infit* et *outfit* basées sur le carré moyen (CM) et triées en ordre croissant avec la méthode 1 et la méthode 2 98

FIGURE 3.7
Distribution des statistiques d'ajustement pour les répondants *infit* et *outfit* basées sur la valeur standardisée (Z) et triées en ordre croissant 99

FIGURE 5.1
Cadre de compétence CanMEDS 144

FIGURE 5.2
Procédure de validation de Kane 151

FIGURE 5.3
Niveaux d'inférence dans le modèle de Kane 154

FIGURE 6.1
Photo annotée pour orienter les apprenants 191

FIGURE 6.2
Exemple d'item 192

FIGURE 6.3
Exemple de réponses fournies aux apprenants (L4) 194

LISTE DES TABLEAUX

TABLEAU 1.1
Comparaison entre la TCT et la modélisation de Rasch. 14

TABLEAU 1.2
Interprétation des indices d'ajustement basés
sur la statistique standardisée 18

TABLEAU 1.3
Résumé de la stratégie d'analyse de la qualité de l'ajustement
retenue pour les sujets et les items 19

TABLEAU 1.4
Résumé de la stratégie d'analyse des échelles
de réponses retenue. 22

TABLEAU 1.5
Résumé de la stratégie d'analyse de la dimensionnalité
et l'indépendance locale retenue 24

TABLEAU 1.6
Résumé de la stratégie d'analyse du fonctionnement différentiel
d'items retenue 26

TABLEAU 1.7
Balises pour l'interprétation de l'indice de séparation 26

TABLEAU 1.8
Résumé des modélisations 29

TABLEAU 1.9
Incidence des modélisations sur le nombre de sujets. 30

TABLEAU 1.10
Items problématiques au regard des statistiques d'ajustement 30

TABLEAU 1.11
Statistiques d'ajustements finaux pour les items avec le modèle RS 31

TABLEAU 1.12
Synthèse du nombre de catégories de réponses utiles, de l'ordonnancement des catégories de réponses et des seuils, ainsi que des regroupements explorés 38

TABLEAU 1.13
Synthèse de la variance expliquée pour les différents modèles 40

TABLEAU 1.14
Synthèse de l'analyse de la dimensionnalité des items 41

TABLEAU 1.15
Corrélations inter-items les plus élevées effectuées sur les résidus standardisés 41

TABLEAU 1.16
Synthèse des résultats des deux tests d'hypothèse utilisés pour la détection de FDI selon le modèle RS 43

TABLEAU 1.17
Indices de fidélité pour les sujets selon les modélisations 46

TABLEAU 1.18
Indices de fidélité pour les items selon les modélisations 46

TABLEAU 2.1
Échelle de communication médecin-patient pour les externes en médecine (ECMP-EM) 60

TABLEAU 2.2
Indices d'ajustement de l'analyse factorielle confirmatoire. 62

TABLEAU 2.3
Saturation et coefficient de détermination des items pour l'analyse factorielle confirmatoire 62

TABLEAU 2.4
Coefficient alpha de Cronbach et statistiques descriptives de l'ECMP-EM 63

TABLEAU 2.5
Corrélations inter-items 63

TABLEAU 2.6
Résultats de l'analyse d'items 64

TABLEAU 2.7
Ajustements itératifs pour ajuster les données au RSM 67

TABLEAU 2.8
Statistiques décrivant la distribution des sujets et des items (en *logit*) 68

TABLEAU 3.1
Estimation des paramètres des items avec la méthode 1 90

TABLEAU 3.2
Estimation des paramètres des items avec la méthode 2 92

TABLEAU 3.3
Indice de séparation et fidélité pour les trois facettes étudiées 101

TABLEAU 3.4
Valeurs corrélationnelles des items basées sur les résidus standardisés 102

TABLEAU 3.5
Résultats de l'analyse en composante principale des résidus standardisés 103

TABLEAU 3.6
Synthèse de l'analyse de la dimensionnalité des items 104

TABLEAU 5.1
Propositions de questions auxquelles doit répondre le chercheur en vue de soutenir l'inférence de notation 163

TABLEAU 5.2
Propositions de questions auxquelles doit répondre le chercheur en vue de soutenir l'inférence de généralisation 165

TABLEAU 5.3
Propositions de questions auxquelles doit répondre le chercheur en vue de soutenir l'inférence d'extrapolation 168

TABLEAU 5.4
Propositions de questions auxquelles doit répondre le chercheur en vue de soutenir l'inférence d'implication 169

LISTE DES SIGLES

ASL	Analyse sémantique latente
CFSM	Communautés francophones en situation minoritaire
CNFS	Consortium national de formation en santé
EACMC	Examen d'aptitude du Conseil médical du Canada
ECMP-EM	Échelle de communication médecin-patient pour les externes en médecine
ECOS	Examen clinique objectif structuré
ECR	Essai contrôlé
FDI	Fonctionnement différentiel d'items
IRMf	Imagerie par résonance magnétique fonctionnelle
NBME	National Board of Medical Examiners
PEG	Project Essay Grade
PRC	Problèmes de raisonnement clinique
TCS	Test de concordance de script
TCT	Théorie classique des tests
TRI	Théorie des réponses aux items
USMLE	United States Medical Licensing Examination

INTRODUCTION

Regards actuels et prospectifs sur la mesure et l'évaluation des compétences en éducation médicale

Eric Dionne et Isabelle Raîche

La mesure et l'évaluation des apprentissages en pédagogie médicale a le vent en poupe depuis maintenant plusieurs années. Le nombre de publications scientifiques, d'ouvrages de vulgarisation ou encore le florilège de conférences scientifiques consacrées généralement à la pédagogie médicale, et particulièrement à l'évaluation des apprentissages en santé, en sont quelques illustrations probantes. On remarque aussi que les facultés d'éducation, du moins au Canada, ne sont plus, comme ce fut le cas pendant longtemps, les seules dépositaires de l'expertise scientifique en mesure et évaluation. En effet, plusieurs experts de la mesure et de l'évaluation, issus par exemple du domaine scolaire, se retrouvent de plus en plus embauchés ou affiliés aux facultés de médecine. Ils sont appelés à mener des travaux d'érudition dans un contexte non plus exclusivement scolaire, mais plutôt relatif à l'éducation médicale. Cette situation fait en sorte que les développements qui marquent la discipline de la mesure et de l'évaluation émanent de plus en plus de l'éducation médicale comme c'est le cas avec la plupart des contributions que compose cet ouvrage collectif.

Ce projet d'ouvrage collectif est né à la suite de la conférence de l'Association pour le développement des méthodologies d'évaluation en éducation (ADMÉÉ), qui se tenait à Gatineau à l'automne 2015, et qui avait pour thème général « L'évaluation des apprentissages complexes ». Durant cet évènement, de nombreux

chercheurs sont venus présenter leurs plus récentes recherches et leurs réflexions concernant l'évaluation des apprentissages dans le domaine de la pédagogie médicale et de la pédagogie des sciences de la santé (pharmacie, sciences infirmières, etc.). Il nous est alors apparu pertinent d'envisager la rédaction d'un ouvrage collectif pour apporter un éclairage sur les plus récents développements en mesure et en évaluation des apprentissages dans ces domaines. De plus, rappelons qu'il existe bien peu d'ouvrages à caractère scientifique qui abordent ce sujet et encore moins en français. Notons également que le présent ouvrage constitue le premier projet d'importance relié au Groupe de recherche interuniversitaire en mesure et évaluation des apprentissages en santé (GRIMEAS) dont plusieurs des auteurs de ce collectif font partie.

En guise d'introduction, il nous apparaît important de situer les thèmes abordés dans les différents chapitres au regard du titre de cet ouvrage qui comporte des mots clés que nous allons maintenant définir : apprentissage, compétence, mesure et évaluation.

Il existe une kyrielle de définitions concernant l'apprentissage et chacune d'elles s'inscrit dans un cadre épistémologique et théorique. Certaines sont campées dans une perspective plus behavioristes, d'autres, plus récentes, renvoient à une conception plus socioconstructiviste. Dans le cadre de cet ouvrage, bien que le fil d'Ariane soit l'évaluation et non l'apprentissage, ces deux dimensions demeurent intimement liées, voire indissociables. Il nous semble donc important de présenter notre conception de l'apprentissage. Pour ce faire, nous citerons la définition proposée dans le *Dictionnaire actuel de l'éducation* de Legendre puisqu'elle reflète bien la position des auteurs de cet ouvrage :

> Processus d'acquisition ou de changement, dynamique et interne à une personne laquelle, mue par le désir et la volonté de développement, construit de nouvelles représentations explicatives cohérentes et durables de son réel à partir de la perception de matériaux, de stimulations de son environnement, de l'interaction entre les données internes et externes au sujet et d'une prise de conscience personnelle (Legendre, 2005, p. 88).

Cette définition s'inscrit résolument dans un cadre socioconstructiviste, mais on y reconnaît également l'influence très didactique. Examinons brièvement certains concepts qui lui sont inhérents. D'abord, *l'apprentissage* est présenté comme un processus. Il peut s'agir d'acquérir (concepts, méthodes, théories, techniques, etc.) de nouvelles informations, mais aussi de s'engager dans un processus de changement. Dans plusieurs autres définitions consultées, l'idée du changement revient souvent. Apprendre, c'est ultimement changer. Un

autre mot clé de cette définition est le vocable *représentation*. Dans un contexte purement didactique, on parlerait davantage de « conception ». Sans entrer dans un débat rhétorique, il faut retenir que l'apprenant va se construire sa propre représentation, sa propre image mentale, de ce qu'il a appris. Les didacticiens ajouteraient que le travail didactique consiste justement à faire évoluer ces conceptions ou ces représentations en posant aux apprenants des défis qui suscitent des conflits cognitifs. Une autre dimension présente dans cette définition concerne la motivation. En effet, pour qu'un changement puisse se réaliser, encore faut-il que l'apprenant veuille changer. Enfin, la dernière partie de la définition fait référence à des aspects bien connus en pédagogie médicale, à savoir les environnements d'apprentissage pris au sens large (laboratoire de simulation, livres de référence, modules de formation, activités d'apprentissage, etc.). Il s'agit du matériel pédagogique et didactique mis à la disposition des apprenants pour favoriser le changement souhaité. La définition que nous venons d'examiner nous rappelle le caractère hautement complexe de l'apprentissage. Cette complexité se transpose, bien évidemment, quand il s'agit de mesurer, observer et, ultimement, évaluer ces apprentissages. Les acteurs impliqués en éducation médicale et, aussi, plus généralement, dans les autres disciplines, notamment en pédagogie de la santé, sont plus que jamais à l'affût de méthodes novatrices pour appréhender cette complexité.

Cette présentation, certainement trop succincte, de l'apprentissage serait, dans le cadre de cet ouvrage, incomplète si nous n'abordions pas la question de la complexité puisqu'il s'agit du thème de ce livre. Toutefois, définir ce qu'est un apprentissage complexe se révèle une entreprise beaucoup plus ardue qu'il n'y paraît. Il serait tentant de définir la complexité des apprentissages en présentant des exemples d'activités pédagogiques qui font appel à des performances complexes comme la simulation immersive ; bien que pratique, cette façon de procéder ne rendrait pas correctement cette idée d'apprentissage complexe. On pourrait également tenter de définir la complexité dans l'optique de l'apprenant. Cette stratégie ne serait pas non plus très fructueuse, car ce qui est complexe pour un apprenant (résident R1) ne le sera pas nécessairement pour un autre (résident R3). Cette notion de complexité, apparemment difficile à circonscrire, fait partie intégrante du concept de compétence qui fait maintenant partie du vocabulaire usuel en éducation médicale. À ce sujet, il suffit de penser au référentiel de compétences CanMeds, développé par le Collège royal des médecins et des chirurgiens du Canada, qui décrit les compétences que les médecins doivent acquérir pour pratiquer leur profession. Dans un essai visant à faire le point sur cette notion de compétence, Hodges (2012) considère qu'elle s'inscrit dans un cadre, qu'il nomme « discours », qui permet de camper cette notion. Il relève

cinq discours dominants : 1) connaissance, 2) performance, 3) psychométrie, 4) réflexion et 5) production. Le premier discours touche à l'accumulation de connaissances, soit ce que nous pouvons associer au pôle érudition dans le CanMeds. La compétence s'inscrit également dans une perspective de performance au sens où les professionnels de la santé doivent agir et poser des questions dans l'action pour le mieux-être de leurs patients. Selon Hodges, le discours psychométrique fait référence à l'exercice de la compétence en vue d'atteindre une norme établie. L'aspect associé avec la réflexion concerne l'aspect métacognitif au sens où le professionnel se doit, par exemple, de s'autoévaluer afin de se remettre en question en vue de chercher constamment à améliorer ses pratiques. Enfin, le dernier discours renvoie à la production, c'est-à-dire au devoir des professionnels en santé de se conformer à de hauts degrés de performance. Dans le cadre de cet ouvrage, les propos des auteurs correspondent principalement aux trois premiers discours. La première partie, consacrée à la mesure, s'inscrit dans un discours davantage psychométrique. Les textes de la deuxième partie, celle de l'évaluation des compétences, touchent, quant à eux, surtout aux quatre autres discours tels que définis par Hodges.

Qu'est-ce que la mesure ? Il s'agit d'un concept qui peut paraître familier, mais qui n'est peut-être pas connu à sa juste valeur. À l'instar de Wilson (2005), nous considérons que le mot « mesure » ne rend pas tout à fait justice au processus qui consiste à assigner des chiffres et des nombres à des observations. Ce dernier propose, dans le cadre de son ouvrage, d'employer le verbe « mesurer » qui suggère à tout le moins une action, voire un processus. Faire état des propriétés métriques d'un instrument consiste à développer un argumentaire basé sur des données empiriques en s'appuyant sur un modèle de mesure pour en estimer la valeur. En effet, un outil peu valide ou peu fiable n'est d'aucune utilité, d'où l'importance indéniable pour les chercheurs, ou toutes personnes intéressées à développer des instruments de mesure ou d'observation, de s'attarder aux modèles de mesure. La première partie de cet ouvrage collectif traite plus particulièrement de la modélisation de la mesure.

Terminons ce tour d'horizon en abordant le dernier concept présenté dans le titre de cet ouvrage : l'évaluation. Comme pour le concept de mesure, il s'agit d'un terme qui nous semble bien familier. Son acception a pourtant bien évolué au fil du temps. Dans les écrits anglo-saxons, on emploie maintenant à profusion le terme *assessment* qui, selon Black et Dylan (2010, p. 82), se définit de la façon suivante :

> L'*assessment* est un terme général qui fait référence à toutes les activités mises de l'avant par l'enseignant ou par les apprenants afin de produire une information qui permettra de fournir une rétroaction permettant de modifier les activités d'enseignement et d'apprentissage (traduction libre).

De cette définition, il faut principalement retenir qu'il s'agit d'un processus, d'une démarche permettant de recueillir des informations pour documenter ou étayer le jugement du professeur ou du mentor qui doit juger de la compétence d'un apprenant. Ce jugement doit s'appuyer sur une démarche rigoureuse et convaincante qui permettra, peu importe le contexte (diagnostique, formatif, sommatif ou encore certificatif) de prendre une décision de nature pédagogique, didactique ou administrative. Dans tous les cas, cette décision devra être appuyée par des éléments de preuve qui touchent autant le produit (la décision) que la démarche ayant permis d'en arriver à cette décision.

Comme nous le mentionnions au début du paragraphe précédent, l'évaluation possède plusieurs acceptions, l'une d'elles la situe dans un contexte non pas micro (p. ex. évaluation d'un apprenant), mais plutôt macro (p. ex. évaluation d'un programme de formation). Le terme français *évaluation* ne permet pas de faire cette distinction, ce qui oblige à le qualifier : évaluation des apprentissages, évaluation institutionnelle ou encore évaluation de programme de formation. Ainsi, le terme anglo-saxon *assessment* peut sembler plus précis puisqu'il se rapporte à la composante micro (*assessment*) et que le terme *program evaluation* vise la composante macro. Cette clarification nous apparaît importante, car le dernier chapitre de cet ouvrage présente une perspective plus macro en abordant la question de l'évaluation de programme dans un contexte d'éducation médicale.

CONTENU DE L'OUVRAGE

Cet ouvrage collectif est organisé en sept chapitres divisés en deux parties. La première, qui regroupe les trois premiers chapitres, présente des textes qui traitent de la mesure des compétences. La deuxième partie comporte, quant à elle, quatre chapitres liés à l'évaluation des compétences.

Le premier chapitre, rédigé par Julie Grondin, Eric Dionne, Jacinthe Savard et Lynn Casimiro, présente, de façon fort détaillée et justifiée, les analyses produites dans le contexte de la mesure de l'offre active de services de santé en français. La modélisation de Rasch a été mise à profit afin de discuter des propriétés métriques de cet instrument. Les auteurs décrivent avec moult détails une méthodologie susceptible d'être employée dans différents contextes d'apprentissage et, entre autres, en éducation médicale pour examiner les propriétés métriques des instruments de mesure développés. Leur exposé permet de mettre en exergue la complexité du processus de validation et l'importance de conserver des traces de cette démarche.

Jean-Sébastien Renaud présente, dans le cadre du chapitre 2, une comparaison entre la théorie classique des tests (TCT), largement utilisée en éducation médicale et en éducation en santé pour étudier les propriétés métriques des instruments, et la modélisation de Rasch. Celle-ci, moins connue, offre des avantages importants comparativement à la TCT. La démonstration qu'il met de l'avant s'appuie sur l'étude des scores issus de l'échelle de communication médecin-patient pour les externes en médecine (ECMP-EM).

Le dernier chapitre de la première partie est rédigé par Eric Dionne, Julie Grondin et Marie-Eve Latreille. Ils présentent les résultats d'une recherche sur les propriétés métriques d'un test de concordance de script (TCS) en s'appuyant sur la modélisation de Rasch. Ils présentent également les conséquences associées au type de modélisation des scores sur le processus d'optimisation du TCS, qui est un instrument de plus en plus développé et utilisé pour mesurer le jugement clinique en éducation médicale.

La deuxième partie de l'ouvrage débute avec un texte de Maxim Morin, André-Philippe Boulais et André F. De Champlain qui œuvrent tous, à titre de psychométriciens, au Conseil médical du Canada. Ils présentent un état des lieux concernant la correction automatisée dans le cadre de tâches d'évaluation authentiques et ouvertes dans le domaine de la santé. En effet, l'approche par compétences étant de plus en plus en vogue dans les programmes de formation en médecine et en santé, il devient impératif de concevoir des situations d'évaluation cohérentes avec cette approche, ce qui pose certains défis, entre autres, pour la correction et la notation. Les auteurs nous renseignent sur les plus récentes avancées dans ce domaine.

Le chapitre 5 est consacré à la question de la validité. Thomas Pennaforte et Nathalie Loye y décrivent une application du modèle de Kane dans le cadre d'une séance de simulation médicale. Après avoir présenté la nature et les caractéristiques de ce modèle, ils expliquent comment, concrètement, il est possible de l'utiliser pour discuter de la validité de l'interprétation des scores.

Le chapitre suivant, le sixième, est l'œuvre d'Isabelle Raîche et de Bernard Charlin. Ils y exposent les résultats de leurs travaux qui visaient à concevoir un outil formatif permettant de développer ou d'améliorer la prise de décision en contexte opératoire. Pour ce faire, ils se sont appuyés sur les concepts actuels de l'évaluation formative et sur l'approche par concordance. Ce chapitre présente également les résultats d'une étude pilote.

Dans le dernier chapitre, Maud Mediell et Eric Dionne abordent d'un point de vue théorique la question de l'évaluation de programme dans le domaine de la santé. Ils s'attardent aux défis actuels auxquels doivent faire face divers acteurs (parties prenantes, praticiens, etc.) au premier chef, les évaluateurs de programmes. En effet, ces derniers sont souvent situés au confluent de différentes traditions (p. ex. épistémologie, évaluation économique, évaluation de programmes), ce qui représente un défi de taille quand il s'agit de rendre compte de programmes qui sont, par définition, souvent complexes à évaluer.

BIBLIOGRAPHIE

Black, P. et D. Wiliam (2010). « Inside the Black Box : Raising Standards through Classroom Assessment », *Phi Delta Kappan Magazine*, *92*(1), p. 81-90, doi : 10.1177/003172171009200119.

Hodges, B.D. (2012). « The Shifting Discourses of Competence », dans B.D. Hodges et L. Lingard (dir.), *The Question of Competence : Reconsidering Medical Education in the Twenty-First Century*, New York : Cornell University Press, p. 14-41.

Legendre, R. (2005). *Dictionnaire actuel de l'éducation*, 3e éd., Montréal : Guérin.

Wilson, M. (2005). *Constructing Measures : An Item Response Modeling Approach*, Mahwah : Lawrence Erlbaum Associates.

PARTIE 1

LA MESURE DES COMPÉTENCES

CHAPITRE 1

Démonstration d'une méthodologie mettant à profit les modèles de Rasch

L'exemple d'une échelle de mesure de l'offre active de services de santé en français[1]

Julie Grondin, Eric Dionne, Jacinthe Savard
et Lynn Casimiro

Il existe peu d'instruments servant à mesurer l'offre active de services sociaux et de santé en français dans les communautés francophones en situation minoritaire (CFSM). Parmi ceux-ci, un seul outil ayant fait l'objet d'une publication scientifique a été construit afin d'évaluer les comportements individuels de l'offre active. Les propriétés métriques de cet outil ont principalement été analysées à l'aide de la théorie classique des tests (TCT). L'un des objectifs de cette étude consistait donc à vérifier dans quelle mesure une modélisation de type Rasch était appropriée pour établir les propriétés métriques de cet instrument. Dans un deuxième temps, nous souhaitions profiter du format du texte (un chapitre de livre) pour rendre cet examen aussi explicite que possible. En effet, l'espace offert pour une publication dans une revue, par exemple, ne permet pas toujours de fournir tous les détails voulus sur la démarche utilisée. Or, durant les analyses,

1. Cette étude a été rendue possible grâce à l'appui financier du Consortium national de formation en santé (CNFS), volet Université d'Ottawa et Secrétariat national, qui sont financés par Santé Canada dans le cadre de la *Feuille de route pour les langues officielles du Canada 2013-2018*. Les points de vue exprimés ici ne reflètent pas nécessairement ceux de Santé Canada.

plusieurs décisions doivent être prises par les analystes et ces décisions peuvent varier d'un analyste à l'autre. Notre deuxième objectif consistait donc à expliciter le plus possible la démarche que nous avons utilisée afin non seulement de la rendre facilement reproductible, mais aussi d'aider toute personne (étudiant ou praticien) à s'initier à une méthodologie mettant à profit une modélisation de type Rasch.

1. LE CONTEXTE

Au Canada, la langue et l'appartenance à un groupe linguistique minoritaire font partie des déterminants sociaux de la santé (Bouchard, Beaulieu et Desmeules, 2012). En effet, il peut être particulièrement difficile de communiquer ses besoins à un professionnel de la santé ou des services sociaux dans une langue avec laquelle on est plus ou moins à l'aise et cela aura un effet sur la qualité et la sécurité des soins qui seront reçus (Bowen, 2015 ; Drolet *et al.*, 2017, à paraître). Dans une situation de vulnérabilité, une personne peut souvent se sentir intimidée et ne pas oser demander l'accès à des soins dans sa langue (Bouchard *et al.*, 2012 ; Drolet *et al.*, 2017, à paraître). Ainsi, en contexte francophone minoritaire, le concept de l'offre active devient un élément central de l'accessibilité à des services en français : « l'offre active peut être considérée comme une invitation, verbale ou écrite, à s'exprimer dans la langue officielle de son choix. L'offre de parler dans la langue officielle de son choix doit précéder la demande de services » (Bouchard *et al.*, 2012, p. 46).

Si ce principe semble simple en théorie, il est parfois difficile à appliquer en pratique par les intervenants bilingues pratiquant dans des milieux anglodominants[2]. Ainsi, offrir des activités d'outillage à l'offre active de services en français et en mesurer les effets constituent des mesures importantes pour améliorer l'accès à des services en français dans les communautés francophones en situation minoritaire (CFSM) (Bouchard, Vézina et Savoie, 2010).

Il existe peu d'instruments servant à mesurer l'offre active de services sociaux et de santé en français dans les CFSM. Parmi ceux-ci, il semble n'y avoir qu'un seul outil ayant fait l'objet d'une publication qui discute de sa validation. Cet outil a été construit afin de permettre à des étudiants ou des intervenants d'évaluer leurs comportements

2. Pour une description détaillée des défis qui peuvent se poser, consulter Bouchard *et al.* (2010) ou Drolet *et al.* (2017, à paraître). Pour un résumé, voir Savard *et al.* (2014, p. 87-89).

individuels d'offre active (Savard *et al.*, 2014 ; Savard *et al.*, 2015). Dans le cadre des travaux liés à la conception et à la validation de cet outil, les données ont principalement été traitées à l'aide de la théorie classique des tests (TCT). Or, cette théorisation, bien qu'intéressante, possède des limites largement documentées dans les écrits scientifiques comme le mentionnent, entre autres, Cano *et al.* (2013). L'un des objectifs de cette étude consistait donc à examiner dans quelle mesure une modélisation basée sur les travaux de Rasch (1960) était appropriée pour examiner les propriétés métriques de cet instrument. En effet, cette modélisation permet, d'une part, de vérifier si les items permettent de produire une échelle de mesure objective à intervalles égaux et, d'autre part, d'examiner les remédiations permettant d'améliorer la qualité objective de la mesure.

Dans un deuxième temps, nous souhaitions profiter du format du texte (un chapitre de livre) pour rendre cet examen aussi explicite que possible. En effet, l'espace offert pour une publication dans une revue, par exemple, ne permet pas toujours de fournir tous les détails voulus sur la démarche utilisée. Or, durant les analyses, plusieurs décisions doivent être prises par les analystes et ces décisions peuvent varier d'un analyste à l'autre. Notre deuxième objectif consistait donc à expliciter le plus possible la démarche que nous avons utilisée afin non seulement de la rendre facilement reproductible, mais aussi d'aider toute personne (étudiant ou praticien) à s'initier à une méthodologie mettant à profit une modélisation de type Rasch.

2. LE CADRE THÉORIQUE

2.1. Les modèles de la famille de Rasch

Les modèles de la famille de Rasch sont des modèles probabilistes qui visent à déterminer si les scores bruts recueillis par un instrument, souvent de nature ordinale, peuvent se situer sur une échelle à intervalles égaux. Le tableau 1.1 montre que, tout comme les modèles issus de la TCT, les modèles de la famille de Rasch s'appuient sur les deux postulats suivants : l'unidimensionnalité et l'indépendance locale. Mais, contrairement à la TCT, lorsque ces deux postulats sont respectés, les données issues de l'analyse de Rasch possèdent une propriété importante : l'invariance. Cela signifie que les paramètres des items sont indépendants des groupes de sujets à qui l'on a administré l'instrument. De plus, il est possible d'espérer obtenir des données pouvant se situer sur une échelle unique à intervalles égaux qui situent autant le niveau d'offre active mesuré par les items que le niveau d'offre active déclaré par les

répondants[3]. Aussi ce type de modélisation s'applique-t-il bien au développement d'une nouvelle échelle de mesure (Tennant et Connaghan, 2007), comme c'est le cas ici pour la mesure de l'offre active.

Tableau 1.1
Comparaison entre la TCT et la modélisation de Rasch

	Théorie classique des tests		Modélisation de type Rasch	
Caractéristique/ principe	**Oui/ Non**	**Conséquence**	**Oui/ Non**	**Conséquence**
Estimation des paramètres, des items et des sujets	Non	Équivalence de la mesure incertaine sur l'échelle.	Oui	Tous les scores, items et sujets sont sur une échelle commune.
Statistique d'ajustement pour les sujets	Non	Pas de statistique d'ajustement pour les sujets.	Oui	Possibilité d'identifier les sujets à soustraire de la modélisation.
Score total est une statistique suffisante (*sufficient statistic*)	Non	Différents vecteurs de réponses peuvent donner le même score total.	Oui	Scores élevés/faibles représentent un construit élevé/faible.
Invariance	Non	Dépendant des groupes de sujets.	Oui	Indépendant des groupes de sujets.
Unidimensionnalité et indépendance locale des items	Oui	Conditions importantes.	Oui	Conditions importantes.

Source : Traduit et adapté de Cano *et al.*, 2013.

2.2. La démarche de modélisation des scores de Tennant et Conaghan

Dans leur article de 2007, Tennant et Conaghan ont relevé sept aspects importants que doit présenter une analyse de type Rasch. Ces aspects forment les assises méthodologiques utilisées pour nos analyses. Bien que nous comptions reprendre et expliciter les recommandations de chacun de ces aspects dans les sections qui suivent, nous en proposons tout de même un sommaire ci-dessous :

1. Fournir une description du modèle choisi et expliciter les raisons ayant motivé le choix du modèle.

3. Le masculin est utilisé dans ce texte dans le seul but d'en alléger la lecture.

2. Effectuer une analyse de la qualité de l'ajustement des données au modèle[4] choisi, tant pour les sujets que pour les items, et une justification des choix (méthode d'analyse, retrait ou non de sujets ou d'items, intervalle d'ajustement[5]).
3. S'assurer d'avoir des échelles de réponses bien ordonnées (c'est-à-dire monotone croissante), effectuer un recodage de celles-ci au besoin (c'est-à-dire penser à regrouper deux ou trois catégories afin d'améliorer l'échelle et de corriger les problèmes trouvés) et expliciter les choix.
4. Fournir une démonstration de la vérification de la condition d'indépendance locale, ainsi que de la vérification de l'indépendance des réponses et de la condition d'unidimensionnalité nécessaires à l'utilisation de ce type de modèle.
5. Effectuer une vérification de la présence ou non d'un fonctionnement différentiel d'items (FDI) et expliciter les actions prises pour les corriger s'il y a lieu.
6. Fournir une description de la qualité de l'échelle de mesure par une bonne mise en correspondance entre les items et les sujets.
7. Fournir une analyse des indices de fidélité pour les sujets et les items.

2.3. Les modèles choisis

Il existe de nombreux modèles (*partial credit*, *rating scale*, etc.) dérivés de la méthodologie initiale développée par Rasch. Une fois le modèle choisi, le chercheur doit réaliser un processus itératif qui consiste à examiner dans quelle mesure le modèle s'applique adéquatement aux données brutes. Concrètement, il s'agit d'examiner les avantages et les limites de différents modèles et de décider lequel semble le plus approprié. Le modèle choisi peut nécessiter certains ajustements (le regroupement de certaines catégories de réponses par exemple). Tennant et Conaghan recommandent donc d'expliciter le choix du modèle et les raisons qui ont conduit à ce choix.

Au regard des items inclus dans le questionnaire de l'offre active que nous avons analysé, il nous semblait que le modèle à crédit partiel (*Partial Credit* ou PC) de Masters (1982) et le modèle *Rating Scale* (RS)

4. Nous n'entrerons pas dans le débat pour déterminer s'il faut discuter de l'ajustement entre le modèle et les données ou entre les données et le modèle.
5. Notons que dans le texte de Tennant et Conaghan, les éléments 2 et 3 sont inversés par rapport à celui du présent texte. Cependant, puisqu'il est d'usage de commencer une analyse de Rasch par l'étude de la qualité de l'ajustement, il nous semblait préférable de présenter les recommandations dans l'ordre où elles sont réalisées.

d'Andrich (1978) étaient les plus appropriés pour modéliser les scores bruts. Compte tenu de l'objectif du présent texte, qui est de rendre explicite toute la démarche utilisée et de la mettre à l'épreuve, il nous apparaissait pertinent d'explorer ces deux modèles afin de pouvoir en comparer les résultats.

2.3.1. *Le modèle* Rating Scale

Le modèle *Rating Scale* s'applique lorsque l'échelle de réponses demeure constante pour tous les items du questionnaire, ce qui est le cas avec celui de l'offre active. L'équation 1 présente l'équation générale de ce modèle.

Équation 1 : modèle *Rating Scale*

$$P_{nix} = \frac{e^{B_n - D_i - F_x}}{1 + e^{B_n - D_i - F_x}}$$

Où P_{nix} représente la probabilité que la personne n avec un niveau d'offre active B_n adhère à la catégorie de réponse x (où $x = 0$ à m-1, pour m catégories de réponses offertes) d'un item i dont le degré de difficulté (à adhérer à un certain niveau d'offre active) est D_i. Le paramètre F_x correspond au point, sur l'échelle de réponses, où la probabilité d'opter pour l'une ou l'autre des catégories est égale.

2.3.2. *Le modèle à crédit partiel (ou* Partial Credit*)*

Bien que les catégories de réponses du questionnaire soient les mêmes pour tous les items, le contexte, le construit mesuré par chaque item et le niveau d'offre active de chaque individu dans ces différents contextes nous apparaissaient une source pouvant faire varier la signification des catégories offertes et, par conséquent, le continuum associé à chacun. Il nous semblait donc pertinent d'expérimenter aussi le modèle à crédit partiel qui permet à l'intervalle entre chaque catégorie de réponses de varier d'un item à l'autre. L'équation 2 présente l'équation du modèle général à crédit partiel également mis à profit dans le cadre de nos analyses. Il s'agit essentiellement d'une équation qui ressemble à l'équation 1, à la différence que l'intervalle entre chaque catégorie de réponses peut varier d'un item à l'autre.

Équation 2 : modèle *Partial Credit*

$$P_{nijk} = \frac{e^{B_n - D_i - F_{ix}}}{1 + e^{B_n - D_i - F_{ix}}}$$

Où P_{nix} représente la probabilité que la personne *n* avec un niveau d'offre active B_n adhère à la catégorie de réponse *x* (où *x* = 0 à *m*-1, pour *m* catégories de réponses offertes) d'un item *i* dont le degré de difficulté (à adhérer à un certain niveau d'offre active) est D_i. Le paramètre F_{ix} correspond au point, pour cet item, où la probabilité d'opter pour l'une ou l'autre des catégories est égale.

2.4. La qualité de l'ajustement des données au modèle

Les statistiques d'ajustement modèle-données nous renseignent sur le degré de concordance entre le modèle théorique choisi et les scores modélisés. La plupart sont basées sur un test du khi carré qui mesure l'association entre l'estimation des paramètres du modèle et du critère (sujets ou items). Plus l'ajustement est adéquat, plus les données s'ajustent bien au modèle et plus la qualité de la mesure s'en trouve bonifiée. Lorsque les ajustements sont peu convaincants, il est possible de changer de modèle afin de déterminer si un autre serait plus adéquat. Deux statistiques ont été utilisées aux fins de nos analyses : la statistique *infit* et la statistique *outfit*.

La statistique *infit* (*inlier fit*) est un indice d'ajustement qui met l'accent sur les réponses inattendues qui sont près du patron de réponses habituel d'un sujet ou d'un item. Par exemple, un item, dont le niveau d'offre active correspond à celui d'un sujet qui n'est pas appuyé par le sujet alors qu'on aurait pu s'attendre à ce qu'il le soit. Il s'en suit que les problèmes décelés grâce à cet indice sont généralement difficiles à diagnostiquer et à corriger.

La statistique *outfit* (*outlier fit*) met l'accent sur les valeurs aberrantes qui s'éloignent du patron de réponses habituel d'un sujet ou d'un item. Par exemple, un sujet dont le niveau d'offre active est faible qui endosse facilement un item difficile à endosser. De façon générale, il s'agit de la première statistique d'ajustement à examiner puisqu'on souhaite généralement éliminer de la modélisation les sujets (ou les items) dont les données sont aberrantes et apportent ainsi peu d'information utile pour construire une mesure de qualité.

Pour procéder à l'analyse de la qualité de l'ajustement, nous avons suivi les recommandations générales formulées par Linacre (2015b) qui suggère d'étudier :

1. les données qui présentent des corrélations négatives ;
2. les problèmes d'ajustement *outfit* avant les *infit* ;
3. les indices d'ajustement basés sur le carré moyen (CM) avant d'étudier la version standardisée (ZSTD) ;

4. les valeurs élevées des indices d'ajustement avant les valeurs plus faibles ou négatives. Il y a effectivement une asymétrie quant à l'implication d'une valeur élevée associée à un problème d'ajustement par rapport à une valeur plus faible : la première présente une plus grande menace pour la validité que la seconde.

Cependant, en ce qui concerne les indices d'ajustement (recommandation 3 de la liste ci-dessus), nous avons décidé de suivre la recommandation de Smith, Schumacker et Busch (1998), ainsi que Smith et Suh (2003), qui stipulent que la version standardisée des indices d'ajustement est plus stable que celle basée sur le carré moyen, pour des tailles d'échantillons différentes, et qu'elle détecte mieux certains problèmes d'ajustement. Le travail d'analyse de la présente étude étant exploratoire, cela nous permettait de conserver plus d'items.

Pour l'interprétation de la version standardisée des indices d'ajustement, il faut savoir que la valeur attendue est de 0,0 (il s'agit d'une statistique normalisée et, donc, centrée et réduite). Une valeur supérieure à 0,0 signifie que les données sont difficiles à prédire parce qu'elles présentent plus de variations que ce qui est attendu par le modèle. Une valeur inférieure à 0,0 signifie que les données sont redondantes et trop facilement prédites par le modèle (Linacre, 2002). Plus précisément, l'interprétation de cette version des indices d'ajustement se fait au regard des balises présentées au tableau 1.2. Nous nous sommes appuyés sur les recommandations de Linacre et avons, dans la plupart des cas, retiré les sujets et les items dont les valeurs de l'ajustement n'étaient pas comprises dans l'intervalle [–2, 2].

Tableau 1.2
Interprétation des indices d'ajustement basés sur la statistique standardisée

Intervalle	Qualité de la mesure
≥ 3,0	Les données† sont très inattendues. Certains facteurs ont pu contrevenir à la mesure (réponses au hasard, inattention, etc.).
2,0 – 2,9	Les données sont difficiles à prédire, et ce, de façon notable.
–1,9 – 1,9	Les données sont raisonnablement prédites.
≥ –2,0	Les données sont trop faciles à prédire. D'autres dimensions (facteurs) ou des problèmes de dépendance locale peuvent contraindre les données dans des patrons de réponses qui les rendent excessivement prévisibles.

† Implicitement ici, il est question des données modélisées.

Source : Traduit et adapté de Linacre, 2002.

Enfin, Rojas Tejada *et al*. (2002) ont étudié deux stratégies pour l'analyse de la qualité de l'ajustement, à savoir commencer par l'analyse de l'ajustement :

1. des items, poursuivre avec celle des sujets, puis terminer avec l'ajustement global des données au modèle ;
2. des sujets, poursuivre avec celle des items, puis terminer avec l'ajustement global des données au modèle.

Leurs résultats ont révélé que la première stratégie permettait de maximiser le nombre de sujets conservés, alors que la seconde permettait de maximiser le nombre d'items conservés. C'est donc la seconde stratégie qui a été retenue pour la présente analyse. En effet, les items (plus que les sujets) constituent l'objet principal de notre analyse puisque nous souhaitons vérifier s'ils permettent de produire une échelle de mesure objective de l'offre active de services en français. Le tableau 1.3 qui suit résume la stratégie d'analyse au regard de la qualité de l'ajustement.

Tableau 1.3
Résumé de la stratégie d'analyse de la qualité de l'ajustement retenue pour les sujets et les items

Étapes†	**Indices**	**Balises de décisions**	
		Optimal	**Problématique**
1	Corrélation point-bisériale	≥ 0	< 0
2	*outfit* standardisé	[–2, 2]	≥ 2
3	*outfit* standardisé	[–2, 2]	≤ –2
4	*infit* standardisé	[–2, 2]	≥ 2
5	*infit* standardisé	[–2, 2]	≤ –2

† Les étapes s'appliquent d'abord sur les sujets, puis sur les items.

2.5. Les échelles de réponses

Lors de la modélisation des données, les résultats obtenus peuvent révéler que l'utilisation des catégories de réponses telles que proposées sur le questionnaire est différente de ce à quoi on se serait attendu. Les répondants peuvent, par exemple, avoir eu du mal à différencier certaines catégories de réponses. Afin d'étudier si ces dernières sont bien ordonnées, Linacre (2015b) propose les recommandations suivantes ; elles sont basées sur l'expérience de différents chercheurs lors de leur étude des catégories de réponses :

1. Il convient d'avoir un minimum de dix observations par catégorie de réponses afin d'assurer une certaine stabilité dans les résultats[6]. Une fréquence moindre pourrait donner des résultats qui ne peuvent pas être reproduits.
2. Il faudrait observer une distribution uniforme des fréquences dans chacune des catégories. Une distribution de fréquence en dents de scie peut indiquer que les catégories sont définies de façon trop étroite et que les répondants ne distinguent pas bien les différentes réponses offertes.
3. La mesure moyenne associée à chacune des catégories de réponses devrait clairement être croissante. Il ne devrait pas y avoir d'inversion. De façon plus précise, Linacre (2004) soutient que la croissance devrait être d'au moins 1 *logit* pour une échelle comprenant cinq catégories de réponses différentes afin d'avoir une bonne distinction entre les catégories, mais de moins de 5 *logit* pour éviter les « points morts » ou les « trous » dans l'échelle de réponses.
4. La mesure moyenne observée pour chacune des catégories de réponses devrait avoir une valeur proche de la mesure moyenne attendue par le modèle.
5. Les données observées dans chacune des catégories devraient présenter une bonne qualité d'ajustement au modèle. La valeur des indices *outfit* basés sur le carré moyen devrait être près de 1,0. Une valeur largement au-dessus de 1,0 est beaucoup plus problématique qu'une valeur largement en dessous de 1,0.

Pour compléter les recommandations de Linacre, nous avons choisi d'analyser les catégories de réponses de façon graphique. Selon Park (2004), l'analyse graphique des catégories de réponses consiste à vérifier si la courbe de probabilité de chacune des catégories de réponses possède un sommet distinct et si les courbes de probabilité

6. Notons que cette recommandation n'est généralement pas problématique pour le modèle RS puisque l'échelle de réponse demeure constante pour tous les items (Linacre, 1994). Par contre, pour le modèle PC, où l'échelle de réponses peut varier d'un item à l'autre, la taille d'échantillon est à surveiller. Linacre précise qu'un échantillon de 50 sujets bien ciblés pourrait permettre d'obtenir des résultats utiles et stables. Blais et Grondin (2010) sont arrivés à une conclusion similaire. Linacre ajoute que 30 sujets pourraient même être suffisants pour des études pilotes bien conçues. Ajoutons qu'il est toujours possible de renforcer les résultats d'une étude comportant un échantillon de petite taille à l'aide de simulations effectuées sur une centaine d'échantillons. Enfin, toujours selon Linacre, l'indice de séparation et l'indice de fidélité (section 4.7) peuvent donner une certaine indication sur la taille d'échantillon. En effet, si la séparation est moindre que 0,3 et la fidélité, inférieure à 0,9, alors la taille d'échantillon ne serait pas assez grande pour confirmer au moins trois niveaux de difficulté (validité de construit) dans les items : faible, moyen et fort.

de l'ensemble des catégories apparaissent comme une suite de collines équidistantes. Si tel est le cas, pour chacune des portions du continuum de l'offre active, une des catégories de réponses aura plus de chances que les autres d'être choisie par les répondants. En revanche, si la courbe de probabilités d'une des catégories de réponses ne possédait pas un sommet distinct qui s'élève au-dessus des courbes de probabilités des catégories adjacentes, cela indiquerait que cette catégorie n'a jamais plus de chances que les autres d'être choisie, et ce, sur tout le continuum.

Lorsque les catégories de réponses présentent des problèmes d'ordonnancement, une solution possible est de regrouper certaines catégories (Linacre, 2015b)[7]. Pour ce faire, il importe de suivre les recommandations suivantes :

1. analyser la définition des différentes catégories. Il ne faut pas regrouper des catégories qui n'ont aucun sens lorsqu'elles sont amalgamées (par exemple « En accord » et « En désaccord ») ;
2. examiner la fréquence d'observation de chacune des catégories. Il est préférable de combiner des catégories de petites fréquences avec des catégories de grandes fréquences plutôt que des catégories de grandes fréquences entre elles ;
3. vérifier la mesure moyenne des différentes catégories. Il est préférable de regrouper des catégories dont la mesure moyenne présente une inversion ou des catégories dont la mesure moyenne est très proche, plutôt que des catégories dont la mesure moyenne est ordonnée et dont les valeurs sont éloignées ;
4. se rappeler que les inversions de catégories peuvent être liées à l'échantillon. Il faut donc rester prudent et éviter de regrouper des catégories qui ne le seraient pas en général ou ne le seront pas dans le futur.

Dans le cadre de nos analyses, les balises du tableau 1.4 nous ont guidés dans nos choix.

7. À ce point-ci, il nous apparaît pertinent de faire une petite remarque. Nous sommes conscients que plusieurs des recommandations que nous avons choisies d'utiliser proviennent du même auteur. Il ne s'agit pas d'un choix politique ou volontaire. Nous avons consulté différents ouvrages liés à la modélisation de Rasch, tels que Bond et Fox (2015), Smith Jr. et Smith (2004) ou Boone, Staver et Yale (2014), pour n'en citer que quelques-uns, et soit ces auteurs se réfèrent au même auteur que nous, soit ils ne formulent aucune recommandation.

Tableau 1.4
Résumé de la stratégie d'analyse des échelles de réponses retenue

Éléments à observer pour chacune des catégories de réponses	Balises de décisions	
	Optimal	**Problématique**
Nombre d'observations	≥ 10	< 10
Distribution de la fréquence des observations	Uniforme	Non uniforme
Mesure moyenne estimée	Croissante et d'une valeur entre 1 et 5 *logit*	Présence d'inversions, points morts ou trous dans l'échelle
Mesure moyenne observée	Proche de la mesure moyenne attendue	Différente de la mesure moyenne attendue
Qualité de l'ajustement	À l'intérieur de l'intervalle choisi	À l'extérieur de l'intervalle choisi†
Courbes de probabilité	Sommets distincts et équidistants	Sommets superposés ou à distances inégales

† La valeur des indices basés sur le carré moyen devrait être près de 1,0. Une valeur largement au-dessus de 1,0 est beaucoup plus problématique qu'une valeur largement en dessous de 1,0. Pour les indices standardisés, ce sont des valeurs comprises dans l'intervalle [–2, 2] qui sont recherchées.

2.6. Les postulats d'unidimensionnalité et d'indépendance locale

La modélisation de type Rasch repose sur deux postulats importants qui la rendent attrayante au regard d'une mesure objective, à savoir l'unidimensionnalité et l'indépendance locale.

2.6.1. *L'unidimensionnalité*

Le premier postulat, l'unidimensionnalité, stipule que les données doivent pouvoir être expliquées par un seul et même facteur (ou trait latent) : le niveau d'offre active (endossé par individu ou mesuré par un item). En effet, si d'autres dimensions influencent les réponses d'un individu au questionnaire, l'estimation des paramètres fournie par le modèle risque d'être déformée. La présence de sous-dimensions, quoique possible, mérite donc d'être étudiée afin d'en examiner les répercussions sur l'échelle de mesure obtenue.

Pour ce faire, il convient de regarder la variance inexpliquée associée aux différents contrastes de l'analyse en composantes principales effectuée sur les résidus standardisés, c'est-à-dire une fois le « facteur Rasch » extrait (Linacre, 2015b). Lorsque la valeur propre des différents contrastes est supérieure à 2, cela indique la présence potentielle de

sous-dimensions mesurées par les items. En effet, une valeur propre équivalente à 2 correspond à un regroupement de deux items, ce qui constitue le nombre minimal d'items pour former une sous-dimension.

Afin de repérer les items qui sont susceptibles de former des sous-dimensions, il convient d'analyser, dans un premier temps, le degré de corrélation de chacun des items avec les différents contrastes de l'analyse en composantes principales. Les items ayant un poids d'au moins 0,40 associé à un facteur sont généralement considérés comme importants. Il faut alors porter une attention à ces items (contenu, choix de mots) afin d'évaluer s'il s'agit réellement de dimensions secondaires ou simplement d'aspects associés à l'offre active (comme le seraient l'addition et la soustraction dans une épreuve d'arithmétique) ou d'un artéfact associé à la formulation des items.

Ensuite, le logiciel Winsteps définit, pour chacun des contrastes de l'analyse en composantes principales, trois groupes d'items : ceux qui corrèlent fortement avec le facteur, ceux qui corrèlent moyennement et ceux qui corrèlent faiblement. Pour chacun de ces groupes, la mesure des répondants est modélisée. Les mesures obtenues pour les différents groupes sont alors mises en relation à l'aide d'une corrélation dite « atténuée », c'est-à-dire que l'erreur de mesure inhérente à la modélisation est corrigée de façon statistique. Si le résultat s'approche de 1,0, alors les items des deux groupes seraient statistiquement les mêmes. Autrement dit, on ne pourrait pas rejeter l'hypothèse nulle voulant que les deux groupes d'items mesurent la même chose, ce qui soutiendrait le postulat d'unidimensionnalité.

Le plus souvent, c'est la valeur de la corrélation entre les mesures des répondants modélisés par les deux groupes d'items les plus opposés, c'est-à-dire ceux qui corrèlent le plus et le moins avec le facteur, qui sera étudiée afin de vérifier si les items mesurent la même chose ou non. L'esquisse suivante, proposée par Linacre (2015b), nous servira à interpréter la force de la corrélation atténuée fournie par le logiciel : en deçà de 0,57, les deux groupes d'items montrent moins de la moitié de leur variance en commun (0,57*0,57 = 32,5 %). Autrement dit, ils sont plus indépendants que dépendants. Au-dessus de 0,71, la mesure des personnes provenant des deux groupes d'items a un peu plus de la moitié de leur variance en commun (50,4 %), de telle sorte qu'ils sont un peu plus dépendants qu'indépendants. Au-dessus de 0,82, ils sont deux fois plus dépendants qu'indépendants (67,2 % de variance commune). Au-dessus de 0,87, ils sont trois fois plus dépendants qu'indépendants (75 % de variance commune).

2.6.2. *L'indépendance locale*

L'indépendance locale est le deuxième postulat associé à ce type de modélisation que l'on doit vérifier. Cette condition requiert que les réponses d'un individu soient statistiquement indépendantes les unes des autres. En effet, si la réponse d'un individu à l'un des items du questionnaire permet de prédire la réponse qu'il fournira à d'autres items de ce questionnaire, c'est qu'il y a un risque que d'autres dimensions (autre que celle visée par le questionnaire) influencent les réponses d'un individu. Encore une fois, l'estimation des paramètres fournie par le modèle risque alors d'être déformée. La présence d'une possible dépendance locale mérite donc d'être étudiée afin d'en examiner l'incidence sur l'échelle de mesure.

Comme pour l'analyse de l'unidimensionnalité, l'étude de l'indépendance locale entre deux items s'effectue sur la matrice des résidus standardisés, c'est-à-dire une fois le facteur Rasch extrait. De façon générale, il faut que la corrélation inter-items effectuée à partir des résidus standardisés soit positive et d'une valeur d'au moins 0,7 avant de commencer à parler de dépendance locale entre les items (Linacre, 2015b).

Le tableau 1.5 présente un résumé de la stratégie que nous avons retenue pour l'étude des postulats d'unidimensionnalité et d'indépendance locale.

Tableau 1.5
Résumé de la stratégie d'analyse de l'unidimensionnalité et l'indépendance locale retenue

Éléments à observer	Balises de décisions	
	Optimal	**Problématique**
Variance inexpliquée associée aux différents contrastes de l'analyse en composantes principales effectuée sur les résidus standardisés.	Valeur propre < 2	Valeur propre ≥ 2
Niveau de corrélation des items avec chacun des contrastes.	< 0,40	≥ 0,40
Corrélation atténuée entre la mesure des sujets associée aux items fortement et faiblement corrélés à chacun des contrastes.	Près de 1 (ou ≥ 0,71)	Près de 0 (ou < 0,71)†
Corrélation inter-items effectuée sur les résidus standardisés.	< 0,70	≥ 0,70

† Une valeur de 0,71 correspond à une variance commune de 50,4 % et donc à des items un peu plus dépendants qu'indépendants (Linacre, 2015b).

Source : Traduit et adapté de Linacre, 2015b.

2.7. Le fonctionnement différentiel d'items (FDI)

Le fonctionnement différentiel d'items (FDI) est une autre menace pour la validité d'une échelle de mesure qu'il importe de surveiller. En effet, si un item présente un potentiel de fonctionnement différentiel, cela signifie que la performance de deux groupes d'individus, qui sont théoriquement équivalents sur le plan de la mesure, diffère. Encore une fois, cela risque d'engendrer du bruit au regard des paramètres estimés par le modèle.

Deux tests d'hypothèses proposés par le logiciel Winsteps nous semblent intéressants pour tenter de déceler la présence ou non d'un biais potentiel dans les items lorsque quatre groupes de sujets sont en présence, comme c'est le cas dans notre étude.

Le premier test est basé sur l'hypothèse qu'un item aurait le même degré de difficulté que son degré moyen de difficulté à travers les différents groupes étudiés. Autrement dit, le degré de difficulté trouvé pour cet item à l'aide d'un groupe de répondants en particulier serait le même que celui trouvé pour l'ensemble des répondants. Pour ce test, deux éléments sont à surveiller selon Linacre (2015b), à savoir : 1) un niveau de signification du test assez petit pour confirmer que la différence trouvée entre les deux mesures (celle d'un groupe par rapport à l'ensemble des groupes) n'est pas simplement le fruit du hasard ; 2) une différence entre les deux mesures qui soit suffisamment grande pour avoir des répercussions majeures sur l'interprétation de la mesure. De façon générale, les balises suivantes sont utilisées pour interpréter les résultats de ce test : une différence d'au moins 0,43 *logit* correspond à un potentiel FDI qualifié de léger à modéré (aussi appelé FDI de catégorie B) ; une différence d'au moins 0,64 *logit* correspond à un potentiel FDI de modéré à grand (FDI de catégorie C). Autrement, le FDI est considéré comme nul ou négligeable.

L'hypothèse du deuxième test stipule que l'item ne possède aucun FDI à travers les différents groupes de sujets. Le résultat est résumé comme une statistique du khi carré pour chacun des items indiquant si le FDI potentiel trouvé est significatif ou non. Une valeur moindre que le seuil généralement fixé à 0,05 révèle la présence d'une différence statistiquement significative entre les différents groupes.

Du point de vue de la mesure, l'objectif d'un questionnaire est généralement d'avoir un ensemble d'items qui visent à mesurer un seul et même trait latent (dans ce cas-ci l'offre active), mais également des items qui permettent d'obtenir, chacun séparément, une information donnée. Inclure un même item deux fois, par exemple, serait redondant et, dans ce cas-ci, créerait aussi un problème de dépendance locale (voir section précédente). Un autre problème associé à la redondance

est qu'elle s'accompagne généralement de « trous » dans l'échelle de mesure. Ainsi, certains niveaux d'offre active ne peuvent être mesurés de façon précise puisqu'aucun item n'est associé à ces niveaux.

Le tableau 1.6 présente un résumé de la stratégie retenue pour l'étude du fonctionnement différentiel des items de cet instrument.

Tableau 1.6
Résumé de la stratégie d'analyse du fonctionnement différentiel d'items retenue

Tests d'hypothèse	**Balises de décisions**	
	Optimal	**Problématique**
Le degré de difficulté moyen d'un item pour un groupe de sujets est le même que pour l'ensemble des sujets.	|FDI| < 0,43 *logit* ou $p \geq 0,05$	|FDI| ≥ 0,64 *logit* et $p < 0,05$ (FDI sévère)
		|FDI| ≥ 0,43 *logit* et $p < 0,05$ (FDI modéré)
L'item ne possède aucun FDI à travers les différents groupes de sujets.	$p \geq 0,05$	$p < 0,05$

FDI = différence entre le degré de difficulté moyen et celui trouvé pour un groupe en particulier ; p = niveau de signification du test d'hypothèse.
Source : Traduit et adapté de Linacre, 2015b.

2.8. Les indices de fidélité

Dans les analyses de Rasch, les statistiques de fidélité ne s'appuient pas sur l'indice alpha de Cronbach comme dans les analyses TCT, mais plutôt sur l'indice de séparation. L'indice de séparation peut s'appliquer autant aux sujets (*person separation*) qu'aux items (*item separation*). La valeur s'étend de 0 à l'infini. Dans tous les cas, une valeur élevée est généralement souhaitable. Le tableau 1.7 présente les valeurs suggérées par Boone, Staver et Yale (2014) concernant l'indice de séparation pour les sujets (*person separation*).

Tableau 1.7
Balises pour l'interprétation de l'indice de séparation

	Indice de séparation des sujets
Acceptable	1,50
Bon	2,00
Excellent	3,00

Source : Boone, Staver et Yale, 2014, p. 231.

L'indice de séparation pour les sujets sert à classer l'échantillon des sujets. Lorsque la valeur est faible, cela veut généralement dire que l'épreuve ne permet pas de distinguer les sujets très performants de ceux qui le sont beaucoup moins. On recommande alors d'ajouter des items (Boone, Staver et Yale, 2014).

L'indice de séparation pour les items sert, quant à lui, à déterminer la hiérarchie entre les items. Boone, Staver et Yale (2014, p. 227) indiquent qu'une valeur inférieure à 3 signifie que l'échantillon de sujets est trop faible pour confirmer la hiérarchie des items. Une valeur de 1,5 est recommandée pour des analyses par sujet et une valeur de 2,5 est minimale pour une analyse par groupe.

3. LA MÉTHODOLOGIE

3.1. Les sujets

Les données analysées ont été recueillies à l'aide d'un questionnaire visant à mesurer les comportements individuels d'offre active de services en français d'intervenants ou de futurs intervenants en santé et en services sociaux qui œuvrent dans des provinces où les francophones sont minoritaires. Trois groupes d'intervenants ont été interrogés : 1) des étudiants en médecine de trois programmes de formation en Ontario et au Nouveau-Brunswick lors d'un stage d'externat de troisième année ; 2) des diplômés des programmes de formation universitaire en santé ou services sociaux appuyés par le Consortium national de formation en santé (CNFS) en Ontario et au Nouveau-Brunswick ; 3) des stagiaires en ergothérapie d'une université en Ontario lors de leur deuxième ou troisième stage. C'est à ce dernier groupe d'étudiants qu'on a demandé de remplir le questionnaire deux fois, après deux périodes de stage différentes. Par conséquent, certains stagiaires auront répondu uniquement en fonction de leur deuxième stage, d'autres uniquement en fonction de leur troisième stage, et certains auront répondu à la suite de chacun des deux stages. Puisque ces étudiants remplissaient le questionnaire en fonction de deux milieux de stage différents, on peut penser que leur appréciation de leurs comportements d'offre active n'est pas la même selon l'expérience clinique vécue. Au total, ce sont les données de ces 152 personnes qui ont été analysées (43 étudiants en médecine, 60 diplômés des programmes du CNFS, 25 stagiaires en ergothérapie lors de leur deuxième stage et 24 stagiaires en ergothérapie lors de leur troisième stage).

3.2. L'instrument

Le questionnaire utilisé est un outil qui a été construit afin d'évaluer les comportements individuels de l'offre active d'intervenants ou de futurs intervenants en santé et services sociaux durant tout le continuum de soin[8]. Ce questionnaire est composé de 23 questions dont dix portent sur les comportements d'offre active liés à l'accueil et à la prise en charge des patients, neuf sur les comportements de l'offre active durant l'intervention auprès des patients et quatre sur les comportements d'offre active durant le soutien et l'aiguillage des patients vers d'autres ressources. Pour la majorité des énoncés, les catégories de réponses offertes consistent en une échelle de fréquence de type Likert à quatre niveaux (4 = Toujours ; 3 = Souvent ; 2 = Rarement ; 1 = Jamais). L'option « Ne s'applique pas » était également offerte puisque certains comportements visés par le questionnaire peuvent ne pas faire partie des tâches de certains répondants (par exemple, l'utilisation d'outils d'évaluation standardisée ou l'aiguillage des patients vers d'autres ressources ne font pas partie de tous les types d'emploi en santé et services sociaux).

3.3. Le logiciel

C'est le logiciel Winsteps 3.90.2.0 (Linacre, 2015a) qui a été mis à profit pour réaliser les analyses.

4. LES RÉSULTATS ET LA DISCUSSION

4.1. Les modèles utilisés

Quatre modélisations ont été effectuées en tout et elles sont résumées au tableau 1.8. Le premier modèle utilisé pour l'analyse des données est le modèle à crédit partiel (PC). Il nous semblait le modèle le plus utile pour obtenir un maximum de détails sur le fonctionnement de chacun des items. En effet, en raison des différents aspects couverts par les questions sur l'offre active, il nous apparaissait probable que l'utilisation et l'interprétation de l'échelle de réponse faite par les répondants puissent varier d'un item à l'autre. Au regard des résultats obtenus à l'aide de cette première modélisation, certains regroupements de catégories se sont révélés nécessaires. Ainsi, une deuxième analyse suivant

8. Pour plus de détails sur le développement de cet outil et sa validation, voir Savard *et al.* (2014, 2015).

le modèle à crédit partiel a été effectuée. Nous avons nommé cette analyse « PC-regroup ». Le détail de ces regroupements est présenté plus loin.

Les résultats de cette deuxième analyse ont montré que, pour un item, un nouveau regroupement de catégories se révélait nécessaire. Une troisième analyse suivant le modèle à crédit partiel a donc été effectuée (PC-regroup2). Enfin, étant donné que dans l'analyse des catégories de réponses item par item, nous étions souvent confrontés à un pourcentage d'utilisation trop faible des différentes catégories de réponses, nous avons décidé d'explorer les résultats d'une analyse suivant la modélisation *Rating Scale* (RS).

Tableau 1.8
Résumé des modélisations

Modélisation	Nom	Explication
1	PC	Pour étudier en détail l'utilisation de l'échelle de réponse.
2	PC-regroup	Pour corriger les problèmes liés à l'utilisation de l'échelle trouvés avec le modèle PC.
3	PC-regroup2	Pour corriger les problèmes restants avec le modèle PC-regroup.
4	RS	Pour s'assurer d'avoir un pourcentage d'utilisation de l'échelle suffisant.

4.2. La qualité de l'ajustement

4.2.1. *Les sujets*

Selon la modélisation utilisée, le processus d'analyse de la qualité de l'ajustement des sujets a nécessité entre 8 et 16 itérations. Le modèle PC est celui qui a permis de conserver le plus de sujets (n = 109) présentant une bonne qualité d'ajustement au modèle[9]. Le modèle RS a permis de conserver 103 sujets, alors que 93 sujets ont pu être conservés à la suite des deux modélisations où des regroupements de catégories de réponses ont été effectués. Le tableau 1.9 résume ces résultats.

9. Aux fins de l'analyse, notons que nous avons choisi de retirer systématiquement les sujets dont les patrons de réponses présentaient une mauvaise qualité d'ajustement au modèle. En effet, une analyse préliminaire des données nous avait amenés à nous interroger sur la façon de répondre de certains participants. N'ayant pas la possibilité de valider les patrons de réponses ou d'interroger les participants, nous avons simplement décidé de les retirer.

Tableau 1.9
Incidence des modélisations sur le nombre de sujets

	PC	PC-regroup	PC-regroup2	RS
Nombre d'itérations	9	16	8	11
Nombre de sujets conservés	109	93	93	103
Nombre de sujets retirés	43	59	59	49

4.2.2. *Les items*

Une fois l'étude de la qualité de l'ajustement des personnes terminée, il convient d'examiner la qualité de l'ajustement des items. Pour le modèle PC, un seul item pose certains problèmes d'ajustement au modèle suivant les balises que nous nous sommes fixées au début de cette section, soit l'item 6. Pour les deux modélisations où des regroupements ont été effectués, c'est au moins trois items qui présentent des problèmes d'ajustement. Notons la récurrence des items 20 et 23. L'item 6 est également problématique autant pour le modèle PC que pour le modèle PC-regroup2.

Pour nos analyses, nous avons choisi de travailler avec les indices d'ajustement standardisés. Toutefois, comme certains auteurs préconisent d'employer les indices basés sur le carré moyen, nous étions curieux de voir si notre analyse aurait été différente si nous avions eu recours à la version basée sur le carré moyen. Les deux dernières lignes du tableau 1.10 présentent donc ces résultats. On note que les items 20 et 23 ressortent encore comme étant problématiques. Il est possible de remarquer une autre occurrence d'items, soit les items 1, 4, 7 et 8; ces items seraient donc à surveiller. Enfin, pour le modèle RS, tous les items présentent une bonne qualité d'ajustement au modèle.

Tableau 1.10
Items problématiques au regard des statistiques d'ajustement

	PC	PC-regroup	PC-regroup2	RS
Corrélation négative	—	—	—	—
***Outfit* standardisé**	—	—	—	—
***Infit* standardisé**	1 (6)†	3 (4-20-23)	4 (6-9-20-23)	—
***Outfit* carré moyen**	1 (19)	5 (1-4-7-8-23)	4 (1-4-7-8)	—
***Infit* carré moyen**	1 (1)	2 (20-23)	2 (20-23)	—

† Les chiffres ou les nombres entre parenthèses font référence aux numéros d'item.

Le tableau 1.11 présente les statistiques d'ajustement finales des items, c'est-à-dire une fois le processus itératif d'ajustement entre les données et le modèle complété. On note que les valeurs minimales et maximales des indices d'ajustement *infit* et *outfit* standardisés (ZSTD) sont tous à l'intérieur de l'intervalle choisi. On peut également remarquer qu'il en est de même pour les statistiques basées sur le carré moyen.

Tableau 1.11
Statistiques d'ajustements finaux pour les items avec le modèle RS

	***infit* CM**	***infit* ZSTD**	***outfit* CM**	***outfit* ZSTD**
Min.	0,69	–1,96	0,67	–1,88
Max.	1,48	1,90	1,37	1,50
Moy.	1,02	–0,01	0,97	–0,10
Écart-type	0,22	1,28	0,21	0,91

4.3. Les catégories de réponses

La figure 1.1 permet de visualiser de quelle façon les répondants ont utilisé les catégories de réponses qui leur étaient proposées suivant la modélisation PC. La position de chacune des catégories est déterminée par la mesure moyenne des personnes ayant opté pour chacune des catégories. Deux items, que nous avons fait ressortir à l'aide de rectangles, présentent un problème d'ordonnancement de leurs catégories. En effet, les catégories 1 et 2 de l'item 19 apparaissent dans un ordre inversé, de même que les catégories 2 et 3 de l'item 22.

La figure 1.2 présente la position, du point de vue métrique, des seuils entre chacune des catégories de réponses suivant le modèle PC. En d'autres mots, il s'agit du point entre deux catégories où une personne a autant de chances d'opter pour l'une ou l'autre des catégories adjacentes. Ainsi, sur la ligne correspondant à chacun des items, on peut voir les chiffres 2, 3 et 4. Le chiffre 2 correspond à la mesure, sur le continuum de l'offre active, où une personne a autant de chances de choisir, pour cet item, la catégorie 1 = Jamais ou 2 = Rarement. Les chiffres 3 et 4 représentent les seuils entre les catégories 2 et 3, pour le premier, et les catégories 3 et 4, pour le dernier. Le chiffre 1 ne figure pas sur le graphique puisqu'il n'y a aucune catégorie précédente (non borné).

Il est possible de noter que dix items ont été encadrés de rectangles. Ces items présentent tous des problèmes d'ordonnancement par rapport à leurs seuils. Par exemple, pour l'item 1, le seuil *4* (correspond au point d'intersection entre les catégories 3 et 4) apparaît avant le seuil *3* et aussi avant le seuil *2*.

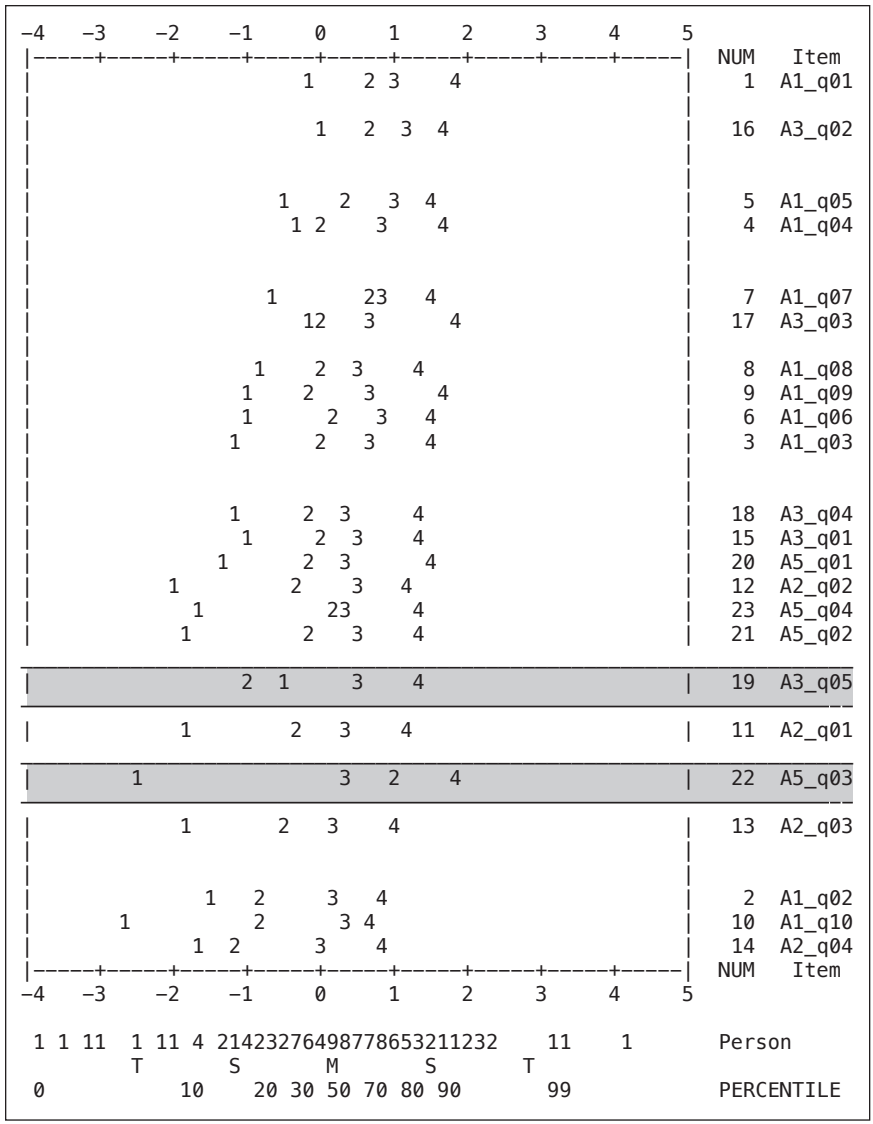


Figure 1.1
Mesure moyenne des sujets ayant opté pour chacune des catégories de réponses suivant le modèle PC[10]

Source : Logiciel Winsteps 3.90.2.0 (Linacre, 2015a).

10. Le logiciel Winsteps fournit les tableaux et les figures dans un format qu'il n'est pas toujours simple, voire possible, de modifier (les titres des axes par exemple). Nous avons choisi de les laisser tels quels.

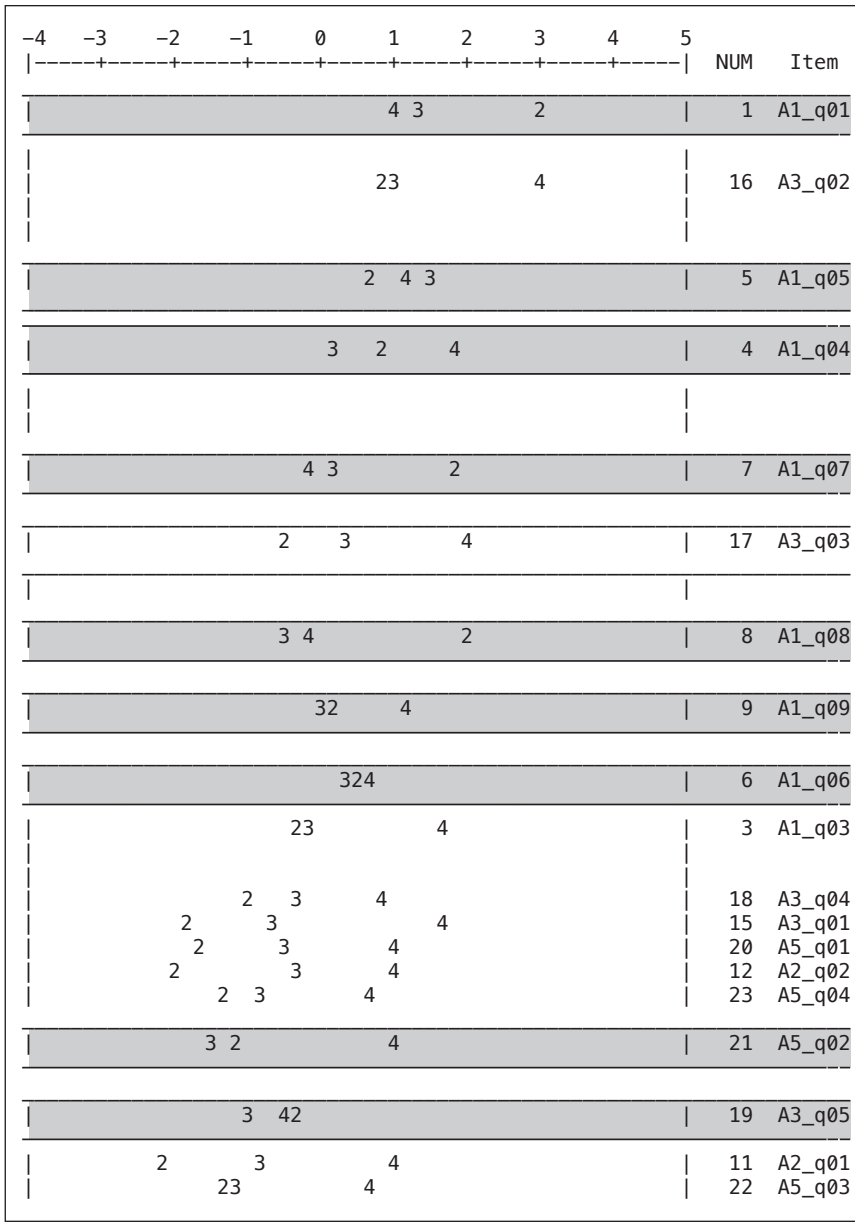


Figure 1.2
Mesure des seuils entre chaque catégorie de réponses selon le modèle PC

Source : Logiciel Winsteps 3.90.2.0 (Linacre, 2015a).

Devant ces premiers résultats, il nous apparaissait pertinent d'interrompre un moment le processus d'analyse choisi afin de revenir au questionnaire et d'y effectuer une analyse qualitative des énoncés proposés, de même que des catégories de réponses offertes pour tenter de trouver l'origine potentielle des problèmes détectés par le modèle PC. Cela nous a conduits à faire trois constats sur la formulation utilisée dans le questionnaire :

1. Il nous semblait que la plupart des items utilisent un verbe d'action pour amener la question, ce qui invite à une réponse du type « oui, je pose cette action » ou « non, je ne la pose pas ». Ainsi, la formulation de la question pourrait créer une rupture sémantique pour le répondant qui, au moment de lire la question, formule une réponse qu'il doit ensuite traduire sous forme de fréquence au moment de choisir une catégorie de réponse. Il pourrait être pertinent de reformuler les questions pour annoncer dès le moment de la lecture de l'énoncé que ce sont des fréquences qui sont attendues dans la réponse, ou bien de modifier les catégories de réponses en fonction des actions visées par les énoncés. Autrement, comment pouvons-nous être certains que les répondants transposeront leur réponse en fréquence de la même façon ? Comment déterminer si poser une action *souvent* consiste à l'accomplir plusieurs fois par jour, par semaine ou par mois ?
2. Certains énoncés d'items nous apparaissaient contenir plusieurs éléments que les répondants doivent considérer avant de pouvoir répondre. Par exemple, « Dans mon milieu de travail, je porte une identification quelconque qui indique que je peux offrir des services en français (p. ex. épinglette) », que doit répondre une personne qui porterait toujours une telle identification, à condition que son milieu de travail lui en fournisse une ? *Jamais* parce que son milieu de travail n'en fournit pas ou *Toujours* parce que cette personne est très favorable à l'offre active ? Si la réponse semble évidente pour certains, comment pouvons-nous nous assurer que tous les répondants répondront de la même façon ?
3. Il nous semblait que les libellés de catégories de réponses choisis pouvaient laisser un grand « trou » au milieu de l'échelle pour les répondants qui font de l'offre active de temps en temps. En effet, pour ces derniers, ni la catégorie de réponse *rarement* ni la catégorie *souvent* ne traduisent bien leur réalité. Par conséquent, ils se doivent de revoir à la baisse ou à la hausse les actions réellement posées d'offre active. Comment savoir alors si tous les répondants auront tendance à répondre de la même façon ? Et

si tous les répondants ne transposent pas leurs réponses de la même façon sur l'échelle, comment pouvons-nous faire une interprétation valide des données recueillies ?

Nous avons jugé bon de présenter ces constats dans le but de susciter la réflexion et peut-être contribuer à l'amélioration du questionnaire. De plus, ils permettent d'apporter un éclairage complémentaire à la méthodologie utilisée dans cette étude et dans la suite de notre analyse. Reprenons donc avec la présentation des résultats de nos quatre modélisations de Rasch.

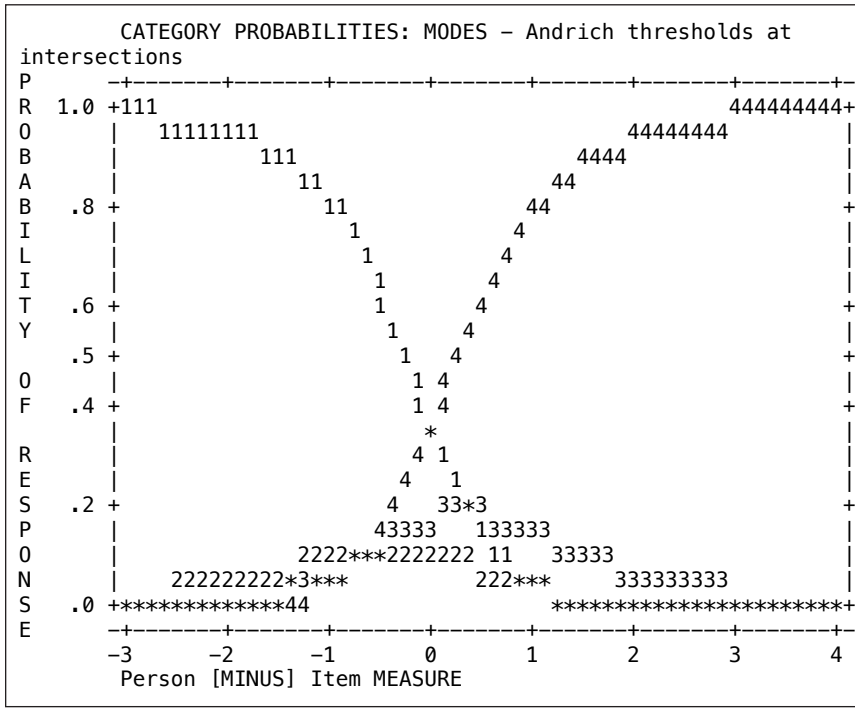


Figure 1.3
Courbes de probabilité des catégories de réponses de l'item 1 selon le modèle PC

Source : Logiciel Winsteps 3.90.2.0 (Linacre, 2015a).

Ainsi, des problèmes d'ordonnancement ayant été détectés, nous avons poursuivi nos analyses en étudiant les catégories de réponses de chacun des items individuellement. À titre d'exemple, la figure 1.3 montre la courbe de probabilité des catégories de réponses de l'item 1. Contrairement à la recommandation de Park (2004), chacune des catégories de réponses ne possède pas un sommet distinct. En effet, on

note que seules les catégories 1 et 4 possèdent un sommet distinct. Les catégories 2 et 3, quant à elles, n'ont jamais plus de chances que les autres d'être choisies par les répondants. Ainsi, par rapport aux quatre catégories de réponses offertes sur le questionnaire, les résultats révèlent que seules deux catégories de réponses sont réellement utiles pour les répondants.

La figure 1.4 présente le sommaire des informations produit par le logiciel Winsteps sur la structure des catégories de réponses de l'item 1. Le pourcentage de réponses observé dans la catégorie 1 est d'environ 83 %, ce qui en fait la catégorie de réponse la plus utilisée. Les autres catégories ont des fréquences d'utilisation de moins de 10 %. Ainsi, le minimum de dix sujets par catégorie recommandé par Linacre (2015b) n'est pas respecté, ni sa recommandation sur la distribution uniforme des fréquences. Pour sa part, la mesure moyenne observée pour chacune des catégories de réponses correspond assez bien à la mesure prédite pour l'échantillon, mais on note que la valeur des seuils n'augmente pas de façon croissante et monotone en fonction de la valeur de la catégorie de réponses. Au contraire, on remarque que le seuil 2 possède une valeur plus élevée que celle du seuil 3.

```
SUMMARY OF CATEGORY STRUCTURE.  Model="R"
FOR GROUPING "0" Item NUMBER: 1  A1_q01

Item DIFFICULTY MEASURE OF 1.80 ADDED TO MEASURES
-------------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT|| ANDRICH |CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||THRESHOLD| MEASURE|
|-------------------+------------+------------++---------+--------|
|  1   1      82  83|  -.11  -.15|  1.95  1.56||  NONE   |(   .65)| 1
jamais
|  2   2       6   6|   .63   .69|   .90   .52||    1.15 |   1.47 | 2
rarement
|  3   3       4   4|  1.02  1.27|  1.13  1.39||    -.41 |   2.09 | 3
souvent
|  4   4       7   7|  1.75  2.00|  1.52  1.27||    -.73 |(  2.96)| 4
toujours
|-------------------+------------+------------++---------+--------|
|MISSING      10   9|  1.38      |            ||         |        |
-------------------------------------------------------------------------
OBSERVED AVERAGE is mean of measures in category. It is not a
parameter estimate.
```

Figure 1.4
Sommaire de la structure des catégories de réponses de l'item 1 selon le modèle PC, tel que produit par le logiciel Winsteps

Source : Logiciel Winsteps 3.90.2.0 (Linacre, 2015a).

Au vu de ces différents résultats, il nous apparaissait judicieux de regrouper les catégories de réponses 2, 3 et 4. En effet, si le regroupement de certaines catégories semble nécessaire, Linacre (2015b) recommande notamment de combiner des catégories ayant peu d'observations, ce qui est le cas pour ces trois catégories. En outre, il convient d'analyser la définition des catégories à regrouper. Ici, ce sont les catégories « rarement », « souvent » et « toujours » qui seraient combinées, alors que la catégorie « jamais » serait conservée telle quelle. Autrement dit, nous nous trouvons à dichotomiser l'item comme si les répondants avaient répondu à l'item par « oui » (rarement, souvent, toujours) ou « non » (jamais).

Une analyse similaire de chacun des items du questionnaire a été effectuée. Le tableau 1.12 présente le détail du nombre de catégories utiles que nous avons trouvées pour les différents items à l'aide de cette analyse. On note que quatre items ne présenteraient que deux catégories de réponses réellement utiles pour les répondants (items 1, 7, 8 et 19). Onze autres items ne révèlent que trois catégories de réponses utiles (items 3 à 6, 9, 13, 14, 16 et 21 à 23). Cela signifie que sur les 23 items analysés, seulement huit affichent une utilisation relativement adéquate des quatre catégories de réponses proposées sur le questionnaire.

Le tableau 1.12 montre également les regroupements de catégories qui nous apparaissaient les plus pertinents à effectuer pour voir s'il nous serait possible d'améliorer nos résultats d'analyse[11]. Ainsi, pour reprendre l'exemple de l'item 1 que nous avons présenté en détail, l'indication « 1(234) » dans la colonne « Regroupement exploré » signifie que nous avons regroupé les catégories de réponses 2 = rarement, 3 = souvent et 4 = toujours en une seule et même catégorie. Pour certains items, seulement deux catégories de réponses ont dû être regroupées. Pour l'item 3 par exemple, ce sont les catégories de réponses 2 et 3 qui ont été regroupées, alors que pour l'item 5, ce sont les catégories 3 et 4. Seul l'item 19 fait exception. En effet, à en juger par les résultats obtenus pour cet item, le nombre de données manquantes était très élevé. Aucun regroupement ne nous permettait de corriger de façon logique les problèmes décelés par l'analyse initiale. C'est pourquoi nous avons décidé de le retirer.

11. Ces regroupements correspondent à l'analyse PC-regroup que nous avons effectuée.

Tableau 1.12
Synthèse du nombre de catégories de réponses utiles, de l'ordonnancement des catégories de réponses et des seuils, ainsi que des regroupements explorés

Nombre d'items	Numéro d'item	Nombre de catégories utiles	Regroupement exploré	Nouveau regroupement exploré
3	1-7-8	2†	1(234)	
1	19	2†	(retiré)	
1	3	3	1(23)4	
3	13-16-23	3	1(23)4	
6	4-6-9-14-21-22	3†	1(23)4	
1	5	3†	12(34)	1(234)
8	2-10-11-12-15-17-18-20	4	Aucun	

† L'item présente aussi un problème d'ordonnancement de ses catégories de réponses ou de ses seuils.

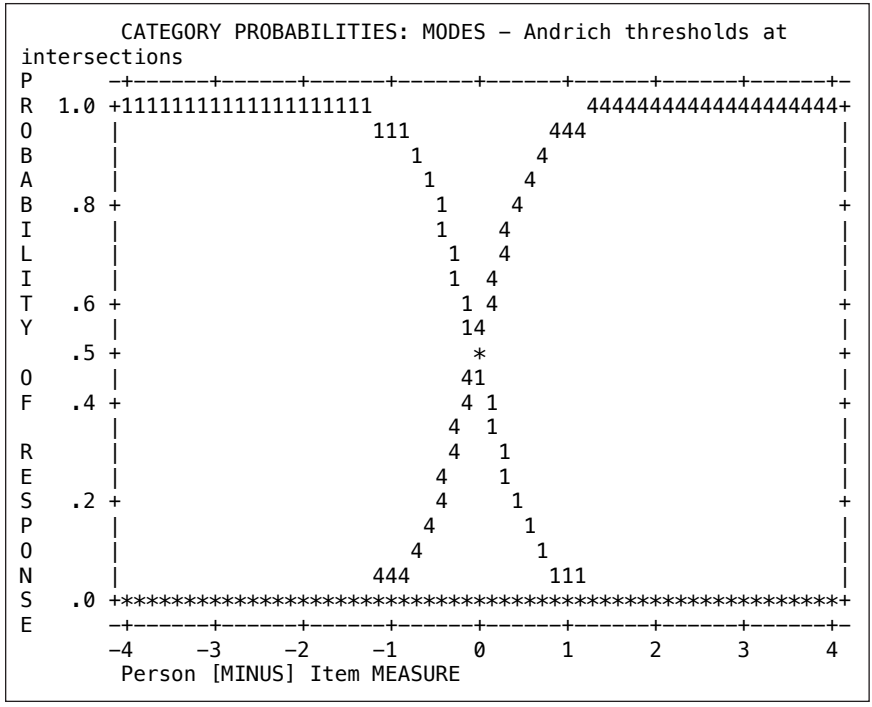


Figure 1.5
Courbes de probabilité des catégories de réponses de l'item 1 selon le modèle PC-regroup2

Source : Logiciel Winsteps 3.90.2.0 (Linacre, 2015a).

Enfin, le tableau 1.12 révèle que pour un des items, l'item 5, le premier regroupement des catégories que nous avons exploré n'a pas corrigé tous les problèmes liés aux catégories de réponses. Nous avons donc dû explorer un nouveau regroupement. L'analyse PC-regroup2 correspond donc à l'analyse PC-regroup, mais où le regroupement de l'item 5 est modifié pour le nouveau regroupement suggéré.

La figure 1.5 illustre les courbes de probabilité des catégories de réponses de l'item 1 finales, c'est-à-dire une fois les regroupements effectués et les données modélisées à nouveau. La figure montre que chacune des catégories de réponses formées possède maintenant un sommet distinct comme recommandé par Park (2004).

Enfin, la figure 1.6 présente le sommaire final des informations produit par le logiciel Winsteps sur la structure des catégories de réponses de l'item 1. On note que les résultats ainsi obtenus suivent maintenant mieux les recommandations formulées par Linacre (2015b).

```
Item DIFFICULTY MEASURE OF 3.01 ADDED TO MEASURES
-------------------------------------------------------------------------
|CATEGORY   OBSERVED|OBSVD SAMPLE|INFIT OUTFIT|| ANDRICH |CATEGORY|
|LABEL SCORE COUNT %|AVRGE EXPECT|  MNSQ  MNSQ||THRESHOLD| MEASURE|
|-------------------+------------+------------++---------+--------|
|  1   1      81  96|   .13   .13|   .98   .59||  NONE   |(  2.21)| 1
non
|  2   2       0   0|            |   .00   .00||  NULL   |   2.78 | 2
|  3   3       0   0|            |   .00   .00||  NULL   |   3.24 | 3
|  4   4       3   4|  3.27  3.31|  1.35   .11||     .00 |(  3.81)| 4
oui
|-------------------+------------+------------++---------+--------|
|MISSING       9  10|  1.51      |            ||         |        |
-------------------------------------------------------------------------
OBSERVED AVERAGE is mean of measures in category. It is not a
parameter estimate.
Unobserved category. Consider: STKEEP=NO
```

Figure 1.6
Sommaire de la structure des catégories de réponses de l'item 1 selon le modèle PC-regroup2, tel que produit par le logiciel Winsteps

Source : Logiciel Winsteps 3.90.2.0 (Linacre, 2015a).

4.4. Les indices sur les postulats d'unidimensionnalité et d'indépendance locale

4.4.1. *L'unidimensionnalité*

De façon générale, les quatre modèles utilisés montrent une variance expliquée par la mesure d'environ 60 % (tableau 1.13), ce qui est bon (Linacre, 2015b). De ce pourcentage, la plus grande proportion

provient des sujets (environ 34 % à 37 %) par rapport aux items qui fournissent une explication pour environ 26 % de la variance si l'on considère le modèle *Rating Scale* (RS), mais seulement 22 % pour le modèle PC-regroup2.

L'analyse en composantes principales (tableau 1.13) indique que, pour tous les modèles explorés, la variance inexpliquée associée aux deux premiers contrastes est plus grande que 2. Le troisième contraste de l'analyse effectuée selon le modèle PC montre aussi une variance inexpliquée plus grande que 2[12]. Or, une valeur propre de 2 correspond à un regroupement de deux items, ce qui constitue le nombre minimal d'items pour former une sous-dimension. Les résultats obtenus révèlent donc la présence potentielle de deux sous-dimensions mesurées par les items.

Tableau 1.13
Synthèse de la variance expliquée pour les différents modèles

		PC	PC-regroup	PC-regroup2	RS
Item retiré		—	19	19	—
Variance expliquée par la mesure		59,0 %	61,2 %	58,9 %	59,0 %
Variance expliquée par les personnes		34,3 %	37,2 %	37,0 %	33,6 %
Variance expliquée par les items		24,7 %	24,0 %	21,9 %	25,5 %
Variance inexpliquée (Valeur propre associée)		41,0 % (23,0)	38,8 % (22,0)	41,1 % (22,0)	41,0 % (23,0)
Variance (valeur propre)	**Contraste 1**	4,5 % (2,52)	4,4 % (2,52)	4,7 % (2,52)	5,2 % (2,92)
	Contraste 2	3,7 % (2,07)	4,0 % (2,25)	4,2 % (2,26)	4,1 % (2,31)
	Contraste 3	3,6 % (2,03)	3,3 % (1,86)	3,6 % (1,92)	3,4 % (1,91)

Le tableau 1.14 affiche les regroupements d'items dont la corrélation associée à chacun des contrastes de l'analyse en composantes principales est d'au moins 0,40. On observe que certains regroupements d'items sont récurrents malgré les différentes modélisations effectuées, par exemple les items 6, 7 et 8, les items 20, 21 et 23, les items 11-12 et 13 ou les items 17 et 18. Toutefois, le tableau indique que pour les différents contrastes produits par le logiciel, la valeur de la corrélation dite atténuée entre le groupe d'items dont la corrélation avec le facteur est la plus forte et celui dont la corrélation avec le facteur est la plus faible est généralement d'au moins 0,72. Selon Linacre (2015b), cela correspondrait à des items deux fois plus dépendants qu'indépendants.

12. Pour tous les modèles, la valeur associée aux contrastes 4 et 5 était inférieure à 2, alors nous avons choisi de ne pas les présenter.

Par conséquent, il s'agirait d'items qui mesurent sensiblement la même chose. Il ne serait donc pas possible de conclure à la présence d'autres dimensions que celle mesurée par le facteur Rasch.

Tableau 1.14
Synthèse de l'analyse de l'unidimensionnalité des items

	Regroupements d'items†		**Corrélation atténuée**	
Contraste	**PC-regroup2**	**RS**	**PC-regroup2**	**RS**
1	(6-7-3-8) et (21-23-20)	(8-7-6) et (21-20-23)	0,8740	0,7344
2	(12-13-11) et (23)	(12-13-11) et (23)	0,7160	0,8156
3	(2-22-5) et (18-8-17)	(17-16-18) et (3)	0,8610	0,7492
4	(15-22-18) et (1)	(5-4) et (18)	(1,00)	0,8007
5	(10-2) et (9)	(19-18) et (7-2)	0,9893	(1,00)

† Pour chaque contraste, deux regroupements d'items sont présentés entre parenthèses. Les premiers correspondent aux items qui corrèlent le plus fortement avec le facteur et dont la valeur de la corrélation est supérieure ou égale à +0,40. Les deuxièmes correspondent aux items qui corrèlent le moins avec le facteur et dont la valeur de la corrélation est inférieure ou égale à –0,40. De plus, il convient de noter que les items sont présentés en ordre décroissant selon leur degré de corrélation avec le facteur.

4.4.2. *L'indépendance locale*

Le tableau 1.15 présente la valeur des corrélations inter-items effectuées sur les résidus standardisés les plus élevées suivant les modèles PC-regroup2 et RS. Aucune paire d'items ne présente une valeur de corrélation plus élevée que 0,70. Ainsi, il n'y aurait pas de dépendance entre les items et le postulat d'indépendance locale semble vérifié.

Tableau 1.15
Corrélations inter-items les plus élevées effectuées sur les résidus standardisés

Modèle	**Corrélation sur les résidus standardisés**	**Items**
PC-regroup2	–0,45	1 et 22
	–0,41	2 et 18
RS	0,53	20 et 21
	0,50	22 et 23
	0,42	12 et 13
	0,41	20 et 23
	0,40	7 et 8

4.5. La détection du FDI

Le tableau 1.16 propose une synthèse des résultats obtenus pour les deux tests d'hypothèse utilisés pour le diagnostic du FDI suivant le modèle RS (seul modèle utilisé pour l'analyse du FDI puisque c'est celui qui affichait la meilleure qualité d'ajustement données-modèle). Les premières colonnes présentent les résultats du premier test, celui basé sur l'hypothèse que le degré de difficulté d'un item devrait être le même pour l'ensemble des groupes que pour un groupe particulier de répondants. Pour ce test, deux éléments sont à surveiller (Linacre, 2015b). Dans un premier temps, la différence entre les deux mesures (celle d'un groupe par rapport à l'ensemble des groupes) doit être suffisamment grande pour avoir un effet sur l'interprétation de la mesure. Les items dont la différence constitue un risque potentiel de FDI de catégorie B et C sont relevés dans le tableau[13]. Huit items (1, 3, 5, 15, 18, 20, 21 et 23) présentent ce risque. Dans un deuxième temps, il faut que le niveau de signification résultant de la comparaison entre les deux mesures soit assez petit pour confirmer que la différence trouvée n'est pas simplement le fruit du hasard. Les différences trouvées sont statistiquement significatives pour seulement quatre des items (5, 15, 18 et 23).

La dernière colonne présente les résultats du deuxième test, soit celui basé sur l'hypothèse que l'item ne possède aucun FDI à travers les différents groupes de sujets. Rappelons que le résultat est résumé comme une statistique du khi carré pour chacun des items indiquant si le FDI potentiel trouvé est important ou non. Une valeur moindre que le seuil généralement fixé à 0,05 révèle la présence d'une différence statistiquement significative entre les différents groupes. Trois items ressortent comme présentant un risque potentiel de FDI suivant cette hypothèse, soit les items 5, 15 et 23.

13. Voir la section 2.7 pour plus de détails.

Tableau 1.16
Synthèse des résultats des deux tests d'hypothèse utilisés pour la détection du FDI selon le modèle RS

	Degré de difficulté					Probabilité
Item	**Global**	**Gr. 1**	**Gr. 2**	**Gr. 3**	**Gr. 4**	**du Khi carré**
1	3,182	4,592 C	3,387	2,442 C	2,866	0,1743
2	–0,848	–1,097	–0,664	–0,631	–1,077	0,4526
3	0,227	0,044	0,044	0,383	0,702 C	0,3100
4	0,812	0,729	0,776	0,878	0,926	0,9590
5	1,054	1,405	0,417 C*	1,486 B	1,286	0,0189*
6	0,402	0,717	0,162	0,280	0,594	0,3816
7	0,545	0,718	0,346	0,411	0,957	0,5488
8	0,312	0,224	0,405	0,390	0,162	0,9206
9	0,283	0,584	–0,103	0,321	0,340	0,3279
10	–1,045	–1,455	–0,995	–0,842	–0,741	0,3124
11	–0,543	–0,543	–0,787	–0,497	–0,233	0,6454
12	–0,412	–0,339	–0,335	–0,505	–0,530	0,9305
13	–0,742	–0,676	–0,653	–0,944	–0,709	0,8773
14	–1,412	–1,302	–1,458	–1,700	–1,227	0,7948
15	–0,276	–0,201	0,221 B*	–0,897 B*	–0,407	0,0458*
16	1,235	1,603	1,418	0,926	0,815	0,2548
17	0,293	–0,052	0,528	0,450	–0,114	0,4613
18	–0,343	–0,761	0,145 B*	–0,502	–0,595	0,1012
19†	–1,042		–1,042			
20	–0,312	–0,513	–0,482	0,617 C	0,169 B	0,4900
21	–0,494	–0,416	–0,793	1,085 C	–1,509 C	0,1793
22	–0,392		–0,330	–0,291	–0,682	0,8946
23	–0,482	–0,813	0,252 C*	–0,291	–1,867 C	0,0350*

Groupe 1 = étudiants en médecine ; groupe 2 = diplômés CNFS ; groupe 3 = ergothérapie stage deux et groupe 4 = ergothérapie stage trois.

C = FDI d'au moins 0,64 *logit* ; B = FDI d'au moins 0,43 *logit* ; * = significatif à 0,05.

† Les cellules vides indiquent qu'il n'y avait aucune donnée recueillie pour ce groupe pour l'item en question.

4.6. La qualité de l'échelle de mesure

La figure 1.7 montre la distribution des sujets et des items le long du continuum de l'offre active (représentation de Wright) suivant le modèle RS. On note que les sujets sont majoritairement centrés autour de zéro, avec une légère asymétrie positive. Ainsi, certains sujets, environ une dizaine, semblent, plus facilement que les autres, acquiescer aux divers énoncés sur l'offre active proposés par le questionnaire.

Du côté des items, on remarque que la plupart d'entre eux sont alignés au centre de la distribution des sujets, voire un peu plus bas. (Notons les M sur l'axe central du graphique qui signifie *moyenne*. La moyenne des items est légèrement plus basse que celle des sujets.) Ainsi, les items sont généralement assez faciles à endosser par les sujets.

On relève également une certaine redondance par rapport à la mesure de certains items. En effet, les items 6, 8, 9 et 17 se trouvent au même endroit le long du continuum de l'offre active. Il en est de même pour les items 12, 15, 18, 20 et 22 ou les items 11, 21 et 23. Les items 10 et 19 présentent aussi un potentiel de redondance. Il conviendrait d'examiner ces items pour vérifier s'ils permettent vraiment de recueillir des informations différentes.

Enfin, il est possible de remarquer de grands espaces vides sur le continuum des items (indiqués par des flèches). Cela signifie qu'il n'existe aucun item à ces endroits pour recueillir de l'information sur l'offre active et pour évaluer de façon plus précise le niveau d'offre active des sujets correspondants. Il conviendrait d'ajouter des items pour obtenir une distribution plus uniforme de la difficulté de ces items le long de l'axe.

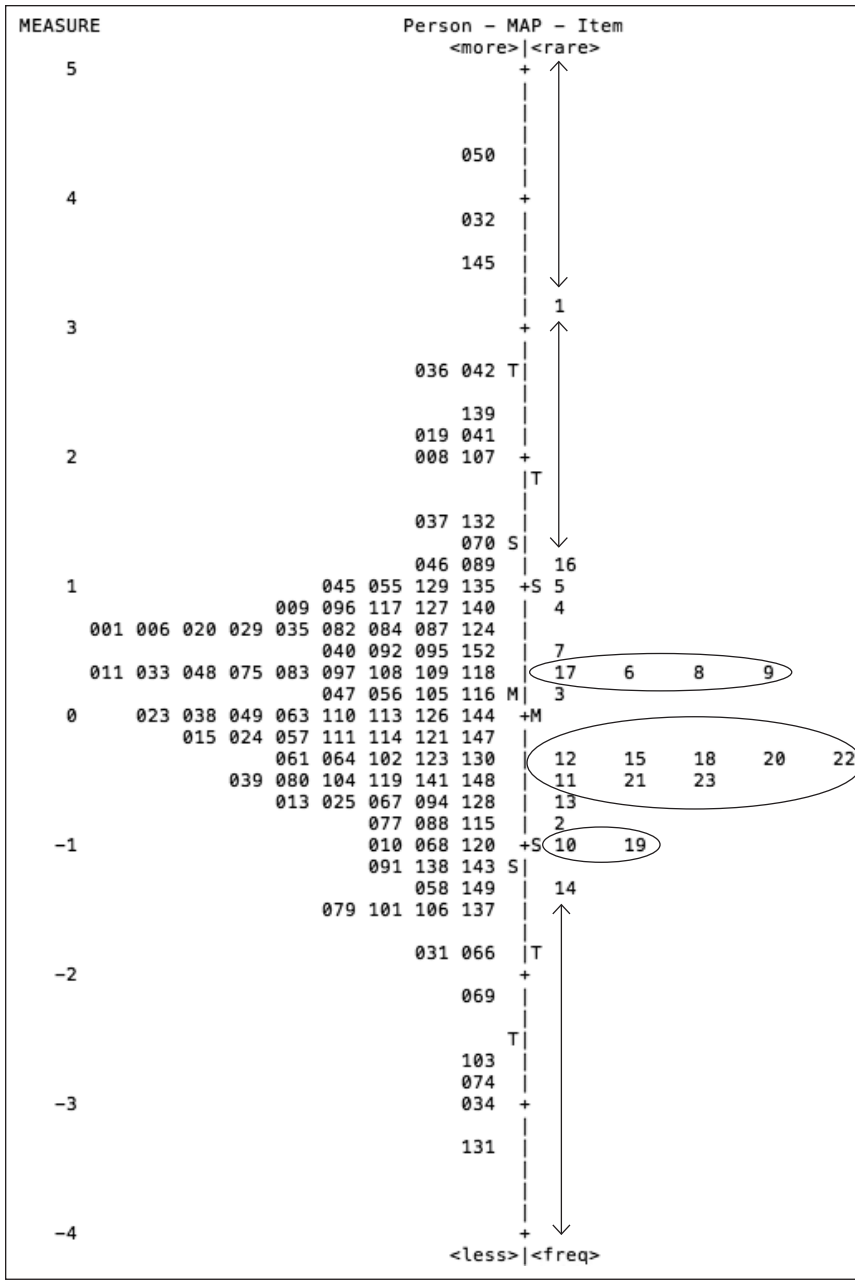


Figure 1.7
Représentation de Wright suivant le modèle RS

Source : Logiciel Winsteps 3.90.2.0 (Linacre, 2015a).

4.7. La fidélité

En ce qui concerne les sujets, le tableau 1.17 montre qu'il y a peu de différences entre les modélisations effectuées. La valeur de l'indice de séparation oscille entre 2,86 pour le modèle RS et 2,99 pour le modèle PC-regroup2. Ce qui, une fois converti en strates, indique qu'il est possible de distinguer environ quatre niveaux d'offre active différents, ce qui est correct. En termes de fidélité (*reliability*), les valeurs oscillent autour de 0,90, ce qui est bon.

Tableau 1.17
Indices de fidélité pour les sujets selon les modélisations

Fidélité pour les sujets	PC	PC-regroup	PC-regroup2	RS
RMSE réel	0,43	0,43	0,42	0,43
RMSE modèle	0,41	0,41	0,41	0,40
S.D. réelle	1,26	1,30	1,26	1,23
S.D. modèle	1,26	1,30	1,27	1,24
Séparation réelle	2,91	2,99	2,99	2,86
Séparation modèle	3,10	3,14	3,13	3,09
Fidélité réelle	0,89	0,90	0,90	0,89
Fidélité modèle	0,91	0,91	0,91	0,91
Corrélation mesure-données brutes	0,79	0,78	0,78	0,77

Pour les items (tableau 1.18), on note également des valeurs assez similaires d'un modèle à l'autre. L'indice de séparation oscille entre 4,26 et 5,07, ce qui signifie qu'il est possible de distinguer un peu plus de six niveaux d'offre active. Pour ce qui est de la fidélité (*reliability*), les valeurs obtenues sont autour de 0,95, ce qui est très bien.

Tableau 1.18
Indices de fidélité pour les items selon les modélisations

Fidélité pour les items	PC	PC-regroup	PC-regroup2	RS
RMSE réel	0,19	0,22	0,22	0,20
RMSE modèle	0,18	0,21	0,20	0,18
S.D. réelle	0,81	1,01	1,00	0,94
S.D. modèle	0,81	1,02	1,01	0,94
Séparation réelle	4,26	4,53	4,56	4,78
Séparation modèle	4,54	4,90	4,97	5,07
Fidélité réelle	0,95	0,95	0,95	0,96
Fidélité modèle	0,95	0,96	0,96	0,96
Corrélation mesure-données brutes	–0,41	–0,31	–0,30	–0,41

CONCLUSION

Le premier objectif de ce texte consistait à vérifier si une modélisation de type Rasch était appropriée pour étudier les propriétés métriques de l'offre active. Suivant les résultats obtenus, il semble que la modélisation de Rasch soit effectivement intéressante pour mesurer les données recueillies à l'aide du questionnaire sur les comportements individuels d'offre active en français. Plus précisément, le modèle à crédit partiel (PC) nous a permis d'en apprendre plus sur le fonctionnement individuel de chacun des items, mais c'est le modèle *Rating Scale* (RS) qui, en définitive, permettait d'obtenir la meilleure qualité d'ajustement entre les données et le modèle.

Notre deuxième objectif était d'expliciter notre démarche d'analyse afin de la rendre facilement reproductible et d'aider toute personne à s'initier à une méthodologie basée sur une modélisation de type Rasch. Pour ce faire, nous avons choisi de suivre les recommandations de Tennant et Conaghan (2007). Il nous est apparu qu'appliquer une telle démarche était utile pour étudier la modélisation des scores bruts et en arriver à produire une échelle de mesure objective de l'offre active de services en français. Mais ce qui nous semble encore plus évident, ce sont toutes les décisions qu'un analyste doit prendre pour y arriver et qui ne sont généralement pas explicitées. Ces décisions doivent être prises tout au long du processus d'analyse. Sans toutes les présenter, voici tout de même quelques exemples.

L'analyse de la qualité de l'ajustement entre les données et le modèle est une étape inhérente à toute modélisation de type Rasch. Or, pour effectuer ces analyses, nous avons décidé de retirer systématiquement les sujets dont les patrons de réponse n'affichaient pas une bonne qualité d'ajustement avec le modèle. La présentation explicite de nos résultats a permis de constater qu'au regard de l'ajustement des sujets, le choix du modèle peut avoir une influence sur le nombre de sujets conservés. Ainsi, selon le contexte entourant un processus d'analyse, un modèle pourrait être préféré à un autre (pour conserver davantage de sujets, par exemple, ou pour assurer la représentativité de l'échantillon). En outre, au lieu de retirer systématiquement le patron de réponses en entier d'un sujet, un autre analyste aurait pu choisir de travailler avec le tableau des réponses inattendues produit par le logiciel Winsteps pour retirer uniquement les réponses qui ne s'ajustent pas bien au modèle. En effet, l'estimation des paramètres d'items ou de sujets suivant la modélisation de Rasch se révèle généralement robuste, même en présence de données manquantes (Smith Jr. et Smith, 2004).

Au regard des catégories de réponses, les résultats ont montré que le nombre de catégories de réponses réellement utiles pouvait varier d'un item à l'autre. Nous avons choisi de procéder à certains regroupements. En effet, notre analyse nous a permis de constater que, pour les répondants, certains items semblent se présenter comme des items dichotomiques (oui/non) plutôt que des items qui s'évaluent en termes de fréquence comme prévu dans le questionnaire. Les regroupements de catégories que nous avons choisi d'effectuer révèlent également qu'il pourrait y avoir confusion au centre de l'échelle pour les répondants puisque pour 10 des 23 items, les problèmes d'ordonnancement des catégories de réponses ou des seuils qui leur sont associés nous ont amenés à regrouper les catégories centrales de l'échelle, soit 2 = rarement et 3 = parfois. Or, bien que de procéder à une telle optimisation de l'échelle de réponses puisse offrir certains avantages, ce ne sont pas tous les chercheurs qui opteront pour ce type de pratique. En effet, certains estimeront qu'il est préférable que l'échelle de réponses soit uniforme (c'est-à-dire équilibrée dans le continuum de réponses offertes, mais aussi constante d'un item à l'autre) autant dans le questionnaire que lors des analyses (Royal *et al*., 2010).

En ce qui concerne la démonstration de la vérification de la condition d'unidimensionnalité, nous avons choisi d'effectuer une analyse en composantes principales sur les résidus standardisés, c'est-à-dire une fois le facteur Rasch extrait. Or, bien que le concept d'unidimensionnalité soit fondamental à ce type de modélisation (postulat de base), la méthode pour vérifier que cette condition est respectée n'est pas définie clairement dans la littérature. En effet, la compréhension des chercheurs en la matière semble encore en développement et, par conséquent, les méthodes proposées ne sont pas toujours uniformes (Nilsson et Tennant, 2011). Pour plusieurs chercheurs, une bonne qualité d'ajustement entre les données et le modèle apparaît comme suffisante pour assurer que la condition d'unidimensionnalité est respectée (Smith, 1996). Exposer de façon explicite l'analyse que nous avons faite nous a permis de constater que, bien que nous ayons obtenu une bonne qualité d'ajustement entre les données et le modèle, certains regroupements d'items sont ressortis de façon récurrente comme représentant potentiellement une sous-dimension, et ce, à travers les différentes modélisations explorées. Une analyse qualitative effectuée sur ces items a semblé corroborer que ces items comportent certaines caractéristiques communes et permis d'amorcer une réflexion pour établir si le questionnaire devait rester tel quel ou être divisé en sections.

En plus de la corrélation entre chacun des items et les différents facteurs de l'analyse en composantes principales, nous avons choisi d'explorer le postulat d'unidimensionnalité en étudiant la force de la

corrélation atténuée entre les paramètres de sujets modélisés par le logiciel à l'aide des deux groupes d'items les plus opposés sur un même contraste, soit ceux qui corrèlent le plus fortement et le plus faiblement avec le facteur. Certains auteurs (voir l'étude de Muller et Roddy, 2009, par exemple), choisissent d'aller encore plus loin que nous dans cette analyse. En effet, il est possible d'utiliser les paramètres de sujets modélisés par le logiciel pour ces deux groupes d'items opposés (fortement et faiblement corrélés) et d'effectuer une série de test *t* de Student afin de déterminer si les différences trouvées entre les mesures produites par les deux groupes sont statistiquement significatives ou non.

Ces exemples nous permettent d'illustrer que les auteurs ne semblent pas toujours s'entendre sur la définition de certains concepts, ni sur la méthode à utiliser pour les vérifier. L'analyste doit donc prendre plusieurs décisions sans avoir une méthode précise à suivre. Par exemple, parmi les diverses méthodes de vérification ou d'indices de détection de la présence ou non de multiples dimensions, Hattie (1985) stipule que trop d'entre elles ont été développées sur une base *ad hoc*, en présentant bien peu de références quant à leur fondement, leur comportement, et peu établissent une comparaison avec d'autres indices existants. Ainsi, il nous semble encore plus approprié d'expliciter davantage la démarche utilisée par un analyste afin d'espérer un jour pouvoir lever le voile sur certains procédés relativement peu détaillés dans la littérature. À notre avis, la démarche proposée par Tennant et Conaghan (2007) permet de tracer un portrait assez complet des éléments qui devraient être discutés lorsqu'une méthodologie basée sur la modélisation de type Rasch est utilisée. Rappelons toutefois que bien qu'une méthodologie exhaustive soit mise en place afin de diagnostiquer les éléments problématiques dans les données recueillies à l'aide d'un questionnaire, le jugement des experts du domaine demeure nécessaire pour décider des changements qu'il conviendrait d'apporter et pour procéder à une nouvelle expérimentation de la version ainsi améliorée.

BIBLIOGRAPHIE

Andrich, D. (1978). « Application of a psychometric rating scale model to ordered categories which are scored with successive integers », *Applied Psychological Measurement*, *2*, p. 581.

Blais, J.-G. et J. Grondin (2010). « L'impact de la formulation des items dans les questionnaires d'enquête : une étude avec le modèle de Rasch pour les données polytomiques », *Mesure et évaluation en éducation*, *33*(2), p. 95-126.

Bond, T.G. et C.M. Fox (2015). *Applying the Rasch Model : Fundamental Measurement in the Human Sciences*, 3e éd., New York : Routledge.

Boone, W.J., J.R. Staver et M.S. Yale (2014). *Rasch Analysis in the Human Sciences*, Dordrecht : Springer Netherlands.

Bouchard, L., M. Beaulieu et M. Desmeules (2012). « L'offre active de services de santé en français en Ontario : une mesure d'équité », *Reflets, revue d'intervention sociale et communautaire*, *18*(2), p. 38-65, doi : 10.7202/1013173ar.

Bouchard, P., S. Vézina et M. Savoie (2010). *Rapport du Dialogue sur l'engagement des étudiants et des futurs professionnels pour de meilleurs services de santé en français dans un contexte minoritaire : Formation et outillage, Recrutement et rétention*, Ottawa : Consortium national de formation en santé.

Bowen, S. (2015). *Impact des barrières linguistiques sur la qualité et la sécurité dans les soins de santé*, <http://santefrancais.ca/wp-content/uploads/SSF-Bowen-S.---tude-Barri--res-linguistiques.pdf>, consulté le 25 avril 2017.

Cano, S.J., A.F. Klassen, A. Scott, P.G. Cordeiro et A.L. Pusic (2013). « Reply : The Rasch model : "Litmus Test" de rigueur for Rating Scales ? », *Plastic and Reconstructive Surgery, 131*(2), p. 286e-288e, doi : 10.1097/PRS.0b013e31828129f4.

Drolet, M., P. Bouchard, J. Savard et M.-J. Laforge (2017, à paraître). « Problématique générale : Enjeux de l'accessibilité et de l'offre active de services sociaux et de santé au sein de la francophonie canadienne en situation minoritaire », dans M. Drolet, P. Bouchard et J. Savard (dir.), *Santé et services sociaux : Accessibilité et offre active en contexte linguistique minoritaire*, Ottawa : Les Presses de l'Université d'Ottawa.

Hattie, J. (1985). « Methodology review : Assessing unidimensionality of tests and items », *Applied Psychological Measurement*, *9*(2), p. 139.

Linacre, J.M. (1994). « Sample size and item calibration [or person measure] stability », *Rasch Measurement Transactions, 7*(4), p. 328, <http://www.rasch.org/rmt/rmt74m.htm>, consulté le 25 avril 2017.

Linacre, J.M. (2002). « What do infit and outfit, mean-square and standardized mean ? », *Rasch Measurement Transactions*, *16*(2), p. 878.

Linacre, J.M. (2004). « Optimizing Rating Scale Category Effectiveness », dans E.V. Smith Jr. et R.M. Smith (dir.), *Introduction to Rasch Measurement : Theory, Models and Applications*, Maple Grove : JAM Press, p. 258-278.

Linacre, J.M. (2012). *Winsteps tutorial 4*, Beaverton, Oregon, Winsteps.com <http://www.winsteps.com/a/winsteps-tutorial-4.pdf>, consulté le 25 avril 2017.

Linacre, J.M. (2015a). *Winsteps® Rasch Measurement Computer Program* (version 3.90.2), Beaverton, Oregon, <http://www.winsteps.com/index.htm>, consulté le 25 avril 2017.

Linacre, J.M. (2015b). *Winsteps® Rasch Measurement Computer Program User's Guide* (version 3.90.2), Beaverton, Oregon, <http://www.winsteps.com/index.htm>, consulté le 25 avril 2017.

Masters, G.N. (1982). « A Rasch Model for Partial Credit Scoring », *Psychometrika, 47*(2), p. 149.

Muller, S. et E. Roddy (2009). « A Rasch analysis of the Manchester foot pain and disability index », *Journal of Foot and Ankle Research*, *2*(29), p. 1-10.

Nilsson, A.L. et A. Tennant (2011). « Past and present issues in Rasch analysis : The functional independence measure (FIM™) revisited », *Journal of Rehabilitation Medecine*, 43, p. 884-891.

Park, T. (2004). « An investigation of an ESL placement test of writing using many-facet Rasch measurement », *Working Papers in TESOL & Applied Linguistics, 4*(1), p. 1-21.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Achievement Tests*, Copenhague : Danish Institute for Educational Research.

Rojas Tejada, A.J., A. Gonzalez Gomez, J.L. Padilla Garcia et C. Perez Melendez (2002). « Two strategies for fitting real data to Rasch polytomous models », *Journal of Applied Measurement, 3*(2), p. 129.

Royal, K.D., A. Ellis, A. Ensslen et A. Homan (2010). « Rating scale optimization in survey research : An application of the Rasch Rating Scale model », *Journal of Applied Quantitative Methods*, *5*(4), p. 607-617.

Savard, J., L. Casimiro, J. Benoît et P. Bouchard (2014). « Évaluation métrologique de la mesure de l'offre active de services sociaux et de santé en français en contexte minoritaire », *Reflets, revue d'intervention sociale et communautaire, 20*(2), p. 83-122, doi : 10.7202/1027587ar.

Savard, J., L. Casimiro, P. Bouchard, J. Benoît, M. Drolet et C. Dubouloz (2015). « Conception d'outils de mesure de l'offre active de services sociaux et de santé en français en contexte minoritaire », *Minorités linguistiques et sociétés*, *6*, p. 131-156.

Smith Jr., E.V. et R.M. Smith (2004). *Introduction to Rasch Measurement : Theory, Models, and Applications*, Maple Grove : JAM Press.

Smith, R.M. (1996). « A comparison of methods for determining dimensionality in Rasch measurement », *Structural Equation Modeling*, *3*(1), p. 25-40.

Smith, R.M., R.E. Schumacker et M.J. Bush (1998). « Using item mean squares to evaluate fit to the Rasch model », *Journal of Outcome Measurement*, *2*, p. 66.

Smith, R.M. et K.K. Suh (2003). « Rasch fit statistics as a test of the invariance of item parameter estimates », *Journal of Applied Measurement, 4*, p. 153.

Tennant, A. et P.G. Conaghan (2007). « The Rasch measurement model in rheumatology : What is it and why use it ? When should it be applied, and what should one look for in a Rasch paper ? », *Arthritis Rheum*, *57*, p. 1358.

CHAPITRE 2

L'analyse psychométrique d'outils d'évaluation en pédagogie des sciences de la santé

Une comparaison des conclusions selon les approches classique et de Rasch

Jean-Sébastien Renaud

L'analyse psychométrique d'un outil d'évaluation peut être réalisée selon diverses approches, la théorie classique des tests étant la plus populaire. Toutefois, l'approche de Rasch, plus récente, présente différents avantages par rapport à l'approche classique et son utilisation pourrait contribuer à l'avancement de l'évaluation des apprentissages complexes et des compétences en éducation médicale et en pédagogie des sciences de la santé. L'objectif de ce chapitre est d'illustrer à l'aide d'un exemple cet apport de l'approche de Rasch en contexte de pédagogie des sciences de la santé. Pour ce faire, nous comparons les résultats de l'analyse psychométrique de l'Échelle de communication médecin-patient pour les externes en médecine (ECMP-EM) selon l'approche classique et celle de Rasch. Nous avons utilisé les évaluations faites à l'aide de l'ECMP-EM de 634 étudiants en stages d'externat entre 2011 et 2013. L'analyse psychométrique de cette grille a été réalisée selon l'approche classique (analyse factorielle confirmatoire, analyse des statistiques d'items, alpha de Cronbach) et selon le Rasch Rating Scale Model (évaluation de la qualité d'ajustement des données au modèle, de l'unidimensionnalité et de l'indépendance locale, examen de l'échelle d'appréciation, coefficient de fidélité Rasch, adéquation entre les distributions des items et des sujets). L'analyse selon l'approche classique montre que les items de l'ECMP-EM

forment un seul facteur, possèdent un bon pouvoir de discrimination et permettent une évaluation fidèle. L'approche de Rasch montre que les items s'ajustent bien au modèle de Rasch, forment un seul facteur et sont indépendants, mais qu'ils ne permettent pas une évaluation fidèle à cause d'une faible adéquation entre les distributions de l'habileté des sujets et de la difficulté des items. L'analyse psychométrique selon l'approche de Rasch peut faire ressortir des problèmes qui ne sont pas détectés en n'utilisant que l'approche classique et peut contribuer à améliorer la qualité de l'évaluation des apprentissages complexes et des compétences en éducation médicale et en pédagogie des sciences de la santé.

1. LE CONTEXTE

La validité est la caractéristique la plus importante d'un outil d'évaluation (Downing et Haladyna, 2009). Les *Standards for Educational and Psychological Assessment* (American Educational Research Association, American Psychological Association et National Council on Measurement in Education, 2014), qui représentent le consensus scientifique actuel en éducation et en psychologie sur les pratiques à privilégier en évaluation, précisent que la validité peut être documentée à partir de cinq types de preuves, soit celles concernant le contenu, le processus de réponse, la structure interne, les relations avec d'autres variables et les conséquences de l'évaluation. En pratique, ce sont les preuves liées au contenu et à la structure interne qui sont les plus souvent utilisées pour documenter la validité d'une évaluation. La validité de contenu est essentielle puisqu'elle assure que ce qui est évalué représente bien les apprentissages visés par l'évaluation (Downing, 2003b).

Les preuves documentant la structure interne cherchent généralement à décrire les relations entre les items et les composantes d'un test de manière à déterminer si elles sont en conformité avec le construit évalué. Par exemple, si un outil vise à porter un jugement global synthétique, il est nécessaire de démontrer que l'ensemble de ses items reflètent bien une seule et même dimension et donc qu'il est justifié de les combiner pour porter ce jugement. Ces preuves de structure interne peuvent être nombreuses, les plus fréquentes étant les données sur la fidélité du test, la difficulté et la discrimination des items, la structure factorielle, l'indépendance locale, l'invariance de paramètres d'items et le fonctionnement différentiel d'items (FDI). Ces preuves de validité ont comme caractéristiques communes de reposer sur l'analyse psychométrique (Downing, 2003b). La démarche d'analyse psychométrique

et le choix des outils statistiques privilégiés pour évaluer la structure interne, et ultimement les conclusions qui en découleront, varieront toutefois selon l'approche théorique utilisée.

Les principales approches théoriques en psychométrie sont la théorie classique des tests (TCT), la théorie de la généralisabilité (GEN), la théorie des réponses aux items (TRI) et les modèles de Rasch, parfois considérés comme un cas spécial de la TRI ou comme une approche distincte de cette dernière (Bond et Fox, 2007). La TCT est l'approche la plus ancienne et la plus utilisée encore aujourd'hui. Cette popularité est probablement attribuable à plusieurs facteurs, dont ceux-ci : le fait d'avoir été la première ; elle est relativement simple et repose sur des analyses statistiques moins sophistiquées ou, du moins, plus connues ; elle est enseignée dans plusieurs programmes universitaires ; les analyses nécessaires sont implémentées dans la majorité des logiciels statistiques populaires ; elle facilite souvent la comparaison avec les études antérieures en raison de son utilisation courante ; les résultats sont faciles à communiquer. Parmi les approches plus modernes, c'est sans doute celle de Rasch qui est la plus répandue (De Champlain, 2010). Cette approche a gagné en popularité, notamment parce qu'elle est relativement plus simple que les autres, qu'elle peut être utilisée avec de plus petits échantillons et qu'elle permet de résoudre de manière élégante une vaste gamme de problèmes de mesure. La TCT et l'approche de Rasch apportent toutes deux un éclairage unique et pertinent sur les données d'évaluation et peuvent être utilisées de manière complémentaire (De Champlain, 2010).

L'approche de Rasch comporte différents avantages par rapport à la TCT (Bond et Fox, 2007 ; Embretson, 1996 ; Masters, 2005), les principaux étant les suivants :

- elle produit une mesure de niveau intervalle ;
- elle fournit une estimation de la précision du score d'habileté de chaque sujet ;
- la difficulté des items est prise en compte dans l'estimation des scores des sujets ;
- elle permet d'évaluer la qualité d'une échelle d'appréciation ou de notation (modèles pour les données polytomiques) ;
- elle utilise la même unité de mesure pour exprimer la difficulté des items et l'habileté des sujets, ce qui permet de positionner les sujets et les items sur un continuum représentant le construit évalué ;
- la difficulté des items ne dépend pas de la distribution de l'habileté de l'échantillon de sujets utilisé pour leur estimation ;
- l'habileté des sujets ne dépend pas de la distribution de la difficulté de l'échantillon d'items utilisé pour leur estimation.

Ces avantages permettent différentes applications, impossibles avec la TCT (Embretson, 1996). Par exemple, il est possible de créer une banque d'items calibrés, c'est-à-dire pour lesquels le degré difficulté est connu avec une précision suffisante, afin de faciliter le développement de plusieurs formes équivalentes d'une même évaluation. De plus, les scores obtenus à ces différentes versions seront directement comparables entre eux, ce qui est particulièrement utile en contexte d'évaluation certificative où la sécurité des questions d'examen est primordiale. Également, en ciblant mieux la difficulté des items selon l'habileté des sujets, il devient possible de créer des évaluations plus courtes, mais tout aussi fiables. Cet avantage est mis à profit dans les tests adaptatifs informatisés. En outre, l'interprétation critériée des résultats est facilitée. En plus de pouvoir comparer les sujets entre eux de manière normative, il est possible de situer leur habileté sur le continuum évalué, et ce, par rapport aux différents items. Le *Berkeley Evaluation & Assessment Research* (BEAR) *Assessment System* (Wilson et Scalise, 2006 ; Wilson et Sloane, 2000) utilise cet avantage pour situer les étudiants sur différents niveaux de développement de compétence suivant leur performance aux évaluations.

Pour ces raisons, plusieurs auteurs avancent que l'approche de Rasch est prometteuse pour l'avancement de l'évaluation en éducation, en psychologie et en santé (Bond et Fox, 2007 ; Engelhard, 2013 ; Salzberger, 2013 ; Tennant, McKenna et Hagell, 2004 ; Wright, 1999). C'est également l'avis de certains auteurs en pédagogie des sciences de la santé (Downing, 2003a ; Tavakol et Dennick, 2013), où l'approche de Rasch gagne en popularité. À titre d'exemple, elle a été utilisée dans le contexte d'entrevues de sélection, plus précisément de mini-entrevues multiples (Roberts *et al*., 2010 ; Sebok, Luu et Klinger, 2014 ; Till, Myford et Dowell, 2013), d'examens écrits (Bhakta *et al*., 2005 ; Yang *et al*., 2011), d'examens cliniques (McManus, Thompson et Mollon, 2006) et d'examens cliniques objectifs structurés (ECOS) (Boone, McWhorter et Seale, 2001 ; Iramaneerat *et al*., 2007).

S'inscrivant dans ce contexte, l'objectif de ce chapitre est d'illustrer à l'aide d'un exemple cet apport de l'approche de Rasch à l'évaluation des compétences et des apprentissages complexes en éducation médicale. Pour ce faire, nous comparons les résultats de l'analyse psychométrique de l'Échelle de communication médecin-patient pour les externes en médecine (ECMP-EM) selon l'approche de la TCT et celle de Rasch. Ce cas permettra de montrer que l'approche de Rasch apporte un éclairage complémentaire pertinent sur les données d'évaluation, que ses conclusions peuvent rejoindre celles de la TCT à certains égards, mais qu'elles peuvent également s'en éloigner.

2. LE CADRE CONCEPTUEL

Dans cette partie du chapitre, nous présentons succinctement les approches de la TCT et de Rasch afin de faciliter la compréhension de l'exemple qui sera traité par la suite. Le lecteur intéressé à en apprendre davantage pourra consulter d'excellents ouvrages sur le sujet (Anastasi et Urbina, 1997 ; Bertrand et Blais, 2004 ; Bond et Fox, 2007 ; Crocker et Algina, 2006 ; De Champlain, 2010 ; DeVellis, 2006 ; Smith et Smith, 2004).

2.1. La théorie classique des tests

Cette description de la TCT est basée sur la présentation qui en est faite dans l'ouvrage de Crocker et Algina (2006, p. 105-130) et de Graham (2006).

Les bases de la TCT ont été jetées par Spearman (1904, 1907, 1913), puis reprises et élaborées par plusieurs autres auteurs (Guilford, 1936 ; Gulliksen, 1950 ; Lord et Novick, 1968 ; Magnusson, 1967). Selon cette théorie, le score d'une personne à un test, appelé score observé (X), est composé de deux éléments : son score vrai (V) et l'erreur de mesure aléatoire (E). Bref, X = V + E. Le score vrai est le score qui représente le plus fidèlement possible la performance d'une personne à un test. Il correspond à l'espérance mathématique du score observé ou, en d'autres termes, au score moyen qu'obtiendrait une personne s'il était possible de lui faire passer le même test un nombre infini de fois. L'erreur est, quant à elle, distribuée normalement et sa valeur aléatoire est tantôt positive, tantôt négative. Sa moyenne tend vers zéro sur un nombre infiniment grand de passations du test. Il est également postulé que l'erreur aléatoire et le score vrai ne sont pas corrélés et que les erreurs aléatoires provenant de différentes passations d'un même test ou d'items d'un même test ne sont pas corrélées.

Un test est considéré comme fidèle si les scores observés se rapprochent du score vrai et, donc, que l'erreur aléatoire est faible. En reprenant l'équation de la TCT, X = V + E, il est en effet possible de constater que plus la valeur de E est petite, plus la valeur de X se rapproche de celle de V. La fidélité peut ainsi être conceptualisée comme la corrélation entre le score observé et le score vrai. De manière plus opérationnelle, elle est définie comme la proportion de la variance des scores à un test causée par les scores vrais des personnes évaluées.

La TCT postule aussi l'unidimensionnalité du test. Ce postulat signifie que les items doivent tous refléter le même score vrai, sans quoi il ne fait aucun sens de les utiliser comme indicateurs d'une même caractéristique évaluée. Ces indicateurs sont néanmoins considérés comme étant imparfaits compte tenu de la présence d'erreurs aléatoires.

L'indépendance locale des items est également exigée. Ce postulat, qui est intimement lié à celui de l'unidimensionnalité, signifie que les réponses aux items sont essentiellement expliquées par le construit évalué et que la variabilité non expliquée est principalement du bruit, c'est-à-dire de l'erreur de mesure aléatoire.

Enfin, il est généralement postulé que la force d'association entre V et X, et donc l'importance de E, peuvent varier pour chaque item (modèle congénérique).

2.2. L'approche de Rasch

Cette description s'appuie sur celles de Bond et Fox (2007), de Pallant et Tennant (2007) et de Schumacker (2004).

L'approche a été développée par Rasch (1960), originalement pour des données dichotomiques. Elle compte aujourd'hui plusieurs modèles et certains permettent de prendre en charge des données polytomiques (Andrich, 1978 ; Masters, 1982). Selon cette approche, le modèle de mesure est un idéal à atteindre permettant l'obtention d'une mesure de niveau intervalle. Bref, les données doivent se conformer à ce modèle de manière satisfaisante, sinon il est impossible d'obtenir une mesure de niveau intervalle du construit d'intérêt. Il devient alors primordial de s'assurer que les données d'évaluation s'ajustent bien au modèle, ce qui est fait à l'aide d'indices de la qualité d'ajustement. Comme pour la TCT, il est postulé que le test est unidimensionnel et que les items respectent l'exigence d'indépendance locale.

Les modèles de mesure issus de cette approche sont des modèles probabilistes s'appuyant sur la distribution logistique. Les items sont décrits par un paramètre de difficulté (δ) et les sujets par un paramètre d'habileté (β). Ces deux paramètres sont exprimés sur la même échelle dont l'unité est le *logit*, qui signifie *log-odds unit* ou, en français, le logarithme naturel du rapport de chances. Par exemple, pour un sujet ayant une probabilité de réussite de 60 %, son niveau d'habileté sera égal au logarithme naturel de 60/40, soit 0,41 *logit*. De même, un sujet ayant une habileté de 0 *logit* aura un rapport de chances de 1 ($\exp(0) = 1$), donc une probabilité de réussite de 50 % et une probabilité d'échec

de 50 %. Cette échelle peut s'étendre de –∞ à +∞, mais se situe le plus souvent entre –4,00 et 4,00. Lorsque les données s'ajustent au modèle, elles forment un continuum gradué représentant le construit évalué où il est possible de positionner à la fois les items et les sujets.

Afin d'illustrer cette approche, nous présentons ici le modèle le plus simple, soit celui visant les données dichotomiques. Ce dernier s'exprime de la manière suivante, où β_n est l'habileté du sujet n et δ_i est la difficulté de l'item i:

$$P_{ni} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}$$

Cette équation signifie que la probabilité que le sujet n réussisse l'item i est une fonction logistique de la différence entre son habileté et la difficulté de l'item i. La probabilité de réussite d'un item dépend donc de sa difficulté et de l'habileté du sujet. Si le paramètre d'habileté d'un sujet est inférieur au paramètre de difficulté d'un item, celui-ci a une probabilité inférieure à 50 % de réussir l'item. Au contraire, si le paramètre d'habileté d'un sujet est supérieur au paramètre de difficulté d'un item, celui-ci a une probabilité supérieure à 50 % de réussir l'item. Dans le cas où le paramètre d'habileté du sujet est égal au paramètre de difficulté d'un item, il a une probabilité de 50 % de réussir cet item.

3. LA MÉTHODOLOGIE

3.1. L'échantillon

Nous avons utilisé les évaluations de la compétence relationnelle médecin-patient des 634 étudiants ayant réalisé au moins un stage d'externat au doctorat en médecine entre 2011 et 2013 à l'Université Laval. Pour les étudiants ayant fait plus d'un stage pendant cette période, nous avons sélectionné aléatoirement un seul de leurs stages afin d'éviter de surestimer la covariance entre les réponses aux items qui pourrait résulter de l'inclusion d'un même étudiant plusieurs fois dans la banque de données. Ainsi, au total, l'échantillon était composé de 634 évaluations de la compétence relationnelle médecin-patient d'étudiants en stages d'externat au doctorat en médecine. Cette banque de données étant anonymisée, les caractéristiques sociodémographiques de ces étudiants n'étaient pas disponibles et ne peuvent être décrites ici.

3.2. L'instrument

L'ensemble des étudiants ont été évalués à l'aide de l'Échelle de communication médecin-patient pour les externes en médecine (ECMP-EM), présentée au tableau 2.1. L'ECMP-EM est composée de cinq items et vise à faire une appréciation synthétique globale et unidimensionnelle de la compétence relationnelle médecin-patient. Cet outil comporte un nombre restreint d'items, car la fiche d'évaluation de stage dont il fait partie doit évaluer plusieurs autres compétences. Plus précisément, cet outil évalue les quatre constituants touchant la communication médecin-patient selon l'approche clinique centrée sur le patient (Stewart *et al.*, 1995), un cadre théorique dominant en matière de soins de santé. Ces quatre constituants sont : explorer la maladie et l'expérience de la maladie vécue par le patient (1 item), comprendre le patient dans sa globalité psychosociale (1 item), s'entendre avec le patient sur le problème, les solutions et le partage de responsabilités (1 item) et établir et développer la relation médecin-patient (2 items). Les cinq items se répondent à l'aide de la même échelle d'appréciation de la performance comprenant quatre options de réponse : 1 = Insuffisant ; 2 = Limite ; 3 = Attendue ; et 4 = Supérieure. Si l'un des items ne s'applique pas au contexte du stage, l'évaluateur a la possibilité de sélectionner « Ne s'applique pas », qui est alors traité comme une donnée manquante. Cette option est toutefois peu fréquente, représentant en moyenne 6,53 % des réponses aux items.

Tableau 2.1
Échelle de communication médecin-patient pour les externes en médecine (ECMP-EM)

Évaluez la performance de cet externe à partir des cinq critères suivants en utilisant l'échelle de performance fournie. La cote « attendue » s'applique à la performance habituelle d'un étudiant en fonction de son niveau de formation.	**Supérieure**	**Attendue**	**Limite**	**Insuffisant**	**Ne s'applique pas**
1 Établit une bonne relation avec le patient relativement à la méthode clinique centrée sur le patient.	4	3	2	1	NSP
2 Explore le vécu émotif du patient dans la vision de la méthode clinique centrée sur le patient.	4	3	2	1	NSP
3 Comprend le patient dans sa globalité (contextes psychosocial et culturel) durant l'entrevue.	4	3	2	1	NSP
4 Vérifie que le patient a une bonne compréhension de son problème.	4	3	2	1	NSP
5 Utilise des attitudes et des stratégies appropriées à la relation thérapeutique avec le patient.	4	3	2	1	NSP

3.3. Les analyses

L'analyse psychométrique selon la TCT, réalisée avec le logiciel SAS version 9.4, a débuté par une analyse factorielle confirmatoire (AFC) afin de vérifier si les données d'évaluation s'ajustent au modèle unidimensionnel postulé pour l'ECMP-EM. L'étape suivante était l'analyse de la cohérence interne à l'aide du coefficient alpha de Cronbach et de la matrice des corrélations inter-items. Enfin, une analyse d'items a été réalisée. Cette dernière a consisté en l'examen du pouvoir de discrimination, estimé à partir de la corrélation item-total corrigée, et des statistiques descriptives de chacun des cinq items de l'ECMP-EM.

L'analyse psychométrique selon l'approche de Rasch a été réalisée à partir du *Rating Scale Model* (RSM) (Andrich, 1978) avec le logiciel Winsteps version 3.91, qui utilise la méthode du maximum de vraisemblance conjoint (*joint maximum likelihood estimation*) pour l'estimation des paramètres. Le RSM permet de modéliser des données d'évaluation utilisant une échelle d'appréciation graduée, comme c'est le cas pour l'ECMP-EM, qui, rappelons-le, utilise une échelle d'appréciation de la performance en quatre points. Notre démarche d'analyse s'est inspirée de celles de Pallant et Tennant (2007), de Tennant et Conaghan (2007), de Loye et Barroso da Costa (2013) et de Linacre (2004, 2014). Il a tout d'abord été nécessaire de vérifier si les conditions d'utilisation du modèle de Rasch étaient remplies, soit l'ajustement des données au modèle, l'unidimensionnalité et l'indépendance locale. Plus précisément, pour la première étape, il s'agissait d'évaluer l'ajustement des données au modèle et de procéder par itérations à des modifications visant à améliorer cet ajustement. Cette étape a été réalisée en s'intéressant à l'ajustement global des données au modèle (khi carré de l'interaction item-trait) ainsi qu'à l'ajustement des items, de l'échelle d'appréciation et des sujets (indices d'ajustement *infit* et *outfit*). La deuxième étape a consisté à s'assurer de l'unidimensionnalité de l'ECMP-EM en procédant à une analyse en composantes principales des résidus. La troisième étape consistait, quant à elle, à vérifier le respect du postulat d'indépendance locale à l'aide de la corrélation entre les résidus standardisés des items. Par la suite, l'adéquation entre les distributions des items et des sujets a été évaluée afin de vérifier si le degré de difficulté des items de l'ECMP-EM était adapté au degré d'habileté des sujets. À noter que le degré d'habileté renvoie à la compétence relationnelle médecin-patient estimée à partir des appréciations des évaluateurs. Enfin, il a été possible d'estimer le coefficient de fidélité Rasch pour la modélisation permettant le meilleur ajustement des données au modèle.

4. LES RÉSULTATS

4.1. L'approche de la théorie classique des tests

4.1.1. *L'analyse factorielle confirmatoire*

L'unidimensionnalité des données à l'ECMP-EM a été évaluée à l'aide d'une analyse factorielle confirmatoire. Cette analyse a été estimée à partir de la matrice des corrélations polychoriques inter-items et de la méthode d'estimation par moindres carrés non pondérés (*unweighted least squares*) compte tenu de la nature ordinale des données (Forero, Maydeu-Olivares et Gallardo-Pujol, 2009 ; Morata-Ramírez et Holgado-Tello, 2013). Les résultats de l'analyse factorielle confirmatoire sont présentés aux tableaux 2.2 et 2.3. Les indices d'ajustement montrent que les données peuvent être modélisées adéquatement à l'aide d'un modèle unidimensionnel selon les barèmes d'interprétation des indices d'ajustement suggérés par Schermelleh-Engel, Moosbrugger et Müller (2003). En effet, les indices GFI (*Goodness of Fit Index*), AGFI (*Adjusted Goodness of Fit Index*) et NFI (*Normed Fit Index*) sont supérieurs à 0,99 et le SRMR (*Standardized Root Mean Square Residual*) est inférieur à 0,03. Par ailleurs, les coefficients de saturation ont des valeurs élevées et supérieures à 0,30 et, selon le coefficient de détermination R^2, cette dimension unique explique entre 54 % et 82 % de la variance des réponses aux items.

Tableau 2.2
Indices d'ajustement de l'analyse factorielle confirmatoire

Indice	Valeur
SRMR	0,018
GFI	0,999
AGFI	0,998
NFI	0,999

Tableau 2.3
Saturation et coefficient de détermination des items pour l'analyse factorielle confirmatoire

Item	Saturation	R^2
1	0,73	0,54
2	0,90	0,82
3	0,85	0,73
4	0,78	0,61
5	0,89	0,79

4.1.2. *L'analyse de la consistance interne*

Le coefficient de fidélité alpha de Cronbach avait une valeur de 0,78 (tableau 2.4). Cette valeur est acceptable dans le contexte de l'ECMP-EM, puisque, d'une part, l'outil d'évaluation n'est composé que de cinq items et que, d'autre part, le jugement final sur la compétence relationnelle médecin-patient d'un étudiant ne repose pas sur une seule évaluation de stage, mais sur l'ensemble des évaluations de stages d'externat (Axelson et Kreiter, 2009). L'examen de la matrice des corrélations entre les items (tableau 2.5) permet également de constater une cohérence entre les items de l'ECMP-EM. Les cinq items corrèlent de manière positive et significative entre eux, la plus faible corrélation ayant une valeur de 0,28 et la plus forte, une valeur de 0,55. De plus, aucun item n'est apparu redondant. En effet, aucune corrélation inter-items n'est supérieure ou égale à 0,70, ce qui signifierait au moins 50 % de variance commune entre deux items.

Tableau 2.4
Coefficient alpha de Cronbach et statistiques descriptives de l'ECMP-EM

Alpha de Cronbach	Erreur type de mesure	Moyenne	É.T.	Min.	Max.
0,78	0,15	3,25	0,31	2,80	4,00

Tableau 2.5
Corrélations inter-items

Item	1	2	3	4	5
1	-	0,38	0,36	0,28	0,35
2		-	0,55	0,45	0,50
3			-	0,40	0,48
4				-	0,46

Note : Toutes les corrélations sont significatives au seuil $p < 0,01$.

4.1.3. *L'analyse d'items*

L'analyse d'items (tableau 2.6) montre que ces derniers discriminent très bien selon les barèmes d'interprétation suggérés par différents auteurs (Ebel et Frisbie, 1991 ; Laveault et Grégoire, 2014 ; Nunnally et Bernstein, 1994 ; Schmeiser et Welch, 2006), la valeur de l'indice de discrimination étant supérieure à 0,40 pour chacun d'eux. L'item 2 apparaît être celui qui discrimine le plus (0,64) et l'item 1 celui qui discrimine le moins (0,44). Les scores moyens aux items se situent entre 3,17 et 3,51, soit entre les options de réponse 3, correspondant à

une performance « Attendue », et 4, correspondant à une performance « Supérieure ». Les items 4 et 5 sont les items les plus difficiles et l'item 1 est le plus facile. La dispersion des scores aux items est faible. Sur une échelle d'appréciation allant de 1 à 4, les minimums et maximums se situent respectivement à 3,00 et 4,00 pour les items 1, 3, 4 et 5 et respectivement à 2,00 et 4,00 pour l'item 2. Toutefois, pour cet item, l'option de réponse 2 n'a été utilisée qu'une seule fois.

Tableau 2.6
Résultats de l'analyse d'items

Item	Discrimination	Moyenne	É.T.	Min.	Max.
1	0,44	3,51	0,50	3,00	4,00
2	0,64	3,18	0,39	2,00	4,00
3	0,60	3,24	0,43	3,00	4,00
4	0,51	3,17	0,38	3,00	4,00
5	0,60	3,17	0,37	3,00	4,00

4.1.4. *La synthèse des résultats selon l'approche de la théorie classique des tests*

Les résultats de l'analyse psychométrique selon l'approche de la TCT démontrent que l'ECMP-EM peut être considérée comme unidimensionnelle, qu'elle possède une fidélité (cohérence interne) acceptable et que ces items discriminent bien entre eux les candidats quant à leur degré de compétence relationnelle médecin-patient en contexte de stage d'externat en médecine. Toutefois, il apparaît que les items sont relativement faciles à endosser, les options de réponse 1 et 2 étant généralement inutilisées. En d'autres termes, les candidats ont presque tous une performance jugée « Adéquate » ou « Supérieure » par les évaluateurs.

4.2. L'approche de Rasch

4.2.1. *L'ajustement des données au modèle*

Dans un premier temps, nous avons évalué l'ajustement des données au modèle et procédé par itérations à des modifications visant à améliorer cet ajustement. Ces itérations sont résumées au tableau 2.7. L'ajustement global des données au modèle initial était adéquat, comme le démontre une interaction item-trait non significative ($\chi^2 = 11051{,}19$; $dl = 1107$; $p = 0{,}8835$). L'ajustement des items a, quant à lui, été évalué

à l'aide des indices *infit mean-squares* et *outfit mean-squares*. Le barème suggéré par Linacre (2014, p. 596) a été utilisé pour interpréter les valeurs de ces indices :

- inférieur à 0,50 : item qui contribue moins à la mesure, mais qui peut le faire sur le plan de la validité de contenu ;
- entre 0,50 et 1,50 : item qui contribue à la mesure ;
- entre 1,50 et 2,00 : item qui ne contribue pas à la mesure, mais qui ne la dégrade pas ;
- plus de 2,00 : item qui détériore la mesure.

Ainsi, des valeurs entre 0,50 et 1,50 ont été considérées comme souhaitables. Les valeurs de l'indice *infit* de l'ECMP-EM étaient adéquates. Quant aux valeurs de l'indice *outfit*, aucune n'était supérieure à 1,50, indiquant que tous les items contribuent à la mesure et qu'aucun ne la détériore. Plus précisément, ces valeurs se situaient entre 0,50 et 1,50 pour les items 1 et 4, mais étaient inférieures à 0,50 pour les items 2, 3 et 5, pour lesquels elles étaient respectivement de 0,43, 0,49 et 0,42.

Dans un deuxième temps, nous avons examiné les indices d'ajustement des options de l'échelle réponse. L'option de réponse 1 n'ayant jamais été observée dans la base de données, seuls des résultats pour les options 2, 3 et 4 étaient disponibles. Les valeurs de l'indice *infit* étaient adéquates pour ces trois options de réponse. Les valeurs de l'indice *outfit* étaient adéquates pour les options 3 et 4, mais pas pour la 2 (*outfit* de 0,00). Comme l'option 2 n'est observée qu'une seule fois dans la base de données et que son indice *outfit* est nul, d'une part, et que l'option 1 n'est pas observée, d'autre part, celles-ci ont été combinées avec l'option 3, ce qui a donné une échelle d'appréciation à deux options, soit 3 et 4. Ce deuxième modèle (tableau 2.7) présentait également un bon ajustement global tout en ayant l'avantage de faire augmenter l'indice de fidélité de 0,00 à 0,34. L'ajustement des items était comparable au regard de l'indice *infit*. Au regard de l'indice *outfit*, les items 2 à 5 avaient des valeurs entre 0,50 et 1,50, mais l'item 1 avait désormais une valeur de 1,74. Les indices d'ajustement des options de réponse étaient, quant à eux, acceptables, soit de 0,99 (option 3) et 0,98 (option 4) pour l'*infit* et de 1,40 (option 3) et 0,98 (option 4) pour l'*outfit*.

Dans un troisième temps, nous avons examiné l'ajustement des sujets aux modèles. Nous avons considéré comme étant problématiques les sujets pour lesquels l'indice *infit* ou *outfit* standardisé était supérieur à 2,50, ce qui correspond à un seuil de signification statistique d'environ 0,01 (Linacre, 2014). Alors que les valeurs non standardisées des indices *infit* et *outfit* permettent de savoir si les données contribuent

ou nuisent à la mesure, leurs valeurs standardisées permettent de vérifier si les données dévient de manière statistiquement significative du modèle (Linacre, 2002), raison pour laquelle ces valeurs ont été utilisées pour l'examen de l'ajustement des sujets au modèle. Cinq sujets ont été jugés problématiques à partir de ce critère. Nous avons alors procédé à quelques itérations en supprimant chaque fois les sujets problématiques jusqu'à ce qu'il n'y en ait plus. Au total, huit sujets ont ainsi été supprimés de la modélisation. Après cette étape, l'ajustement global du modèle était toujours adéquate (voir tableau 2.7, modèle 3), comme le démontre l'interaction item-trait non significative (χ^2 = 1155,56 ; *dl* = 1404 ; *p* = 1,000). L'ajustement des options de l'échelle d'appréciation l'était également, les indices *infit* et *outfit* se situant tous deux entre 0,77 et 1,08. Quant à l'ajustement des items, l'indice *infit* indiquait un ajustement adéquat des items, c'est-à-dire entre 0,50 et 1,50. Du côté de l'indice *outfit*, il se situait entre 0,50 et 1,50 pour les items 1, 2, 3 et 5, mais avait une valeur de 1,65, c'est-à-dire supérieur à 1,50 pour l'item 4. Puisqu'une valeur d'*outfit* entre 1,50 et 2,00 peut être interprétée comme ne contribuant pas à la mesure, mais ne la dégradant pas non plus, nous avons décidé de conserver cet item. Deux raisons expliquent ce choix. Premièrement, l'outil comprend déjà peu d'items et en retirer un affecterait la validité de contenu. Deuxièmement, un problème d'*outfit* nuit moins à la qualité de la mesure qu'un problème d'*infit* (Bond et Fox, 2007 ; Linacre, 2002). Par conséquent, nous avons considéré cette modélisation comme notre modèle final, c'est-à-dire celui s'ajustant le mieux au modèle de Rasch.

4.2.2. *La dimensionnalité et l'indépendance locale*

L'analyse factorielle confirmatoire présentée précédemment dans les résultats selon la TCT permettait de conclure que l'ECMP-EM peut être considérée comme unidimensionnelle. Néanmoins, avec l'approche de Rasch, il est suggéré de réaliser une analyse en composantes principales (ACP) des résidus afin de vérifier si des dimensions moins importantes que la première, qui correspond à la dimension évaluée par l'ECMP-EM, pourraient avoir un poids suffisant pour devoir être considérées. Selon cette méthode, si l'une des dimensions résiduelles a une valeur propre (*eigenvalue*) de 2,00 ou plus, cela indique la présence d'une autre dimension d'importance (Linacre, 2014). Pour l'ECMP-EM, le poids de la première dimension résiduelle était de 1,56, dénotant que la variance résiduelle n'est pas expliquée par une autre dimension d'importance et que cet outil peut être considéré unidimensionnel.

Relativement au postulat d'unidimensionnalité, nous avons vérifié celui de l'indépendance locale entre les items. Un problème de dépendance locale peut être révélé par une corrélation entre les résidus standardisés de deux items de 0,70 ou plus (en valeur absolue), indiquant au moins 50 % de variance commune entre les résidus de ces items (Linacre, 2014). Pour l'ECMP-EM, les valeurs de cette corrélation variaient entre –0,13 et –0,47, indiquant l'absence de dépendance locale entre les items.

Tableau 2.7
Ajustements itératifs pour ajuster les données au RSM

Description du modèle	**Ajustement global**	**Ajustement des items**		**Fidélité**
		Infit **Min.-Max.**	***Outfit*** **Min.-Max.**	
(1) Modèle initial	$\chi^2 = 1051{,}1851$; $dl = 1107$; $p = 0{,}8835$	0,87-1,18	0,42-0,93	0,00
(2) Modèle après regroupement des options de réponse 1, 2 et 3	$\chi^2 = 1206{,}3109$; $dl = 1463$; $p = 1{,}0000$	0,85-1,05	0,75-1,74	0,34
(3) Modèle après retrait des sujets s'ajustant moins bien†	$\chi^2 = 1106{,}2136$; $dl = 1346$; $p = 1{,}0000$	0,86-1,24	0,61-1,65	0,36

† Modèle final ; *Infit* : *Infit mean square* ; *Outfit* : *Outfit mean square* ; Min. : Minimum ; Max. : Maximum.

4.2.3. *L'adéquation entre la difficulté des items et l'habileté des sujets et la fidélité de la mesure*

Nous avons ensuite examiné l'adéquation entre la distribution de la difficulté des items et celle de l'habileté des sujets. Une bonne adéquation entre ces deux distributions signifie que le degré de difficulté des items est adapté à l'habileté des sujets, ce qui permet d'évaluer avec une certaine précision les degrés observés d'habileté. Par exemple, des items trop faciles ou trop difficiles pour les sujets évalués ne permettraient pas de distinguer ces derniers entre eux au regard de leur habileté. Le tableau 2.8 présente les statistiques descriptives des distributions des sujets et des items. L'examen des valeurs de ces statistiques montre que la distribution de l'habileté des sujets est plus étendue que celle des items. L'habileté des sujets s'étend de –4,36 *logit* à 3,43 *logit* (É.T. = 2,44), alors que la difficulté des items s'étend de –3,40 *logit* à 1,25 *logit* (É-T. = 1,95). De plus, les moyennes des distributions des sujets (M = –2,09 *logit*) et des

items (M = 0,00 *logit*) sont éloignées et, en moyenne, les items ont un degré de difficulté élevé pour le groupe évalué (une fois les options de réponse 1, 2 et 3 regroupées pour ajuster les données au modèle).

Tableau 2.8
Statistiques décrivant la distribution des sujets et des items (en *logit*)

	Moyenne	É.T.	Min.	Max.	Étendue
Sujets	–2,09	2,44	–4,36	3,43	7,79
Items	0,00	1,95	–3,40	1,25	4,65

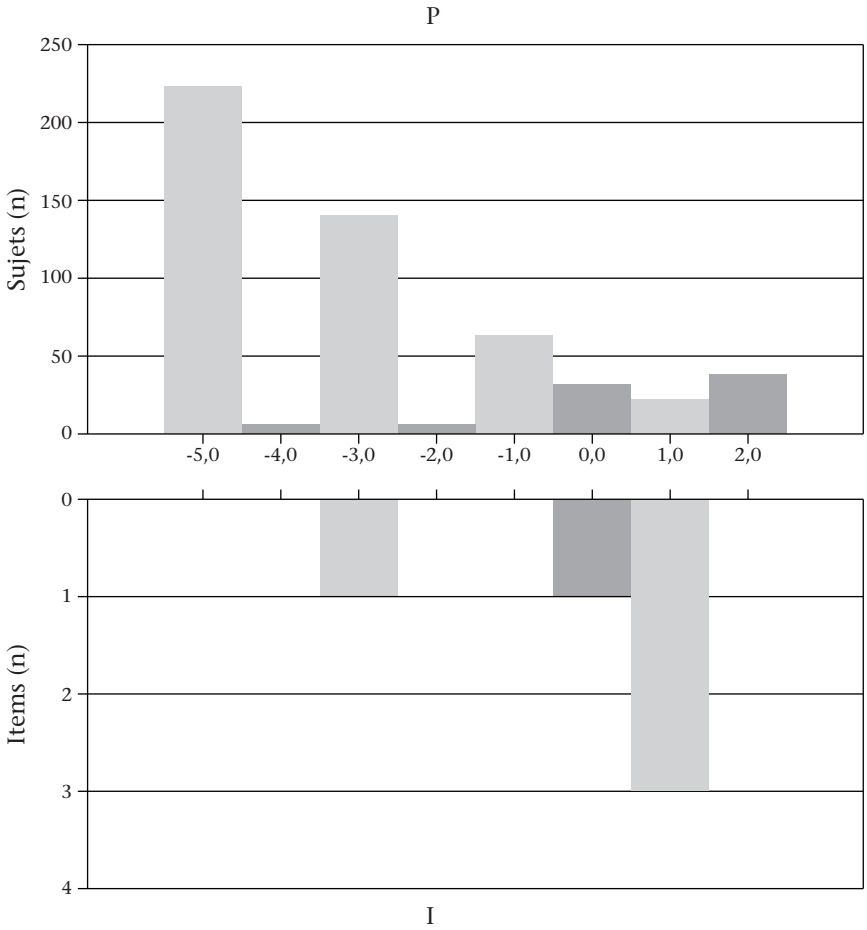


Figure 2.1
Adéquation entre la distribution des sujets (histogramme du haut) et des items (histogramme du bas)

De manière complémentaire, la figure 2.1 présente les histogrammes de distribution des sujets et des items et permet de visualiser l'adéquation entre ces deux distributions. Cette figure amène à constater trois choses. Premièrement, comme les sujets ont une distribution plus étendue que celle des items, l'ECMP-EM ne peut pas différencier entre eux les nombreux sujets ayant une habileté autour de –4,00 *logit* (environ 225 sujets), ainsi que ceux ayant une habileté située au-delà de 1,50 *logit*. Deuxièmement, la distribution des items affiche un vide entre –2,5 et 1,5 *logit*, c'est-à-dire qu'aucun item ne se retrouve dans cette partie de la distribution des habiletés, alors que plusieurs sujets ont une habileté se situant autour de –2,00. Troisièmement, un seul item se retrouve dans la partie de la distribution où se situent la majorité des sujets, c'est-à-dire entre –4,50 et –0,50. En résumé, les distributions des sujets et des items ont des formes presque inverses. La plupart des items de l'ECMP-EM (quatre sur cinq) se situent à l'endroit sur le continuum où il y a moins de sujets et les sujets sont plus nombreux dans la zone où il y a un seul ou aucun item. Le coefficient de fidélité Rasch de l'ECMP-EM reflète ces observations. En effet, sa valeur de 0,36, qui correspond à un indice de séparation de 0,75, démontre une faible précision de la mesure de l'habileté des sujets. Selon ces valeurs, les résultats ne permettent pas de distinguer des strates de candidats ayant des performances différentes de manière statistiquement significative.

4.2.4. *La synthèse des résultats selon l'approche de Rasch*

Quelques ajustements ont été nécessaires pour que les données d'évaluation de l'ECMP-EM s'ajustent adéquatement au modèle de Rasch, plus précisément au RSM : regroupement des options 1, 2 et 3 de l'échelle d'appréciation et suppression de huit sujets. Après ces ajustements, l'approche de Rasch a permis de constater que l'ECMP-EM formait un outil unidimensionnel sans problème de dépendance locale entre les items. Cependant, la distribution de la difficulté des items de l'ECMP-EM n'est pas bien ajustée à la distribution de l'habileté des sujets. Devant ce résultat, force est de constater que l'ECMP-EM ne permet pas de faire une évaluation fidèle des sujets quant à leur compétence relationnelle médecin-patient.

CONCLUSION

L'objectif de ce chapitre était d'illustrer à l'aide d'un exemple l'apport de l'approche de Rasch en contexte d'évaluation des apprentissages complexes et des compétences en éducation médicale. Les résultats des analyses psychométriques de l'ECMP-EM selon les approches de la TCT et de Rasch démontrent bien comment chacune amène à employer des

démarches d'analyse et des outils statistiques différents, d'une part, et, d'autre part, l'apport complémentaire pertinent de l'approche de Rasch. L'avantage le plus évident dans cet exemple est que les sujets et les items peuvent être positionnés sur le même continuum à partir d'une unité de mesure de niveau intervalle, le *logit*. Dans le cas de l'ECMP-EM, cela a permis de déceler une faible précision de l'estimation de l'habileté des sujets, et ce, malgré une cohérence interne acceptable de l'outil d'évaluation.

En effet, l'estimation de la fidélité Rasch a jeté un éclairage différent sur la fiabilité des résultats d'évaluation. Dans le contexte de la TCT, le coefficient alpha de Cronbach est acceptable et incite à conclure que les scores aux items sont cohérents et que les résultats d'évaluation sont reproductibles. Le coefficient de fidélité Rasch est, quant à lui, faible et suggère que l'ECMP-EM ne permet pas de situer de manière fiable les sujets sur le continuum évalué. Cette différence est attribuable au fait que ces deux coefficients diffèrent quelque peu dans leur manière d'estimer la fidélité (Schumacker, 2004 ; Smith, 2004).

Dans le contexte de la TCT, le coefficient alpha de Cronbach se base sur la force des corrélations entre les items, ou cohérence interne, pour estimer la reproductibilité des scores au test. Plus cette covariabilité entre les items représente une grande part de la variabilité totale du test, plus ce coefficient aura une valeur élevée. Dans l'approche de Rasch, le coefficient de fidélité est estimé à partir des erreurs standards des estimations des paramètres d'habileté des sujets. Ces erreurs standards représentent la précision avec laquelle l'habileté de chaque sujet a été estimée à l'aide du modèle de Rasch à partir des données d'évaluation. Cette précision dépend notamment de l'adéquation entre les distributions des items et des sujets (Linacre, 2014), qui est plutôt faible dans le cas de l'ECMP-EM. En effet, plusieurs sujets se situaient aux extrêmes de l'échelle en *logit*, là où l'erreur de mesure est grande (Linacre, 1997). D'une part, un effet de plafond est observé, c'est-à-dire qu'un bon nombre de candidats ont une habileté supérieure à ce que l'ECMP-EM est capable d'évaluer. D'autre part, un effet de plancher a été créé en regroupant les options d'évaluation 1 à 3 pour ajuster les données au modèle. À la suite de cette modification, plusieurs sujets se retrouvaient avec une habileté inférieure à ce que l'ECMP-EM était capable d'évaluer. En somme, le coefficient de fidélité Rasch a permis de tenir compte de la faible adéquation entre la difficulté des items et l'habileté des sujets dans l'estimation de la fidélité de la mesure de l'habileté des sujets.

Deux hypothèses permettent d'expliquer les résultats observés à la suite de l'analyse de Rasch. La première est que l'ECMP-EM n'est pas fiable parce qu'elle n'est pas suffisamment sensible pour

détecter les candidats ayant une compétence relationnelle médecin-patient inférieure au niveau attendu. La seconde hypothèse est que les candidats ont effectivement presque tous atteint ou dépassé le niveau attendu de compétence et que cette compétence est relativement similaire d'un candidat à l'autre. Dans ce cas, si l'objectif est uniquement de repérer les étudiants faibles relativement à cette compétence et non de mesurer avec précision le degré de compétence des étudiants jugés adéquats, la précision de l'ECMP-EM ne serait pas aussi problématique. D'autres indicateurs de la fidélité seraient toutefois également à considérer, par exemple la stabilité temporelle et l'accord inter-juges.

S'il était jugé important de mesurer avec plus de précision la compétence relationnelle médecin-patient des étudiants, deux avenues seraient à considérer. L'une d'elles consisterait à ajouter des items, notamment dans les zones de la distribution des sujets qui sont peu ou pas couvertes par les items actuels. L'autre avenue serait d'envisager de modifier ou de changer l'échelle d'appréciation de manière à mieux différencier les sujets entre eux. En effet, seulement deux des quatre options de réponse sont utilisées, ce qui revient à avoir des items dichotomiques et, en règle générale, les items dichotomiques contribuent moins à la précision de la mesure de l'habileté des sujets que les items polytomiques (Bond et Fox, 2007). Cette faible dispersion des scores sur l'échelle d'appréciation pourrait avoir un lien avec le libellé de l'option 3 « Attendue ». Il est probable qu'un large spectre de performances en communication médecin-patient soit classé par les évaluateurs comme entrant dans la catégorie « Attendue ». Pour faire une analogie avec l'échelle de notation en pourcentage, il est possible que l'ensemble des performances situées entre le seuil traditionnel de réussite (60 %) et 85 % soient considérées comme attendues, alors que celles situées entre 86 % et 100 % soient classées comme supérieures (option de réponse 4). Et si la quasi-totalité des étudiants ont une compétence relationnelle médecin-patient minimalement acceptable, presque aucun d'entre eux n'obtient la cote 2 « Limite » et encore moins la cote 1 « Insatisfaisant ». Pour corriger cette situation et favoriser une meilleure dispersion des candidats sur l'échelle d'appréciation, la catégorie « Attendue » pourrait être fragmentée en deux ou trois catégories plus fines. Une échelle d'appréciation différente pourrait aussi être testée, par exemple : 1 = Insatisfaisant ; 2 = Bon ; 3 = Très bon ; 4 = Excellent, ou encore 1 = Jamais ; 2 = Rarement ; 3 = La plupart du temps ; 4 = Toujours. Une échelle descriptive (Scallon, 2004), où chaque niveau de l'échelle d'appréciation est décrit à l'aide d'éléments observables, pourrait également être envisagée. Une telle échelle aurait l'avantage d'obliger à préciser les comportements observables reflétant

les différents niveaux de performance pour chacun des cinq items de l'ECMP-EM. De plus, entraîner les évaluateurs à l'utilisation de cette échelle descriptive pourrait améliorer la fidélité de l'évaluation.

Enfin, l'apport de l'approche de Rasch variera inévitablement selon les données d'évaluation analysées. Les conclusions peuvent être les mêmes qu'avec la TCT et n'apporter qu'un angle d'analyse différent (Courville, 2004 ; Fan, 1998). Même lorsque c'est le cas, le fait d'arriver aux mêmes conclusions avec un deuxième cadre d'analyse vient renforcer ces dernières. Dans l'exemple présenté ici, certaines conclusions demeurent les mêmes peu importe l'approche utilisée. Les deux démarches d'analyse amènent notamment à conclure que l'ECMP-EM est unidimensionnelle et qu'uniquement deux des quatre options de l'échelle d'appréciation sont utilisées, soit les options 3 et 4. Il arrive également que l'analyse psychométrique selon l'approche de Rasch permette, comme nous l'avons démontré, de relever des problèmes qui seraient passés inaperçus ou d'aller plus loin qu'en n'utilisant que la TCT. Dans d'autres études, l'approche de Rasch a parfois permis de mettre en relief des problèmes dans l'ordre des options de l'échelle d'appréciation (Andrich et Styles, 2004 ; Franchignoni *et al.*, 2011) ou de fidélité et d'unidimensionnalité de la mesure (Waugh et Chapman, 2005), ainsi que de raffiner des outils de mesure (Nijsten, Unaeze et Stern, 2006).

En conclusion, lors de la validation d'un outil d'évaluation des apprentissages complexes et des compétences en éducation médicale, l'analyse psychométrique selon l'approche de Rasch peut faire ressortir des problèmes qui ne sont pas détectés par la seule approche classique. On gagne donc à l'utiliser de manière complémentaire à la TCT lorsque les données le permettent, c'est-à-dire lorsqu'elles s'ajustent bien au modèle et qu'elles répondent à ses conditions d'utilisation.

BIBLIOGRAPHIE

American Educational Research Association, American Psychological Association et National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*, Washington : American Educational Research Association.

Anastasi, A. et S. Urbina (1997). *Psychological Testing*, 7e éd., Upper Saddle River : Prentice Hall.

Andrich, D. (1978). « A rating formulation for ordered response categories », *Psychometrika*, *43*(4), p. 561-573.

Andrich, D. et I. Styles (2004). *Final Report on the PSYCHOMETRIC ANAlysis of the Early Development Instrument (EDI) using the Rasch Model : A Technical Paper Commissioned for the Development of the Australian Early Development Instrument (AEDI)*, Murdoch University, Australia.

Axelson, R.D. et C.D. Kreiter (2009). « Reliability », dans S.M. Downing et R. Yudkowsk (dir.), *Assessment in Health Professions Education*, New York : Routledge, p. 57-73.

Bertrand, R. et J.-G. Blais (2004). *Modèles de mesure*, Québec : Presses de l'Université du Québec.

Bhakta, B., A. Tennant, M. Horton, G. Lawton et D. Andrich (2005). « Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education », *BMC Medical Education*, *5*(1), p. 1-13, doi : 10.1186/1472-6920-5-9.

Bond, T.G. et C.M. Fox (2007). *Applying the Rasch Model : Fundamental Measurement in the Human Sciences*, 2e éd., Mahwah : Lawrence Erlbaum Associates Publishers.

Boone, W.J., A.G. McWhorter et N.S. Seale (2001). « Purposeful Assessment Techniques (PAT) applied to an OSCE-Based Measurement of Competencies in a pediatric dentistry curriculum », *Journal of Dental Education*, *65*(11), p. 1232-1237.

Courville, T.G. (2004). *An Empirical Comparison of Item Response Theory and Classical Test Theory Item/Person Statistics*, Thèse de doctorat, Texas A&M University, College Station.

Crocker, L.M. et J. Algina (2006). *Introduction to Classical and Modern Test Theory*, Manson : Cengage Learning.

De Champlain, A.F. (2010). « A primer on classical test theory and item response theory for assessments in medical education », *Medical Education*, *44*(1), p. 109-117, doi : 10.1111/j.1365-2923.2009.03425.x.

DeVellis, R.F. (2006). « Classical test theory », *Medical Care Research and Review*, *44*(11), suppl. 3, p. S50-S59.

Downing, S.M. (2003a). « Item response theory : Applications of modern test theory in medical education », *Medical Education*, *37*(8), p. 739-745, doi : 10.1046/j.1365-2923.2003.01587.x.

Downing, S.M. (2003b). « Validity : On the meaningful interpretation of assessment data », *Medical Education*, *37*(9), p. 830-837, doi : 10.1046/j.1365-2923.2003.01594.x.

Downing, S.M. et T.M. Haladyna (2009). « Validity and Its Threats », dans S.M. Downing et R. Yudkowsk (dir.), *Assessment in Health Professions Education*, New York : Routledge, p. 21-55.

Ebel, R.L. et D.A. Frisbie (1991). Essentials of Educational Measurement, 5e éd., Englewood Cliffs : Prentice-Hall.

Embretson, S.E. (1996). « The new rules of measurement », *Psychological Assessment*, *8*(4), p. 341-349, doi : 10.1037/1040-3590.8.4.341.

Engelhard, G. (2013). *Invariant Measurement : Using Rasch Models in the Social, Behavioral, and Health Sciences*, New York : Routledge.

Fan, X. (1998). « Item response theory and classical test theory : An empirical comparison of their item/person statistics », *Educational and Psychological Measurement*, *58*(3), p. 357-381, doi : 10.1177/0013164498058003001.

Forero, C.G., A. Maydeu-Olivares et D. Gallardo-Pujol (2009). « Factor analysis with ordinal indicators : A Monte Carlo study comparing DWLS and ULS estimation », *Structural Equation Modeling : A Multidisciplinary Journal*, *16*(4), p. 625-641, doi : 10.1080/10705510903203573.

Franchignoni, F., G. Ferriero, A. Giordano, F. Sartorio, S. Vercelli et E. Brigatti (2011). « Psychometric properties of QuickDASH–A classical test theory and Rasch analysis study », *Manual Therapy*, *16*(2), p. 177-182, <http://www.mskscienceandpractice.com/article/S1356-689X(10)00174-8/fulltext>, consulté le 25 avril 2017.

Graham, J.M. (2006). « Congeneric and (essentially) tau-equivalent estimates of score reliability : What they are and how to use them », *Educational and Psychological Measurement*, *66*(6), p. 930-944, doi : 10.1177/0013164406288165.

Guilford, J.P. (1936). *Psychometric Methods*, 1re éd., New York et Londres : McGraw-Hill Book Company, inc.

Gulliksen, H. (1950). *Theory of Mental Tests*, New York et Londres : John Wiley & Sons, Inc.

Iramaneerat, C., R. Yudkowsky, C.M. Myford et S.M. Downing (2007). « Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement », *Advances in Health Sciences Education*, *13*(4), p. 479-493, doi : 10.1007/s10459-007-9060-8.

Laveault, D. et J. Grégoire (2014). *Introduction aux théories des tests en psychologie et en sciences de l'éducation*, 3e éd., Bruxelles : De Boeck.

Linacre, J.M. (1997). « KR-20/Cronbach Alpha or Rasch Person Reliability : Which Tells the "Truth" ? », *Rasch Measurement Transactions*, *11*(3), p. 580-581.

Linacre, J.M. (2002). « What do Infit and Outfit, Mean-square and Standardized mean ? », Rasch Measurement Transactions, 16(2), p. 878.

Linacre, J.M. (2004). « Optimizing Rating Scale Category Effectiveness », dans E.V. Smith et R.M. Smith (dir.), *Introduction to Rasch Measurement : Theory, Models, and Applications*, Maple Grove : JAM Press, p. 258-278.

Linacre, J.M. (2014). *A User's Guide to WINSTEPS MINISTEPS Rasch-Model Computer Programs – Program Manual 3.81.0*, Chicago : MESA, <http://www.winsteps.com/winman/copyright.htm>, consulté le 25 avril 2017.

Lord, F.M. et M.R. Novick (1968). *Statistical Theories of Mental Test Scores*, Reading : Addison-Wesley Pub. Co.

Loye, N. et C. Barroso da Costa (2013). « Hiérarchiser les besoins de diagnostic en mathématique en FP à l'aide d'un modèle de Rasch », *Mesure et évaluation en éducation*, *36*(2), p. 59-85.

Magnusson, D. (1967). *Test Theory*, Reading : Addison-Wesley Pub. Co.

Masters, G.N. (1982). « A Rasch model for Partial Credit scoring », *Psychometrika*, *47*(2), p. 149-174, doi : 10.1007/bf02296272.

Masters, G.N. (2005). « Objective Measurement », dans S. Alagumalai, D.D. Curtis et N. Hungi (dir.), *Applied Rasch Measurement : A Book of Exemplars*, Dordrecht : Springer, p. 15-26.

McManus, I., M. Thompson et J. Mollon (2006). « Assessment of examiner leniency and stringency ("hawk-dove effect") in the MRCP (UK) clinical examination (PACES) using multi-facet Rasch modelling », *BMC Medical Education*, *6*(1), p. 1-22, doi : 10.1186/1472-6920-6-42.

Morata-Ramírez, M. et F. Holgado-Tello (2013). « Construct validity of Likert scales through confirmatory factor analysis : A simulation study comparing different methods of estimation based on Pearson and polychoric correlations », *International Journal of Social Science Studies*, *1*(1), p. 54-61.

Nijsten, T., J. Unaeze et R.S. Stern (2006). « Refinement and reduction of the impact of Psoriasis Questionnaire : Classical Test Theory vs. Rasch analysis », *British Journal of Dermatology*, *154*(4), p. 692-700, doi : 10.1111/j.1365-2133.2005.07066.x.

Nunnally, J. et L. Bernstein (1994). *Psychometric Theory*, 3e éd., New York : McGraw-Hill.

Pallant, J.F. et A. Tennant (2007). « An introduction to the Rasch measurement model : An example using the Hospital Anxiety and Depression Scale (HADS) », *British Journal of Clinical Psychology*, *46*(1), p. 1-18, doi : 10.1348/014466506X96931.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhague : Danish Institute for Educational Research.

Roberts, C., I. Rothnie, N. Zoanetti et J. Crossley (2010). « Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview ? », *Medical Education*, *44*(7), p. 690-698, doi : 10.1111/j.1365-2923.2010.03689.x.

Salzberger, T. (2013). « Attempting measurement of psychological attributes », *Frontiers in Psychology*, *4*, p. 75, doi : 10.3389/fpsyg.2013.00075.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*, Saint-Laurent : Éditions du Renouveau pédagogique Inc.

Schermelleh-Engel, K., H. Moosbrugger et H. Müller (2003). « Evaluating the fit of structural equation models : Tests of significance and descriptive goodness-of-fit measures », *Methods of Psychological Research Online*, *8*(2), p. 23-74.

Schmeiser, C.B. et C.J. Welch (2006). « Test Development », dans R.L. Brennan (dir.), *Educational Measurement*, 4e éd., Westport : Praeger Publishers, p. 307-353.

Schumacker, R. (2004). « Rasch Measurement : The Dichotomous Model », dans E.V. Smith Jr. et R.M. Smith (dir.), *Introduction to Rasch Measurement*, Maple Grove : JAM Press, p. 226-257.

Sebok, S.S., K. Luu et D.A. Klinger (2014). « Psychometric properties of the multiple mini-interview used for medical admissions : Findings from generalizability and Rasch analyses », *Advances in Health Sciences Education*, *19*(1), p. 71-84, doi : 10.1007/s10459-013-9463-7.

Smith, E.V. (2004). « Evidence for the reliability of measures and validity of measure interpretation : A Rasch measurement perspective », dans E.V. Smith Jr. et R.M. Smith (dir.), *Introduction to Rasch Measurement*, Maple Grove : JAM Press, p. 93-122.

Smith, E.V. et R.M. Smith (2004). *Introduction to Rasch Measurement : Theory, Models, and Applications*, viii, Maple Grove : JAM Press.

Spearman, C. (1904). « The proof and measurement of association between two things », *The American Journal of Psychology*, *15*(1), p. 72-101, doi : 10.2307/1412159.

Spearman, C. (1907). « Demonstration of formulae for true measurement of correlation », *The American Journal of Psychology*, *18*(2), p. 161-169, doi : 10.2307/1412408.

Spearman, C. (1913). « Correlations of sums or differences », *British Journal of Psychology, 1904-1920*, *5*(4), p. 417-426, doi : 10.1111/j.2044-8295.1913.tb00072.x.

Stewart, M., J.B. Brown, W.W. Weston, I.R. McWhinney, C.L. McWilliam et T.R. Freeman (1995). « Patient-centered medicine : Transforming the clinical method », *British Medical Journal*, *311*(7019), p. 1580.

Tavakol, M. et R. Dennick (2013). « Psychometric evaluation of a knowledge based examination using Rasch analysis : An illustrative guide. AMEE Guide No. 72 », *Medical Teacher*, *35*(1), p. e838-e848, doi : 10.3109/0142159X.2012.737488.

Tennant, A. et P.G. Conaghan (2007). « The Rasch measurement model in rheumatology : What is it and why use it ? When should it be applied, and what should one look for in a Rasch paper ? », *Arthritis Care & Research*, *57*(8), p. 1358-1362, doi : 10.1002/art.23108.

Tennant, A., S.P. McKenna et P. Hagell (2004). « Application of Rasch analysis in the development and application of quality of life instruments », *Value in Health*, *7*, p. S22-S26, doi : 10.1111/j.1524-4733.2004.7s106.x.

Till, H., C. Myford et J. Dowell (2013). « Improving student selection using Multiple Mini-Interviews with Multifaceted Rasch modeling », *Academic Medicine*, *88*(2), p. 216-223, doi : 10.1097/ACM.0b013e31827c0c5d.

Waugh, R.F. et E.S. Chapman (2005). « An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement : What is the difference ? Which method is better ? », *Journal of Applied Measurement*, *6*(1), p. 80-99.

Wilson, M. et K. Scalise (2006). « Assessment to improve learning in higher education : The BEAR assessment system », *Higher Education*, *52*(4), p. 635-663, doi : 10.1007/s10734-004-7263-y.

Wilson, M. et K. Sloane (2000). « From principles to practice : An embedded assessment system », *Applied Measurement in Education*, *13*(2), p. 181-208, doi : 10.1207/s15324818ame1302_4.

Wright, B.D. (1999). « Common sense for measurement », *Rasch Measurement Transactions*, *13*, p. 704-705.

Yang, S.-C., M.-Y. Tsou, E.-T. Chen, K.-H. Chan et K.-Y. Chang (2011). « Statistical item analysis of the examination in anesthesiology for medical students using the Rasch model », *Journal of the Chinese Medical Association*, *74*(3), p. 125-129, <http://dx.doi.org/10.1016/j.jcma.2011.01.027>, consulté le 25 avril 2017.

CHAPITRE 3

Exploration des scores à un test de concordance de script sous la loupe de la modélisation de Rasch

Eric Dionne, Julie Grondin et Marie-Eve Latreille

Le test de concordance de script (TCS) est un instrument de plus en plus utilisé depuis les années 1990 pour développer et mesurer le raisonnement clinique chez les étudiants en médecine. De nos jours, son développement et son utilisation visent à évaluer le raisonnement clinique chez bon nombre d'étudiants qui se destinent à une profession dans le secteur de la santé (pharmacie, sciences infirmières, physiothérapie, etc.). Dans un TCS, il n'y a pas clairement de bonne ou de mauvaise réponse. En effet, le score attribué aux répondants est fonction de la distribution des scores d'un panel d'experts ayant répondu préalablement au TCS. Plus la réponse d'un sujet « concorde » avec la réponse modale du panel d'experts, plus le sujet aura un score élevé. À l'inverse, plus son score s'éloigne de la réponse modale, plus son score sera faible. La structure des scores dépend donc considérablement de la façon dont les experts auront eux-mêmes répondu aux questions. Mentionnons que les études psychométriques liées au TCS s'appuient généralement sur la théorie classique des tests. Dans le cadre de l'étude retenue, nous avons mis à profit le modèle à crédit partiel de Rasch (Masters, 1982) afin d'étudier un TCS développé en sciences infirmières (Latreille, 2012). Nous souhaitions répondre aux deux questions de recherche suivantes : 1) Est-ce que les scores à un TCS se situent sur une

échelle à intervalles égaux ? 2) Existe-t-il une différence à employer le modèle de Rasch (1960) plutôt que la théorie classique des tests, dans le cadre d'un processus de validation d'un TCS ? Pour réaliser cette étude, nous avons fait une analyse secondaire des données. Les données initiales proviennent de la recherche de Latreille (2012) qui a développé, passé et expérimenté un TCS en sciences infirmières dans le domaine de la pédiatrie. Le panel d'experts était composé de 15 répondants et l'épreuve a été proposée à 70 étudiantes en sciences infirmières. La théorie classique des tests (TCT) a été utilisée pour déterminer les propriétés psychométriques de cet instrument. Dans le cadre de nos travaux, nous avons donc réanalysé les données, mais, en mettant cette fois à profit la modélisation de Rasch pour vérifier si les items problématiques étaient les mêmes en utilisant cette dernière plutôt que la TCT.

1. LA PROBLÉMATIQUE

Les médecins ainsi que les professionnels de la santé doivent, tous les jours, prendre de nombreuses décisions dans un contexte à enjeux critiques puisque ultimement c'est de la santé des patients dont il s'agit. Ces décisions doivent souvent être prises rapidement et s'appuyer sur un processus reconnu de résolution de problème. Thivierge *et al.* (2005) dégagent d'ailleurs deux caractéristiques principales associées à la prise de décision : 1) elle s'exerce au terme d'une démarche de résolution de problème complexe ou ambiguë ; et 2) elle a lieu dans un contexte d'incertitude. Comme ils le soulignent, cette situation s'applique non seulement aux médecins, mais aussi à tous les professionnels qui doivent prendre des décisions en mettant en relation des données de différentes natures, en faisant appel à des connaissances antérieures qu'ils doivent transférer et utiliser pour exercer leur jugement clinique. Ce contexte est souvent caractérisé par l'incertitude, car les données auxquelles ils ont accès sont souvent limitées. Elles peuvent être, par exemple, valides, invalides, incomplètes ou encore biaisées, ce qui peut évidemment avoir une incidence considérable sur le jugement clinique. La théorie des scripts, développée initialement en psychologie cognitive, est un cadre théorique souvent cité pour expliquer le mode de raisonnement des médecins lorsqu'ils ont, par exemple, à poser un diagnostic. Il va sans dire que la théorie des scripts peut également expliquer le mode de pensée des autres professionnels de la santé. Ce texte n'a

pas pour objectif de décrire les fondements de la théorie des scripts, mais le lecteur désireux d'en savoir plus à ce sujet peut consulter, entre autres, l'article de Charlin, Tardif et Boshuizen (2000).

C'est dans ce contexte que le test de concordance de script (TCS) a été développé initialement dans les années 1990 afin de développer et de mesurer le raisonnement clinique chez les futurs médecins en situation d'incertitude. Il s'agit, en quelque sorte, de présenter une situation problème authentique susceptible d'être vécue par les futurs cliniciens dans le cadre de leur pratique. De nos jours, son développement et son utilisation visent à évaluer le raisonnement clinique chez bon nombre d'étudiants qui se destinent à une profession dans le secteur de la santé (médecine, sciences infirmières, physiothérapie, etc.), mais on le retrouve également en formation continue. Essentiellement, il s'agit d'un outil composé de quatre sections. Examinons la figure 3.1 pour voir de quoi il en retourne.

1. Scénario lié à la douleur

Zachary est un garçon de 6 ans pesant 20 kg qui a subi la veille une appendicectomie. Il se plaint d'une douleur de 9 sur 10 dans la région abdominale. Il a reçu son médicament contre la douleur deux heures auparavant. Le médecin a prescrit une dose de morphine 2-4 mg IV Q2H PRN.

2. Si vous pensez à… (hypothèse)	**3. Et que vous constatez…**
Administrer la dose de morphine 4 mg IV pour réduire la douleur abdominale de Zachary.	Zachary s'endort dans son lit. Son rythme respiratoire (RR) est à 12 respirations/min alors que normalement il se situe à 20 lorsqu'il est endormi et à 24 lorsqu'il est éveillé.

4. L'intervention (administrer la dose de morphine) est…

	–2	Fortement contre-indiqué ou même néfaste
	–1	Contre-indiqué
	0	Ni plus ni moins indiqué
	+1	Indiqué
	+2	Fortement indiqué ou même absolument nécessaire

Figure 3.1
Exemple d'une vignette traduite du TCS construit par Latreille

Source : Latreille, 2012.

La première section, intitulée ici « Scénario lié à la douleur », est composée d'un cas clinique fournissant aux répondants un certain nombre d'informations (âge du patient, principaux symptômes, etc.). On donne les informations de nature clinique qui permettront aux répondants de prendre connaissance du scénario. La deuxième

section, intitulée ici « Si vous pensez à… », propose, quant à elle, une hypothèse de jugement clinique qui peut s'opérationnaliser, par exemple, par un geste clinique (administrer ou non un médicament, demander un bilan sanguin, etc.). La troisième section, nommée « et que vous constatez », présente une conséquence liée à l'hypothèse précédemment émise (à la suite de l'administration du médicament, vous observez que…, le bilan sanguin démontre…, etc.). Il s'agit d'une résultante hypothétique liée au geste posé précédemment. La dernière section (l'intervention est…) permet aux répondants de se prononcer sur la valeur de l'hypothèse émise. Autrement dit, dans quelle mesure cette hypothèse, liée au raisonnement clinique, permet-elle de résoudre le problème du patient ? L'option choisie par le répondant devient ainsi la réponse à la question liée au scénario. Dans la plupart des cas, l'échelle de réponse est de type Likert avec cinq options de réponses. Par exemple, ici, l'échelle va de –2 (fortement contre-indiqué) à +2 (fortement indiqué). Une option neutre (0) est aussi généralement offerte aux répondants. L'une des particularités du TCS, et qui le rend à notre avis fort intéressant du point de vue de la mesure, est qu'il n'y a pas clairement de bonne ou de mauvaise réponse, comme c'est le cas dans un test à choix multiples. En effet, le répondant choisit d'abord l'une des cinq catégories de réponses. Par la suite, ce choix est comparé aux choix d'un panel d'experts ayant répondu préalablement au TCS. Plus la réponse d'un sujet « concorde » avec la réponse modale du panel d'experts, plus le sujet aura un score élevé. À l'inverse, plus son score s'éloigne de cette réponse modale, plus son score sera faible. Autrement dit, plus un répondant sélectionne une catégorie choisie aussi par la majorité des experts, plus il obtiendra un score élevé. La structure des scores dépend donc considérablement de la façon dont les experts auront eux-mêmes répondu aux questions. Le TCS permet, en quelque sorte, de déterminer le niveau d'accord entre le répondant et un groupe d'experts plutôt que de comparer le choix de ce répondant à une bonne réponse.

Plusieurs variables peuvent affecter les qualités métriques du TCS. Nous pouvons mentionner, à titre d'exemple, la composition du panel et le nombre d'experts (Gagnon *et al.*, 2011), le degré d'accord des experts (Blais *et al.*, 2011), le nombre de vignettes et le nombre de questions (Gagnon *et al.*, 2009), la méthode d'attribution des scores (Bland, Kreiter et Gordon, 2005 ; Charlin *et al.*, 2002 ; Wilson, Pike et Humbert, 2014), l'effet des réponses au hasard (Vanbelle *et al.*, 2007) et le nombre de catégories de réponses (Fournier, Demeester et Charlin, 2008). Une variable moins étudiée dans le cadre du TCS concerne les étiquettes des catégories de réponses qui semble,

dans le cadre de questionnaires, être une variable à ne pas négliger (Bradburn et Sudman, 1979 ; Hofmans *et al.*, 2007 ; Schwarz, 1995). Bien que certaines études se soient intéressées aux effets variables que nous venons de mentionner, il reste qu'elles sont relativement peu documentées.

Les études visant à discuter des propriétés métriques liées au TCS s'appuient majoritairement sur la théorie classique des tests (Lineberry, Kreiter et Bordage, 2013). Selon ces auteurs, la grande majorité des chercheurs (34/41) qui présentent des résultats sur la fidélité de leur test ne s'appuient que sur le coefficient de Cronbach ou le coefficient KR20. Or, l'indice alpha de Cronbach est biaisé en raison de la non-linéarité des scores bruts[1] à partir desquels l'indice est calculé (Boone, Staver et Yale, 2014, p. 223). L'indice est quasi toujours surestimé laissant croire aux concepteurs que le test possède de meilleures propriétés métriques qu'il n'en a en réalité (Linacre, 1997). Une autre critique importante adressée à l'étude des propriétés métriques des TCS touche à la fidélité inter-juges qui est, toujours selon ces mêmes auteurs, presque jamais rapportée. À notre connaissance, seule l'étude de Blais *et al.* (2011) s'est intéressée à cet aspect. Dans pratiquement toutes les études que nous avons consultées (Carrière *et al.*, 2009 ; Chang *et al.*, 2014 ; Deschênes *et al.*, 2011 ; Lambert *et al.*, 2009 ; Petrucci *et al.*, 2013), une partie importante des preuves scientifiques reposent sur les différences statistiquement significatives entre les groupes de répondants (experts, praticiens ou étudiants). Il va de soi qu'il s'agit de données importantes, mais la validité des interprétations des scores à un TCS ne peut reposer essentiellement sur ces différences, d'autant plus qu'il est parfaitement logique que les scores des experts soient plus élevés que ceux des praticiens et que les scores de ces derniers soient, à leur tour, plus élevés que ceux des étudiants. C'est seulement dans les cas où cette logique ne serait pas respectée qu'il faudrait sonner l'alarme. Dans la mesure où les scores se distribuent logiquement selon le niveau d'expertise, on peut difficilement affirmer quoi que ce soit puisque ce sont des résultats attendus. Il faut donc développer d'autres preuves scientifiques afin de remettre en question les propriétés métriques du TCS.

Dans le cadre de l'étude retenue, nous avons mis à profit le modèle à facettes de Rasch pour étudier les scores à un TCS développé en sciences infirmières (Latreille, 2012). Nous souhaitions

1. Les scores bruts représentent, selon Salkind (2013, p. 80, traduction libre), « les données initialement recueillies et non transformées au moyen d'un instrument de mesure ». Les scores bruts n'ont pas nécessairement les attributs d'une mesure objective.

répondre aux deux questions de recherche suivantes : 1) Est-ce que les scores à un TCS se situent sur une échelle à intervalles égaux ? 2) Quels sont les effets, sur le processus d'optimisation du TCS, de recourir à la modélisation de Rasch plutôt qu'à la théorie classique des tests ?

2. LE MODÈLE DE RASCH

Les qualités métriques d'un instrument peuvent être discutées et analysées en s'appuyant sur différents modèles ou théories dont les plus connues sont certainement la théorie classique des tests, la théorie de la généralisabilité, l'analyse factorielle, la théorie des réponses aux items (TRI) et la modélisation de Rasch[2]. Cette dernière est particulièrement intéressante, car le but ultime est de vérifier si les scores bruts, obtenus à l'aide d'un instrument de mesure, peuvent être situés sur une échelle à intervalles égaux. En effet, les chercheurs et les praticiens supposent souvent que les données brutes peuvent être raisonnablement considérées sur une échelle à intervalles égaux sans en faire la démonstration alors que ces données brutes se situent généralement sur une échelle ordinale. Autrement dit, la nature des scores et l'échelle qui la sous-tend ne sont pas toujours bien définies. C'est d'ailleurs le cas en ce qui concerne la nature des scores obtenus avec le TCS. En effet, les scores obtenus pourraient aussi bien se situer sur une échelle ordinale que sur une échelle à intervalles égaux, d'où la nécessité de conduire une recherche comme celle-ci afin d'obtenir un éclairage basé sur des données empiriques. Les catégories de réponses généralement rencontrées avec le TCS (totalement contre-indiqué, contre-indiqué, plus ou moins indiqué, indiqué et totalement indiqué) s'apparentent souvent à celles utilisées, par exemple, pour mesurer l'opinion (items ou échelle de type Likert). Afin d'amener des preuves supplémentaires concernant la validité des inférences à partir des scores à un TCS, il nous apparaissait pertinent de nous attarder aux propriétés métriques des échelles utilisées dans le cadre d'un tel format d'instrument.

La modélisation de Rasch a été développée par Georg Rasch (1960) il y a maintenant plus de cinquante ans. Les applications de cette modélisation trouvent écho dans de très nombreuses disciplines (sciences infirmières, médecine, éducation, marketing, etc.). Il s'agit essentiellement d'une modélisation probabiliste qui met en

2. Le lecteur peut consulter plusieurs ouvrages de référence autant en français (Bertrand et Blais, 2004 ; Penta, Arnould et Decruynaere, 2005) qu'en anglais (Crocker et Algina, 2008 ; Embretson et Reise, 2000) sur ces théories ou modèles.

relation, sur une échelle de mesure à intervalles égaux, la position de la difficulté d'un item et la position de l'habileté d'un sujet. Cette modélisation possède, comme le soulignent Cano *et al*. (2013), de nombreux avantages par rapport à la théorie classique des tests, par exemple : 1) la possibilité d'estimer les paramètres des items et des répondants et de les situer sur une même échelle ; 2) les scores élevés ou faibles reflètent en totalité un construit élevé ou faible ; 3) l'invariance des paramètres des items ; et 4) la possibilité de documenter l'adéquation données-modèle. En revanche, la modélisation de Rasch suppose le respect de certaines conditions ou de certains postulats comme l'indépendance locale et l'unidimensionnalité. Des preuves raisonnables du respect de ces deux conditions devraient être présentées pour justifier le recours au modèle de Rasch.

L'équation suivante présente la formulation mathématique du modèle de Rasch le plus simple, soit le modèle relatif aux items dichotomiques, où P_i représente la probabilité de réussir à l'item i, θ, la compétence de jugement clinique et b_i, la difficulté associée à l'item i.

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

Dans le cadre de notre étude, nous avons utilisé le modèle à trois facettes de Rasch qui permet de tenir compte des variables contextuelles.

$$P_{nijk} = \frac{e^{(B_n - D_i - C_j - F_k)}}{1 + e^{(B_n - D_i - C_j - F_k)}}$$

Où P_{nijk} représente la probabilité que le sujet n pour l'item i ayant une formation j choisisse la catégorie k. B_n représente l'estimation de la compétence de jugement clinique du sujet n sur l'échelle de mesure, D_i, la position de l'item i, C_j, la position de la formation (étudiante, praticienne) et F_k, la difficulté de passer de la catégorie k–1 à la catégorie k.

Le modèle de Rasch[3] repose sur de solides fondements théoriques. Pour pouvoir en tirer tous les bénéfices, il importe cependant de documenter deux postulats, à savoir l'unidimensionnalité et l'indépendance locale. Dans le cas du TCS, il est loin d'être assuré que

3. Dans ce texte, nous emploierons l'expression « modèle de Rasch » même si, pour être plus rigoureux, nous devrions parler des modèles de la famille de Rasch. En effet, il existe de nombreuses variantes (*rating scale*, crédit partiel, multidimensionnel, etc.) du modèle original que Rasch a proposé en 1960.

ces deux conditions soient satisfaites compte tenu du format du test. Briggs et Wilson (2004, p. 323) définissent la dimensionnalité par le nombre de traits latents qui sont, statistiquement et conceptuellement parlant, bien définis dans un test. Quand ce dernier est clairement unidimensionnel, on ne retrouve ainsi qu'un seul trait latent. Dans le cas qui nous occupe, nous considérons que le trait latent est le raisonnement clinique qui permet d'expliquer la performance des répondants à chacun des items. Quand le construit mesuré est multidimensionnel, cela suppose que deux ou plusieurs traits latents expliquent la réussite aux items du test par le répondant. La plupart des modèles de Rasch, sauf ceux, du reste, peu utilisés qui sont nettement multidimensionnels, supposent l'unidimensionnalité du construit mesuré. Précisons que l'unidimensionnalité est aussi une condition de l'utilisation de la TCT. Quant à lui, le concept d'indépendance locale est inhérent à la nature probabiliste des modèles de Rasch. La probabilité de réponse à un item ne peut être biaisée par la réponse à un autre item. Si tel était le cas, le modèle de Rasch souffrirait d'une lacune sévère. Comme le soulignent Penta, Arnould et Decruynaere (2005), il ne doit pas exister de liens entre les réponses des sujets aux différents items qui leur ont été soumis. Dans le cas du TCS, le respect du postulat d'indépendance locale pourrait être problématique puisque plusieurs items (généralement trois ou quatre) sont liés à un seul et même scénario. À cet égard, le TCS ressemble davantage à un *testlet*, c'est-à-dire à un regroupement d'items associés à un même construit mesuré et considérés comme un tout (Wainer *et al.*, 2000, p. 125). Les résultats de la présente étude pourraient permettre de jeter un certain éclairage sur cette question. Comme nous le mentionnions précédemment, des preuves raisonnables doivent permettre de discuter de ces deux postulats, mais d'autres statistiques doivent aussi être prises en compte, comme nous le verrons dans la section suivante.

3. LA MÉTHODOLOGIE

Cette partie vise à exposer le cadre méthodologique qui a servi à baliser nos travaux. Plus précisément, nous présenterons les participants de cette étude, la méthode de transformation des scores que nous avons utilisée, les méthodes de détermination des scores sur lesquelles nous nous sommes appuyés ainsi que le plan d'analyse suivi.

3.1. Les participants

Pour réaliser cette étude, nous avons réanalysé les données d'une recherche menée par Latreille en 2012 qui a développé, administré et expérimenté un TCS en sciences infirmières dans le domaine de la pédiatrie. Le panel d'experts était composé de 15 répondants (n = 15) et l'épreuve a été passée à 70 répondants (n = 70, soit 40 praticiens et 30 étudiants) en sciences infirmières.

3.2. La transformation des scores

Les données utilisées dans le cadre de cette recherche ont été obtenues à partir d'un test comportant 15 cas cliniques et 45 items. Il s'agit de données dites optimisées obtenues après avoir éliminé les items dont les indices de corrélation item-total étaient trop faibles (≤ 0,10) et ceux que les experts suggéraient d'éliminer. Ce processus permettrait, en principe, d'augmenter la fidélité du test tout en réduisant le nombre d'items nécessaires pour circonscrire le construit (Wilson *et al.*, 2014). Ainsi, la moitié des cas cliniques (30) et la moitié des items initialement développés (90) ont été abandonnés. Lors du traitement des données initiales[4], la théorie classique des tests (TCT) a été utilisée (cohérence interne, corrélations items-total, difficulté, etc.) pour cerner les propriétés psychométriques de l'instrument.

3.3. Les méthodes de détermination des scores

Les scores bruts obtenus avec le TCS se déclinaient sur une échelle en cinq catégories (–2, –1, 0, +1, +2). Pour des raisons de commodité, nous avons réalisé une première opération de transformation (–2, –1, 0, +1, +2 → 1, 2, 3, 4, 5) sur les scores bruts obtenus auprès des experts, des praticiens et des étudiants. Par la suite, il fallait choisir une méthode permettant d'accorder un score par item aux répondants en fonction du vecteur réponse des experts. Nous rappelons au lecteur qu'à ce stade du processus il n'y a pas encore de « bonne réponse » et que cette dernière doit être déterminée en fonction des réponses des experts. En examinant les écrits scientifiques, nous remarquons qu'il existe une dizaine de méthodes servant à déterminer les scores pour chacun des items à un TCS (Bland *et al.*, 2005 ; Lemay, Donnon et Charlin, 2010 ; Wilson *et al.*, 2014). Dans le cadre de cette étude, nous avons d'abord analysé les scores en nous

4. Nous rappelons au lecteur qu'il s'agit ici d'une analyse de données secondaires. Nous rapportons les décisions méthodologiques prises lors du traitement primaire des données.

appuyant sur la méthode des scores combinés développée initialement par Norman (1985) et reprise dans la plupart des articles qui traitent du TCS malgré certaines critiques (Lineberry *et al.*, 2013). Il s'agit de la méthode la plus couramment utilisée dans les écrits scientifiques rapportant les propriétés psychométriques de TCS au cours des dernières années. Cette méthode permet d'accorder un score maximal à un sujet qui choisit la réponse modale (la plus fréquemment choisie par les experts) alors qu'un score partiel est assigné si la réponse s'éloigne de la réponse modale. À titre d'exemple, pour la question 1, un répondant obtenait le score maximal (1 point) s'il choisissait la catégorie –2 ; 0,5 point pour la catégorie –1 ; 0,25 point pour la catégorie 0 ; 0,13 point pour la catégorie +1 ; et, finalement, 0 point s'il choisissait la catégorie +2. Autrement dit, plus la réponse s'éloignait de la réponse modale, plus le score obtenu était faible. Puisque le logiciel FACETS ne permet pas d'insérer des scores décimaux, les scores ont donc été transformés en pourcentages dans l'exemple que nous venons de présenter : 100 %, 50 %, 25 %, 13 % et 0 %. Les résultats peu concluants obtenus avec cette méthode au moyen du modèle de Rasch nous ont incités à nous raviser et à explorer d'autres méthodes de détermination des scores. Dans le cadre de cette étude, nous présentons les résultats obtenus à partir de deux méthodes (M1 et M2). La première (M1) consiste à fusionner les catégories situées aux extrémités (–2 et –1 ainsi que +1 et +2) de l'échelle de réponses et de ne pas accorder de crédit partiel. Il s'agit donc de scores dichotomiques obtenus à partir d'une échelle comportant trois catégories de réponses. La seconde méthode (M2) s'appuie sur une échelle ayant cinq catégories de réponses et qui n'accorde pas, elle non plus, de crédit partiel pour une réponse qui ne correspondrait pas à la réponse modale.

3.4. Le déroulement des analyses

Dans le cadre de nos travaux, nous avons donc réanalysé les données, mais cette fois-ci, en mettant à profit la modélisation de Rasch plutôt que la TCT pour vérifier si les items problématiques étaient les mêmes. Les logiciels Winsteps (Linacre, 2016a) et FACETS (Linacre, 2015) ont été employés pour réaliser les analyses touchant les 15 cas cliniques et les 45 items résultant du processus d'optimisation. Les analyses ont été réalisées en suivant les propositions de Tennant

et Conaghan (2007)[5] qui suggèrent, entre autres, d'examiner : 1) les modèles étudiés ; 2) l'ordonnancement et le nombre de catégories ; 3) la qualité de l'échelle de mesure ; 4) l'ajustement données-modèle ; 5) les indices sur le respect de l'indépendance locale (matrice de corrélation des résidus) et de la dimensionnalité et 6) les indices de fidélité. Le plan d'analyse suit donc les prescriptions de ces auteurs ainsi que les recommandations de Smith, Linacre et Smith (2003). La présentation des résultats, à la section suivante, tient compte de l'ensemble de ces recommandations ou suggestions.

4. LES RÉSULTATS ET LA DISCUSSION

La présente section expose les résultats de notre étude. Nous discuterons tour à tour des modèles étudiés, de l'échelle de mesure, des statistiques d'ajustement, des indices de fidélité, de l'analyse des résidus et de la dimensionnalité. Nous terminerons en comparant les résultats obtenus avec la TCT et ceux produits à partir de la modélisation de Rasch.

4.1. Les modèles étudiés

Il n'y a pas de démarche standardisée ni une seule « bonne méthode » lorsqu'on modélise les scores à un test. Certains chercheurs soutiennent qu'il faut ajuster le modèle aux données alors que d'autres considèrent qu'il faut plutôt ajuster les données au modèle. Nous n'entrerons pas dans ce débat dans le cadre de cet écrit. Compte tenu du caractère exploratoire de cette recherche, nous avons opté pour une approche pragmatique en réalisant de nombreuses modélisations et en les comparant. Il n'était pas possible, ici, de décrire toutes les modélisations expérimentées. Il faut cependant signaler que nous avons vérifié que la méthode de détermination des scores choisie avait un effet notable sur la qualité de la mesure. Ce résultat, en lui-même, n'est pas étonnant, mais ce qui l'est, c'est que cette méthode est couramment utilisée sans que les chercheurs la remettent en question ou s'interrogent sur son influence sur la qualité de la mesure. Comme nous le mentionnions précédemment, la méthode des scores combinés a donné de piètres résultats en employant le modèle de Rasch. À titre d'exemple, la variance

5. Le chapitre intitulé « Démonstration d'une méthodologie mettant à profit les modèles de Rasch : l'exemple d'une échelle de mesure de l'offre active de services de santé en français », soit le premier chapitre du présent ouvrage, détaille les propositions de ces auteurs.

expliquée par le modèle était extrêmement faible, les scores des répondants étaient regroupés dans un intervalle restreint de –1 à +1, sans compter plusieurs problèmes d'ajustement. Nous avons donc décidé de ne pas recourir à cette méthode bien qu'elle soit la plus couramment rapportée dans les écrits scientifiques. En soi, il s'agit déjà d'un résultat important à nos yeux, compte tenu de l'emploi quasi généralisé de cette méthode dans l'étude du TCS.

Plus précisément, nous avons examiné : 1) un modèle à trois facettes comportant trois catégories de réponses (M1) et 2) un modèle à trois facettes comportant cinq catégories de réponses (M2). Le premier modèle étudié comportait trois facettes (item, sujet et expérience) et trois catégories de réponses (–1, 0, +1). En effet, nous avons regroupé les catégories –2 et –1 ainsi que les catégories +1 et +2 pour former une échelle ayant trois catégories de réponses (–2, –1), 0, (+1, +2) → (–1, 0, +1). Le second modèle comportait les mêmes facettes (item, sujet et expérience), mais avec une échelle de réponses en cinq catégories (–2, –1, 0, +1, +2). Nous souhaitions vérifier s'il y avait des avantages à conserver les catégories situées aux extrémités de l'échelle qui permettent, normalement, d'affirmer si une hypothèse est seulement probable ou très probable. Pour ces deux modèles, nous avons traité de façon dichotomique les scores. Un répondant obtenait un score de « 1 » s'il choisissait la réponse modale et un score de « 0 » s'il choisissait une autre réponse. Nous sommes conscients que cela constitue une limite à cette recherche, car la plupart des études consultées accordent un crédit partiel pour une réponse, par exemple, adjacente à la réponse modale. Cependant, nos travaux nous laissent penser que cette méthode altère sérieusement la qualité de la mesure avec le modèle de Rasch. Nous avons donc choisi de privilégier la qualité de la mesure au détriment, d'une certaine façon, de la tendance lourde et pas toujours bien documentée à utiliser la méthode des scores combinés. Puisqu'il s'agit d'une recherche exploratoire, nous avons aussi décidé de voir s'il était possible d'obtenir une haute qualité de mesure avec des scores dichotomisés plutôt que polychotomisés.

4.2. Les échelles de mesure

Le tableau 3.1 présente les estimations des paramètres des items avec la méthode 1 (M1). Nous avons choisi de présenter ces résultats en conservant tous les répondants et tous les items puisque les statistiques d'ajustement sont satisfaisantes, comme nous le verrons dans la section suivante. Par ailleurs, le logiciel n'a pas réussi à estimer les paramètres

pour l'item 27 puisque tous les sujets ont fourni la réponse optimale pour cet item. L'échelle s'étend de –2,36 à 3,34 en excluant l'estimation arbitraire de l'item 27.

La figure 3.2 présente, quant à elle, la représentation de Wright en estimant les paramètres à partir des scores calculés avec la méthode 1. Dans un premier temps, on constate que les répondants sont tous regroupés dans un intervalle allant de 0,0 à 2,5. On constate également que les praticiennes et les étudiantes se situent presque au même endroit sur l'échelle, soit tout près de 0. Dans un second temps, on peut voir que la position des items est assez concentrée sur l'échelle de mesure. En effet, la position de la difficulté des items se situe essentiellement entre –2,5 et +2,0. Il n'y a que l'item 44 qui se situe en dehors de cet intervalle. On remarque un trou important dans l'échelle de mesure entre les valeurs +1,8 et +3,4. Il y a une douzaine de répondants qui se trouvent dans cet intervalle, mais aucun item ne possède un indice de difficulté équivalent. De façon analogue, la moitié des items sont situés dans une région de l'échelle où aucun répondant n'apparaît. De toute évidence, il faudrait ajouter des items plus difficiles afin d'obtenir un indice de difficulté qui correspond mieux à l'habileté des répondants.

Le tableau 3.2 présente les estimations des paramètres des items avec la méthode 2 (M2). On constate que le logiciel FACETS a pu estimer tous les paramètres, y compris ceux qui se rapportent à l'item 27. La position des items se situe entre –2,77 et +1,77. Cet intervalle est restreint de plus d'un *logit* (1,16) par rapport à la modélisation 1. Par ailleurs, l'erreur standard de mesure est en moyenne plus faible (0,27) pour la modélisation 2 que pour la modélisation 1 (0,36).

La figure 3.3 présente, quant à elle, la représentation de Wright en estimant les paramètres à partir des scores calculés avec la méthode 2. Dans un premier temps, on constate que les répondants sont tous regroupés dans un intervalle allant environ de –1,4 à +0,6. On constate également, comme c'était le cas avec la méthode 1, que les praticiennes et les étudiantes se situent presque au même endroit sur l'échelle (tout près de 0). Dans un second temps, on peut voir que la position des items est aussi, comme c'était le cas avec la M1, assez concentrée sur l'échelle de mesure. En effet, la position de la difficulté des items se situe essentiellement entre –2,0 et +2,0. Il n'y a que les items 27 et 41 qui se situent en dehors de cet intervalle. On remarque un trou important dans l'échelle de mesure, soit entre les valeurs –1,8 et –2,6, et plusieurs zones où l'ajout d'items serait nécessaire entre 0,0 et –1,5. Il nous apparaît encore plus important d'ajouter des items dans cette zone de l'échelle de mesure puisque près de la moitié des répondants s'y trouvent.

Tableau 3.1
Estimation des paramètres des items avec la méthode 1

Item	Score brut	Nombre de répondants	Mesure	ES Modèle	*Infit* CM	*Outfit* CM	*Infit* Z	*Outfit* Z	Corrélation PTMEA
1	60	70	–0,59	0,35	1,03	1,01	0,20	0,15	0,13
2	47	70	0,55	0,26	1,03	1,02	0,36	0,22	0,19
3	44	70	0,75	0,26	1,07	1,11	0,85	1,08	0,12
4	32	70	1,50	0,25	1,09	1,10	1,37	1,31	0,11
5	66	70	–1,62	0,52	0,97	0,69	0,07	–0,47	0,25
6	48	70	0,48	0,27	1,01	1,04	0,16	0,38	0,21
7	63	70	–1,00	0,40	0,94	0,80	–0,09	–0,45	0,30
8	66	70	–1,62	0,52	0,99	0,95	0,14	0,07	0,14
9	34	70	1,37	0,25	1,02	1,01	0,31	0,19	0,24
10	66	70	–1,62	0,52	0,99	1,02	0,14	0,21	0,13
11	66	70	–1,62	0,52	0,95	0,59	0,03	–0,72	0,31
12	40	70	1,00	0,25	1,11	1,12	1,58	1,43	0,07
13	39	70	1,07	0,25	1,04	1,03	0,60	0,48	0,20
14	47	70	0,55	0,26	1,03	1,01	0,32	0,14	0,20
15	52	70	0,18	0,28	0,93	0,85	–0,52	–0,90	0,38
16	30	70	1,62	0,25	0,93	0,93	–0,97	–0,93	0,38
17	43	70	0,81	0,25	1,04	1,05	0,51	0,55	0,18
18	36	70	1,25	0,25	0,94	0,93	–0,98	–1,02	0,37
19	36	70	1,25	0,25	1,04	1,06	0,63	0,84	0,19
20	67	70	–1,93	0,59	1,05	1,38	0,26	0,75	–0,07
21	29	70	1,68	0,25	1,05	1,05	0,75	0,62	0,17
22	54	70	0,01	0,29	0,89	0,79	–0,73	–1,13	0,45
23	61	70	–0,71	0,36	1,01	0,97	0,11	0,00	0,17

24	27	70	1,81	0,25	1,00	1,01	0,04	0,13	0,25
25	65	70	–1,38	0,47	0,99	0,87	0,10	–0,13	0,18
26	48	70	0,48	0,27	1,05	1,13	0,55	1,00	0,12
27†	70	70	–3,69	0,00	1,00	1,00	0,00	0,00	0,00
28	63	70	–1,00	0,40	0,99	0,92	0,07	–0,08	0,20
29	47	70	0,55	0,26	0,99	1,00	–0,09	0,01	0,26
30	67	70	–1,93	0,59	0,97	0,67	0,11	–0,38	0,24
31	45	70	0,68	0,26	1,04	1,06	0,44	0,60	0,17
32	68	70	–2,36	0,72	1,01	0,88	0,24	0,11	0,08
33	66	70	–1,62	0,52	1,00	1,03	0,16	0,24	0,11
34	61	70	–0,71	0,36	0,90	0,66	–0,31	–1,11	0,42
35	43	70	0,81	0,25	1,05	1,05	0,67	0,54	0,17
36	59	70	–0,47	0,33	1,01	1,03	0,11	0,19	0,17
37	60	70	–0,59	0,35	0,96	0,81	–0,09	–0,61	0,30
38	54	70	0,01	0,29	0,86	0,75	–0,91	–1,34	0,49
39	65	70	–1,38	0,47	0,95	0,73	0,01	–0,49	0,28
40	58	70	–0,36	0,32	1,00	1,10	0,09	0,48	0,17
41	68	70	–2,36	0,72	0,98	0,62	0,19	–0,29	0,21
42	31	70	1,56	0,25	1,05	1,04	0,75	0,61	0,18
43	56	70	–0,16	0,31	1,00	0,99	0,08	0,02	0,20
44	9	70	3,34	0,36	1,02	1,16	0,15	0,57	0,11
45	28	70	1,75	0,25	1,04	1,05	0,57	0,63	0,18
Moyenne	50,32	70	0,00	0,36	1,00	0,96	0,18	0,08	0,21
Écart-type	14,65	0,00	1,35	0,13	0,05	0,16	0,52	0,67	0,11
Maximum	68	70	3,34	0,72	1,11	1,38	1,58	1,43	0,49
Minimum	9	70	–2,36	0,25	0,86	0,59	–0,98	–1,34	–0,07

† L'estimation de la mesure a donc été arbitrairement fixée à –3,69 et les statistiques d'ajustement ont été fixées à 1,00 pour le carré moyen et à 0 pour la valeur standardisée.

Tableau 3.2
Estimation des paramètres des items avec la méthode 2

Item	Score brut	Nombre de répondants	Mesure	ES Modèle	*Infit* CM	*Outfit* CM	*Infit* Z	*Outfit* Z	Corrélation PTMEA
1	36	70	–0,16	0,25	0,90	0,89	–1,92	–1,89	0,44
2	22	70	0,72	0,26	1,12	1,18	1,14	1,43	–0,06
3	24	70	0,59	0,26	1,09	1,11	1,06	1,02	0,02
4	32	70	0,08	0,25	1,03	1,02	0,54	0,35	0,17
5	53	70	–1,30	0,28	1,06	1,09	0,46	0,58	0,07
6	20	70	0,87	0,27	1,01	1,05	0,09	0,39	0,17
7	24	70	0,59	0,26	1,00	1,01	0,05	0,15	0,20
8	56	70	–1,56	0,30	0,97	0,89	–0,13	–0,50	0,30
9	21	70	0,79	0,27	0,96	0,94	–0,32	–0,45	0,29
10	50	70	–1,06	0,27	0,96	0,96	–0,28	–0,28	0,30
11	50	70	–1,06	0,27	0,83	0,80	–1,44	–1,50	0,56
12	26	70	0,46	0,25	1,10	1,14	1,33	1,50	–0,01
13	22	70	0,72	0,26	0,96	0,95	–0,35	–0,39	0,29
14	30	70	0,21	0,25	1,05	1,07	0,87	1,05	0,11
15	26	70	0,46	0,25	0,99	0,99	–0,13	–0,09	0,24
16	24	70	0,59	0,26	0,88	0,85	–1,40	–1,47	0,47
17	10	70	1,77	0,35	1,06	1,21	0,35	0,80	–0,03
18	30	70	0,21	0,25	0,93	0,94	–1,11	–0,84	0,36
19	19	70	0,94	0,27	1,14	1,26	1,11	1,64	–0,14
20	45	70	–0,72	0,26	0,96	0,95	–0,38	–0,43	0,30
21	22	70	0,72	0,26	1,06	1,08	0,57	0,68	0,09
22	25	70	0,52	0,26	0,89	0,84	–1,40	–1,65	0,47
23	26	70	0,46	0,25	1,04	1,04	0,54	0,47	0,13

24	21	70	0,79	0,27	1,08	1,14	0,74	1,05	0,02
25	26	70	0,46	0,25	0,91	0,88	–1,13	–1,32	0,41
26	17	70	1,10	0,28	0,95	0,91	–0,30	–0,46	0,30
27	65	70	–2,77	0,47	0,92	0,70	–0,08	–0,59	0,36
28	58	70	–1,75	0,32	0,97	0,89	–0,10	–0,39	0,28
29	30	70	0,21	0,25	1,01	1,00	0,15	–0,04	0,21
30	36	70	–0,16	0,25	1,06	1,06	1,02	1,01	0,11
31	29	70	0,27	0,25	1,04	1,04	0,59	0,60	0,14
32	47	70	–0,85	0,26	0,93	0,91	–0,64	–0,75	0,36
33	42	70	–0,53	0,25	0,96	0,96	–0,51	–0,46	0,30
34	48	70	–0,92	0,26	1,00	1,02	0,01	0,20	0,21
35	18	70	1,02	0,28	1,00	1,06	0,03	0,45	0,17
36	45	70	–0,72	0,26	0,97	0,94	–0,38	–0,55	0,31
37	44	70	–0,65	0,25	1,03	1,03	0,33	0,40	0,17
38	47	70	–0,85	0,26	1,01	1,00	0,09	0,03	0,21
39	36	70	–0,16	0,25	0,92	0,91	–1,45	–1,51	0,39
40	30	70	0,21	0,25	1,08	1,07	1,31	1,05	0,06
41	64	70	–2,57	0,43	0,96	0,85	0,01	–0,27	0,25
42	24	70	0,59	0,26	1,07	1,12	0,83	1,08	0,04
43	14	70	1,35	0,30	1,00	0,98	0,08	–0,03	0,18
44	26	70	0,46	0,25	1,03	1,05	0,43	0,60	0,14
45	23	70	0,65	0,26	1,06	1,08	0,65	0,73	0,08
Moyenne	32,96	70	0,00	0,27	1,00	1,00	0,02	0,03	0,21
Écart-type	13,84	0	0,99	0,04	0,07	0,11	0,80	0,90	0,15
Maximum	65	70	1,77	0,47	1,14	1,26	1,33	1,64	0,56
Minimum	10	70	–2,77	0,25	0,83	0,70	–1,92	–1,89	–0,14

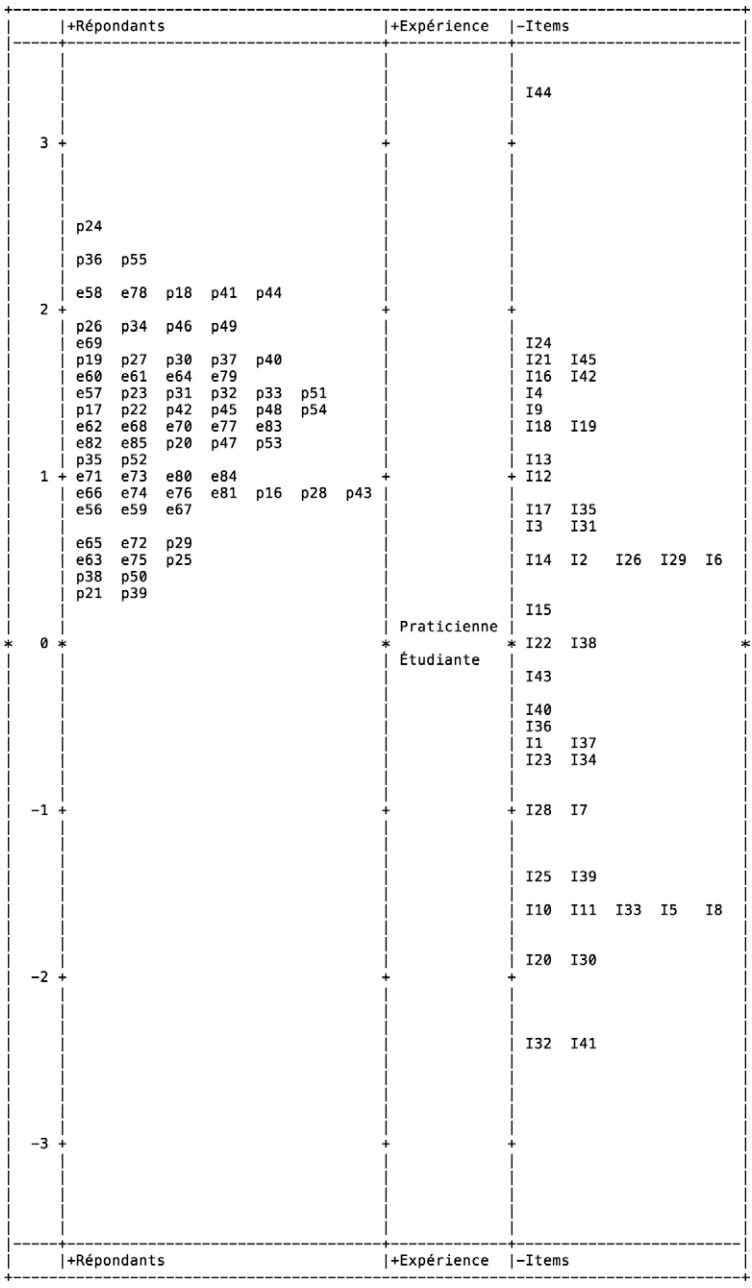


Figure 3.2
Position des répondants (e : étudiante ; p : praticienne), du niveau d'expérience et des items sur l'échelle de mesure avec la méthode 1

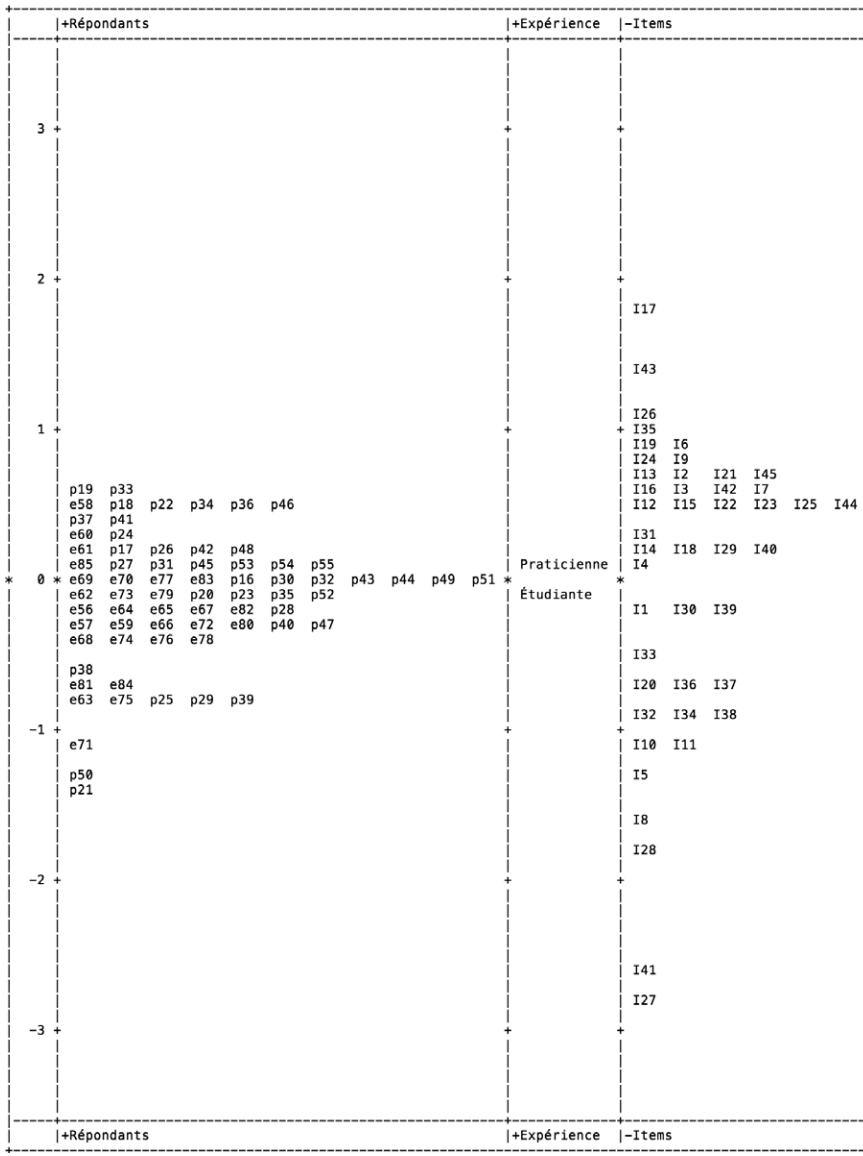


Figure 3.3
Position des répondants (e : étudiante ; p : praticienne), du niveau d'expérience et des items sur l'échelle de mesure avec la méthode 2

4.3. Les statistiques d'ajustement pour les items et les répondants

La figure 3.3 illustre la distribution des statistiques d'ajustement *infit* et *outfit* basées sur le carré moyen pour les 45 items du TCS avec la méthode 1 et la méthode 2. Les valeurs ont été classées en ordre croissant afin d'en faciliter la lecture. Linacre et Wright (1994) suggèrent différents intervalles de valeurs acceptables selon le type de test. Dans le cas d'un test à choix multiples administré dans un contexte à enjeux critiques (il s'agit du type de test qui ressemble, à notre avis, le plus au TCS), ils précisent que les valeurs de la statistique *infit* et *outfit* basées sur le carré moyen devraient être comprises entre 0,8 et 1,2. Nous rappelons au lecteur qu'une valeur équivalente à 1,0 pour cette statistique témoigne d'un ajustement parfait. L'intervalle 0,8–1,2 représente une proposition stricte. Il n'y a pas de seuils clairement définis où l'on peut affirmer que les données et le modèle ne s'ajustent pas adéquatement. Dans le cadre de notre étude, nous avons examiné l'intervalle strict tel que défini précédemment (0,8–1,2), mais nous avons également jugé correcte une valeur comprise entre 0,5 et 1,7. Cet intervalle correspond, selon Linacre et Wright, à une étendue acceptable pour un instrument qui collige des données sur des observations cliniques. Puisque le TCS est un instrument qui partage des points communs autant avec le test à choix multiples administré dans un contexte à enjeux critiques qu'avec un instrument qui collige des données sur des observations cliniques, nous nous sommes référés aux intervalles correspondants. Les valeurs que nous avons obtenues sont effectivement comprises à l'intérieur de l'intervalle strict 0,8–1,2. En effet, les valeurs basées sur le carré moyen de la statistique *infit* varient entre 0,86 et 1,11 dans le cas de la méthode 1 et entre 0,83 et 1,14 dans le cas de la méthode 2. En ce qui concerne la statistique *outfit* basée sur le carré moyen, les valeurs se situent dans l'intervalle compris entre 0,59 et 1,38 pour la méthode 1 et entre 0,70 et 1,26 pour la méthode 2. En ce qui concerne la statistique *outfit*, l'ajustement est légèrement meilleur avec la méthode 2 qu'avec la méthode 1. Pour M1, huit valeurs sont inférieures à 0,8 et une valeur est supérieure à 1,2. Pour M2, une seule valeur est inférieure à 0,8 et deux seulement sont supérieures à 1,2.

Nous avons également examiné les mêmes statistiques (*infit* et *outfit*) mais, cette fois, en nous basant sur la valeur standardisée. Linacre (2002) suggère quelques balises afin d'aider les chercheurs à interpréter cette statistique. Il soutient que les valeurs comprises dans un intervalle allant de –1,9 à 1,9 démontrent un ajustement adéquat. La figure 3.4 présente les résultats du calcul de ces statistiques pour les 45 items du test et triées en ordre croissant. Pour la statistique *infit*, les valeurs varient de –0,98 à 1,58 avec la méthode 1 et de –1,92 à 1,33 avec la méthode 2.

Pour les statistiques *outfit*, elles varient de –1,34 à 1,43 avec la méthode 1 et de –1,89 à 1,64 avec la méthode 2. Ainsi, toutes les valeurs obtenues s'inscrivent à l'intérieur de l'intervalle prescrit par Linacre.

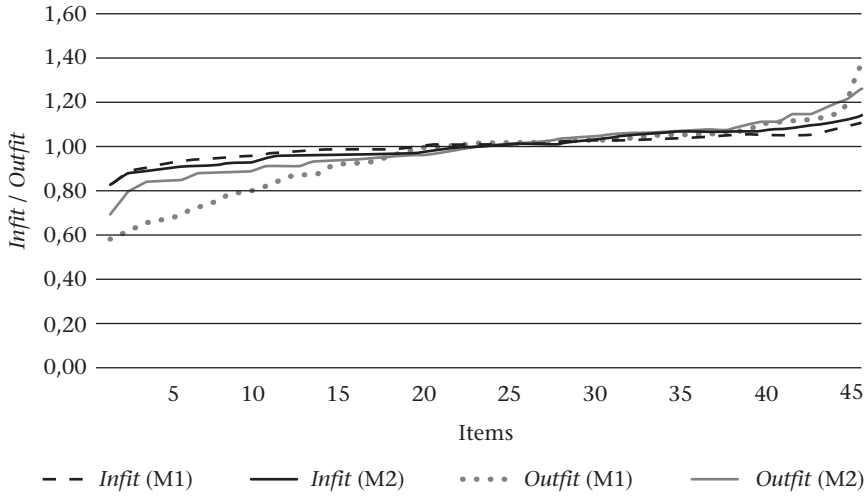


Figure 3.4
Distribution des statistiques d'ajustement *infit* et *outfit* basées sur le carré moyen (CM) pour les items et triées en ordre croissant avec la méthode 1 et la méthode 2

Pour les répondants, la statistique *infit* basée sur le carré moyen varie de 0,71 à 1,60 avec la méthode 1 et de 0,78 à 1,37 avec la méthode 2. En ce qui concerne la statistique *outfit* basée elle aussi sur le carré moyen, elle varie de 0,45 à 1,81 avec la méthode 1 et de 0,72 à 1,56 avec la méthode 2. Les résultats présentés à la figure 3.5 montrent que les valeurs associées à la statistique *infit* sont majoritairement comprises dans l'intervalle 0,8 à 1,2. Nous remarquons que les valeurs en deçà de 0,8 concernent seulement 10 répondants pour la méthode 1 et seulement un répondant avec la méthode 2. Les valeurs supérieures à 1,20 visent, quant à elles, seulement 11 répondants avec la méthode 1 et huit répondants avec la méthode 2. Il y a donc environ 28,6 % des données relatives aux répondants (21/70) qui ne sont pas comprises dans l'intervalle strict espéré pour la méthode 1 et 12,9 % (9/70) avec la méthode 2. Quoi qu'il en soit, les valeurs qui ne sont pas comprises dans cet intervalle sont, somme toute, acceptables, puisqu'elles demeurent toutes, sans exception, dans l'intervalle 0,5–1,7. Étant donné qu'il s'agit d'un TCS nouvellement développé, nous considérons que les données et le modèle s'ajustent de façon satisfaisante.

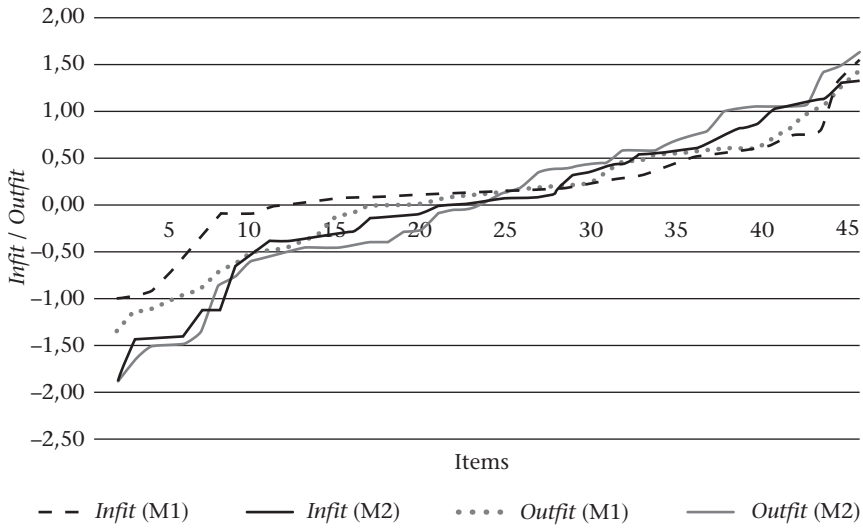


Figure 3.5
Distribution des statistiques d'ajustement *infit* et *outfit* basées sur la valeur standardisée (Z) pour les items et triées en ordre croissant avec la méthode 1 et la méthode 2

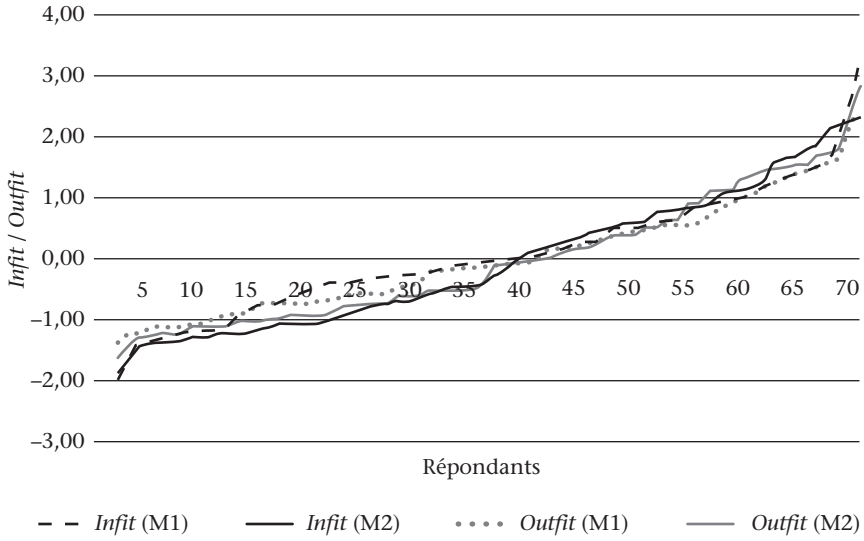


Figure 3.6
Distribution des statistiques d'ajustement pour les répondants *infit* et *outfit* basées sur le carré moyen (CM) et triées en ordre croissant avec la méthode 1 et la méthode 2

De façon analogue à ce que nous avons présenté pour les items, la figure 3.6 présente les résultats standardisés pour les statistiques d'ajustement des répondants. Pour la statistique *infit*, il n'y a que quatre valeurs qui ne se situent pas dans l'intervalle –1,9 à +1,9 avec la méthode 1. Pour ce qui est de la méthode 2, il n'y a que cinq valeurs qui sont à l'extérieur de cet intervalle. Pour la statistique *outfit*, il n'y a que deux valeurs supérieures à 1,9, autant pour la méthode 1 que pour la méthode 2.

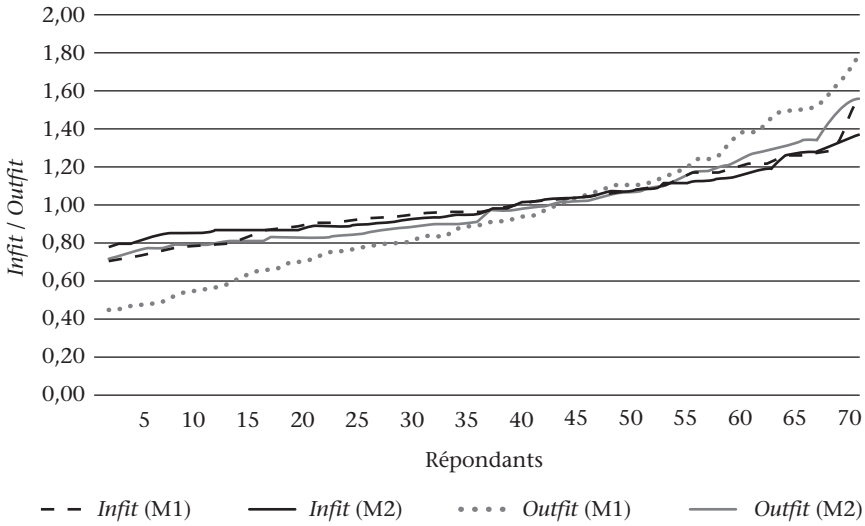


Figure 3.7
Distribution des statistiques d'ajustement pour les répondants *infit* et *outfit* basées sur la valeur standardisée (Z) et triées en ordre croissant

Dans l'ensemble, autant pour les items que pour les répondants, nous remarquons que les statistiques d'ajustement sont très bonnes et s'inscrivent à l'intérieur des intervalles suggérés dans les écrits scientifiques. Ce constat s'applique aussi bien aux résultats obtenus avec la méthode 1 qu'à ceux obtenus avec la méthode 2.

4.4. Les indices de fidélité

La présente section est consacrée aux indices de fidélité. Dans la théorie classique des tests, l'indice légendaire alpha de Cronbach est omniprésent et abondamment rapporté dans les écrits scientifiques. Les chercheurs qui s'appuient sur la modélisation de Rasch délaissent cette statistique en raison, comme nous le relevions précédemment, du risque

de surestimation de la fidélité liée à la nature non linéaire des scores bruts. Mallinson, Stelmack et Velozo (2004) ont aussi montré les limites de l'alpha de Cronbach lorsqu'il s'agit de réduire le nombre d'items dans une épreuve. Le tableau 3.3 présente les principaux indices associés à la fidélité des deux ensembles de données que nous avons étudiés. Nous présentons les bornes inférieures (min.) et supérieures (max.) pour ces indices[6]. Comme le soulignent Boone, Staver et Yale (2014), lorsque l'indice de séparation des items est trop faible (<3), cela est souvent le signe que l'échantillon de répondants est trop petit pour s'assurer de la validité des estimations de la difficulté des items. Dans le cas qui nous occupe, nous voyons que l'indice de séparation des items (*item separation*) se situe à 3,36 pour la valeur minimale et à 3,40 pour la valeur maximale pour le modèle 1 et entre 3,39 et 3,43 pour le modèle 2, ce qui est acceptable. L'échantillon de répondants permettrait de bien estimer les paramètres des items. En ce qui concerne l'indice de séparation des répondants, nous remarquons que les valeurs sont faibles. Pour le modèle 1, les valeurs associées sont 0,85 et 0,86 alors que pour le modèle 2, les valeurs correspondantes sont 0,84 et 0,85. Duncan *et al*. (2003) proposent une balise inférieure se situant à 1,5. Les valeurs que nous avons obtenues sont, à cet égard, très faibles (<1) et confirment l'interprétation visuelle des données illustrées par la représentation de Wright (figures 3.2 et 3.3), à savoir que les deux modèles étudiés ne permettent pas de bien distinguer les répondants manifestant un faible raisonnement clinique de ceux qui en démontrent un bon. En effet, sur les figures 3.2 et 3.3, il est possible de voir que les répondants sont tous concentrés dans une portion limitée de l'échelle bornée entre 0 et 2,5 dans le cas de la première modélisation et entre –1,5 et 0,75 pour la seconde modélisation. Un indice de séparation des répondants d'environ 0,8 permettrait de distinguer 2 à 3 niveaux de répondants, ce qui est peu dans le contexte d'utilisation du TCS. L'indice de séparation pour la facette expérience se situe, quant à elle, entre 1,61 et 2,49 pour le modèle 1 et entre 1,86 et 2,82 pour le modèle 2. Comme nous pouvons le constater, ces valeurs suggèrent qu'il est possible de discriminer les deux groupes de répondants (praticiennes et étudiantes) sur l'échelle de mesure. Le faible indice de fidélité des répondants présenté dans le tableau 3.1 confirme les résultats obtenus avec l'indice de séparation. Les indices de fidélité pour les items et l'expérience sont satisfaisants (>0,70) et autorisent à penser que l'échantillon de répondants est adéquat pour l'estimation des paramètres des modèles.

6. Le logiciel FACETS nomme la borne inférieure *population* et la borne supérieure *sample*.

Tableau 3.3
Indice de séparation et fidélité pour les trois facettes étudiées

		Séparation		Fidélité	
		Min.	**Max.**	**Min.**	**Max.**
Items	M1	3,36	3,40	0,92	0,92
	M2	3,39	3,43	0,92	0,92
Répondants	M1	0,85	0,86	0,42	0,43
	M2	0,84	0,85	0,41	0,42
Expérience	M1	1,61	2,49	0,72	0,86
	M2	1,86	2,82	0,78	0,89

4.5. L'indépendance locale

L'analyse des résidus standardisés présentée au tableau 3.4, c'est-à-dire une fois le facteur de Rasch extrait, révèle deux corrélations inter-items ayant une valeur d'au moins 0,40 (ou de –0,40 ou moins) pour la méthode 1. Il s'agit de la paire d'items 4 et 6 qui affiche une corrélation de –0,44 et des items 34 et 41 qui affichent une corrélation de 0,42. Il convient de souligner que les items 34 et 41 n'appartiennent pas au même scénario clinique alors que c'est le cas pour les items 4 et 6. De façon générale, pour considérer que deux items sont à risque de présenter un problème de dépendance locale, il devrait y avoir une corrélation positive d'au moins 0,70. En effet, rappelons que la variance commune entre deux items dont la corrélation est de 0,40 est de 0,40*0,40 soit 16 %, ce qui veut dire qu'il y a 84 % de différence entre ces items. Les items présentent donc plus de différences que de ressemblances. Ainsi, une corrélation de 0,40 est considérée comme un indice de faible dépendance inter-items (Linacre, 2016b). De plus, une corrélation négative est généralement à l'opposé de la définition habituellement donnée à un problème de dépendance locale. Les résultats ne semblent donc pas révéler de problème de dépendance locale avec la méthode 1. Pour la méthode 2, ce sont les deux paires d'items 4 et 36, et 26 et 38 qui montrent une corrélation, en absolu, de plus de 0,40 (avec une valeur de –0,41). Encore une fois, étant donné la valeur négative et inférieure à 0,70 de la corrélation, les résultats ne semblent pas poser un problème de dépendance locale entre les items avec la méthode 2.

Tableau 3.4
Valeurs corrélationnelles des items basées sur les résidus standardisés

Modèle 1		**Modèle 2**	
Corrélation	**Items**	**Corrélation**	**Items**
0,42	34-41	0,37	36-41
0,39	17-35	0,33	27-32 / 24-45
0,35	38-39†	0,31	8-18 / 30-36 / 18-31 / 1-15
0,33	3-19 / 24-45 / 11-39	–0,41	4-36 / 26-38
0,32	17-36 / 6-17	–0,38	1-2†
0,30	36-41	–0,36	19-29
0,29	22-38 / 6-21	–0,35	27-44
–0,44	4-6†	–0,34	23-45
–0,38	3-17	–0,33	8-24 / 6-41
–0,36	1-44 / 3-18	–0,32	12-22
–0,32	16-29	–0,31	14-16 / 9-26
–0,29	2-44 /14-45 / 12-22 / 14-38	–0,30	20-40

† Items qui appartiennent à un même scénario.

4.6. La dimensionnalité

Nous avons aussi examiné si le postulat d'unidimensionnalité semblait être respecté. Les résultats, présentés au tableau 3.5, montrent que la variance expliquée par la mesure est de 24,9 % pour la méthode 1 et de 19,2 % pour la méthode 2, ce qui est très faible (Linacre, 2016b). De ce pourcentage, la plus grande proportion de la variance provient des items (19,3 % [M1] et 16,2 % [M2]) et pour les répondants, c'est 5,6 % dans le cas de M1 et 3,1 % dans le cas de M2. Ces résultats sont assez cohérents avec la concentration des sujets que nous avions notée dans les figures 3.2 et 3.3. Ces résultats laissent également sous-entendre que les scores du TCS pourraient être influencés par différents facteurs et non pas seulement par l'habileté des répondants. Selon Linacre (2003), il est souvent difficile de déterminer avec précision les raisons qui pourraient expliquer un si faible pourcentage de variance expliquée par la mesure. Selon ce dernier, trois raisons pourraient être à l'origine de cette situation : 1) l'estimation de l'habileté des répondants et l'estimation de la difficulté des items ne sont pas précises ; 2) les scores des répondants sont aléatoires, mais demeurent alignés sur les prédictions du modèle ; et 3) les scores des répondants ne sont pas prédits par le modèle. Dans le cas qui nous occupe, compte tenu du fait que les statistiques d'ajustement sont très bonnes, ces hypothèses nous

semblent peu probables pour expliquer les résultats que nous avons obtenus. Dans une série de simulations avec des items dichotomiques, Linacre (2008) a relevé un lien étroit entre, d'une part, l'étendue de la distribution de l'estimation de la difficulté des items et l'étendue de la distribution de l'estimation de l'habileté des répondants et, d'autre part, le pourcentage de variance expliquée par le facteur Rasch. Autrement dit, le pourcentage de variance expliquée serait moindre dans le cas où les distributions sont moins étendues. C'est une hypothèse qui serait plus probable étant donné que les estimations, tant de la difficulté des items que de l'habileté des sujets, sont concentrées comme l'illustre la représentation de Wright.

Suivant les résultats de l'analyse en composantes principales effectuée sur les résidus standardisés, on note que pour la méthode 1, la valeur propre associée aux cinq contrastes analysés par le logiciel Winsteps est supérieure à 2 (tableau 3.5). Elle l'est également pour les trois premiers contrastes de la méthode 2. Cela révèle la présence potentielle de regroupements de deux, trois ou peut-être quatre items qui pourraient former des sous-dimensions (facteurs). En effet, une valeur propre d'au moins 2 correspond à un regroupement d'au moins deux items, ce qui constitue le nombre minimal d'items pour former une sous-dimension (Linacre, 2016b).

Tableau 3.5
Résultats de l'analyse en composantes principales des résidus standardisés

Variance expliquée et valeurs propres	**Modèle 1**	**Modèle 2**
Mesure	24,9 % (14,58)	19,2 % (10,70)
Répondants	5,6 % (3,30)	3,1 % (1,70)
Items	19,3 % (11,28)	16,2 % (9,00)
Inexpliquée	75,1 % (44,00)	80,8 % (45,00)
Pourcentage de variances inexpliquées		
Contraste 1	6,0 % (3,51)	6,3 % (3,52)
Contraste 2	5,0 % (2,92)	4,9 % (2,75)
Contraste 3	4,4 % (2,60)	4,4 % (2,46)
Contraste 4	4,3 % (2,53)	
Contraste 5	3,8 % (2,23)	

Le tableau 3.6 dresse la liste, pour chacun des contrastes, des items qui corrèlent le plus et de ceux qui corrèlent le moins avec le facteur. Notons que seuls les items dont la corrélation, en valeur absolue, est d'au moins 0,40 sont présentés dans ce tableau, puisque ce sont

généralement ceux qui sont considérés comme importants (Linacre, 2016b). Il est ainsi possible de repérer différents groupes d'items pouvant former des sous-dimensions à la mesure du raisonnement clinique à l'aide du TCS à l'étude. Les items 16 et 18, 32 et 33, 38 et 39 ou 37 et 38 pourraient former des paires d'items à surveiller puisqu'ils appartiennent à un même scénario. Plus encore, les items 37, 38 et 39 sont liés à un même scénario. Peut-être conviendrait-il de leur porter attention.

Tableau 3.6
Synthèse de l'analyse de la dimensionnalité des items

Contrastes	Méthode 1			Méthode 2		
	Groupe d'items†		Corrélation‡	Groupe d'items		Corrélation
	≥ 0,40	≤ –0,40		≥ 0,40	≤ –0,40	
1	21-6-17-13-36	3	–0,7542	20-27-32*-33*	14	–0,3293
2	2-24-19	26	–0,8012	24-45	23-36-40	–0,8788
3	22-29-34-38*-39*	3-14	–0,3425	8-18	2-37*-38*	–0,4435
4	18-39*-16•-38*	41	–0,2581	—	—	—
5	5-25	—	(–1,00)	—	—	—

† Regroupements d'items selon la corrélation avec le facteur. Les items sont présentés en ordre décroissant selon la force de leur corrélation avec le facteur.

‡ La corrélation atténuée est une corrélation ajustée par des opérations statistiques de façon à réduire l'effet de l'erreur de mesure. Cette corrélation est établie entre la mesure estimée pour les sujets à l'aide de tous les items qui corrèlent le plus avec le facteur (ce qui inclut les items dont la corrélation est d'au moins +0,40) et ceux qui corrèlent le moins (incluant les items dont la corrélation est moindre que –0,40). Les parenthèses indiquent que la corrélation atténuée ne pouvait pas être estimée.

• et * Items qui appartiennent à un même scénario.

Le tableau 3.6 montre également la valeur de la corrélation atténuée entre les items qui corrèlent le plus avec le facteur et ceux qui corrèlent le moins. Lorsque le résultat s'approche de 1,0, les items des deux groupes seraient statistiquement les mêmes et il ne serait ainsi pas possible de rejeter l'hypothèse que les deux groupes d'items mesurent la même chose (Linacre, 2016b). Pour la méthode 1, la valeur des deux premiers contrastes semble assez élevée (>0,70) pour que les items puissent mesurer un seul et même construit. Pour ce qui est du troisième et quatrième contraste, par contre, la valeur de la corrélation atténuée est, en valeur absolue, moindre que 0,57. Suivant les balises proposées par Linacre (2016b) pour l'interprétation de cette corrélation, une valeur moindre que 0,57 révèle que les items montrent moins de la moitié de leur variance en commun (0,57*0,57 = 32,5 %). En d'autres mots, ces items sont plus indépendants que dépendants, ce qui signifie qu'ils ne mesureraient pas la même chose. Ainsi, les

paires d'items 38 et 39, ainsi que 16 et 18 sont donc susceptibles de former une sous-dimension. Pour la méthode 2, la corrélation atténuée du deuxième contraste est assez élevée pour que les items puissent mesurer un seul et même construit ; il n'en est pas de même pour les contrastes 1 et 3. Il en ressort que les items 32 et 33 qui appartiennent à un même scénario, de même que les items 37 et 38, pourraient être à surveiller davantage. Dans le cas où ce TCS serait réutilisé, il faudrait examiner avec attention ces regroupements d'items.

4.7. La comparaison des modèles

La finalité d'une telle entreprise de recherche demeure de savoir dans quelle mesure un modèle permet de mieux appréhender le réel. L'un des buts de cette recherche était d'examiner dans quelle mesure les scores à un TCS pouvaient se situer sur une échelle à intervalles égaux. Nos résultats ont présenté essentiellement les fruits de nos analyses en mettant à profit le modèle de Rasch. Nous rappelons au lecteur qu'il peut consulter les travaux de Latreille (2012) concernant les résultats des analyses avec la TCT. Dans le cadre de nos analyses, nous avons constaté que tous les items analysés avec le modèle à trois facettes et comportant trois catégories semblaient adéquats. Nous en sommes arrivés à la même conclusion avec le modèle à trois facettes comportant cinq catégories de réponses. Nous n'avons aucune raison, basée sur la mesure, d'éliminer ne serait-ce qu'un seul item. Dans l'ensemble, aucun des deux modèles ne s'est clairement démarqué de l'autre. Les résultats sont comparables entre les deux modélisations et il ne semble pas y avoir de différence à utiliser une échelle de trois catégories plutôt qu'une échelle comportant cinq catégories de réponses dans le cas où les items sont dichotomisés. Par ailleurs, malgré le processus d'optimisation, les analyses de la TCT tendent à montrer que neuf items (2-6-12-17-19-24-30-35-42) pourraient encore être éliminés en nous basant sur les critères (p. ex. les valeurs corrélationnelles) assurant, en principe, l'optimisation du TCS. Comme nous pouvons le constater, les conséquences associées à l'utilisation d'un modèle de mesure (TCT ou Rasch) plutôt qu'un autre peuvent être importantes. Dans tous les cas, nous souhaitons éliminer les items qui apportent peu d'information et conserver ceux qui, au contraire, fournissent une information importante et riche pour la bonification de l'instrument. Comme nous le mentionnions au début de ce chapitre, les propriétés psychométriques du TCS sont largement rapportées sous l'angle de la théorie classique des tests. Notre étude, sans tirer de conclusions définitives sur la question, nous révèle qu'il serait judicieux d'étudier les propriétés psychométriques en mettant à profit d'autres modélisations statistiques des scores. Il serait alors possible d'avoir une idée plus juste des propriétés de cet instrument.

CONCLUSION

Dans le cadre de cette recherche, nous nous sommes intéressés aux deux questions de recherche suivantes : 1) Est-ce que le score à un TCS se situe sur une échelle à intervalles égaux ? 2) Quels sont les effets, sur le processus d'optimisation du TCS, de recourir à la modélisation de Rasch plutôt qu'à la théorie classique des tests ? Avant de répondre à la première question, signalons qu'à notre connaissance aucune étude ne s'était, jusqu'à maintenant, intéressée à la modélisation des scores des répondants obtenus avec un TCS en mettant à profit le modèle de Rasch. Les résultats que nous avons obtenus nous laissent penser que le modèle de Rasch pourrait être adéquat pour situer sur une échelle à intervalles égaux la position de la difficulté des items et la position de l'habileté des répondants. À titre d'exemple, mentionnons les statistiques d'ajustement qui se sont révélées étonnamment bonnes surtout si l'on considère que l'instrument développé par Latreille n'en est qu'à une première expérimentation. Les indices de fidélité, pour la plupart, étaient adéquats même si ceux associés aux répondants étaient faibles. En revanche, la variance expliquée par la mesure demeure relativement faible. Il faudra mener d'autres études semblables à celle-ci avant de se prononcer avec conviction sur le caractère objectif de la mesure avec un TCS.

Parallèlement à d'autres études qui s'intéressent aux scores obtenus avec d'autres TCS, nous pouvons ajouter que des études de simulation seraient également pertinentes afin de mettre réellement à l'épreuve le modèle de Rasch dans toutes les situations qu'offre le TCS. Cette étude ne représente qu'une modeste tentative pour mieux circonscrire les propriétés psychométriques du TCS. Quoi qu'il en soit, cette étude est, à notre connaissance, la première à mettre à profit la modélisation de Rasch afin de mieux comprendre les propriétés des scores des répondants à un TCS. Les chercheurs qui s'intéressent aux propriétés psychométriques du TCS devraient examiner les scores non seulement sous la loupe de la théorie classique, mais également en s'appuyant sur différentes modélisations comme celle de Rasch. Les écrits scientifiques tendent à montrer que le TCS est pratiquement toujours examiné sous l'angle restrictif de la TCT. Cette étude illustre l'existence d'autres indicateurs pouvant être utilisés afin d'analyser les propriétés du TCS sous d'autres angles.

En ce qui concerne la seconde question de recherche, nos travaux semblent démontrer que l'optimisation d'un TCS serait améliorée en utilisant la modélisation de Rasch en permettant, entre autres, de repérer les items réellement problématiques tout en conservant des items

utiles qui seraient trop rapidement abandonnés à la seule lumière de la TCT. Il faut aussi retenir que des différences importantes peuvent s'observer selon le modèle de mesure utilisé.

Nous tenons également à souligner les piètres résultats obtenus quand nous avons essayé de modéliser les scores au moyen de la méthode des scores combinés. Des critiques sévères ont été formulées à l'égard de cette méthode et nos travaux incitent à la prudence avec l'emploi de cette méthode, du moins lorsque les considérations métriques sont un enjeu de première importance. D'autres études devront être menées afin de vérifier si les résultats obtenus dans le cadre de cette recherche sont généralisables à d'autres TCS.

Il ne faut pas s'étonner que les propriétés psychométriques du TCS ne soient pas parfaites surtout quand l'instrument n'a été expérimenté qu'à une seule occasion. En effet, les construits complexes et les compétences ne sont, encore aujourd'hui, que bien peu étudiés. L'évaluation de ces construits demeurent un défi colossal qui exigera de raffiner autant les méthodes d'analyse que les instruments eux-mêmes.

BIBLIOGRAPHIE

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

Blais, J.-G., B. Charlin, J. Grondin, C. Lambert, N. Loye et R. Gagnon (2011). « Estimation du degré d'accord entre des experts lors du calibrage d'un test de concordance de script avec le modèle à facettes de Rasch », dans G. Raîche, K. Paquette-Côté et D. Magis (dir.), *Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation. Volume 2*, Québec : Presses de l'Université du Québec, p. 139-161.

Bland, C.A., D.C. Kreiter et A.J. Gordon (2005). « The psychometric properties of five scoring methods applied to the Script Concordance Test », *Academic Medicine*, *80*(4), p. 395-399.

Boone, W.J., J.R. Staver et M.S. Yale (2014). *Rasch Analysis in the Human Sciences*, Dordrecht : Springer Netherlands.

Bradburn, N.M. et S. Sudman (1979). *Improving Interview Method and Questionnaire Design : Response Effects to Threatenning Questions in Survey Research*, San Francisco : Jossey-Bass

Briggs, D.C. et M. Wilson (2004). « An Introduction to Multidimensional Measurement Using Rasch Models », dans E.V.J. Smith et R.M. Smith (dir.), *Introduction to Rasch Measurement*, Maple Grove : JAM Press, p. 322-341.

Cano, S.J., A.F. Klassen, A. Scott, P.G. Cordeiro et A.L. Pusic (2013). « Reply : The Rasch model : "Litmus Test" de rigueur for Rating Scales ? », *Plastic and Reconstructive Surgery*, *131*(2), p. 286e-288e, doi : 10.1097/PRS.0b013e31828129f4.

Carrière, B., R. Gagnon, B. Charlin, S. Downing et G. Bordage (2009). « Assessing clinical reasoning in pediatric emergency medicine : Validity evidence for a Script Concordance Test », *Annals of Emergency Medicine*, *53*(5), p. 647-652, doi : 10.1016/j.annemergmed.2008.07.024.

Chang, P.T., M.D. Kessler, M. B. McAninch, V.D. Fein, V.D.J. Scherzer, V. E. Seelbach et V.M. Pusic (2014). « Script Concordance Testing : Assessing residents' clinical decision-making skills for infant lumbar punctures », *Academic Medicine*, *89*(1), p. 128-135, doi : 10.1097/ACM.0000000000000059.

Charlin, B., M. Desaulniers, R. Gagnon, D. Blouin et C. Van Der Vleuten (2002). « Comparison of an Aggregate Scoring Method With a Consensus Scoring Method in a Measure of Clinical Reasoning Capacity », *An International Journal*, *14*(3), p. 150-156, doi : 10.1207/S15328015TLM1403_3.

Charlin, B., J. Tardif et H. Boshuizen (2000). « Scripts and medical diagnostic knowledge : Theory and Applications for clinical reasoning instruction and research », *Academic Medicine*, *75*(2), p. 182-190.

Crocker, L.M. et J. Algina (2008). *Introduction to Classical and Modern Test Theory*, Mason : Cengage Learning.

Deschênes, M.-F., B. Charlin, R. Gagnon et J. Goudreau (2011). « Use of a Script Concordance Test to Assess Development of Clinical Reasoning in Nursing Students », *Journal of Nursing Education*, *50*(7), p. 381-387, doi : 10.3928/01484834-20110331-03.

Duncan, P.W., R.K. Bode, S.M. Lai et S. Perera (2003). « Rasch analysis of a new stroke-specific outcome scale : The stroke impact scale », *Archives of Physical Medicine and Rehabilitation*, *84*(7), p. 950-963, doi : 10.1016/S0003-9993(03)00035-2.

Embretson, S.E. et S.P. Reise (2000). *Item Response Theory for Psychologists*, Mahwah : Lawrence Erlbaum Associates.

Fournier, J.P., A. Demeester et B. Charlin (2008). « Script Concordance Tests : Guidelines for construction », *BMC Medical Informatics and Decision Making*, *8*(18), p. 18.

Gagnon, R., B. Charlin, C. Lambert, B. Carriere et C. Van der Vleuten (2009). « Script Concordance Testing : More Cases or More Questions ? », *Advances in Health Sciences Education*, *14*(3), p. 367-375, doi : 10.1007/s10459-008-9120-8.

Gagnon, R., S. Lubarsky, C. Lambert et B. Charlin (2011). « Optimization of answer keys for Script Concordance Testing : Should we exclude deviant panelists, deviant responses, or neither ? », *Advances in Health Sciences Education*, *16*(5), p. 601-608, doi : 10.1007/s10459-011-9279-2.

Hofmans, J., P. Theuns, S. Baekelandt, O. Mairesse, N. Schillewaert et W. Cools (2007). « Bias and changes in perceived intensity of verbal qualifiers effected by scale orientation », *Survey Research Methods*, *1*(2), p. 97-108.

Lambert, C., R. Gagnon, D. Nguyen et B. Charlin (2009). « The script concordance test in radiation oncology : Validation study of a new tool to assess clinical reasoning », *Radiation Oncology*, *4*, p. 4-7, doi : 10.1186/1748-717X-4-7.

Latreille, M.-E. (2012). *Évaluation du raisonnement clinique d'étudiantes et d'infirmières dans le domaine de la pédiatrie, à l'aide d'un test de concordance de script*, M.A. : Université d'Ottawa.

Lemay, J.-F., T. Donnon et B. Charlin (2010). « The reliability and validity of a Paediatric Script Concordance Test with medical students, paediatric residents and experienced paediatricians », *Canadian Medical Education Journal*, *1*(2), p. e89-e95.

Linacre, J.M. (1997). « KR-20/Cronbach Alpha or Rasch Person Reliability : Which tells the "Truth" ? », *Rasch Measurement Transactions*, *11*(3), p. 580-581.

Linacre, J.M. (2002). « What do Infit and Outfit, Mean-square and Standardized mean ? », *Rasch Measurement Transactions*, *16*(2), p. 878.

Linacre, J.M. (2003). « Data variance : Explained, modeled and empirical », *Rasch Measurement Transactions*, *17*(3), p. 942-943.

Linacre, J.M. (2008). « Variance in data explained by Rasch measures », *Rasch Measurement Transactions*, *22*(1), p. 1164.

Linacre, J.M. (2015). *FACETS Computer Program for Many-Facets Rasch Measurement* (Version 3.71.4), Beaverton, OR : Winsteps.com, <http://www.winsteps.com/index.htm>, consulté le 25 avril 2017.

Linacre, J.M. (2016a). *Winsteps® Rasch Measurement Computer Program*, Beaverton, OR : Winsteps.com, <http://www.winsteps.com/index.htm>, consulté le 25 avril 2017.

Linacre, J.M. (2016b). *Winsteps® Rasch Measurement Computer Program User's Guide* (Version 3.92.0), Beaverton, OR : Winsteps.com, <http://www.winsteps.com/index.htm>, consulté le 25 avril 2017.

Linacre, J.M. et B.D. Wright (1994). *Reasonable Mean-Square Fit Values*, *8*(3), p. 370, <http://www.rasch.org/rmt/rmt83b.htm>, consulté le 25 avril 2017.

Lineberry, M., C.D. Kreiter et G. Bordage (2013). « Threats to validity in the use and interpretation of script concordance test scores », *Medical Education*, *47*(12), p. 1175-1183, doi : 10.1111/medu.12283.

Mallinson, T., J. Stelmack et C. Velozo (2004). « A comparison of the separation ratio and coefficient α in the Creation of minimum item sets », *Medical Care*, *42*(1), suppl. 1, p. 17-24.

Masters, G. N. (1982). « A Rasch Model for Partial Credit Scoring », *Psychometrika*, *47*(2), 149.

Norman, G.R. (1985). « Objective measurement of clinical performance », *Medical Education*, *19*(1), p. 43-47, doi : 10.1111/j.1365-2923.1985.tb01137.x.

Penta, M., C. Arnould et C.I. Decruynaere (2005). *Développer et interpréter une échelle de mesure : applications du modèle de Rasch*, Sprimont, Belgique : Mardaga.

Petrucci, A.M., T. Nouh, M. Boutros, R. Gagnon et S.H. Meterissian (2013). « Assessing clinical judgment using the Script Concordance test : The importance of using specialty-specific experts to develop the scoring key », *The American Journal of Surgery*, *205*(2), p. 137-140, doi : 10.1016/j.amjsurg.2012.09.002.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Achievement Tests*, Copenhague : Danish Institute for Educational Research.

Salkind, N.J. (2013). *Tests & Measurement for People Who (Think They) Hate Tests & Measurement*, 2e éd., Los Angeles : SAGE.

Schwarz, N. (1995). « What respondents learn from questionnaires : The survey interview and the logic of conversation », *International Statistical Review/ Revue internationale de statistique*, *63*(2), p. 153-168, doi : 10.2307/1403610.

Smith, R.M., J.M. Linacre et E.V. Smith Jr. (2003). « Guidelines for Manuscripts », *Journal of Applied Measurement*, *4*, p. 198-204.

Tennant, A. et P.G. Conaghan (2007). « The Rasch Measurement Model in rheumatology : What is it and why use it ? When Should it be applied, and what should one look for in a Rasch paper ? », *Arthritis Rheum*, *57*, p. 1358.

Thivierge, R., D. Kazi-Tani, R. Gagnon et B. Charlin (2005). « Le test de concordance comme outil d'évaluation en ligne du raisonnement des professionnels en situation d'incertitude », *Revue internationale des technologies en pédagogie universitaire*, *2*(2), p. 22-33.

Vanbelle, S., V. Massart, D. Giet et A. Albert (2007). « Test de concordance de script : un nouveau mode d'établissement des scores limitant l'effet du hasard », *Pédagogie médicale*, *8*(2), p. 71-81.

Wainer, H., N.J. Dorans, D. Eignor, R. Flaugher, B.F. Green, F.J. Mislevy et D. Thissen (2000). *Computerized Adaptive Testing : A Primer*, 2e éd., Mahwah : Lawrence Erlbaum Associates.

Wilson, A., G. Pike et A. Humbert (2014). « Analyzing Script Concordance Test scoring methods and items by difficulty and type », *An International Journal*, *26*(2), p. 135-145, doi : 10.1080/10401334.2014.884464.

PARTIE 2

L'ÉVALUATION DES COMPÉTENCES

CHAPITRE 4

Les avancées technologiques, les enjeux et les défis de la notation automatisée en éducation dans le domaine de la santé

Maxim Morin, André-Philippe Boulais
et André F. De Champlain

L'adoption de l'approche par compétences dans plusieurs programmes de formation professionnelle dans le domaine de la santé a suscité, au cours des dernières années, une réflexion accrue sur les dispositifs d'évaluation disponibles. Au cœur de ces réflexions, il y a un intérêt grandissant pour le développement et l'utilisation de tâches d'évaluation plus authentiques et plus ouvertes. La correction de telles tâches représente néanmoins un défi pour l'évaluateur : elle requiert d'ordinaire plus de temps et de ressources que la correction des réponses collectées au moyen d'instruments d'évaluation traditionnels. Pour plusieurs raisons (logistiques, financières…), ces ressources sont souvent limitées. La correction automatisée est une solution prometteuse pour réduire le temps consacré à l'évaluation des apprentissages, en soumettant à un algorithme informatique des opérations qui auraient autrement été réalisées par des humains. Les progrès de la correction automatisée au cours des dernières années, et ce, notamment dans le champ de l'éducation obligatoire, sont impressionnants. Ce chapitre a pour objectif d'exposer les récentes avancées technologiques de ce domaine et d'explorer dans quelle mesure elles peuvent contribuer à l'évaluation en éducation dans le domaine

de la santé. Seront ensuite discutés les enjeux potentiels et les défis qui guettent l'introduction de telles technologies dans les pratiques d'évaluation en salle de classe ou à grande échelle.

L'adoption de l'approche par compétences dans plusieurs programmes de formation professionnelle du domaine de la santé a suscité, au cours des dernières années, une réflexion accrue sur les dispositifs d'évaluation disponibles. Au cœur de ces réflexions, on peut déceler un intérêt croissant pour l'évaluation d'une multitude d'habiletés telles que le professionnalisme, la communication, le raisonnement clinique ou la gestion des patients, et ce, à l'aide de tâches d'évaluation plus authentiques, plus complexes et plus ouvertes. La correction de tâches complexes pose néanmoins un défi de taille : elle requiert d'ordinaire plus de temps et de ressources (logistiques, financières…) que la correction des réponses collectées au moyen d'instruments d'évaluation traditionnels (Bennett, 1993) en plus d'ajouter de nouvelles sources potentielles d'erreurs de mesure, soit des incohérences dans les procédures de correction.

Dans ce contexte, la recherche de solutions pour minimiser les coûts associés à la correction de tâches complexes est susceptible de favoriser leur insertion dans les épreuves d'évaluation. La notation automatisée est une stratégie particulièrement prometteuse pour réduire ces coûts. En soumettant à un algorithme informatique des opérations qui auraient autrement été réalisées par des humains, elle offre la possibilité de réduire certains fardeaux, comme le travail de correction des enseignants, ou de recourir à des outils et stratégies pédagogiques qui n'auraient pas vu le jour autrement, comme les systèmes tutoriels intelligents. L'instantanéité de la correction ouvre également la porte à de nouvelles modalités de passation comme le testing adaptatif pour des examens comportant des items à réponse construite (Mislevy *et al*., 2010) ou encore la production immédiate de rétroactions (Attali et Powers, 2010). La capacité de l'ordinateur à collecter, stocker et traiter une masse d'information lui permet ainsi de réaliser des tâches qui seraient trop complexes ou fastidieuses pour des humains.

Les progrès de la notation automatisée au cours des dernières années, et ce, notamment pour l'évaluation de la compétence d'expression écrite, sont impressionnants. Ce chapitre a pour objectif, d'une part, d'exposer les récentes avancées technologiques de ce domaine et d'explorer dans quelle mesure elles peuvent contribuer à l'évaluation en éducation dans le domaine de la santé. D'autre part, il traitera d'enjeux potentiels et de défis qui guettent l'introduction de telles technologies dans les pratiques d'évaluation en salle de classe ou à grande échelle.

Le chapitre est divisé en six sections. La section 1 établit quelques définitions. Les sections 2 à 4 donnent un portrait des avancées technologiques dans le domaine de la notation automatisée. Plus précisément, la section 2 retrace quelques moments clés de son histoire. La section 3 décrit les différentes approches de conception d'un engin de notation automatisée et la section 4 offre une description détaillée de quatre exemples de systèmes de notation automatisée conçus pour l'évaluation des apprentissages dans le domaine de la santé. Ensuite, la section 5 fait état d'enjeux et défis à relever relatifs à la notation automatisée. Enfin, la dernière section conclut ce texte en proposant quelques perspectives d'avenir.

1. LES DÉFINITIONS

Ce chapitre s'intéresse à ce qui est communément appelé l'*automated scoring*. Pour Williamson, Bejar et Mislevy (2006, p. 2), l'*automated scoring* désigne « tout mécanisme informatisé qui évalue les qualités de performances ou de productions[1] ». La recherche d'une expression équivalente en français n'est pas aisée compte tenu du caractère polysémique du mot *scoring*. En effet, ce terme peut faire référence à deux démarches connexes en évaluation des apprentissages : la correction et la notation. Selon Morissette (1993), la correction est une démarche qui s'effectue au regard de la tâche alors que la notation est une démarche qui s'effectue à l'échelle de l'épreuve. Plus précisément, la correction est « le processus d'attribution d'un résultat littéral ou plus souvent numérique, habituellement par l'enseignant, à chacune des tâches réalisées par chaque étudiant » (Morissette, 1993, p. 275) et la notation est le processus de synthèse des résultats d'un examen ou d'une épreuve sous forme d'une valeur numérique ou littérale de manière à exprimer l'appréciation de la performance de l'étudiant par rapport aux objectifs du programme.

Ainsi, le terme *automated scoring* désigne en français tantôt la correction automatisée, tantôt la notation automatisée. Cela dit, pour ne pas confondre la correction automatisée et la correction automatique, qui, elle, désigne la correction orthographique, syntaxique et grammaticale de textes, seule l'expression *notation automatisée* sera utilisée dans ce chapitre.

1. Traduction libre de : « *any computerized mechanisms that evaluates qualities of performances or work products* ».

2. LES REPÈRES HISTORIQUES

Les débuts de la notation automatisée remontent à l'époque où les premiers lecteurs optiques de marque ont fait leur apparition. Le premier engin, la IBM 805 Test Scoring Machine, a été conçu dans les années 1930, soit à peu près au même moment où les questions à choix multiple (QCM) ont gagné en popularité pour l'évaluation de masse (Kuklick, 1987). De tels appareils sont toujours en service de nos jours.

Pendant plusieurs décennies, les QCM ont dominé l'industrie du test, aux dépens d'autres modalités de passation comme les questions à réponse élaborée (QRE). Alors même que les QRE étaient très peu populaires dans les années 1960, Ellis Page (1966) a entrepris un projet de recherche très innovateur, en se penchant sur la notation automatisée de productions écrites. En combinant des techniques issues des domaines de la linguistique informatique et de l'intelligence artificielle, son équipe et lui ont mis au point le premier engin de notation, nommé Project Essay Grade (PEG). Malgré les moyens limités de l'époque, ce système permettait de produire des scores qui s'accordaient très fidèlement avec ceux attribués par des correcteurs humains. L'approche du PEG est relativement simple : construire un ensemble de variables textuelles qu'un ordinateur puisse reconnaitre et décoder, puis déterminer lesquelles prédisent les notes décernées par des correcteurs expérimentés. Les résultats très encourageants de Page n'ont pourtant pas suscité un grand intérêt à cette époque, et ce, notamment en raison de défis d'ordre technologique. Comme Page (2003) le rappelle, les ordinateurs personnels n'étaient pas encore accessibles au moment de ses recherches. Les productions écrites, saisies sur support papier-crayon, devaient être retranscrites sur des cartes perforées à 80 colonnes avant d'être analysées par le PGE. Il a donc fallu attendre plusieurs années, soit jusqu'au début des années 1990, avant que l'invention de Page renaisse et que certaines organisations, comme l'Educational Testing Service (ETS), commencent à s'intéresser plus sérieusement à la notation automatisée de productions écrites.

N'empêche que la période des années 1970 et 1980 a été marquée par l'arrivée d'une première génération d'évaluation assistée par ordinateur, à la suite de la commercialisation de l'ordinateur personnel. Même si cette première génération se limitait presque uniquement à la présentation de questions à choix multiple (Bennett et Bejar, 1998), l'apport de l'ordinateur était non négligeable. Il a permis d'éliminer certaines opérations couramment réalisées par des humains, comme la correction de questions à réponses choisies et de diversifier les modalités de passation des épreuves (intégration d'éléments multimédias, ordonnancement aléatoire des questions, passation de tests adaptatifs, etc.).

La seconde génération d'évaluation assistée par ordinateur s'est, quant à elle, ouverte à l'évaluation de tâches plus complexes. La recherche et le développement des solutions pour automatiser la notation de réponses complexes et multidimensionnelles ont emboité le pas très rapidement. Dès les années 1990, les premiers engins de notation automatisée ont vu le jour dans les opérations courantes des organisations de *testing*. Le National Board of Medical Examiners (NBME) a été l'une des premières organisations à proposer une solution en développant une stratégie pour évaluer les performances de médecins soumis à des scénarios de simulation assistée par ordinateur (Clauser, Margolis *et al.*, Clyman et Ross, 1997). L'utilisation du logiciel E-rater® pour l'évaluation des productions écrites du Graduate Management Admissions Test® a été un autre moment marquant dans l'histoire de la notation automatisée. L'émergence de tels systèmes n'aurait pu être possible sans les progrès en matière d'environnements complexes d'évaluation assistée par ordinateur (Williamson, Bejar et Mislevy, 2006), de techniques de traitement automatique du langage naturel (TALN) et des méthodes d'apprentissage automatique.

Aujourd'hui, l'intérêt, tant du public que du secteur commercial, pour les engins de notation automatisée a explosé. Trois compétitions internationales ont d'ailleurs été consacrées au développement d'engins de notation (Dzikovska *et al.*, 2013 ; Hewlett Fondation, 2012 ; Shermis, 2014). L'évaluation dans le domaine de la formation en santé n'est pas en reste : plusieurs initiatives ont vu le jour au cours des dernières décennies dont les quatre applications décrites à la section 4.

En somme, il y a eu de nombreux progrès depuis la mise en marché des lecteurs optiques de marque pour la correction de questions à réponse choisie. D'une part, l'évolution constante de nouvelles technologies offre de nombreuses solutions pour évaluer les apprentissages à l'aide de questions à réponses construites et de tâches plus réalistes, complexes, flexibles et interactives (Blais, 2009 ; Williamson, Bejar et Mislevy, 2006). D'autre part, plusieurs technologies sont maintenant disponibles pour automatiser la notation des productions résultant de tâches ouvertes et multidimensionnelles. Le reste de ce chapitre portera sur les avancées technologiques les plus récentes, et particulièrement sur la notation des réponses complexes.

3. LA CONCEPTION DES ENGINS

Les engins de notation automatisée sont des technologies hautement spécialisées et conçues pour répondre aux besoins propres d'un programme d'évaluation des apprentissages. Les approches adoptées pour

concevoir un tel engin varient grandement selon les construits[2] évalués, les tâches d'évaluation et leurs modalités de passation. Cette section vise à présenter un aperçu global des différentes stratégies et outils disponibles lors de la conception et de la mise au point de tels engins.

Dans la pratique, un engin de notation automatisé est un module qui s'intègre à un système déjà existant d'évaluation assistée par ordinateur (Bennett et Bejar, 1998 ; Mislevy, Steinberg, Almond et Lukas, 2006). Selon Mislevy *et al.* (2006), ce module peut réaliser soit la correction des tâches prises individuellement, soit la notation de l'épreuve, soit une combinaison des deux approches précédentes. Toutefois, ils notent que la majorité des recherches se sont concentrées sur la correction des tâches jusqu'à présent. La notation d'une épreuve dans un environnement d'évaluation complexe a attiré beaucoup moins l'attention. En effet, et ce, bien que les épreuves soient composées de tâches plus complexes, elles sont encore assemblées de manière très linéaire et notées conformément aux modèles de mesure traditionnels comme ceux de la théorie classique des tests ou de la théorie de la réponse à l'item. La notation dans des environnements d'évaluation plus riches, c'est-à-dire des environnements où l'on évalue plusieurs habiletés des candidats et où les tâches ne sont pas toutes présentées dans la même séquence, est plutôt rare (voir l'exemple dans Williamson, Bejar et Mislevy, 2006). C'est pourquoi le reste de ce chapitre se limitera aux processus reliés à la notation à l'échelle de la tâche.

En évaluation des apprentissages, l'élaboration d'un modèle de notation est essentiellement un effort d'opérationnalisation du construit, au cours duquel ce dernier est décortiqué en éléments observables puis reconstruit sous forme d'une représentation numérique ou qualitative de la performance de l'étudiant. Le processus de déconstruction a pour but d'objectiver les manifestations observables représentatives du construit, tandis que le processus de construction se veut un moyen d'établir les critères de notation permettant de représenter la performance ou la production du candidat sous forme d'une variable quantitative ou qualitative (Bennett et Bejar, 1998 ; Mislevy *et al.*, 2006). Les deux sous-sections suivantes examinent de plus près ces deux processus.

2. En mesure et évaluation, un construit est une abstraction qui caractérise un objet ou un attribut, comme un attribut psychologique (Netemeyer, Bearden et Sharma, 2003). La notion de construit est souvent associée à celle de variable latente.

3.1. L'identification de manifestations observables

L'opérationnalisation du construit commence par l'identification des attributs observables, c'est-à-dire des caractéristiques de la performance ou de la production du candidat qui reflètent ce qui est évalué. Un attribut peut être plusieurs choses : la présence (ou l'absence) d'un certain concept dans une réponse écrite, le nombre d'interventions bénéfiques pour un patient dans une simulation clinique, la quantité de tumeurs retirées dans une simulation d'intervention chirurgicale, etc. Dans un système d'évaluation informatisée, l'identification des attributs requiert un effort considérable pour mettre en relation les données collectées et le construit. Pour les tâches complexes en particulier, les données sont de nature variée (clics de souris, suivi oculaire, saisie de textes, déplacement de manettes électroniques, temps) et les ensembles de données sont naturellement éparses et multidimensionnelles. Dans ce contexte, les données brutes collectées par l'environnement d'évaluation informatisé ne sont pas toujours propices à la correction et à la notation. Il n'est donc pas rare que ces données soient remaniées et réorganisées afin qu'elles soient plus signifiantes pour la machine. Les données textuelles non structurées, c'est-à-dire celles qui sont saisies directement par l'étudiant, en sont un bon exemple. La machine ne peut pas lire un texte comme un humain. Le texte doit au préalable être décomposé en une ou plusieurs représentations (lexicales, syntaxiques, sémantiques, etc.) par l'entremise de diverses techniques de traitement automatique du langage naturel (TALN). Ce traitement des données textuelles est primordial pour garantir la représentation fidèle des réponses par le modèle de notation.

D'ordinaire, le niveau de granularité de la définition des attributs n'est pas le même selon que cette définition sert à la correction par des humains ou à la correction par la machine. Prenons l'exemple d'une question à réponse construite courte dont la réponse hypothétique acceptée est « ictère du nouveau-né ». En toute vraisemblance, un correcteur expérimenté, sans même les lui donner, accepterait des réponses dont le sens est similaire, comme « ictère du nourrisson », « jaunisse néonatale » ou encore « le bébé a une jaunisse ». En revanche, l'ordinateur est incapable de procéder à un tel traitement s'il ne reçoit pas des instructions précises et claires (Carr et Xi, 2010). Pour cet exemple hypothétique, un système très simple chercherait seulement des mots clés, par exemple « ictère » ou « jaunisse ». Un engin hypothétique un peu plus sophistiqué tenterait de déterminer les séquences consécutives de deux ou trois mots qui correspondent à la bonne réponse. Après avoir éliminé les mots vides comme « du » ou « une », le système chercherait des combinaisons valides comme « jaunisse » suivi de « néonatale » ou « nourrisson ». Ces séquences consécutives

de *n* mots sont couramment appelées des *n*-grammes[3] en TALN et forment l'un des attributs les plus simples, mais parmi les plus utilisés dans la correction de réponses textuelles courtes.

Les tâches plus complexes requièrent généralement une plus grande variété d'attributs et à certaines occasions, le construit ne peut être représenté par des variables qui sont directement codifiables en termes informatiques. Ces variables doivent alors être représentées par d'autres variables qui permettent de les approximer (Attali et Burstein, 2006 ; Page et Petersen, 1995). La longueur des phrases, la rareté des mots, la présence de marqueurs de relations ne sont que quelques exemples des dizaines, voire des centaines d'attributs qui composent les modèles de notation de productions écrites. L'objectif est de déterminer un ensemble d'attributs qui offre une représentation suffisante et adéquate du construit.

Bref, l'ordinateur impose des contraintes logiques dans la manière de représenter les données et les attributs de la tâche qui doivent être prises en considération lors de l'identification des manifestations observables. Comme nous le verrons à la section 5.2, ces contraintes ne doivent pas pour autant déformer la définition du construit.

3.2. L'établissement des critères de notation

La seconde étape de l'élaboration d'un modèle de notation consiste à associer les observations, catégorisées en attributs observables, aux niveaux de performance de la tâche. Les approches servant à modéliser les critères de notation et les niveaux de performances peuvent être regroupées en deux familles : les méthodes d'explicitation des critères par des experts et les méthodes statistiques (Burrows, Gurevych et Stein, 2015).

La première famille de méthode met à profit l'expérience et le savoir des experts de contenu en vue de produire des patrons de réponses (voir, par exemple, Rosé *et al.*, 2003) ou des arbres de décisions permettant de distinguer chaque niveau de performance. Dans l'exemple précédent, les experts de contenu fourniraient probablement tous les synonymes de « ictère » et de « nouveau-né » et établiraient les règles de composition des réponses à partir de ces mots.

3. Par exemple, la réponse « ictère du nourrisson » est représentée par trois unigrammes (ictère, du, nourrisson), deux bigrammes (ictère du, du nourrisson) et un trigramme (ictère du nourrisson).

Cette stratégie a globalement le mérite d'établir des règles claires et aisément compréhensibles par les concepteurs du modèle de notation. Toutefois, elle se heurte rapidement à un problème d'extensibilité : le nombre de patrons et de critères croît rapidement selon le nombre d'attributs observables (voir l'exemple dans Clauser, Margolis *et al.*, 1997). Il a vraisemblablement un seuil au-delà duquel il devient quasiment impossible de représenter toutes les réponses possibles ; les règles deviennent très complexes et perdent peu à peu de leur compréhensibilité. De plus, l'élaboration des règles et critères est un travail exigeant qui requiert très souvent une expertise multidisciplinaire, à la fois dans le domaine de connaissances visé et en linguistique informatique (Pulman et Sukkarieh, 2005). Cela dit, l'approche d'explicitation des critères par des experts de contenu demeure attrayante pour certains types de tâches ou certains contextes d'évaluation, et ce, notamment lorsqu'il est impossible de collecter des données préalablement annotées par des humains (voir l'exemple dans Braun, Bejar et Williamson, 2006), comme cela est nécessaire pour la famille des méthodes statistiques.

La deuxième approche diffère de la précédente tant du point de vue méthodologique que conceptuel. Les méthodes statistiques visent à inférer les critères et les règles, soit à partir d'annotations de productions des candidats, soit à partir d'annotations de productions d'experts. Autrement dit, ces approches tentent d'émuler le raisonnement des correcteurs ou des experts à l'aide de méthodes statistiques (Braun *et al.*, 2006). Dans l'exemple hypothétique de l'ictère du nouveau-né, un groupe de correcteurs serait appelé à corriger un échantillon de réponses, dans un premier temps, puis la modélisation statistique servirait à déterminer les variables expliquant les relations significatives entre les notes et les réponses, dans un deuxième temps.

Une panoplie de méthodes statistiques ont été utilisées pour l'élaboration de modèles, parmi les plus populaires on retrouve les méthodes de régression (Attali et Burstein, 2006 ; Clauser *et al.*, 1995), l'analyse sémantique latente (Landauer, Laham et Foltz, 2003) ou les techniques d'apprentissage automatique (Latifi *et al.*, 2016). Toutes ces techniques ont en commun ceci : une fois que les règles d'évidences sont inférées, elles peuvent être réutilisées pour noter de nouveaux ensembles de réponses.

Les approches statistiques sont attrayantes pour plusieurs raisons. En comparaison des méthodes d'explicitation des règles par des experts, elles sont susceptibles d'être moins influencées par les caractéristiques idiosyncrasiques de chaque correcteur humain. Meehl (1954) a depuis longtemps mis en évidence que des procédures de prédiction

statistique, lorsqu'utilisées adéquatement, peuvent faire mieux que le jugement d'expert quand vient le temps de fournir un diagnostic clinique pour un patient. Ces procédures parviennent à pondérer plus fidèlement les attributs qui servent à modéliser les niveaux de performance. En revanche, les résultats des modèles statistiques ne sont pas toujours simples à interpréter, et ce, surtout lorsque le nombre de variables explicatives explose (parfois des centaines ; Attali et Burstein, 2006 ; Elliot, 2003) ou lorsque le modèle tient compte de relations non linéaires entre les variables. La modélisation a ainsi tendance à agir comme une boite noire : elle prédit très bien les performances des candidats, mais le processus pour y arriver ne s'explique pas. Il devient alors difficile de mettre en relation les variables explicatives d'un modèle statistique et le modèle de la tâche ou du construit (Margolis et Clauser, 2006).

Depuis quelques années, des solutions hybrides ont émergé (voir l'exemple à la section 4.2). En combinant les deux méthodes d'établissement de critères, elles visent à mieux faire valoir la participation des experts de contenu dans le processus d'élaboration d'un modèle de notation. L'introduction de méthodes statistiques dans les approches d'explicitation des critères a l'avantage de réduire l'effort des experts de contenu et vice versa, la validation des attributs ou des critères ajoute une certaine crédibilité à la démarche statistique.

4. DES EXEMPLES DE STRATÉGIES DE NOTATION

Au cours des années, quelques systèmes de notation automatisée ont été mis au point pour des tâches complexes dans le domaine de la formation en santé. Quatre exemples illustrant diverses techniques de notation automatisée sont présentés dans les sous-sections suivantes.

4.1. L'évaluation de la prise en charge de patients à l'aide de simulation

La simulation est une stratégie d'apprentissage et d'évaluation fortement préconisée et encouragée dans le milieu de la formation en santé (Lane, Slavin et Ziv, 2001 ; Ziv *et al.*, 2006). Elle offre aux étudiants l'occasion de développer ou de manifester une grande diversité d'habiletés et de compétences pratiques sans compromettre la sécurité des patients (Khan, Pattison et Sherwood, 2011). La simulation peut prendre différentes formes selon les outils (p. ex. la simulation assistée par ordinateur) ou les stratégies privilégiées (p. ex. le recours à un patient simulé).

Dans les années 1990, le NBME (National Board of Medical Examiners) a introduit la simulation assistée par ordinateur dans la troisième partie de l'United States Medical Licensing Examination® (USMLE). Cet examen est un préalable pour l'obtention d'une licence de pratique médicale aux États-Unis sans supervision. Dans cette épreuve, le candidat est soumis à une série de scénarios au cours desquels il doit fournir des soins à des patients virtuels. Chaque scénario commence avec une présentation du cas, l'anamnèse et un relevé des signes vitaux. Ensuite, le candidat a la responsabilité de procéder à certaines interventions médicales (ordonner des tests médicaux, placer le patient en observation, effectuer des examens physiques, déplacer le patient vers une autre unité d'intervention, etc.) tout en contrôlant la progression du temps simulé. La situation du patient évolue ainsi de manière dynamique selon des contraintes préréglées dans le scénario et selon les interventions du médecin. Tout au long de la tâche, le système enregistre automatiquement ces actions dans une liste d'interventions médicales.

Au cours d'un programme de recherche qui s'est échelonné sur plusieurs années, Clauser et son équipe ont étudié attentivement les qualités psychométriques des scénarios de simulation, et ce, au moyen de deux méthodes de notation : par modélisation statistique (Clauser *et al*., 1995) et par explicitation des critères par des experts de contenus (Clauser, Rose *et al*., 1997). Des synthèses de résultats de ce programme de recherche ont aussi fait l'objet de publications scientifiques depuis (p. ex. Margolis et Clauser, 2006). Comme première stratégie de notation, Clauser et son équipe recourent à la régression linéaire multiple pour différencier les actions ordonnées par le médecin selon qu'elles sont bénéfiques ou potentiellement risquées pour la santé du patient. Pour y arriver, deux comités de médecins-experts participent à l'élaboration du modèle de notation. Le premier groupe a pour tâche de catégoriser toutes les actions qu'un médecin peut commander lors d'un scénario de simulation clinique. Plus particulièrement, ils doivent classer les actions bénéfiques selon trois niveaux (appropriée mais d'importance modérée, importante pour des soins optimaux, essentielle pour des soins adéquats) et les actions risquées selon trois niveaux (non indiquée mais minimalement intrusive ou risquée, non indiquée et risquée ou intrusive, non indiquée et extrêmement risquée). Lorsque les experts de contenu ne s'entendent pas sur le bien-fondé d'une action, celle-ci est catégorisée comme neutre.

Un second comité de médecins, indépendant du premier, a le mandat de réviser la liste d'interventions médicales de chaque candidat à l'examen et de la noter selon une échelle d'appréciation

numérique de 1 à 9[4]. Ces notes servent alors de variable dépendante dans un modèle de régression linéaire multiple. Sept variables indépendantes sont définies. Six d'entre elles représentent le dénombrement des actions dans chacune des catégories d'actions risquées ou bénéfiques, alors que la dernière correspond au temps requis avant de réaliser l'ensemble des actions bénéfiques essentielles. Les coefficients de régression définis lors de cette démarche peuvent ensuite servir à produire les scores de nouvelles listes d'interventions.

Dans une deuxième série de recherches, Clauser et son équipe ont exploré une méthode de notation basée sur l'explicitation de critères par des experts. Cette approche se divise en trois étapes. Dans un premier temps, un groupe d'experts de contenu est invité à répondre au cas comme le ferait un candidat à l'examen, puis à discuter du problème médical et de la solution. Dans un deuxième temps, les experts doivent formaliser les critères de correction en indiquant les actions visées et le temps alloué qui correspondent à chacune des notes sur une échelle d'appréciation de 1 à 9. Ces critères sont particuliers à chaque cas. La dernière étape a pour visée le perfectionnement des critères.

Une comparaison des deux méthodes dans une série de recherches subséquentes a révélé que, même si elles prédisent assez précisément les notes attribuées par des experts (Clauser, Rose *et al.*, 1997), les valeurs de la corrélation entre les notes générées automatiquement et les notes attribuées par les experts sont généralement plus élevées pour la méthode de régression (Clauser, Margolis *et al.*, 1997) et que les scores produits par la méthode de régression sont plus généralisables que ceux de la méthode d'explicitation des critères par les experts, et presque aussi généralisables que ceux provenant d'un groupe de quatre correcteurs (Clauser, Swanson et Clyman, 1999). En somme, les résultats de ces études suggèrent que, d'un point de vue psychométrique, la méthode de régression est préférable à la méthode basée sur les règles d'experts, du moins dans le contexte des tâches de simulation clinique assistée par ordinateur.

Même s'ils ont été réalisés au cours des années 1990, les travaux de Clauser et ses collaborateurs attirent toujours l'attention dans le milieu du *testing*. Ils ont été parmi les premiers à se pencher sur la notation automatisée de scénarios de simulation et à démontrer la pertinence d'un algorithme de notation basé sur la régression linéaire. Leur approche offrait une option des plus intéressantes pour les programmes d'évaluation à grande échelle qui sont en quête de solutions de rechange

4. Il n'y a pas de descripteur associé à chaque niveau de l'échelle d'appréciation.

aux approches traditionnelles pour évaluer les habiletés des candidats, tout en respectant les contraintes opérationnelles. D'ailleurs, le NBME utilise toujours cette méthode de notation.

4.2. L'évaluation de la prise de notes cliniques

Très proactif, le NBME a aussi entrepris la recherche et le développement de méthodes de notation automatisée des notes cliniques rédigées lors d'une autre épreuve de l'USMLE, soit l'examen des habiletés cliniques, partie 2. Cette épreuve simule des rencontres avec des patients standardisés pour évaluer les habiletés d'étudiants en médecine à recueillir les informations des patients, à effectuer des examens physiques et à communiquer les résultats aux patients et aux collègues. À la fin de chaque rencontre d'une durée de 15 minutes, le candidat a 10 minutes de plus pour remplir une note clinique dans laquelle il collige par écrit les résultats de l'examen médical, les diagnostics potentiels et les interventions recommandées. Diverses composantes des notes sont évaluées par des correcteurs à l'aide d'échelles d'appréciation globales et analytiques.

À ce jour, deux méthodes ont été développées et mises à l'essai par le NBME afin de corriger automatiquement les notes d'observation. La première méthode, décrite par Swygert *et al.* (2003), emploie l'analyse sémantique latente (ASL) en combinaison avec la régression linéaire multiple, pour prédire les notes décernées à une échelle d'appréciation globale. Dans le cadre de leur recherche, ils ont étudié les notes rédigées par les candidats à quatre cas cliniques. Chaque note clinique a été évaluée par un groupe de trois correcteurs à qui l'on a demandé de donner une appréciation selon une échelle globale et cinq échelles particulières (investigation diagnostique, entretien, examen physique, diagnostic différentiel, communication écrite). Les résultats globaux ont été prédits à l'aide de deux approches : 1) en prédisant directement la note globale à l'aide de l'ASL ; et 2) en prédisant la note globale sous forme d'une régression linéaire multiple des notes prédites par l'ASL de chacune des échelles particulières.

La corrélation entre les scores prédits et les scores décernés par les correcteurs a été calculée pour ces deux approches. Afin de refléter les conditions opérationnelles de l'examen, les estimations de ces corrélations sont comparées à la corrélation entre la note globale décernée par un seul correcteur et la note moyenne du groupe de correcteurs. Selon les chercheurs, les résultats de cette première étude sont assez encourageants. L'algorithme de notation basée sur la régression (0,64 – 0,77) prédit mieux la note globale moyenne que

le modèle ASL seul (0,36 – 0,60) ou même que les notes individuelles des correcteurs (0,59 – 0,73). Le recours aux modèles ASL pour les échelles d'appréciation spécifiques est donc bénéfique pour la modélisation des notes globales. Cela dit, les résultats montrent que les modèles de notation automatisée ne performent pas de manière similaire selon les cas cliniques étudiés ; la contribution relative des échelles particulières varie grandement d'un cas à l'autre.

La seconde méthode conçue par le NBME est décrite dans une demande de brevet d'invention (Mitkov *et al.*, 2014). Elle consiste à prédire les notes attribuées par les correcteurs humains à partir d'une stratégie séquentielle de notation semi-automatisée. Dans un premier temps, l'engin extrait une multitude de caractéristiques linguistiques des notes cliniques, puis il relève les plus discriminantes d'entre elles et les regroupe selon diverses dimensions d'analyses. Ces caractéristiques linguistiques et leur catégorisation sont ensuite présentées à des experts de contenu qui ont pour tâche de déterminer lesquelles devront être prises en considération dans la stratégie finale. Enfin, les critères de notation sont établis à l'aide d'une ou plusieurs méthodes statistiques (régression linéaire ou méthodes d'apprentissage automatique).

Les résultats de la mise à l'essai de cette seconde stratégie de notation indiquent qu'en moyenne les corrélations entre les résultats générés automatiquement et les notes produites par les humains ($r = 0,47$) sont comparables aux corrélations des notes données par les humains ($r = 0,47$), dans la mesure où les experts de contenu révisent le contenu des listes des caractéristiques linguistiques et des catégorisations utilisées pour la notation. En effet, sans l'intervention des humains, les algorithmes de notation automatisée obtiennent de bien moins bons résultats ($r = 0,27$).

La représentation linguistique des notes cliniques est un défi de taille. À la différence des productions écrites, le vocabulaire et le style d'écriture des notes cliniques varient grandement, en raison de l'usage d'abréviations, de termes médicaux spécialisés ou courants, en plus de la présence de fautes d'orthographe. À cela s'ajoute le défi des paraphrases ou d'implication linguistique, c'est-à-dire du remplacement de termes ou d'expressions dont le sens est similaire à ceux de la clé de réponse. Pour y arriver, l'engin doit donc incorporer une multitude de filtres et de techniques de TALN de manière à ce que la stratégie se montre robuste par rapport aux perturbations des données. La robustesse du traitement des réponses est un enjeu fondamental pour toute démarche d'évaluation du contenu des réponses textuelles (Carr, 2008 ; Carr et Xi, 2010 ; Leacock et Chodorow, 2003). Tout

compte fait, les résultats plus ou moins satisfaisants des deux solutions présentées pour l'évaluation des notes cliniques expliquent pourquoi ces méthodes ne sont pas encore utilisées de manière opérationnelle.

4.3. L'évaluation de la prise de décisions cliniques

L'examen d'aptitude du Conseil médical du Canada, partie I (EACMC, partie I) est un examen informatisé bilingue d'une journée qui évalue les connaissances, les aptitudes et les attitudes cliniques des diplômés qui demandent à être admis à des programmes de formation clinique postdoctorale sous supervision au Canada. Cet examen comporte un segment qui est consacré à l'évaluation de la prise de décisions cliniques, lequel est constitué de questions à choix multiples élargis (souvent plus de dix choix) et de questions à courtes réponses écrites. Le Conseil médical du Canada explore depuis quelques années diverses solutions pour faciliter la correction des réponses écrites. À l'heure actuelle, la correction des réponses se fait par un procédé de double correction : par exemple, si les deux premiers correcteurs ne s'entendent pas sur la note décernée, un troisième expert tranche. Dans ce contexte, les ressources nécessaires pour corriger les quelques centaines de milliers de réponses, produites chaque session d'examen, sont colossales.

Dans une première étude exploratoire, Latifi et ses collaborateurs (Gierl *et al.*, 2014 ; Latifi *et al.*, 2016) ont étudié la notation automatisée des questions à réponse construite selon une approche statistique, et ce, pour un échantillon de six questions. Les attributs sont relevés à l'aide d'une représentation relativement simplifiée des réponses, sous forme de *n*-grammes, et les notes dichotomiques sont prédites à l'aide de diverses techniques d'apprentissage automatique supervisé. Comme il est de pratique courante en TALN (traitement automatique du langage naturel), des représentations et des modèles de notation distincts ont été construits pour les réponses des francophones ou anglophones. Dans l'ensemble, les résultats de cette recherche sont très encourageants : les notes attribuées par les engins de notation sont comparables à celles des correcteurs. Comme rapporté dans Latifi *et al.* (2016), les pourcentages d'accord parfait varient entre 88,1 et 98,4 %, et les statistiques de kappa de Cohen[5], entre 0,62 et 0,95. Ces valeurs des coefficients kappa, en particulier, sont considérées comme étant substantielles (0,61 – 0,80) ou presque parfaites (0,81 – 1,00) selon les seuils établis par Landis et Koch (1977), et ce, pour les six items.

5. La statistique kappa de Cohen est une mesure d'accord inter-juges qui tient compte de la proportion d'accord qui est due seulement au hasard.

Dans une étude subséquente, Morin, Boulais et De Champlain. (2016) ont étendu la méthode proposée par Latifi et ses collaborateurs à l'ensemble des questions à réponses courtes administrées lors d'une session d'examen de l'EACMC, partie I. Outre de généraliser l'approche à un ensemble plus grand de questions (N = 70), cette nouvelle étude avait pour objectif d'évaluer l'incidence de cette stratégie de notation sur l'estimation des habiletés des candidats et sur la consistance des résultats d'échec ou réussite à l'examen. À la différence de la recherche précédente qui comportait uniquement des questions notées de manière dichotomique, près de 20 % des questions de cette seconde étude étaient notées sur des échelles comportant plus de deux catégories de réponses.

Quoique toujours très encourageants, les résultats de l'étude de Morin *et al.* indiquent que ce ne sont pas tous les items qui satisfont aux critères de qualité énoncés par Latifi *et al.* Les performances des modèles pour les réponses en français et des modèles pour les questions notées selon des échelles à plus de deux catégories sont, de façon générale, moins élevées que celles des autres conditions. Deux hypothèses peuvent expliquer ces résultats. D'une part, il est probable que les modèles en français soient moins stables que ceux en anglais. Ils sont entraînés sur de petits corpus (moins de 200 observations) et les ensembles de réponses des candidats francophones sont beaucoup plus homogènes que ceux du groupe de candidats anglophones. Une variabilité de bonnes et mauvaises réponses aide à construire un bon modèle prédictif. D'autre part, les méthodes d'apprentissage automatique utilisées lors de l'élaboration des modèles ne sont pas nécessairement adaptées à la prédiction selon une échelle ordinale (0, 1, 2, etc.). Ce sont naturellement des méthodes de classification nominale ; elles ne tiennent donc pas compte de la progression ordinale entre les scores. Cela dit, il a été observé que, malgré les performances modestes de certains modèles de notation à l'échelle de la tâche, celles-ci ont une incidence négligeable sur les résultats finaux et particulièrement sur le résultat de réussite ou d'échec à l'examen. Les questions à réponse courte représentent de fait seulement 10 % du nombre total de questions à l'examen.

L'étude de Morin *et al.* (2016) illustre le défi de généraliser une méthodologie de notation automatisée. L'élaboration de modèles est une entreprise coûteuse, surtout lorsque chaque modèle, pour chaque tâche d'évaluation, doit être élaboré et révisé par des experts de contenu. L'application à l'aveugle d'une même méthodologie peut être une opération tout aussi risquée, surtout lorsque les enjeux de

l'évaluation sont grands. Une analyse détaillée des sources d'erreurs des modèles peut néanmoins offrir des pistes de solutions pour améliorer une version ultérieure de la stratégie de notation.

4.4. L'évaluation des habiletés techniques de chirurgiens

Depuis quelques décennies, des robots sont utilisés pour assister le travail du chirurgien dans une multitude d'interventions dites minimalement invasives. La manipulation à distance de bras robotisés confère au chirurgien un contrôle accru des instruments chirurgicaux. L'arrivée de telles technologies dans la pratique médicale requiert toutefois une certaine appropriation de l'outil et le développement d'habiletés fondamentales de manipulation du robot (Buchs *et al.*, 2013). Il y a donc un intérêt à introduire ces outils dans le curriculum de formation des futurs chirurgiens.

Le simulateur est une solution simple pour développer ces habiletés et non invasive pour le patient. Les tâches de simulation de chirurgie assistée par robot sont particulièrement attrayantes dans un contexte de formation, parce qu'elles enregistrent une foule de données, telles que la durée de la tâche, le nombre de collisions entre les instruments, l'usage de force excessive, l'économie des mouvements, les instruments hors du champ de vision ou la maitrise de l'espace de travail (Brinkman *et al.*, 2013 ; Gomez, Willis et Van Sickle, 2015). À ces indicateurs génériques peuvent s'ajouter des mesures spécifiques à la tâche comme la perte de sang, le nombre de vaisseaux sanguins blessés, la proportion de tissus cancéreux retirés, les dommages aux tissus sains, etc. (Gomez *et al.*, 2015). Ces indicateurs sont difficilement mesurables autrement.

Les études des propriétés psychométriques des modèles d'évaluation des habiletés techniques des chirurgiens sont encore à l'étape exploratoire. La définition du ou des construits mériterait d'ailleurs quelques précisions. Les scores globaux produits automatiquement par les simulateurs sont définis comme une somme pondérée des différents indicateurs génériques ou particuliers (Brinkman *et al.*, 2013). Dans les écrits recensés, ni les coefficients de pondération ni la démarche pour les obtenir ne sont divulgués, ce qui rend difficile l'appréciation de la validité de cette démarche d'évaluation. Quelques recherches montrent d'ailleurs que ce ne sont pas tous ces indicateurs qui discriminent les performances des chirurgiens (Gélinas-Phaneuf *et al.*, 2014 ; Gomez *et al.*, 2015 ; Perrenot *et al.*, 2012). Qui plus est, certains indicateurs particuliers (p. ex. la perte de sang) sont déjà difficiles à évaluer par des humains, et peuvent donc poser un défi

supplémentaire pour la notation automatisée. Ces résultats partagés soulèvent des doutes quant à la méthode de constitution des scores et, éventuellement, quant à l'interprétation de ceux-ci.

À ce jour, les stratégies d'interprétation des scores sont encore très limitées. La plateforme Mscore™ est l'une des rares à proposer une solution, en comparant les résultats des apprenants (sur l'échelle globale et les échelles particulières) aux résultats moyens d'un groupe de plus de 100 chirurgiens experts[6]. Cependant, cette stratégie d'interprétation normative ne permet pas de relever avec précision ce qui caractérise les performances des experts.

Il va sans dire que, devant cette insuffisance de données probantes, il est très difficile d'évaluer les qualités psychométriques des tâches d'évaluation de simulation de chirurgie assistée par robot. Les connaissances concernant l'usage de ces simulateurs vont probablement s'accroître au fur et à mesure que ces instruments hautement spécialisés deviendront plus accessibles dans les milieux médicaux.

4.5. La synthèse

Le choix des quatre exemples présentés ci-dessus n'est pas anodin. Ils offrent une vue d'ensemble des méthodologies de notation automatisée de tâches complexes et ouvertes. Le programme de recherche de Clauser et ses collaborateurs (Clauser, Margolis *et al.*, 1997 ; Clauser, Rose *et al.*, 1997 ; Clauser *et al.*, 1995 ; Clauser *et al.*, 1999) est d'ailleurs l'un des rares à avoir comparé, toutes disciplines confondues, des méthodes basées sur la régression et une approche fondée sur l'explicitation des critères. Les trois premiers exemples font tous usage, à un moment ou un autre, de méthodes statistiques. Ces stratégies s'appliquent bien lorsqu'il y a déjà un corpus de productions annotées par des correcteurs, comme c'est le cas pour les grandes organisations de *testing*, car celles-ci ont accès à d'imposantes banques de données. L'approche d'évaluation à interprétation normative, basée sur les performances d'experts est, quant à elle, une pratique moins fréquente. Elle ne nécessite pas d'annoter les productions préalablement, mais elle souffre néanmoins des mêmes limites que la plupart des méthodes statistiques : elles expliquent mal, de manière conceptuelle, ce qui distingue une mauvaise performance d'une bonne performance. Dans ce cas, il serait intéressant de construire un modèle théorique de performance de chirurgiens experts. Nous reviendrons sur l'idée d'un modèle théorique de l'expertise, plus loin à la section 5.1.

6. <http://www.mimicsimulation.com/products/dv-trainer/mscore-evaluation>.

Dans un autre ordre d'idées, il est d'intérêt de remarquer que seulement deux des quatre stratégies discutées ont fait la transition d'un projet de recherche et développement à un programme opérationnel. Les deux projets de correction de réponses textuelles, qui se heurtent à des difficultés de traitement de la langue, n'ont pas encore franchi cette étape. D'une part, le vocabulaire et le style d'écriture des notes cliniques ou des réponses courtes sont bien différents de ce qui est attendu dans une production écrite de sorte que les techniques de TALN doivent y être adaptées. Qui plus est, le nombre de tâches administré chaque année dans les épreuves de l'USMLE, partie 2 ou de l'EACMC, partie I, est assez grand, ce qui requiert de construire et de valider un très grand nombre de modèles de notation. Cela est encore plus compliqué lorsque les procédés d'identification des attributs et d'assemblage des critères varient d'une tâche à l'autre. Autant les travaux de Mitkov *et al.* (2014) que ceux de Morin *et al.* (2016) ont mis en relief les défis que pose l'opérationnalisation de tels procédés. Dans tel cas, une meilleure compréhension des sources d'erreur est une étape nécessaire en vue de rationaliser ces démarches et, éventuellement, de les entreprendre.

En somme, les quatre exemples ci-dessus font état de différents progrès et de travaux de recherche prometteurs, mais aussi de défis propres à l'évaluation en formation à la santé. Il reste encore beaucoup de chemin à parcourir avant d'exploiter le plein potentiel de ces stratégies de notation. Pour l'instant, ce sont essentiellement les grandes organisations en évaluation des apprentissages en médecine qui ont investi dans ces projets. Trois des quatre exemples ont été financés par le National Board of Medical Examiners (NBME) ou le Conseil médical du Canada, deux organismes ayant pour mandat d'évaluer les connaissances et habiletés de milliers de médecins chaque année. Les investissements de départ étant considérables, il faut comprendre la méfiance d'autres organisations. De nouvelles avancées offriront peut-être un accès plus facile à ces technologies et, dans le futur, nous aurons la chance de voir apparaître des projets de notation automatisée dans d'autres domaines de formation en santé.

5. LES ENJEUX ET DÉFIS

La notation automatisée de tâches complexes est un domaine de recherche actif qui a déjà mené à plusieurs réalisations réussies. Il n'en demeure pas moins que ce domaine en émergence comporte plusieurs enjeux et pose bien des défis. Quatre d'entre eux sont discutés dans les sous-sections suivantes. Il s'agit de divers enjeux et défis qui touchent autant aux aspects conceptuels, méthodologiques et techniques de la notation automatisée.

5.1. L'opérationnalisation du construit

Comme on l'a vu à la section 3, l'automatisation du processus de correction oblige très souvent à redéfinir les attributs observables et les critères de notation. Ces transformations ne doivent pourtant pas dénaturer la définition du construit (Williamson, Bejar et Hone, 1999) ; il y a là un enjeu de validité[7]. Que ce soit pour la notation automatisée ou quelque autre forme de notation, les modalités d'évaluation doivent assurer une représentation pertinente et suffisante du construit. À cet égard, certains pièges sont à éviter lorsqu'on entreprend de concevoir et de mettre au point une stratégie de notation automatisée.

La sous-représentation du construit est l'un de ces pièges. En dérivant les critères de notation d'information statistique comme la force de la relation (ou de la corrélation) entre les manifestations des attributs et les scores à prédire, les approches statistiques sont particulièrement exposées à ce piège. Une corrélation élevée entre les attributs et le trait mesuré ne garantit pas pour autant la validité de l'opération de mesure (Bejar, 2011). La longueur des textes en est un bon exemple. Cet attribut contribue de manière statistiquement significative à la prédiction des scores décernés aux productions écrites, mais il est peu pertinent à l'égard d'un modèle de la compétence à écrire. C'est pourquoi, dans les stratégies de notation de productions écrites, les qualités intrinsèques de la compétence à écrire (p. ex. l'organisation du texte) sont mises en lien avec les variables qui servent à approximer ces qualités (p. ex. la présence et le nombre d'organisateurs textuels). Cela est important lors de la validation du modèle.

Le recours à une théorie forte, que ce soit un modèle cognitif de l'apprenant ou un modèle cognitif du correcteur, est une solution attrayante pour prévenir un glissement entre le construit et son opérationnalisation (Chung et Baker, 2003). Shermis *et al.* (2010) remarquent toutefois que cette approche est encore peu ou pas utilisée de nos jours. Il semble en effet y avoir une préférence pour les modèles corrélationnels. Ces derniers sont plus simples à implanter et prédisent relativement bien les résultats des candidats. À l'heure actuelle, les engins de notation automatisée sont utilisés principalement pour prédire un score final et c'est peut-être pourquoi il y a peu d'intérêt à ajouter un degré de complexité à la modélisation des réponses.

7. Une description exhaustive d'une démarche de validation dans ce contexte dépasse largement le cadre de ce chapitre. Le lecteur est invité à consulter les articles de Clauser, Kane et Swanson (2002), de Williamson, Xi et Breyer (2012) ou de Ramineni et Williamson (2013) pour un exposé complet de stratégies d'évaluation des qualités psychométriques de tâches d'évaluation intégrant la notation automatisée.

5.2. Le glissement lors de la correction humaine

Le qualificatif « automatisée » dans le terme « notation automatisée » porte parfois à confusion. Rien n'est en fait pleinement automatisé. Le jugement humain joue un rôle central dans l'élaboration d'une stratégie de notation automatisée. En particulier, les méthodes statistiques misent beaucoup sur un échantillon d'annotations de réponses par des correcteurs expérimentés pour construire un modèle de notation. Dans un tel cas, la qualité de la correction est une condition sine qua non à l'élaboration de ces modèles (Quinlan, Higgins et Wolff, 2009). Une correction humaine déficiente peut devenir une menace pour la validité d'une démarche de notation assistée par ordinateur.

Un premier piège potentiel lors de l'évaluation à l'aide de correcteurs est de se contenter d'une définition plus ou moins rigoureuse du construit, sachant que l'humain, qui est capable de gérer un certain degré d'incertitude et d'ambiguïté, comblera les lacunes du système d'évaluation (Carr et Xi, 2010). Une définition rigoureuse, pertinente et suffisante du construit est un enjeu fondamental à toute démarche d'évaluation. Ce collectif de textes l'illustre d'ailleurs : c'est un défi récurrent de l'évaluation des compétences complexes dans le domaine de la formation à la santé.

La tension entre les critères de fidélité et de validité est à l'origine d'un second piège qui peut avoir des répercussions néfastes sur la validité du construit. Une trop grande attention portée à la fidélité des scores, qui s'incarne entre autres par l'appareillage imposant de la correction (p. ex. des grilles d'appréciation, double correction, arbitrage en cas de désaccords), peut nuire à la représentativité du construit (Bejar, Williamson et Mislevy, 2006). Cela est d'autant plus vrai lorsque la tâche de correction est colossale : le contexte peut alors encourager l'usage de raccourcis mentaux, voire un certain relâchement de la part du correcteur, ce qui compromet la validité de l'évaluation (Bejar *et al.*, 2006). Le processus de collecte de données auprès des correcteurs et des experts de contenu doit garantir l'obtention de données de qualité, qui reflètent elles aussi le construit évalué.

5.3. L'accès à des outils d'élaboration des critères

Quelle que soit la méthode employée pour élaborer les modèles de notation, les experts de contenus doivent être bien encadrés et soutenus dans leur travail. En raison du caractère très technique de l'encodage des modèles de notation, il est généralement nécessaire de concevoir des outils informatiques conviviaux pour accompagner les experts de contenus dans leur tâche. Ces experts ont rarement les

habiletés informatiques nécessaires à la programmation des modèles de notation. Des logiciels spécialisés peuvent ainsi les guider dans les différentes tâches de correction d'échantillon de réponses, d'identification des attributs, des critères de notation ou de validation des modèles de notation. Quelques exemples d'outils pour les experts de contenu peuvent être trouvés dans les écrits suivants : Leacock et Chodorow, 2003 ; Pulman et Sukkarieh, 2005 ; Sukkarieh, Pulman et Raikes, 2003 ; Sukkarieh et Stoyanchev, 2009. La mise au point de ces outils est une étape nécessaire pour mettre en œuvre une stratégie de notation.

5.4. La robustesse des techniques de TALN

Enfin, comme il a été vu tout au long du chapitre et particulièrement dans les exemples d'évaluation de notes cliniques (section 4.2) ou d'évaluation de la prise de décisions cliniques (section 4.3), le traitement automatique des réponses textuelles est un défi de taille pour plusieurs engins de notation automatisée. Les techniques de TALN ont fait des progrès considérables au cours des dernières décennies, mais elles ne sont pas parfaites pour autant[8]. Parmi les outils disponibles, plusieurs, dont les analyseurs syntaxiques et les correcteurs automatiques de l'orthographe, sont construits à partir de documents annotés, de lexiques ou de dictionnaires non spécialisés qui ne reflètent pas suffisamment le sous-langage propre aux domaines de la santé. Il y a donc un exercice important à faire pour développer ces outils et les partager, de manière à rendre les modèles de notation encore plus robustes aux perturbations des données. Cela dit, la robustesse des modèles de notation est un enjeu qui va bien au-delà de l'évaluation en formation à la santé.

CONCLUSION

Les progrès technologiques des dernières années, notamment dans certaines sciences contributives comme les sciences informatiques ou la linguistique informatique, ont catalysé les développements en matière de notation automatisée si bien qu'il est maintenant possible de corriger des tâches dont la complexité dépasse de loin celle de questions d'examen traditionnel.

Ce chapitre a présenté les avancées technologiques, les enjeux et les défis de la notation automatisée dans un contexte d'évaluation des apprentissages en formation de la santé. Plus particulièrement, la

8. Il convient de noter que plusieurs techniques du TALN sont intrinsèquement de nature statistique et stochastique.

section 3 a dressé un portrait général des méthodologies de notation automatisée en insistant sur les interactions entre le modèle conceptuel du construit évalué et les principales composantes d'un engin de notation automatisée. La technologie ne peut pas en effet se substituer à une conception rigoureuse d'un système de notation automatisée. La conception d'un engin exige des efforts et des ressources substantiels afin de s'assurer que les cotes ou les notes prédites reflètent effectivement les performances du candidat. L'implantation d'une stratégie de notation automatisée ne peut se faire au détriment d'une évaluation de qualité.

À la section 4, les quatre exemples tirés du domaine de l'évaluation des apprentissages en formation de la santé illustrent à la fois les réussites de certains projets, comme l'approche d'évaluation de la prise en charge de patients à partir de simulation assistée par ordinateur, et les difficultés auxquelles se heurtent d'autres projets de recherche dans le domaine de la notation automatisée. Les progrès se font à petits pas et d'autres réussites seront vraisemblablement recensées dans les prochaines années. La section 5, quant à elle, a mis en lumière quatre enjeux et défis à relever : les deux premiers concernent la validité de l'évaluation, alors que les deux derniers touchent aux aspects méthodologiques et techniques de l'élaboration de modèles de notation. Ce qu'il faut principalement en retenir, c'est que l'instauration d'une stratégie de notation automatisée oblige à un effort considérable et à un travail rigoureux qui sollicite des expertises dans plusieurs domaines de connaissance. Un effort bipartite de conceptualisation de l'objet d'évaluation et de programmation est nécessaire pour arriver à concevoir un engin qui produit des résultats satisfaisants. L'expert de contenu joue d'ailleurs un rôle capital dans cette démarche.

Un tour d'horizon sur la notation automatisée ne serait pas complet sans proposer quelques perspectives d'avenir. Dans un premier temps, force est de constater qu'il reste encore beaucoup de chemin à parcourir avant que les techniques et technologies discutées dans ce chapitre soient plus accessibles. Si l'implantation d'une stratégie de notation automatisée est une entreprise coûteuse, au regard des ressources humaines et financières, les gains potentiels sont toutefois bien réels. Sans surprise, les programmes d'évaluation à grande échelle sont les premiers à s'investir dans ce terreau fertile en raison des bénéfices que peuvent procurer ces technologiques à court ou moyen terme. Trois des quatre projets cités à la section 4 sont financés et menés par des organisations qui administrent des épreuves à forts enjeux en vue de la certification des médecins. L'intégration de ces nouvelles technologies dans l'évaluation à faible enjeu, comme dans l'évaluation pour l'apprentissage ou l'évaluation en salle de classe, se fait beaucoup plus lentement. Les intervenants

et organisations de premières lignes que sont les enseignants, les formateurs et les institutions scolaires sont très peu exposés à ces technologies. Prenons l'exemple de la compétence en expression écrite. Pour un enseignant en salle de classe, l'accès à un logiciel de notation automatisée pourrait lui permettre de multiplier les pratiques d'écriture de ses élèves, sans nécessairement augmenter sa tâche de correction.

Dans un deuxième temps, il apparaît tout à fait raisonnable de penser que ces technologies de notation automatisée puissent être employées à des fins formatives ou diagnostiques (Attali et Powers, 2010 ; Nielsen, Ward et Martin, 2009). Tout naturellement, les engins peuvent être conçus pour prédire des observations plus fines qu'une note globale et même produire des rétroactions. Dans ce cas, il faut toutefois s'assurer que les preuves recueillies et les modèles correspondent effectivement aux rétroactions que veut offrir le système. Ces rétroactions doivent être associées à de nouvelles catégories de réponses (Attali *et al.*, 2008) qui doivent être encodées dans les annotations des productions des étudiants (Jordan et Mitchell, 2009). Cela dit, une fois ces règles encodées, elles sont disponibles pour corriger un nombre quasi illimité de réponses.

Ce dernier défi quant à l'usage de la notation automatisée à des fins d'évaluation formative remet en avant-plan le travail colossal qui est exigé des experts de contenu pour développer des modèles de notation. Une question se pose : le développement de stratégies de notation automatisée peut-il moins dépendre de la participation des experts de contenu ? L'émergence d'approches hybrides, combinant les méthodes statistiques et les méthodes d'explicitation des critères, est une réponse partielle à cette question. Des études plus récentes vont plus loin et cherchent des solutions pour réduire encore plus, voire éliminer, l'intervention humaine. Jadidinejad et Mahmoudi (2014) ont étudié l'usage des données de Wikipédia pour construire les clés de réponses à des questions à réponse construite, plutôt que de faire appel à des experts de contenu pour définir les bonnes réponses. Même si les résultats de leur étude indiquent qu'il est encore trop tôt pour voir apparaître de telles méthodes, il s'agit là d'une perspective fort intéressante. De telles approches pourraient être généralisées à d'autres sources de données externes (des textes, des bases de données, etc.).

En fin de compte, la notation automatisée de tâches complexes est encore à ses balbutiements. Plusieurs techniques et technologies ont déjà fait leurs preuves, il suffit de les utiliser de façon créative, mais réfléchie, pour multiplier les occasions d'automatiser la correction et la

notation de tâches d'évaluation. Certes, cette automatisation n'est pas impérative, mais elle constitue un outil supplémentaire dans la boite à outils de tous ceux qui sont appelés à évaluer les apprentissages.

BIBLIOGRAPHIE

Attali, Y. et J. Burstein (2006). *Automated Essay Scoring With E-Rater® V.2.*, <https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1650>, consulté le 25 avril 2017.

Attali, Y. et D. Powers (2010). « Immediate feedback and opportunity to revise answers to open-ended questions », *Educational and Psychological Measurement*, *70*(1), p. 22-35.

Attali, Y., D. Powers, M. Freedman, M. Harrison et S. Obetz (2008). *Automated Scoring Of Short-Answer Open-Ended GRE® Subject Test Items ETS Research Report Series*, Princeton : Educational Testing Service.

Bejar, I.I. (2011). « A validity-based approach to quality control and assurance of automated scoring », *Assessment in Education : Principles, Policy & Practice*, *18*(3), p. 319-341.

Bejar, I.I., D. Williamson et R.J. Mislevy (2006). « Human Scoring », dans D.M. Williamson, R.J. Mislevy et I.I. Bejar (dir.), *Automated Scoring of Complex Tasks in Computer-Based Testing*, New York : Routledge, p. 15-47.

Bennett, R.E. (1993). « Toward intelligent assessment : An integration of constructed-response testing, artificial intelligence, and model-based measurement », dans N. Frederiksen, R.J. Mislevy et I.I. Bejar (dir.), *Test Theory for a New Generation of Tests*, New York : Routledge, p. 99-124.

Bennett, R.E. et I.I. Bejar (1998). « Validity and automad scoring : It's not only the scoring », *Educational Measurement : Issues and Practice*, *17*(4), p. 9-17.

Blais, J.-G. (dir.). (2009). *Évaluation des apprentissages et technologies de l'information et de la communication : enjeux, applications et modèles de mesure*, Québec : Presses de l'Université Laval.

Braun, H., I.I. Bejar et D.M. Williamson (2006). « Rule based methods for automated scoring : Application in a licensing context », dans D.M. Williamson, R.J. Mislevy et I.I. Bejar (dir.), *Automated Scoring of Complex Tasks in Computer Based Testing*, New York : Routledge, p. 83-122.

Brinkman, W.M., J.-M. Luursema, B. Kengen, B.M.A. Schout, J.A. Witjes et R.L. Bekkers (2013). « Da Vinci skills simulator for assessing learning curve and criterion-based training of robotic basic skills », *Urology*, *81*(3), p. 562-566, <http://www.goldjournal.net/article/S0090-4295(12)01280-0/fulltext>, consulté le 25 avril 2017.

Buchs, N.C., F. Pugin, F. Volonté et P. Morel (2013). « Learning tools and simulation in robotic surgery : State of the art », *World Journal of Surgery*, *37*(12), p. 2812-2819, doi : 10.1007/s00268-013-2065-y.

Burrows, S., I. Gurevych et B. Stein (2015). « The eras and trends of automatic short answer grading », *International Journal of Artificial Intelligence in Education*, *25*(1), p. 60-117.

Carr, N.T. (2008). *Decisions about Automated Scoring : What they Mean for our Constructs, Communication présentée à Towards Adaptive CALL : Natural Language Processing For Diagnostic Language Assessment*, Ames : Iowa State University.

Carr, N.T. et X. Xi (2010). « Automated scoring of short-answer reading items : Implications for constructs », *Language Assessment Quarterly*, *7*(3), p. 205-218.

Chung, G.K. et E.L. Baker (2003). « Issues in the reliability and validity of automated scoring of constructed responses », dans M.D. Shermis et J.C. Burstein (dir.), *Automated Essay Scoring : A Cross-Disciplinary Perspective*, New York : Routledge, p. 21-36.

Clauser, B.E., M.T. Kane et D.B. Swanson (2002). « Validity issues for performance-based tests scored with computer-automated scoring systems », *Applied Measurement in Education*, *15*(4), p. 413-432.

Clauser, B.E., M.J. Margolis, S.G. Clyman et L.P. Ross (1997). « Development of automated scoring algorithms for complex performance assessments : A comparison of two approaches », *Journal of Educational Measurement*, *34*(2), p. 141-161.

Clauser, B.E., L.P. Rose, S.G. Clyman, K.M. Rose, M.J. Margolis, R.J. Nungester et G.L. Malakoff (1997). « Development of a scoring algorithm to replace expert rating for scoring a complex performance-based assessment », *Applied Measurement in Education*, *10*(4), p. 345-358.

Clauser, B.E., R.G. Subhiyah, R.J. Nungester, D.R. Ripkey, S.G. Clyman et D. McKinley (1995). « Scoring a performance-based assessment by modeling the judgments of experts », *Journal of Educational Measurement*, *32*(4), p. 397-415.

Clauser, B.E., D.B. Swanson et S.G. Clyman (1999). « A comparison of the generalizability of scores produced by expert raters and automated scoring systems », *Applied Measurement in Education*, *12*(3), p. 281-299.

Dzikovska, M.O., R.D. Nielsen, C. Brew, C. Leacock, D. Giampiccolo, L. Bentivogli et H.T. Dang (2013). *SemEval-2013 Task 7 : The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge*, Communication présentée à la Second Joint Conference on Lexical and Computational Semantics, 14-15 juin, Atlanta, Géorgie.

Elliot, S. (2003). « IntellimetricTM : From here to validity », dans M.D. Shermis et J.C. Burstein (dir.), *Automated Essay Scoring : A Cross-Disciplinary Perspective*, New York : Routledge, p. 67-81.

Gélinas-Phaneuf, N., N. Choudhury, A.R. Al-Habib, A. Cabral, E. Nadeau, V. Mora, V. Pazos, R. DiRaddo et R.F. Del Maestro (2014). « Assessing performance in brain tumor resection using a novel virtual reality simulator », *International Journal of Computer Assisted Radiology and Surgery*, *9*(1), p. 1-9.

Gierl, M.J., S. Latifi, H. Lai, A.P. Boulais et A. Champlain (2014). « Automated essay scoring and the future of educational assessment in medical education », *Medical Education*, *48*(10), p. 950-962.

Gomez, P.P., R.E. Willis et K.R. Van Sickle (2015). « Development of a virtual reality robotic surgical curriculum using the da Vinci Si surgical system », *Surgical Endoscopy*, *29*(8), p. 2171-2179, doi : 10.1007/s00464-014-3914-y.

Hewlett Fondation (2012). *Automated Student Assessment Prize : Phase Two – Short Answer Scoring*, <https://www.kaggle.com/c/asap-sas>, consulté le 25 avril 2017.

Jadidinejad, A.H. et F. Mahmoudi (2014). « Unsupervised short answer grading using spreading activation over an associative network of concepts », *Canadian Journal of Information and Library Science*, *38*(4), p. 287-303.

Jordan, S. et T. Mitchell (2009). « e-Assessment for learning ? The potential of short-answer free-text questions with tailored feedback », *British Journal of Educational Technology*, *40*(2), p. 371-385.

Khan, K., T. Pattison et M. Sherwood (2011). « Simulation in medical education », *Medical Teacher*, *33*(1), p. 1-3, doi : 10.3109/0142159X.2010.519412.

Kuklick, H. (1987). « The testing movement and its founders : Psychological testing and american society, 1890-1930 », *Science*, *237*(4820), p. 1358-1359.

Landauer, T.K., D. Laham et P. Foltz (2003). « Automatic essay assessment », *Assessment in Education : Principles, Policy & Practice*, *10*(3), p. 295-308.

Landis, J. et G.G. Koch (1977). « The measurement of observer agreement for categorical data », *Biometrics*, *33*(1), p. 159-174.

Lane, J.L., S. Slavin et A. Ziv (2001). « Simulation in medical education : A review », *Simulation & Gaming*, *32*(3), p. 297-314, doi : 10.1177/104687810103200302.

Latifi, S., M.J. Gierl, A.-P. Boulais et A.F. de Champlain (2016). « Using automated scoring to evaluate written responses in English and French on a high-stakes clinical competency examination », *Evaluation & The Health Professions*, *39*(1), p. 100-113.

Leacock, C. et M. Chodorow (2003). « C-rater : Automated scoring of short-answer questions », *Computers and the Humanities*, *37*(4), p. 389-405.

Margolis, M. et B.E. Clauser (2006). « A regression-based procedure for automated scoring of a complex medical performance assessment », dans D.M. Williamson, R.J. Mislevy et I.I. Bejar (dir.), *Automated Scoring Of Complex Tasks In Computer-Based Testing*, New York : Routledge, p. 123-166.

Meehl, P.E. (1954). *Clinical Versus Statistical Prediction : A Theoretical Analysis and a Review of the Evidence*, Minneapolis : University of Minnesota.

Mislevy, R.J., I.I. Bejar, R.E. Bennett, G.D. Haertel et F.I. Winters (2010). « Technology supports for assessment design », dans P. Peterson, E. Baker et B. McGraw (dir.), *International Encyclopedia of Education*, vol. 3, New York : Elsevier Science, p. 56-65.

Mislevy, R.J., L.S. Steinberg, R.G. Almond et J.F. Lukas (2006). « Concepts, terminology, and basic models of evidence-centered design », dans D.M. Williamson, R.J. Mislevy et I.I. Bejar (dir.), *Automated Scoring Of Complex Tasks In Computer-Based Testing*, New York : Routledge, p. 15-47.

Mitkov, R., L.A. Ha, R.J. Evans, G.C. Marsic, S. Baldwin, B. Clauser et R.J. Nungester (2014). *US, Brevet No. 20140272832*, United States Patent and Trademark Office.

Morin, M., A.-P. Boulais et A.F. de Champlain (2016). *Automated Marking of Written Response Items in a National Medical Licensing Examination*, Communication présentée au National Council on Measurement in Education, 7-11 avril, Washington, DC.

Morrissette, D. (1993). *Les examens de rendement scolaire*, 3e éd., Québec : Presses de l'Université Laval.

Netemeyer, R.G., W.O. Bearden et S. Sharma (2003). *Scaling Procedures : Issues And Applications*, Thousand Oaks : Sage Publications.

Nielsen, R.D., W. Ward et J.H. Martin (2009). « Recognizing entailment in intelligent tutoring systems », *Natural Language Engineering*, *15*(4), p. 479-501.

Page, E.B. (1966). « The imminence of… grading essays by computer », *The Phi Delta Kappan*, *47*(5), p. 238-243.

Page, E.B. (2003). « Project Essay Grade : PEG », dans M.D. Shermis et J.C. Burstein (dir.), *Automated Essay Scoring : A Cross-Disciplinary Perspective*, New York : Routledge, p. 39-50.

Page, E.B. et N.S. Petersen (1995). « The computer moves into essay grading : Updating the ancient test », *Phi Delta Kappan*, *76*(7), p. 561.

Perrenot, C., M. Perez, N. Tran, J.-P. Jehl, J. Felblinger, L. Bresler et J. Hubert (2012). « The virtual reality simulator dV-Trainer® is a valid assessment tool for robotic surgical skills », *Surgical Endoscopy*, *26*(9), p. 2587-2593, doi : 10.1007/s00464-012-2237-0.

Pulman, S.G. et J.Z. Sukkarieh (2005). *Automatic Short Answer Marking*, Communication présentée à Proceedings of the Second Workshop on Building Educational Applications Using NLP, juin, Ann Arbor, Michigan.

Quinlan, T., D. Higgins et S. Wolff (2009). *Evaluating the Construct Coverage of the E-Rater® Scoring Engine, ETS Research Report Series*, Princeton : Educational Testing Service.

Ramineni, C. et D.M. Williamson (2013). « Automated essay scoring : Psychometric guidelines and practices », *Assessing Writing*, *18*(1), p. 25-39.

Rosé, C.P., A. Roque, D. Bhembe et K. VanLehn (2003). *A Hybrid Approach to Content Analysis for Automatic Essay Grading*, Communication présentée à la Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 27 mai-1er juin, Edmonton, Canada.

Shermis, M.D. (2014). « State-of-the-art automated essay scoring : Competition, results, and future directions from a United States demonstration », *Assessing Writing*, *20*, p. 53-76.

Shermis, M.D., J. Burstein, D. Higgins et K. Zechner (2010). « Automated essay scoring : Writing assessment and instruction », *International Encyclopedia of Education*, *4*, p. 20-26.

Sukkarieh, J.Z., S.G. Pulman et N. Raikes (2003). *Auto-Marking : Using Computational Linguistics to Score Short, Free Text Responses*, Communication présentée à la Annual Conference of the International Association for Educational Assessment (IAEA), Manchester, UK.

Sukkarieh, J.Z. et S. Stoyanchev (2009). *Automating Model Building in C-Rater*, Communication présentée aux Proceedings of the 2009 Workshop on Applied Textual Inference, 6 août, Suntec, Singapour.

Swygert, K., M. Margolis, A. King, T. Siftar, S. Clyman, R. Hawkins et B. Clauser (2003). « Evaluation of an automated procedure for scoring patient notes as part of a clinical skills examination », *Academic Medicine*, *78*(10), p. S75-S77.

Williamson, D.M., R.G. Almond, R.J. Mislevy et R. Levy (2006). « An application of Bayesian networks in automated scoring of computerized simulation tasks », dans D.M. Williamson, R.J. Mislevy et I.I. Bejar (dir.), *Automated Scoring of Complex Tasks in Computer-Based Testing*, New York : Routledge, p. 201-258.

Williamson, D.M., I.I. Bejar et A.S. Hone (1999). « "Mental model" comparison of automated and human scoring », *Journal of Educational Measurement*, *36*(2), p. 158-184.

Williamson, D.M., I.I. Bejar et R.J. Mislevy (2006). « Automated scoring of complex tasks in computer-based testing : An introduction », dans D.M. Williamson, R.J. Mislevy et I.I. Bejar (dir.), *Automated Scoring of Complex Tasks in Computer-Based Testing*, New York : Routledge, p. 1-13.

Williamson, D.M., X. Xi et F.J. Breyer (2012). « A framework for evaluation and use of automated scoring », *Educational Measurement : Issues and Practice*, *31*(1), p. 2-13.

Ziv, A., P.R. Wolpe, S.D. Small et S. Glick (2006). « Simulation-based medical education : An ethical imperative », *Simulation in Healthcare*, *1*(4), p. 252-256, doi : 10.1097/01.sih.0000242724.08501.63.

CHAPITRE 5

Une approche pragmatique de validation en éducation médicale

L'application du modèle de Kane à un outil d'évaluation du raisonnement clinique

Thomas Pennaforte et Nathalie Loye

L'évaluation des compétences est un élément central en éducation médicale. Elle permet d'émettre des jugements et de prendre des décisions en fonction des performances de l'apprenant. En cela, l'analyse des processus décisionnels liés à la façon dont les compétences sont jugées est indispensable. Il semble cependant exister un décalage entre le développement des outils d'évaluation, largement abordé dans la littérature, et les efforts visant à valider les scores produits, peu rapportés. Le modèle de Kane (2006, 2013) est né d'un désir de simplification et d'opérationnalisation du concept de validité. Son applicabilité à l'éducation médicale a été récemment décrite (Clauser et al.*, 2008 ; Hawkins* et al.*, 2010 ; Schuwirth et Van der Vleuten, 2012 ; Cook* et al.*, 2015 ; Hatala* et al.*, 2015). L'accent ne porte plus sur l'outil lui-même, mais sur les interprétations que l'on fait des scores qu'il produit. Pour cela, Kane considère le processus de validation comme une accumulation de preuves structurées selon quatre niveaux d'inférence : la notation, la généralisation, l'extrapolation et l'interprétation. Ce texte est le fruit d'une réflexion visant à décrire les principes de la démarche de validation selon le modèle de Kane et à les illustrer dans le contexte de la pédagogie médicale. Pour cela, nous prenons comme exemple le développement d'un outil*

évaluant le raisonnement clinique et qui consiste à intégrer des questions évaluatives dans une séance de simulation médicale. Le processus de validation est structuré par les différentes inférences de Kane, en lien avec la collecte d'une variété d'éléments de preuve destinés à soutenir la qualité des interprétations.

L'évaluation objective est essentielle en éducation médicale. Cela est devenu encore plus explicite depuis l'adoption par les institutions de formation médicale de l'approche par compétences, formalisée depuis plus de 30 ans par le cadre CanMEDS et qui regroupe sept compétences exigibles des médecins en exercice (figure 5.1) (Frank, 2005).

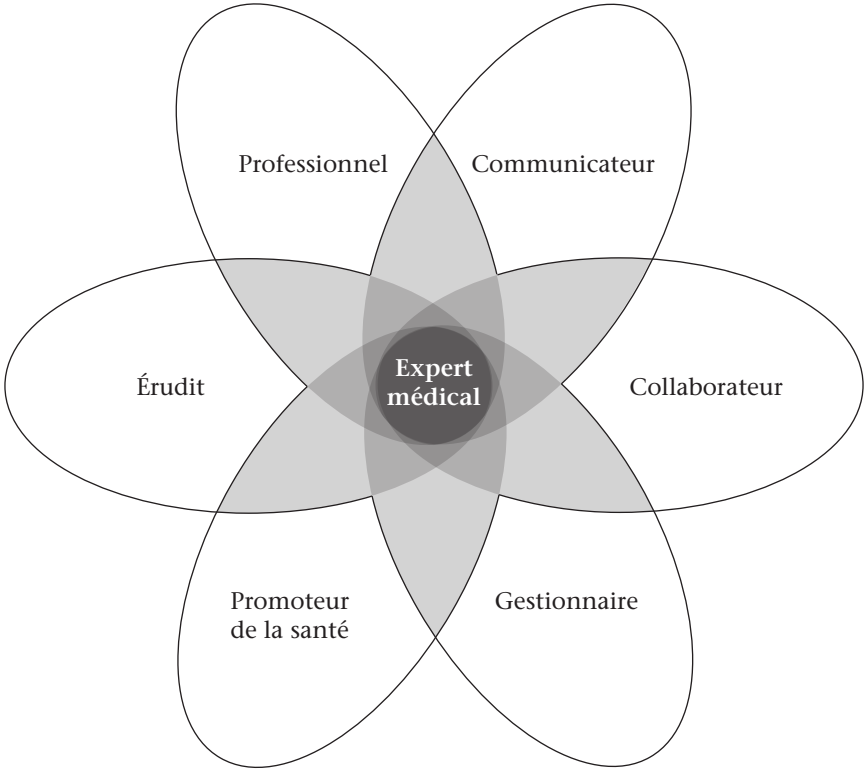


Figure 5.1
Cadre de compétences CanMEDS

Source : © Le Collège royal des médecins et chirurgiens du Canada, 2015, <http://rcpsc.medical.org/canmeds>. Reproduit avec autorisation.

La pertinence de l'évaluation des compétences nécessite que deux types d'exigences soient satisfaits.

Premièrement, il convient que l'objet soit précisément défini avant de développer un instrument pour son évaluation, puis de l'administrer. En d'autres termes, l'évaluation doit répondre aux caractéristiques de la compétence qu'elle tente de mesurer. En l'absence de modèle cognitif précis sur lequel se base le cadre CanMEDS, nous retenons la définition de la compétence proposée par Tardif (2006, p. 22) : « un savoir-agir complexe prenant appui sur la mobilisation et la combinaison d'une variété de ressources internes et externes à l'intérieur d'une famille de situations ». Pour l'auteur, la compétence se distingue par son caractère « intégrateur » (elle fait appel à une multitude de ressources de nature variée), « combinatoire » (elle s'appuie sur une combinaison différenciée de ressources, ce qui permet de résoudre différents problèmes de la même famille de situations concernée), « développemental » (elle se développe tout au long de la vie), « contextuel » (elle est mise en œuvre à partir de contextes particuliers qui orientent l'action) et « évolutif » (elle est conçue afin d'intégrer de nouvelles ressources et de nouvelles situations sans être dénaturée). Une telle définition fournit explicitement, ou implicitement, des indications quant aux conditions dont on doit tenir compte au moment de la planification des dispositifs d'évaluation qui entendent accorder une place centrale au concept de compétence (Nguyen et Blais, 2007).

Deuxièmement, l'évaluation doit désormais s'entendre comme une démarche globale dont l'intérêt porte non plus sur les propriétés du test en lui-même, mais plutôt sur la qualité des décisions prises à la suite du test. Elle est d'ailleurs en général présentée comme un processus, ou démarche évaluative, qui va de la planification des situations d'évaluation jusqu'à la communication des résultats en passant par la collecte des données et leur interprétation, le jugement et la décision (Laurier, Tousignant et Morissette, 2005).

En éducation médicale, Schuwirth et Van der Vleuten (2012) ont rapporté le décalage qui existait entre l'abondante littérature traitant du développement d'outils d'évaluation et celle – embryonnaire – traitant des efforts visant à valider les interprétations des scores obtenus. Dans ce contexte, l'apport des sciences de l'éducation semble essentiel, notamment par l'émergence de plusieurs modèles de validation « unifiés » depuis la fin du XX^{e} siècle. Les modèles récents mettent ainsi un accent sur la qualité de la prise de décision après avoir pris connaissance des résultats de l'évaluation. Parmi eux, le modèle de Kane (2006, 2013) souhaite rendre « abordable » le processus de validation au moyen d'une démarche « pas à pas » de collecte de preuves soutenant la prise de décision. Une place grandissante a récemment été accordée à ce modèle dans la littérature médicale, l'angle abordé étant

son applicabilité « théorique » en éducation médicale (Hawkins *et al.*, 2010 ; Clauser *et al.*, 2012 ; Schuwirth et Van der Vleuten, 2012 ; Cook *et al.*, 2015 ; Hatala *et al.*, 2015).

Dans la lignée de ces écrits, nous retenons le raisonnement clinique comme compétence à évaluer pour y articuler notre réflexion sur les principes de la démarche de validation de Kane. En effet, force est de constater que les principaux outils spécialement développés pour évaluer le raisonnement clinique des médecins – les problèmes à éléments clés (PEC) (Page et Bordage, 1995), les tests de concordance de script (TCS) (Charlin *et al.*, 2000) et les problèmes de raisonnement clinique (PRC) (Groves, Scott et Alexander, 2002) – ne répondent pas à la définition de la compétence qu'ils évaluent. De plus, ils ne considèrent pas l'ensemble des processus décrits dans la théorie mixte du raisonnement clinique, qui rallie pourtant la plupart des chercheurs en éducation médicale (une description détaillée de cette théorie est proposée dans une section ultérieure). Ensuite, l'évaluation qu'ils proposent est décontextualisée ; elle est habituellement réalisée en salle de classe et ne reflète pas l'environnement dans lequel le raisonnement clinique est quotidiennement activé. Enfin, l'évaluation est souvent ponctuelle et ne s'inscrit pas dans le parcours de développement d'une compétence.

Par ailleurs, la validation de tels outils se limite la plupart du temps à la mesure de la fiabilité, à la mise en évidence de la validité de contenu et à la capacité discriminante de l'instrument en fonction du niveau de formation. Elle ne semble donc pas prendre en compte les avancées réalisées en mesure et évaluation, et notamment l'utilisation des modèles « unifiés ». Cette limite concerne de nombreux domaines, mais est particulièrement présente en éducation médicale. Cela peut s'expliquer par la complexité du concept de validité, concept dont la définition a souvent été débattue, ne fait pas consensus et dont l'opérationnalité n'est que très récente.

C'est sur cette dernière constatation que se fonde ce texte, puisqu'il a pour objectif de familiariser le chercheur en éducation médicale à la démarche de validation décrite par Kane, en proposant des illustrations concrètes afin de lui donner les outils nécessaires à son appropriation dans le cadre de sa propre recherche. Pour cela, nous l'aborderons en trois parties.

Dans un premier temps, nous présenterons les grandes lignes du modèle de Kane en nous appuyant sur un exemple concret. Dans un deuxième temps, nous décrirons les grands principes du modèle de Kane. Dans un troisième et dernier temps, nous détaillerons pour chaque niveau d'inférence de Kane les méthodes disponibles pour la

collecte de preuves et illustrerons concrètement leur application par la démarche que nous prévoyons suivre pour valider l'outil d'évaluation du raisonnement clinique que nous proposons.

1. LE PRÉTEXTE À L'UTILISATION DU MODÈLE DE KANE : LA VALIDATION D'UN OUTIL D'ÉVALUATION DU RAISONNEMENT CLINIQUE

Nous souhaitons illustrer l'opérationnalisation du modèle de Kane par un exemple concret. Pour ce faire, nous présenterons un exemple d'une démarche de collecte de preuves qu'un chercheur pourrait adopter en vue de soutenir la validation d'un outil d'évaluation.

1.1. L'assise théorique de notre exemple

La théorie du processus double fait de plus en plus consensus afin d'expliquer la nature des processus qui sont mis en œuvre lorsque le médecin est confronté à une situation de résolution de problème (Faucher *et al.*, 2016). Selon cette théorie issue de la psychologie cognitive (Evans, 2008 ; Wason et Evans, 1975) et récemment appliquée au raisonnement clinique (Eva, 2005 ; Pelaccia *et al.*, 2011), raisonner repose sur l'interaction de deux types de processsus cognitifs. L'un est automatique (ou « type 1 ») et l'autre, analytique (ou « type 2 »). Très schématiquement, le type 1, souvent qualifié de « rapide » et « intuitif », renvoie au dernier stade du développement de l'expertise, où le médecin compare inconsciemment une nouvelle situation à celles qu'il a mémorisées depuis le début de sa pratique. Le type 2, plus « lent » et « réfléchi », correspond à la démarche dite « hypothéticodéductive », où le médecin confronte ses hypothèses diagnostiques initiales aux données recueillies lors de l'interrogatoire, de l'examen clinique ou d'explorations complémentaires. Certains auteurs (Goel et Dolan, 2003 ; Durning *et al.*, 2015) ont relevé une activation de zones cérébrales distinctes en fonction du type de processus activé. La nature des interactions entre les processus de type 1 et de type 2 est débattue et a donné lieu à différents modèles explicatifs, dont une synthèse a été récemment proposée par Custers (2013).

1.2. Le problème à résoudre

L'évaluation du raisonnement clinique doit donc tenir compte de la complexité des processus cognitifs qui interviennent lors de son activation ainsi que des facteurs contextuels qui en influencent l'interaction

(Pelaccia *et al.*, 2011). Ce lourd cahier des charges explique probablement pourquoi il n'existe pas d'outil d'évaluation de référence actuellement disponible (Ilgen *et al.*, 2012).

Parmi les instruments spécialement développés en vue d'une évaluation formelle, notre choix se porte sur deux d'entre eux, car ils présentent un intérêt particulier. Les TCS évaluent l'interprétation d'une nouvelle donnée dans un contexte d'incertitude (Charlin, Tardif et Boshuizen, 2000). Ils sont largement diffusés au travers des curriculums médicaux, tant pour l'évaluation pour les non-diplômés que pour l'évaluation en spécialité (Lubarsky *et al.*, 2011 ; Charlin *et al.*, 2010 ; Piovezan *et al.*, 2012 ; Nouh *et al.*, 2012 ; Duggan et Charlin, 2012 ; Goos *et al.*, 2016). Les PRC sont moins diffusés, mais présentent une approche pertinente de l'évaluation du raisonnement clinique : ils évaluent les habiletés de génération d'hypothèses, d'identification et d'interprétation de données initialement recueillies (Groves *et al.*, 2002).

Comme beaucoup d'autres non mentionnés ici, ces outils « écrits » offrent divers avantages, dont leur objectivité, une meilleure standardisation et un échantillonnage possible au travers d'une large variété de pathologies cliniques. Ils comportent cependant trois limites majeures.

Premièrement, souvent utilisés à grande échelle, par exemple pour toute une promotion d'étudiants en médecine, ils ne permettent pas que l'évaluation se déroule dans les services hospitaliers et nécessitent plutôt de réquisitionner une salle pour l'examen. Le contexte de cette évaluation ne correspond donc pas à l'environnement dans lequel le futur médecin aura l'occasion d'activer son raisonnement clinique.

Deuxièmement, ils ne permettent d'évaluer que le type 2 du raisonnement clinique, en raison de leur conception (les PRC et les TCS ont été développés pour évaluer différentes étapes de la démarche hypothéticodéductive) et de leur modalité d'application (les conditions des examens permettent à l'étudiant de répondre à une série de questions dans un temps déterminé, ce qui lui permet de passer la totalité de l'examen de manière *raisonnablement* réflexive et non intuitive). À ce propos, l'augmentation des performances en raisonnement clinique a été liée à l'absence de perturbateurs environnementaux (Ilgen *et al.*, 2012). On peut imaginer que cela passe par l'activation préférentielle du type 2, notamment en raison du caractère aseptisé de l'environnement de la salle d'examen.

Enfin, et ce point est particulièrement important dans le cadre de ce texte, les arguments utilisés pour soutenir leur validité sont faibles, et reposent uniquement sur les caractéristiques liées au test (la fidélité, la validité de construit et la validité de contenu), que ce soit pour les

PRC (Groves *et al.*, 2002 ; Groves *et al.*, 2013) ou pour les TCS, dont la littérature est plus fournie (Sibert *et al.*, 2002 ; Fournier, Demeester et Charlin, 2008 ; Gagnon *et al.*, 2009 ; Dory *et al.*, 2012. ; Lubarsky *et al.*, 2011). Ces études, auxquelles le lecteur peut se référer, ne s'intéressent qu'aux propriétés psychométriques du test et ne prennent pas en compte l'ensemble de la démarche de validation.

1.3. L'idée de développement

Il convient donc d'utiliser les outils d'évaluation décrits précédemment (TCS et PRC) et dont les qualités existent, mais au sein d'un contexte d'évaluation plus proche de la réalité et dont le contrôle pourrait permettre d'activer préférentiellement le type 1 ou le type 2 du raisonnement clinique.

La simulation médicale semble en mesure de répondre aux exigences d'un tel contexte d'évaluation. Il s'agit d'une approche pédagogique qui utilise un matériel particulier (mannequin ou patient simulé) et le place dans un environnement contrôlé afin de reproduire une grande variété de situations réalistes (Granry et Moll, 2012). Une séance s'articule classiquement autour des trois phases suivantes : le *briefing*, la mise en situation et le *debriefing*. La mise en situation a pour objectif de reproduire le mieux possible le contexte de soins dans lequel le participant évoluera plus tard, ce qui n'est pas possible avec d'autres options telles que la présentation d'une vignette clinique photographique ou vidéographique. Le *debriefing* est, quant à lui, considéré comme « l'élément pédagogique clé », car il facilite la rétention des données acquises par les interactions avec l'éducateur ou avec l'équipe soignante (Raemer *et al.*, 2011).

L'évaluation du raisonnement clinique par la simulation médicale a été rapportée soit par l'intermédiaire d'une grille de *debriefing* adaptée, soit en couplant la simulation à une évaluation réalisée après la séance (Kuiper *et al.*, 2008 ; Dreifuerst, 2012 ; Amin et Friedmann, 2013 ; Lusk et Fater, 2013 ; Fida et Kassab, 2015). Cependant, évaluer le raisonnement clinique après la mise en situation place le participant dans des conditions de confort différentes de celles de l'expérience vécue, et n'explore qu'une réflexion rétrospective sur l'action. De plus, on sait qu'après une expérience stressante, le participant peut soit oublier, soit modifier son processus cognitif en fonction de l'évolution de la situation vécue (Cheng *et al.*, 2014).

En revanche, l'intégration de l'évaluation au cours de brèves interruptions de la mise en situation pourrait présenter deux avantages majeurs. Tout d'abord, elle permettrait d'évaluer le raisonnement

clinique du participant en immersion, c'est-à-dire tel qu'il est activé dans la pratique réelle (en effet, on estime à 10 à 20 par heure le nombre des interruptions en médecine d'urgence) (Chisholm *et al.*, 2011). Ensuite, elle permettrait de distinguer l'activation du type 1 ou du type 2 en faisant varier les conditions de l'environnement. Une étude pilote réalisée en marge de cette étude (Pennaforte *et al.*, 2016) a montré la faisabilité d'une telle approche et son acceptabilité par les étudiants.

2. LE MODÈLE DE KANE

La description du modèle de Kane est relativement récente, mais le concept de validité est débattu depuis la deuxième moitié du XIX^{e} siècle. Sans renier les apports de ses prédécesseurs, Kane apporte une pierre à l'édifice d'une nouvelle compréhension de la validation. Une caractéristique importante est la structuration cohérente de l'argumentation supportant l'interprétation des scores du test.

2.1. Les principes généraux

Le modèle de Kane (2006, 2013) est une structure d'argumentation interprétative selon quatre niveaux d'inférence : notation, généralisation, extrapolation, implication. Chaque inférence doit contribuer à valider les conclusions concernant un candidat sur la base des résultats de son évaluation. Pour ce faire, Kane explique comment et pourquoi les arguments doivent être construits et défendus. La vision est plus souple et plus « sur mesure » que celle du modèle de Messick (1989) dont il est dérivé : le besoin de preuves dépend ici uniquement de l'utilisation et de l'interprétation du test que l'on veut en faire. Si celle-ci n'est pas spécialement ambitieuse, seules quelques preuves et analyses seront requises. Cela souligne l'importance de définir avant toute chose le contexte de l'évaluation, car l'importance accordée au type de preuves à collecter variera selon la visée de l'évaluation, formative ou certificative, et selon les enjeux découlant du résultat de l'évaluation, élevés ou non.

L'évaluation « programmatique » décrite par Schuwirth et Van der Vleuten (2012), et parfois utilisée en éducation médicale, partage des similitudes avec la démarche de validation de Kane. En effet, ces auteurs considèrent qu'elle doit tenir compte de multiples sources de données provenant de la combinaison de divers instruments. Par conséquent, aucun instrument n'est supérieur à un autre, mais chacun a ses propres forces et faiblesses. Il doit ainsi être utilisé avec une variété

d'autres méthodes dans l'évaluation en fonction de la composante particulière d'un programme en médecine. Partant de la souplesse du choix des outils d'évaluation, la mesure de la qualité des différents outils devrait être « sur mesure ». Sans véritablement aller jusqu'à le mettre en pratique, ces auteurs ont cependant fourni une éloquente présentation de la façon dont le modèle de Kane pouvait être appliqué à l'évaluation en éducation médicale. Pour ce faire, ils ont présenté différentes méthodes pouvant soutenir chaque niveau d'inférence, et qui seront reprises dans la section 3.

2.2. La définition des termes

La figure 5.2 représente une schématisation du modèle de Kane.

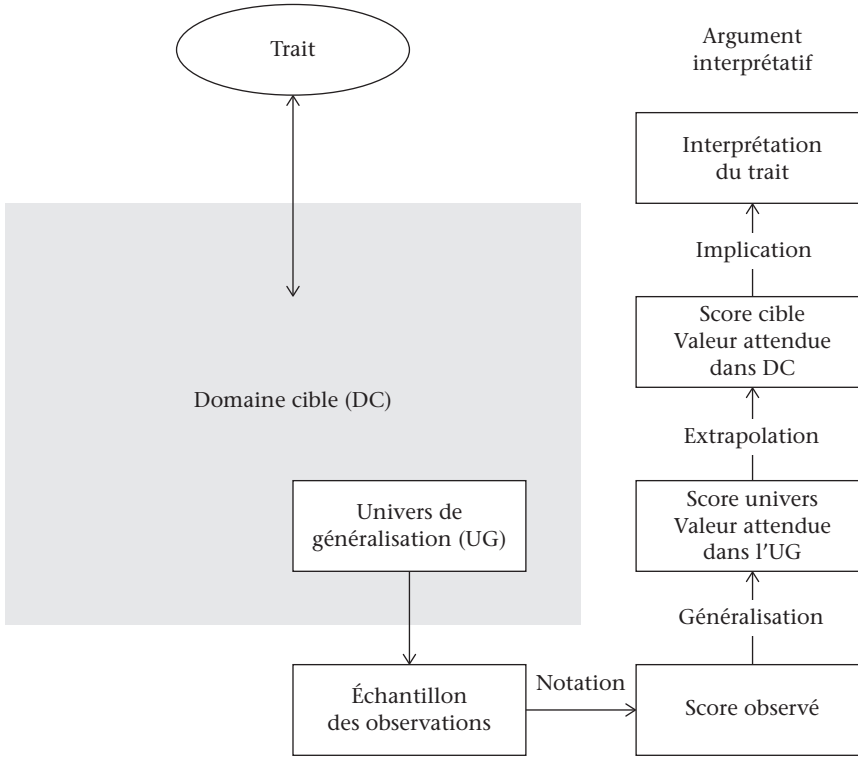


Figure 5.2
Procédure de validation de Kane

Source : Adapté de Kane, 2006, p. 33.

Kane définit le trait comme une disposition à agir d'une certaine façon en réponse à certains types de stimuli ou de tâches, dans certaines circonstances. Le trait est un attribut unidimensionnel en tant que déterminant du succès pour tous les items d'un test, même si son évaluation implique une combinaison de composantes pratiques. On peut s'attendre à ce que des personnes qui présentent un trait fort réussissent bien dans des situations ou des tâches liées à ce trait. Dans le cadre de l'évaluation des compétences, on peut rapprocher le terme de trait de la compétence étudiée. Ici, le raisonnement clinique en est un exemple. Il ne peut pas être caractérisé d'unidimensionnel à proprement parler, mais comporte plutôt une combinaison de composantes pratiques et évaluables. Comme nous le verrons plus loin, c'est notamment le cas de la collecte de données, de la génération d'hypothèse, de l'interprétation des données et de l'intuition. L'évaluation de chacune de ces composantes permettra d'approcher de façon concrète une certaine mesure du trait.

Le domaine cible est un ensemble d'observations liées au trait. Il représente les manifestations de ce trait dans les divers contextes dans lesquels il est impliqué. L'objectif n'est pas de définir un domaine bien ordonné ou un domaine qui soit facile à évaluer, mais d'établir la gamme des observations associées avec la compétence ou le trait d'intérêt. La médecine d'urgence en est un exemple.

Le score cible correspond à la valeur attendue d'une personne dans le domaine cible, c'est-à-dire à la performance attendue d'une personne à l'ensemble des observations pouvant être associées à ce trait. Par exemple, le score cible pourrait être une estimation du niveau de la compétence « raisonnement clinique » dans le domaine de la médecine d'urgence.

L'univers de généralisation est un échantillon du domaine cible à partir duquel sont tirées les observations qui constitueront l'évaluation (par exemple les items d'un examen). Idéalement, il faudrait former un échantillon aléatoire ou représentatif des observations à partir du domaine cible. Cela permettrait de généraliser le score observé dans cet échantillon au score attendu dans le domaine cible. En réalité, et notamment pour l'évaluation d'une compétence, il est habituellement impossible de former un tel échantillon. La gamme des observations incluses dans l'évaluation est donc plus restreinte que celle du domaine cible. Conséquemment, les observations incluses dans l'évaluation du trait sont typiquement tirées d'un sous-ensemble du domaine cible, souvent un très petit sous-ensemble. L'univers de généralisation pour l'évaluation du trait peut ainsi être limité aux réponses à des questions objectives portant sur une ou plusieurs situations particulières.

Par exemple, l'univers de généralisation peut être représenté par une spécialité bien définie de la médecine d'urgence, soit les soins intensifs néonataux. Dans ce contexte, on peut se demander si le score observé reflète avec justesse le score cible, sachant que le domaine cible inclut une bien plus grande variété d'observations que celles incluses dans l'univers de généralisation. La représentativité du score observé par rapport au score cible est ainsi un questionnement à la base de la validation de l'argument interprétatif pour l'interprétation d'un trait.

L'argument de validité, non représenté ici, guide la collecte et l'interprétation des preuves de validité. L'objectif d'un argument est de convaincre que tant le processus ayant permis d'obtenir le score que l'interprétation même du score d'un test sont justifiables. Il est ainsi rare qu'un seul élément de preuve soit incontestable au point de suffire à lui seul. Au contraire, l'argument consiste habituellement en de multiples éléments de preuve, incomplètes individuellement, mais collectivement suffisantes pour convaincre de l'utilité et de la pertinence du test. Cette description peut donner l'impression que le chercheur a une grande latitude dans son choix et dans son utilisation des arguments. En réalité, chaque argument doit être choisi avec soin dans une perspective stratégique de façon à veiller à ce qu'il fournisse une preuve optimale de validité. La définition de l'argument de validité s'effectue en différentes étapes. Pour reprendre l'exemple de l'évaluation d'une compétence telle que le raisonnement clinique, le chercheur doit tout d'abord prendre en considération la décision finale qu'il attend de l'interprétation des résultats du test (par exemple « l'évaluation aura-t-elle lieu dans un but formatif ou certificatif ? », « quels candidats doivent réussir cette évaluation du raisonnement clinique ? », « quels sont ceux qui doivent s'améliorer ? »), ainsi qu'une interprétation proposée qui appuierait cette décision (par exemple « des scores élevés reflètent une bonne compétence en raisonnement clinique en soins intensifs néonataux »). Ensuite, le chercheur doit identifier les principales hypothèses (par exemple « la régulation des processus du raisonnement clinique dépend de tel ou tel facteur »), déductions et inférences (par exemple à partir de quelques observations de soins intensifs néonataux jusqu'au domaine de la médecine d'urgence). Celles-ci seront associées à l'interprétation et à l'utilisation du test, et sont appelées par Kane « arguments d'interprétation et d'utilisation ». Le chercheur développe alors un plan pour tester ces hypothèses et ces déductions. Enfin, guidé par ce plan, il recueille un certain nombre de preuves empiriques provenant de sources multiples et les organise dans un argument de validité.

2.3. Le principe des inférences

La structuration de la collecte d'arguments de preuve se fait selon quatre niveaux d'inférence : de l'observation au score (inférence de notation), du score au score univers (inférence de généralisation), du score univers au domaine cible (inférence d'extrapolation) et du domaine cible au construit (inférence d'implication) (figure 5.3).

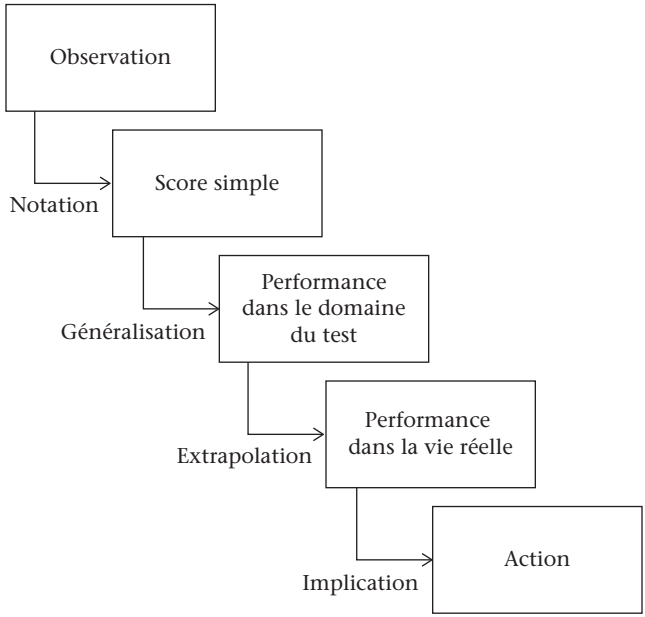


Figure 5.3
Niveaux d'inférence dans le modèle de Kane

Source : Traduit et adapté de Kane (2006).

Les principes de ces quatre niveaux d'inférence sont exposés ci-après, et les méthodes utilisées pour les soutenir seront détaillées dans la section 3.

2.3.1. *De l'observation au score « observé » : l'inférence de notation*

Le score d'un test est fortement influencé par la méthode ayant permis sa construction et par les modalités de sa notation. Dans cette optique, avant de recueillir des preuves pour appuyer les inférences d'un test dans un processus de validation, on devrait logiquement commencer par s'intéresser aux éléments du test en eux-mêmes. En réalité,

beaucoup de discussions sur la validité d'un test portent sur des questions psychométriques telles que la reproductibilité des scores et des décisions, la présence ou l'absence de biais statistiques dans les résultats des tests, et ainsi de suite (Schuwirth et Van der Vleuten, 2012), *a posteriori* et une fois les données collectées et analysées.

L'objectif de cette première inférence est en amont de s'intéresser à la méthode de construction et de notation des tests, et de vérifier que les données issues du test ont été collectées de façon appropriée.

2.3.2. *Du score « observé » au score « univers » : l'inférence de généralisation*

Autant une évaluation des connaissances peut espérer circonscrire un domaine avec un nombre défini de questions, autant il existe en théorie un nombre illimité de questions que nous pourrions créer ou sélectionner pour évaluer le domaine d'une compétence testée. Par exemple, nous pourrions utiliser 4, 12 ou un grand nombre de stations de tâches d'un examen clinique objectif structuré (ECOS) et pourrions modifier les particularités de chaque scénario ou les modalités de notation de nombreuses façons (Cook *et al.*, 2015).

L'objectif de cette deuxième inférence est d'étudier dans quelle mesure il y a corrélation entre le score « observé » et le score « univers » inconnu, mais correspondant théoriquement à l'ensemble de tous les items représentant le domaine d'intérêt. On cherche à répondre à la question suivante : Dans quelle mesure les éléments du test dans notre échantillon (les questions, les cas, les stations, les observations, les portfolios, etc.) représentent-ils tous les items théoriquement possibles dans l'univers de l'évaluation ?

2.3.3. *Du score « univers » au score « cible » : l'inférence d'extrapolation*

Les évaluateurs s'intéressent rarement à la capacité qu'ont les participants à répondre à des questions à choix multiples ou à interagir avec un patient standardisé. Ce qui les intéresse, ce sont plutôt leurs connaissances de base, leurs compétences en résolution de problème, leur jugement clinique, leurs habiletés relationnelles (Clauser *et al.*, 2012). Les scores obtenus après un test n'apportent que des preuves indirectes sur le domaine d'intérêt, même s'ils sont reproductibles. Seule l'extrapolation peut permettre de mesurer ou du moins d'anticiper la performance du participant dans le monde réel, c'est-à-dire auprès du patient.

L'objectif de cette troisième inférence est d'étudier la corrélation entre le score univers et le score cible dans le domaine d'intérêt ; en d'autres termes, l'extrapolation nous emmène de l'univers du test vers le monde réel. Cette inférence est une phase difficile du processus de validation, car elle repose sur un cadre théorique moins bien développé que pour l'inférence de généralisation. Elle est pourtant tout aussi essentielle : un score hautement reproductible, mais qui mesure faussement le trait a peu de valeur. À l'inverse, un score semblant mesurer adéquatement le trait ne dispense pas de faire une collecte de preuves, même si cette « validité de façade » pourrait soutenir une acceptabilité politique, voire légale, de l'évaluation (Clauser, Margolis et Case, 2006).

2.3.4. *Du score « cible » au trait : l'inférence d'implication*

Cette dernière inférence doit répondre à la question fondamentale de l'utilité du test, c'est-à-dire de la pertinence des décisions qui en découleront. À ce propos, Kane déclare (2006, p. 20, traduction libre) : « Il est généralement inapproprié de présumer que les preuves à l'appui d'une interprétation particulière des résultats des tests justifient automatiquement une utilisation des scores ». Il ajoute : « Une procédure de décision qui ne peut pas atteindre ses objectifs, ou qui présente un coût trop élevé, est susceptible d'être abandonnée, même si elle est basée sur une information parfaitement exacte » (Kane, 2006, p. 20, traduction libre). En d'autres termes, même si nous mesurons le trait correctement, cela ne signifie pas que cette information sera utile (et qu'il sera économiquement rentable de le documenter).

L'objectif de cette quatrième inférence est d'évaluer les conséquences ou les effets de l'évaluation sur les participants examinés et sur les autres parties prenantes et la société au sens large. Pour cela, le chercheur doit s'assurer de la crédibilité de la prise de décision à la suite du test et il doit analyser les conséquences attendues ou inattendues du test sur son utilisateur.

Il importe de rappeler ici que la démarche par inférence ne peut être réalisée que si le trait et l'argument de validité ont été correctement définis. Les caractéristiques du trait doivent pouvoir répondre aux questionnements propres à chaque niveau. Pour satisfaire à l'inférence de notation, la définition du trait à évaluer doit être précise. Toutefois, une définition trop précise risque de restreindre l'évaluation aux seuls aspects pouvant être facilement définis, ce qui peut avoir des répercussions négatives sur l'extrapolation d'une performance du score. Pour satisfaire à l'inférence de généralisation, les manifestations du trait doivent être reproductibles. Comme nous le verrons plus tard, le choix des méthodes utilisées dépendra du caractère homogène

ou hétérogène du trait. L'inférence d'extrapolation, quant à elle, ne pourra être soutenue que si le trait présente des caractéristiques observables ; une analyse de pratique peut, dans ce cas, servir à collecter les informations nécessaires à la définition du trait.

La détermination de l'argument de validité, et notamment de la finalité formative ou certificative de l'outil d'évaluation, va permettre à Kane de privilégier l'inférence de généralisation ou d'extrapolation. À ce propos, le chercheur note ceci :

> Nous pouvons renforcer l'extrapolation au détriment de la généralisation en faisant aussi bien que possible des tâches d'évaluation représentant le domaine cible, ou plutôt renforcer la généralisation au détriment de l'extrapolation en utilisant un nombre plus grand d'évaluations standardisées (Kane, 2013, p. 21, traduction libre).

Par exemple, si l'évaluation est utilisée pour prendre des décisions à enjeux élevés (examens certificatifs par exemple), les inférences de notation et de généralisation sont cruciales, car elles soutiennent l'équité et la normalisation de l'évaluation. Si, au contraire, l'évaluation a une visée formative, l'inférence d'extrapolation peut avoir plus de poids (puisque les informations issues de l'observation de la performance fournissent des indices sur une performance clinique observable).

2.3.5. *La finalité de la démarche*

Il devient clair que si le score numérique (ou le commentaire) généré à la suite d'un test (ou d'une entrevue) a une certaine valeur en lui-même, la véritable raison pour laquelle nous réalisons le test, c'est pour émettre un jugement et prendre une décision au sujet du candidat. Quelle que soit la modalité de l'évaluation, un jugement sera rendu (par exemple, conforme, à améliorer, au-delà des attentes) à un certain moment et aboutira à une prise de décision (par exemple, accepté, refusé, admissible à l'oral) (Driessen *et al.*, 2006 ; Larsen, Butler et Roediger, 2008 ; Kuper *et al.*, 2007). Chaque prise de décision peut avoir d'importantes conséquences dans la vie de cet apprenant et, dans le domaine médical, pour le patient qu'il soignera plus tard. La prise de décision doit donc être hautement justifiée par une série d'arguments soutenant les étapes de la validation, dont le modèle de Kane opérationnalise la collecte.

En fin de compte, la validation selon Kane représente tout le processus de collecte de preuves pour soutenir la démarche évaluative. En l'absence d'éléments de preuve suffisants, on pourra aboutir à la remise en question du test, à sa modification, voire à son élimination, même s'il semblait *a priori* pertinent. De plus, on pourra préciser le contexte d'utilisation du test pour lequel il est le plus adapté.

3. L'APPLICATION DU MODÈLE DE KANE POUR SOUTENIR LE PROCESSUS DE VALIDATION DE L'OUTIL D'ÉVALUATION DU RAISONNEMENT CLINIQUE : UNE DÉMARCHE PRAGMATIQUE

Intéressons-nous de plus près maintenant à la démarche employée dans le cadre du processus de validation selon Kane. Elle est ici représentée selon trois grandes étapes : 1) la définition du trait, 2) la définition de l'argument de validité et 3) la collecte d'éléments de preuve soutenant les inférences de notation, de généralisation, d'extrapolation et d'implication. Les grandes lignes étant posées dans la section 2, il sera désormais possible d'illustrer chacune des étapes de la démarche. Précisons que le trait étudié est, dans notre exemple, le raisonnement clinique, le domaine cible, la médecine d'urgence et l'univers de généralisation, les soins intensifs néonataux.

3.1. Première étape : la définition du trait

Le raisonnement clinique doit être défini précisément en tenant compte de l'assise théorique que nous avons choisie et des exigences de chacune des inférences. On peut proposer ici la définition suivante : une compétence centrale de l'exercice médical faisant intervenir des processus de pensée automatiques et réflexifs dont l'interaction dépendra du contexte dans lequel ils sont activés, dans l'objectif de prendre une décision diagnostique et de prise en charge appropriée des situations particulières de la vie quotidienne. Pour soutenir l'inférence de notation, on pourra s'assurer par des entretiens que tous les évaluateurs ont bien compris cette définition et ont la même conception de la performance qu'ils évaluent. Par ailleurs, la complexité des processus mis en jeu et l'influence d'éléments précis (contexte, situations) supposent une certaine hétérogénéité. Il faudra en tenir compte lors du choix des méthodes soutenant l'inférence de généralisation. Le caractère réaliste des mises en situation aidera, quant à lui, à soutenir l'inférence d'extrapolation. Enfin, la place centrale du raisonnement clinique au sein du cadre CanMeds indique l'intérêt de son évaluation pour la société (inférence d'implication).

3.2. Deuxième étape : la définition de l'argument de validité

L'évaluation sera ici formative. D'abord, parce qu'on privilégie la rétroaction individuelle après avoir relevé une faille dans le raisonnement clinique d'un participant donné. En outre, un test discriminant le « bon » et le « mauvais » raisonnement clinique dans un but certificatif

exigerait de le soumettre à grande échelle afin d'avoir des mesures solides. Or, les investissements humains et financiers requis par la simulation ne le permettent pas, du moins dans un premier temps.

La finalité formative de l'évaluation demande de récolter un maximum d'éléments de preuve soutenant l'extrapolation en cherchant à évaluer le raisonnement clinique sous des angles variés. En revanche, le soutien de l'inférence de généralisation sera moins essentiel.

3.3. Troisième étape : la collecte d'éléments de preuve

Plusieurs auteurs en éducation médicale ont proposé un éventail de méthodes en vue de collecter des éléments de preuve soutenant chaque niveau d'inférence du modèle de Kane. Celles-ci ont notamment été utilisées en tant que « grilles de lecture » pour vérifier la qualité de la validation d'outils d'évaluation tels que le *Mini-Clinical Evaluation Exercise* ou « mini-CEX » (Hawkins *et al.*, 2010), l'*Objective Structured Assessment of Technical Skills* (Hatala *et al.*, 2015), le *Prostate Specific Antigen* ou une évaluation qualitative basée sur des entretiens avec des résidents en médecine interne (Cook *et al.*, 2015). Après avoir exposé une liste non exhaustive de méthodes disponibles pour chaque niveau d'inférence, nous proposerons une liste de questions que le chercheur peut se poser en vue de soutenir la validation de l'outil d'évaluation qu'il développe.

3.3.1. *Les éléments de preuve soutenant l'inférence de notation*

Les éléments de preuve soutenant l'inférence de notation concernent les règles de construction des tests, les modalités de l'évaluation et la collecte et l'analyse des données.

Les règles de construction du test sont conçues pour optimiser la probabilité qu'un étudiant qui maitrise le sujet réponde correctement et qu'un autre qui ne le maitrise pas réponde incorrectement (Downing, 1997), en évitant d'introduire des biais dans la manière de poser les questions. En d'autres termes, elles servent à réduire au minimum la probabilité qu'un étudiant donne un résultat « faux positif » ou « faux négatif » en utilisant, par exemple, des stratégies de réponses (choisir la réponse la plus longue, capitaliser sur les sujets de prédilection de l'évaluateur, etc.). Certaines d'entre elles sont exposées ci-dessous.

– Les items du test doivent répondre à une définition précise et étroitement liée à la définition du trait. Ils doivent être représentatifs du domaine cible dans une perspective de généralisation, et permettre de bâtir des échelles d'évaluation solides

(Hatala *et al.*, 2015). Nous devrions soumettre préalablement les questions à un panel d'experts en évaluation, en néonatologie et en simulation.

– La standardisation de l'environnement du test doit être assurée, particulièrement dans les contextes certificatifs. Elle permettra de maximiser les chances que les données concernent des participants soumis à des conditions identiques de test. Dans le cas des évaluations en salle de classe, des facteurs tels que l'heure de l'examen, le temps alloué, les conditions d'assise, la température, la luminosité, la longueur des questions, la taille de l'écran ou la qualité d'impression doivent être contrôlés. Ce « contrôle de qualité » pourra être documenté et permettra à l'évaluateur d'avoir un degré de confiance élevé en ses notations. Pour une évaluation basée sur un patient standardisé, la standardisation peut être plus difficile à obtenir, car elle implique les facteurs humains liés au patient. Deux patients standardisés, aussi bien entraînés qu'ils soient, ne pourront jouer leur rôle que de façon *quasiment* identique pour une même situation, de même qu'un patient standardisé ne pourra jouer son rôle que d'une façon *quasiment* identique dans deux situations. Cela pose le problème de la cohérence intra-patient et inter-patient. Pour une simulation sur mannequin, la standardisation de l'évaluation est facilitée par l'absence de facteurs humains ; cela permet de s'affranchir des problèmes liés à la variation des performances des patients. À l'extrême, il pourrait être possible de tout standardiser lors d'une séance de simulation, afin que seules les performances du participant contribuent à la variance des résultats du test. Cette standardisation maximale se ferait cependant au détriment du réalisme concernant les relations médecin-patient (de nombreux participants trouvent déjà artificiel le fait d'interroger et d'examiner un mannequin) et pourrait affaiblir l'extrapolation du score au trait tel qu'il existe dans la réalité.
– Le contrôle de la sécurité du test doit permettre de s'assurer de l'absence de tricherie, *a fortiori* pour les évaluations à enjeux élevés. Dans cette perspective, il importe de ne pas réutiliser des items d'un test pour un autre test si le participant a accès à des ressources externes entre ses deux passations. La réutilisation d'un item est d'autant plus fréquente que la base de données est informatisée ; l'encryptage des items permet alors d'accroître la sécurité. De même, l'absence d'accès à des ressources électroniques par le participant durant l'examen est indispensable. Dans le cas d'évaluations basées sur des patients standardisés ou sur des mannequins, il faut s'assurer que les participants ne

communiquent pas entre eux (ou avec les patients standardisés), la connaissance préalable d'une situation pouvant augmenter artificiellement les performances des participants. Pour y parvenir, nous prévoyons faire passer le test sur une période limitée et de contrôler l'accès à des ressources externes.

– La crédibilité du test doit se baser sur une argumentation théorique. Ici, les questions de types TCS et PRC explorent différentes étapes de la démarche hypothéticodéductive, elle-même largement rapportée dans la littérature depuis sa première description par Elstein, Shulman et Sprafka (1978). De plus, le test sera basé sur l'hypothèse que les connaissances et le jugement sont indispensables pour la prise de décision en pratique (Clauser *et al.*, 2012).
– La faisabilité du test doit être assurée. Une inférence valide ne peut être faite que si l'outil est suffisamment convivial et facile d'utilisation. Lors de la construction du test, on doit prendre en compte son utilisation future et veiller à ce que le participant soit complètement à l'aise avec l'instrument. La formulation et la présentation des questions doivent donc être réfléchies. Dans notre exemple, différentes mises à l'essai permettront de vérifier la faisabilité de l'évaluation auprès d'étudiants sélectionnés.
– Le mode de sélection et la formation des évaluateurs doivent également être anticipés. Le choix de ceux-ci doit garantir un degré de sévérité aussi comparable que possible dans l'évaluation. Le travail de sélection peut être fait en amont par le développeur du test et devra reposer sur une enquête approfondie sur l'aptitude des évaluateurs, voire sur des mises à l'essai préalables pour s'assurer d'une forte corrélation inter-évaluateurs. La formation des évaluateurs est également essentielle, afin qu'ils soient à l'aise avec les objectifs de l'évaluation et les modalités de notation. On remarque, par exemple, que la tolérance des évaluateurs est moins élevée lorsque le mini-CEX est administré en contexte de recherche que lorsqu'il est appliqué en pratique (Hawkins *et al.*, 2010). La tendance des évaluateurs à surestimer les résultats lorsque l'implication clinique est plus importante peut s'expliquer par un défaut dans leur formation.

Les modalités d'évaluation doivent aussi avoir été considérées. Tout comme la construction du test, elles doivent respecter un certain nombre de principes.

– L'évaluation doit être faisable. Un test compliqué à noter (par exemple une grille constituée de 60 items) accaparera la charge cognitive de l'observateur, qui devra trouver comment

gérer l'outil au lieu d'observer la performance du participant. De plus, la notation sera d'autant plus compliquée à réaliser que la corrélation inter-items sera forte. Voilà une autre limite du mini-CEX (Hawkins *et al.*, 2010). Les grilles d'évaluation devront comporter différentes dimensions bien distinctes et permettre d'établir clairement les forces et faiblesses des participants. Dans notre exemple, les questions de type TCS seront associées à une définition précise des attentes en fonction des niveaux des échelles de type Likert, et les questions de type PRC, à une définition précise des descriptions des diagnostics et de la collecte des données attendus.

– La modalité de notation doit être cohérente avec certaines caractéristiques liées au construit évalué. L'application d'une pénalité pour l'hésitation en est un bon exemple (Muijtjens *et al.*, 1999). Si l'on considère que le test tente de capturer les connaissances de l'étudiant, la prise en compte de l'hésitation (« je ne sais pas ») est une source d'erreur dans l'inférence de l'observation au score. Si, au contraire, on considère que le test cherche à évaluer les connaissances que l'élève est prêt à utiliser, l'hésitation assumée par le candidat pourrait bien être considérée comme une source pertinente de variance du trait. Dans le cas du raisonnement clinique, il faudra prendre en considération les différences qui existent entre experts et novices pour l'évaluation de la collecte des données. Les travaux de recherches réalisés par Elstein *et al.* (1978) ont été les premiers à montrer que les experts avaient besoin de moins d'informations que les novices pour prendre des décisions concernant un diagnostic ou un traitement. Cette particularité devra être intégrée afin de ne pas sous-estimer les performances des experts dans la collecte des données.
– Les méthodes de notations sont à déterminer avec soin. Les plus compliquées ne sont pas intrinsèquement meilleures que les simples notations dichotomiques « 1-0 ». Bien qu'elles introduisent plus de variance, il ne s'agit pas forcément d'une variance reliée au trait (Swanson, Grosso et Norcini, 1987). Par exemple, toutes les études concernant le TCS, sauf une (Kelly, Durning et Denton, 2012), utilisent une échelle de Likert comportant cinq choix (–2 : élimine, –1 : n'est pas en faveur, 0 : n'influence pas, +1 : en faveur, +2 : confirme). Pourtant, Lineberry, Kreiter et Bordage (2013, 2014) ont récemment rapporté les limites de l'utilisation d'une telle échelle à la faveur d'une échelle proposant trois choix. Selon ces auteurs, l'utilisation de cinq choix est associée à une tendance qu'ont

les experts à ne pas concevoir de questions aux réponses « extrêmes » (–2 : élimine et +2 : confirme) et permet aux utilisateurs d'adapter leurs stratégies de réponses. Les réponses autour de la moyenne sont ainsi souvent associées au maximum de points, ce phénomène étant diminué lors de l'utilisation d'une échelle avec trois choix.

Enfin, la collecte et l'analyse des données doivent avoir été soigneusement réalisées. L'utilisation des moyennes, écarts-types et autres tests statistiques est classique pour convertir objectivement un grand nombre de données quantitatives en scores. La preuve de la qualité de la conversion des données en score repose alors sur l'application correcte des statistiques descriptives et sur l'exactitude des calculs. Le recours au seul jugement d'expert est aussi possible pour les données qualitatives. Dans ce cas, la preuve sera subjective et basée sur l'expertise des examinateurs et leur formation à porter un jugement sur les manifestations observables. Cette étape est essentielle pour passer des résultats à l'interprétation des résultats d'un test. Dans notre cas, un contrôle de la qualité des données quantitatives sera réalisé par un statisticien.

Le tableau 5.1 résume les notions abordées plus haut en proposant des questions que le chercheur doit se poser en vue de soutenir la première inférence du modèle de Kane.

Tableau 5.1
Propositions de questions auxquelles doit répondre le chercheur en vue de soutenir l'inférence de notation

Objectifs	**Questions**
Définir les conditions de construction du test	• Les items du test couvriront-ils le domaine d'intérêt ? • La formulation et la présentation des items seront-elles standardisées ? • L'environnement sera-t-il contrôlé ? • Les règles de sécurité seront-elles respectées ? • Les items du test seront-ils crédibles ? • L'utilisation de l'évaluation par le participant sera-t-elle aisée ? • La sélection des évaluateurs sera-t-elle justifiée ? • Les évaluateurs recevront-ils une formation ?
Définir les modalités de l'évaluation	• L'évaluation sera-t-elle facile à réaliser par l'évaluateur ? • Comment la notation sera-t-elle établie ? • La méthode de notation sera-t-elle adaptée à l'évaluation du trait ?
Définir le mode de collecte et d'analyse des données	• Des mesures de contrôle de qualité des données seront-elles prévues ? • La fidélité inter-évaluateurs sera-t-elle évaluée ? • Quelles analyses descriptives ou analytiques seront prévues ?

3.3.2. *Les éléments de preuve soutenant l'inférence de généralisation*

Les éléments de preuve soutenant l'inférence de généralisation concernent la représentativité des observations et la reproductibilité du test.

La représentativité des observations est quelquefois très théorique. Il faudrait idéalement démontrer que l'échantillon des questions ou des observations est représentatif de l'univers dans lequel le score doit être généralisé. Si l'univers est considéré comme homogène, l'échantillonnage doit être suffisamment large pour s'affranchir de toutes les sources indésirables de variance. En revanche, si l'univers est hétérogène, l'échantillonnage doit être tel que tous les aspects de l'univers soient inclus dans l'échantillon. En pratique, cela est très difficile, et c'est le cas pour une spécialité médicale qui est reconnue pour être hétérogène et une compétence telle que le raisonnement clinique qui est propre à chaque cas (Eva, 2003). Les méthodes pour assurer un échantillonnage approprié pourraient donc plutôt inclure un plan de test validé par des experts qu'un échantillonnage aléatoire pour aider à la sélection systématique d'éléments. Le concept de saturation peut être d'une grande aide, car il permet de répondre à la question : Une nouvelle observation apporte-t-elle des informations importantes par rapport à celles déjà recueillies ? Cela ressemble à la question que se posent quotidiennement les médecins : Une investigation supplémentaire changera-t-elle le diagnostic ou la prise en charge d'une situation donnée ?

La reproductibilité du test désigne la probabilité d'obtenir des scores similaires si l'on utilise un nouvel échantillon de situations. Elle fait appel à l'application de certains modèles dont la nature dépend encore une fois du caractère homogène ou hétérogène de l'univers du score.

À ce titre, le domaine de la mesure offre diverses manières de modéliser les données selon leur dimensionnalité, selon la théorie choisie (théorie classique des tests ou théorie de réponse à l'item) ou encore selon la nature du lien entre le score observé et le trait d'intérêt. Le chercheur peut pour cela consulter les travaux de différents auteurs (voir, par exemple, Laveault, 2002 ; Ricketts, 2009 ; Couturat, 2012).

De même que pour l'inférence précédente, le tableau 5.2 propose une liste de questions que le chercheur doit se poser en vue de soutenir la deuxième inférence du modèle de Kane.

Tableau 5.2
Propositions de questions auxquelles doit répondre le chercheur en vue de soutenir l'inférence de généralisation

Objectifs	Questions
Soutenir la représentativité de l'échantillon d'observations	L'univers du score sera-t-il homogène ? Un plan d'échantillonnage sera-t-il prévu ?
Soutenir la reproductibilité de l'évaluation	Le trait est-il homogène ? Quelles sources de variance liées ou non au trait seront considérées ?

3.3.3. *Les éléments de preuve soutenant l'inférence d'extrapolation*

Les éléments de preuve soutenant l'inférence d'extrapolation concernent l'authenticité du test et la corrélation entre le score du test et les scores provenant d'autres évaluations du même trait.

L'authenticité du test vise à assurer que le score produit par le test est transposable dans la vie réelle ; elle nécessite des items et un environnement réalistes.

- Le réalisme des items du test fait appel à l'expérience du ou de ses concepteurs. En médecine, le test doit être cohérent avec les situations auxquelles l'utilisateur pourrait être confronté dans son exercice futur, il est donc important que le chercheur conserve des liens avec son exposition clinique. On pourrait, par exemple, demander à des médecins de différents niveaux de formation de verbaliser une tâche à réaliser. Cela pourrait aussi permettre d'évaluer la pertinence du poids des items du score pour la pratique réelle (Cook *et al.*, 2015).
- Le test doit aussi permettre de discriminer les utilisateurs selon leur expérience, comme cela s'observe dans la pratique réelle. Ce peut être particulièrement intéressant dans le cadre du raisonnement clinique, où certains auteurs ont rapporté une prépondérance du raisonnement de type 1 chez les « experts » (Norman, Young et Brooks, 2007). Généralement, les études sur l'expertise comparent la performance des novices à celle des experts (Chi, 2006). Cela suppose de définir un seuil entre « novices » et « experts ». Cependant, Patel, Groen et Norman (1991) ont rapporté le caractère trop général d'une telle classification pour appréhender le développement d'une compétence. Dans le cadre du raisonnement clinique, Schmidt et Rikers (2007) considèrent en outre que le développement de l'expertise est un processus continu, de la première année de médecine jusqu'à la fin de la carrière. En l'absence de seuil

identifiable, on pourra donc observer la tâche réalisée par des praticiens de différents niveaux de compétence afin d'en déduire les items à sélectionner dans le test ou la pondération de leur notation.

– L'authenticité de l'environnement permet de limiter les sources de variances « artificielles » reliées aux conditions de l'examen (Clauser *et al.*, 2012). Il est toujours bon de tenter de les contrôler, même lorsque la visée de l'évaluation est formative, mais surtout lorsqu'elle est certificative. Par exemple, un examen à enjeu élevé peut comporter un nombre important de questions dans un temps limité pour des raisons organisationnelles. Dans ce cas, la performance peut être influencée par les conditions dans lesquelles elle a été mesurée (un étudiant peut moins bien réussir son examen s'il manque de temps). La simulation médicale permet, pour sa part, de recréer un environnement authentique et dont les conditions sont de surcroît contrôlées. Comme nous l'avons vu, la standardisation maximale des conditions pourrait théoriquement permettre de réduire au maximum toutes sources de variances « artificielles » et de n'étudier – dans notre cas – que le raisonnement clinique du participant. En réalité, les modalités de notation pourraient tout de même interférer sur l'extrapolation à la vie réelle. Par exemple, certaines listes de contrôles peuvent échouer à capter certaines compétences subtiles de l'entrevue qui peuvent faciliter la collecte des données. De même, la connaissance par le participant du fait que ses performances seront évaluées peut changer sa stratégie avec le patient afin de maximiser les points obtenus. Tout cela explique qu'un environnement aussi authentique qu'il soit ne pourra jamais constituer une preuve de validité en lui-même.

La concordance entre le résultat au test avec celui d'autres évaluations du même trait est utilisée en l'absence d'évaluation de référence, comme c'est le cas pour le raisonnement clinique. Si des informations provenant de diverses sources ou éléments d'évaluation sont combinées (par exemple une station d'un ECOS sur l'examen du genou et un examen écrit qui met l'accent sur l'anatomie du genou), la triangulation des informations devient plus appropriée pour l'argument, en permettant l'exhaustivité de la récolte d'informations (Van der Vleuten *et al.*, 2010 ; Govaerts *et al.*, 2011). Eu égard au raisonnement clinique, il faut donc chercher d'autres prises d'informations que les TCS et PRC, déjà intégrés dans la simulation que nous proposons. Plus précisément, il faut trouver d'autres formes d'évaluation des types 1 et 2 en cohérence avec l'ancrage théorique que nous en avons retenu.

- Un exemple est la verbalisation des participants. Il s'agit d'une approche souvent rapportée dans la littérature concernant le raisonnement clinique (Banning, 2008 ; Pottier *et al.*, 2010 ; Durning *et al.*, 2012 ; Roberts, 2013 ; Forsberg *et al.*, 2014 ; Burbach, Barnason et Thompson, 2015). Elle permet au participant d'expliciter les processus mis en œuvre lors de l'activation de son raisonnement clinique et à l'évaluateur de l'appréhender. Cette verbalisation se produit le plus souvent lors d'entrevues semi-structurées dont l'analyse qualitative fournit des données intéressantes sur les processus cognitifs engagés.
- Un autre exemple est l'analyse du parcours visuel. Elle s'intéresse au comportement visuel des participants et postule que la quantification du regard vers une zone d'intérêt peut différencier les médecins selon leur niveau de formation. Elle utilise des caméras de réflexion qui mesurent des reflets cornéens émis en lumière infrarouge. Il faut en premier lieu déterminer des zones particulières de l'environnement (par exemple une anomalie sur une radiographie), puis l'analyse quantifiera la pose du regard du participant sur cette zone et donnera des informations sur les zones du champ visuel inspectées. Cela permet d'obtenir des indices au sujet de l'information qui a été incluse dans le processus de prise de décision. Blondon, Wipfli et Lovis (2015) ont récemment présenté une revue systématique de cet outil pour l'évaluation du raisonnement clinique. Dans l'ensemble, les études ont montré que les experts avaient une meilleure précision de diagnostic, avaient besoin de moins de temps pour repérer le premier élément anormal de l'environnement et ensuite le regardaient plus souvent. On peut aussi imaginer que la mesure de l'errance visuelle pourrait permettre de repérer les étudiants moins expérimentés ou « paniqués ».
- Enfin, l'analyse de l'activité des zones cérébrales sollicitées dans le raisonnement clinique fait appel à l'imagerie par résonance magnétique fonctionnelle (IRMf). Elle peut confirmer les postulats théoriques impliquant les processus cognitifs de types 1 et 2, voire apporter de nouvelles hypothèses sur leur interaction. Goel *et al.* (1998 ; Goel et Dolan, 2004) ont été les premiers à montrer que les processus « inductifs » et « déductifs » étaient caractérisés par une activation de zones distinctes du cortex préfrontal (CPF). D'autres auteurs ont par la suite confirmé le lien entre le raisonnement clinique et le CPF (Lee *et al.*, 2007 ; Arden *et al.*, 2010 ; Beer, Lombardo et Bhanji, 2010 ; Liu *et al.*, 2012 ; Hruska *et al.*, 2015 ; Durning *et al.*, 2015 ; Chang *et al.*, 2016), apportant une validité que

Durning et ses collaborateurs (2015) nomment « biologique » à l'évaluation du raisonnement clinique. Même s'il est évident que le coût, les impératifs organisationnels et le caractère anxiogène de l'IRMf la rendent impraticable pour l'évaluation quotidienne du raisonnement clinique, les données que cet examen procure permettent de consolider la triangulation des éléments de preuve liés à l'extrapolation.

Comme pour les inférences précédentes, le tableau 5.3 propose une liste de questions que le chercheur doit se poser en vue de soutenir la troisième inférence du modèle de Kane.

Tableau 5.3
Propositions de questions auxquelles doit répondre le chercheur en vue de soutenir l'inférence d'extrapolation

Objectifs	Questions à se poser
Réduire les sources de variances « artificielles »	• Comment s'assurera-t-on que les items et l'environnement sont réalistes ? • La notation permettra-t-elle de discriminer les niveaux de formation ?
Définir la corrélation entre le test et d'autres méthodes d'évaluation du même trait	• Comment mettra-t-on en relation les résultats des autres sources d'évaluation du même trait ?

3.3.4. *Les éléments de preuve soutenant l'inférence d'implication*

Les éléments de preuves soutenant l'inférence d'implication concernent la crédibilité de la prise de décision et les conséquences de l'évaluation.

La crédibilité de la prise de décision (par exemple succès ou échec) ne peut être acquise si elle est argumentée. Cela est encore plus important si l'enjeu de l'évaluation est élevé. Dans ce cas, le score de passage doit être également justifié. Par exemple, un score devant distinguer les novices des experts en fonction de leur habileté à collecter des données doit s'appuyer sur une solide théorie du développement de l'expertise. Dans le cas contraire, l'argument sur lequel repose l'interprétation du score sera faible.

Les conséquences de l'évaluation sur le participant doivent avoir été anticipées. Ce type d'argument de preuve, pourtant essentiel, n'est jamais retrouvé dans les articles publiés en éducation médicale. Cook *et al*. (2015) proposent plutôt différentes pistes de réflexion.

– La façon la plus simple de recueillir des données sur les conséquences de l'évaluation serait d'offrir l'évaluation à certains candidats mais pas à d'autres, et de comparer les résultats.

– Également, on pourrait étudier les retombées volontaires et involontaires du test, par la quantité et la qualité des commentaires reçus, la durée et le coût de la formation, le taux d'abandon, le niveau de stress, les autres mesures de performance à court et long terme, l'influence sur les évaluateurs, et les effets sur les soins aux patients. Ces études sont toutefois difficiles à conduire et dépassent la portée de la plupart des chercheurs.
– Enfin, on pourrait inviter les participants ayant échoué à appliquer des stratégies de remédiation et à suivre l'évolution de leurs résultats.

Le tableau 5.4 propose une liste de questions que le chercheur doit se poser en vue de soutenir la quatrième inférence du modèle de Kane.

Tableau 5.4
Propositions de questions auxquelles doit répondre le chercheur en vue de soutenir l'inférence d'implication

Objectifs	**Questions à se poser**
Soutenir la crédibilité des interprétations du test	• Un score de passage devra-t-il être déterminé, et si oui, comment le justifier ? • Les interprétations du test seront-elles soutenues par une théorie ?
Soutenir les conséquences du test	• Des actions réalisées à la suite du test seront-elles envisagées ? • Quelles seront les conséquences attendues ou inattendues du test ? • Des études de suivi seront-elles prévues ? • Des entrevues avec les participants seront-elles prévues ? • Des études d'incidence seront-elles prévues ?

CONCLUSION

La validation de l'évaluation doit s'entendre comme un processus visant à soutenir l'interprétation qu'on en fait. L'intérêt du modèle de Kane réside dans son opérationnalisation. Il propose une démarche structurée de collecte d'arguments de preuves variées tout en étant suffisamment souple pour s'adapter aux besoins du chercheur. Son utilisation nécessite au préalable une planification cohérente de la finalité de l'évaluation, car c'est elle qui dictera la collecte de preuves. Il pourra arriver que les preuves soient en contradiction, au sein d'une même inférence ou entre diverses inférences. Encore une fois, la détermination préalable de l'argument de validation permettra d'établir lesquelles prioriser.

Ce texte avait pour objectif de fournir au chercheur les outils nécessaires pour appliquer le modèle de Kane à la validation de l'outil d'évaluation qu'il étudie.

Notre réflexion sert également de base à une méthodologie que nous souhaitons mettre éventuellement à l'épreuve dans le cadre d'une recherche qui viserait à construire un outil d'évaluation du raisonnement clinique dans un contexte de simulation médicale. Dans ce contexte, nous prévoyons définir précisément le raisonnement clinique et la finalité de l'évaluation que nous proposons. Ensuite, nous constituerons un panel d'experts afin de contrôler la qualité des mises en situation et des tests. Cela permettra de satisfaire à l'inférence de notation. Notre capacité à recruter des participants et l'utilisation de modèles probabilistes soutiendront la généralisabilité des performances de l'outil. Les contraintes liées à cette inférence doivent cependant être ici relativisées par le caractère formatif de l'évaluation que nous développons : la généralisabilité de l'outil est pour nous moins essentielle que la rétroaction personnalisée qu'il permettra. À l'inverse, l'extrapolation à la vie réelle des résultats observés sera déterminante pour une compétence telle que le raisonnement clinique. L'utilisation de séances de simulations médicales comme contexte d'évaluation permet d'ores et déjà d'optimiser le réalisme des situations. De plus, nous comptons y intégrer d'autres méthodes d'évaluation du raisonnement clinique et corréler leurs résultats à ceux de notre outil. Enfin, nous pourrons, dans un second temps, étendre cette recherche à l'étude du suivi des participants et de la répercussion de l'évaluation sur leurs performances futures. Pris dans leur ensemble, ces divers éléments de preuve pourront apporter des arguments de poids quant à l'utilité de l'outil d'évaluation que nous proposons pour le participant, pour le formateur et pour la société idéalement.

Fort de cet exemple, nous espérons que le chercheur pourra maintenant s'approprier le modèle de Kane en fonction des objectifs de sa recherche, dont l'essence tient avant tout dans la cohérence de son argumentation.

BIBLIOGRAPHIE

Amin, M.R. et D.R. Friedmann (2013). « Simulation-based training in advanced airway skills in an otolaryngology residency program », *Laryngoscope*, *123*(3), p. 629-634, doi : 10.1002/lary.23855.

Arden, R., R.S. Chavez, R. Grazioplene et R.E. Jung (2010). « Neuroimaging creativity : A psychometric view », *Behavioural Brain Research*, *214*(2), p. 143-156, doi : 10.1016/j.bbr.2010.05.015.

Banning, M. (2008). « The think aloud approach as an educational tool to develop and assess clinical reasoning in undergraduate students », *Nurse Educator Today*, *28*(1), p. 8-14, doi : 10.1016/j.nedt.2007.02.001.

Beer, J.S., M.V. Lombardo et J.P. Bhanji (2010). « Roles of medial prefrontal cortex and orbitofrontal cortex in self-evaluation », *Journal of Cognitive Neuroscience*, *22*(9), p. 2108-2119, doi : 10.1162/jocn.2009.21359.

Blondon, K., R. Wipfli et C. Lovis (2015). « Use of eye-tracking technology in clinical reasoning : a systematic review », *Studies in Health Technology and Informatics*, *210*, p. 90-94.

Burbach, B., S. Barnason et S.A. Thompson (2015). « Using "think aloud" to capture clinical reasoning during patient simulation », *The International Journal of Nursing Education Scholarship*, *12*, doi : 10.1515/ijnes-2014-0044.

Chang, H.J., J. Kang, B.J. Ham et Y.M. Lee (2016). « A functional neuroimaging study of the clinical reasoning of medical students », *Advances in Health Sciences Education. Theory and Practice*, doi : 10.1007/s10459-016-9685-6.

Charlin, B., R. Gagnon, S. Lubarsky, C. Lambert, S. Meterissian, C. Chalk et C. van der Vleuten (2010). « Assessment in the context of uncertainty using the script concordance test : More meaning for scores », *Teaching and Learning in Medicine*, *22*(3), p. 180-186. doi :10.1080/10401334.2010.488197.

Charlin, B., L. Roy, C. Brailovsky, F. Goulet et C. Van der Vleuten, C. (2000). « The Script Concordance test : A tool to assess the reflective clinician », *Teaching and Learning in Medicine*, *12*(4), p. 189-195, doi : 10.1207/S15328015TLM1204_5.

Charlin, B., J. Tardif et H.P. Boshuizen (2000). « Scripts and medical diagnostic knowledge : Theory and applications for clinical reasoning instruction and research », *Academic Medicine*, *75*(2), p. 182-190.

Cheng, A., W. Eppich, V. Grant, J. Sherbino, B. Zendejas et D.A. Cook (2014). « Debriefing for technology-enhanced simulation : A systematic review and meta-analysis », *Medical Education*, *48*(7), p. 657-666, doi : 10.1111/medu.12432.

Chi, M.T.H. (2006). « Two approaches to the study of experts' characteristics », dans K.A. Ericsson, N. Charness, P. Feltovich et R.R. Hoffman (dir.), *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge : Cambridge University Press, p. 21-30.

Chisholm, C.D., C.S. Weaver, L. Whenmouth et B. Giles (2011). « A task analysis of emergency physician activities in academic and community settings », *Annals of Emergency Medicine*, *58*(2), p. 117-122, doi : 10.1016/j.annemergmed.2010.11.026.

Clauser, B.E., J.M. Margolis et S.M. Case (2006). « Testing for licensure and certification in the professions », dans R.L. Brennan (dir.), *Educational measurement*, Westport : Praeger Publishers.

Clauser, B.E., M.J. Margolis, M.C. Holtman, P.J. Katsufrakis et R.E. Hawkins (2012). « Validity considerations in the assessment of professionalism », *Advances in Health Sciences Education. Theory and Practice*, *17*(2), p. 165-181, doi : 10.1007/s10459-010-9219-6.

Cook, D.A., R. Brydges, S. Ginsburg et R. Hatala (2015). « A contemporary approach to validity arguments : A practical guide to Kane's framework », *Medical Education*, *49*(6), p. 560-575, doi : 10.1111/medu.12678.

Couturat, P. (2012). *Troubles de l'acquisition des coordinations à l'école maternelle : validation d'une échelle d'hétéroévaluation*, Montpellier, Université Paul Valery-Montpellier III.

Custers, E.J. (2013). « Medical education and cognitive continuum theory : An alternative perspective on medical problem solving and clinical reasoning », *Academic Medicine*, *88*(8), p. 1074-1080, doi : 10.1097/ACM.0b013e31829a3b10.

Dory, V., R. Gagnon, D. Vanpee et B. Charlin (2012). « How to construct and implement script concordance tests : Insights from a systematic review », *Medical Education*, *46*(6), p. 552-563, doi : 10.1111/j.1365-2923.2011.04211.x.

Downing, S. (1997). « Test item development : Validity evidence from quality assurance procedures », *Applied Measurement in Education*, *10*, p. 61-82.

Dreifuerst, K.T. (2012). « Using debriefing for meaningful learning to foster development of clinical reasoning in simulation », *The Journal of Nursing Education*, *51*(6), p. 326-333, doi : 10.3928/01484834-20120409-02.

Driessen, E.W., K. Overeem, J. Van Tartwijk, C.P. Van der Vleuten et A.M. Muijtjens (2006). « Validity of portfolio assessment : Which qualities determine ratings ? », *Medical Education*, *40*(9), p. 862-866, doi : 10.1111/j.1365-2929.2006.02550.x.

Duggan, P. et B. Charlin (2012). *Summative assessment of 5th year medical students' clinical reasoning by Script Concordance Test : requirements and challenges*, *BMC Medical Education*, *12*(29), doi :10.1186/1472-6920-12-29.

Durning, S.J., M.E. Costanzo, T.J. Beckman, A.R. Artino Jr., M.J. Roy, C. Van der Vleuten et L. Schuwirth (2015). « Functional neuroimaging correlates of thinking flexibility and knowledge structure in memory : Exploring the relationships between clinical reasoning and diagnostic thinking », *Medical Teacher*, p. 1-8, doi : 10.3109/0142159x.2015.1047755.

Durning, S.J., J. Graner, A.R. Artino Jr., L.N. Pangaro, T. Beckman, E. Holmboe et L. Schuwirth (2012). « Using functional neuroimaging combined with a think-aloud protocol to explore clinical reasoning expertise in internal medicine », *Military Medicine*, *177*(9), suppl., p. 72-78.

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

Eva, K.W. (2003). « On the generality of specificity », *Medical Education*, *37*(7), p. 587-588.

Eva, K.W. (2005). « What every teacher needs to know about clinical reasoning », *Medical Education*, *39*(1), p. 98-106, doi : 10.1111/j.1365-2929.2004.01972.x.

Evans, J.S. (2008). « Dual-processing accounts of reasoning, judgment, and social cognition », *Annual Review of Psychology*, *59*, p. 255-278, doi : 10.1146/annurev.psych.59.103006.093629.

Faucher, C., T. Pelaccia, M. Nendaz, M. Audétat et B. Charlin (2016). « Un professionnel de santé qui résout efficacement les problèmes : le raisonnement clinique », dans T. Pelaccia et J. Tardif (dir.), *Comment [mieux] former et évaluer les étudiants en médecine et en sciences de la santé ?*, Louvain-la-Neuve : De Boeck Supérieur, p. 33-44.

Fida, M. et S.E. Kassab (2015). « Do medical students' scores using different assessment instruments predict their scores in clinical reasoning using a computer-based simulation ? », *Advances in Medical Education and Practice*, *6*, p. 135-141, doi : 10.2147/amep.s77459.

Forsberg, E., K. Ziegert, H. Hult et U. Fors (2014). « Clinical reasoning in nursing, a think-aloud study using virtual patients : A base for an innovative assessment », *Nurse Educator Today*, *34*(4), p. 538-542, doi : 10.1016/j.nedt.2013.07.010.

Fournier, J.P., A. Demeester et B. Charlin (2008). « Script concordance tests : Guidelines for construction », *BMC Medical Informatics and Decision Making*, *8*, p. 18, doi : 10.1186/1472-6947-8-18.

Frank, J.R. (2005). *Le cadre de compétences CanMEDS 2005 pour les médecins. L'excellence des normes, des médecins et des soins*, <http://chirurgie.umontreal.ca/wp-content/uploads/sites/20/CanMEDS.pdf>, consulté le 25 avril 2017.

Gagnon, R., B. Charlin, C. Lambert, B. Carriere et C. Van der Vleuten (2009). « Script concordance testing : More cases or more questions ? », *Advances in Health Sciences Education. Theory and Practice*, *14*(3), p. 367-375, doi : 10.1007/s10459-008-9120-8.

Goel, V. et R.J. Dolan (2003). « Reciprocal neural response within lateral and ventral medial prefrontal cortex during hot and cold reasoning », *Neuroimage*, *20*(4), p. 2314-2321.

Goel, V. et R.J. Dolan (2004). « Differential involvement of left prefrontal cortex in inductive and deductive reasoning », *Cognition*, *93*(3), p. B109-B121, doi : 10.1016/j.cognition.2004.03.001.

Goel, V., B. Gold, S. Kapur et S. Houle (1998). « Neuroanatomical correlates of human reasoning », *Journal of Cognitive Neuroscience*, *10*(3), p. 293-302.

Goos, M., F. Schubach, G. Seifert et M. Boeker, M. (2016). « Validation of undergraduate medical student script concordance test (SCT) scores on the clinical assessment of the acute abdomen », *BMC Surgery*, 16(1), p. 57, doi : 10.1186/s12893-016-0173-y.

Govaerts, M.J., L.W. Schuwirth, C.P. Van der Vleuten et A.M. Muijtjens (2011). « Workplace-based assessment : Effects of rater expertise », *Advances in Health Sciences Education. Theory and Practice*, *16*(2), p. 151-165, doi : 10.1007/s10459-010-9250-7.

Granry, J.C.M. et M.C. Moll (2012). *État de l'art (national et international) en matière de pratiques de simulation dans le domaine de la santé*, <http://www.has-sante.fr/portail/>, consulté le 25 avril 2017.

Groves, M., M.L. Dick, G. McColl et J. Bilszta (2013). « Analysing clinical reasoning characteristics using a combined methods approach », *BMC Medical Education*, *13*(1), p. 144, doi : 10.1186/1472-6920-13-144.

Groves, M., I. Scott et H. Alexander (2002). « Assessing clinical reasoning : A method to monitor its development in a PBL curriculum », *Medical Teacher*, *24*(5), p. 507-515, doi : 10.1080/01421590220145743.

Hatala, R., D.A. Cook, R. Brydges et R. Hawkins (2015). « Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS) : A systematic review of validity evidence », *Advances in Health Sciences Education. Theory and Practice*, *20*(5), p. 1149-1175, doi : 10.1007/s10459-015-9593-1.

Hawkins, R.E., M.J. Margolis, S.J. Durning et J.J. Norcini (2010). « Constructing a validity argument for the mini-Clinical Evaluation Exercise : A review of the research », *Academic Medicine*, *85*(9), p. 1453-1461, doi : 10.1097/ACM.0b013e3181eac3e6.

Hogarth, R.M. (2001). *Educating Intuition*, Chicago : University of Chicago Press.

Hruska, P., K.G. Hecker, S. Coderre, K. McLaughlin, F. Cortese, C. Doig, T. Beran, B. Wright et O. Krigolson (2015). « Hemispheric activation differences in novice and expert clinicians during clinical decision making », *Advances in Health Sciences Education. Theory and Practice*, doi : 10.1007/s10459-015-9648-3.

Ilgen, J.S., A.J. Humbert, G. Kuhn, M.L. Hansen, G.R. Norman, K.W. Eva et J. Sherbino (2012). « Assessing diagnostic reasoning : A consensus statement summarizing theory, practice, and future needs », *Academic Emergency Medicine*, 19(12), p. 1454-1461, doi : 10.1111/acem.12034.

Kane, M. (2006). « Validation », dans R.L. Brennan (dir.), *Educational Measurement*, 4e éd., Westport : American Council on Education and Praeger, p. 17-64.

Kane, M.T. (2013). « Validating the interpretations and uses of test scores », *Journal of Educational Measurement*, *50*(1), p. 1-73.

Kelly, W., S. Durning et G. Denton (2012). « Comparing a script concordance examination to a multiple-choice examination on a core internal medicine clerkship », *Teaching and Learning in Medicine*, *24*(3), p. 187-193, doi : 10.1080/10401334.2012.692239.

Kuiper, R., C. Heinrich, A. Matthias, M.J. Graham et L. Bell-Kotwall (2008). « Debriefing with the OPT model of clinical reasoning during high fidelity patient simulation », *The International Journal of Nursing Education Scholarship*, *5*, article 17, doi : 10.2202/1548-923X.1466.

Kuper, A., S. Reeves, M. Albert et B.D. Hodges (2007). « Assessment : Do we need to broaden our methodological horizons ? », *Medical Education*, *41*(12), p. 1121-1123, doi : 10.1111/j.1365-2923.2007.02945.x.

Larsen, D.P., A.C. Butler et H.L. Roediger. (2008). « Test-enhanced learning in medical education », *Medical Education*, *42*(10), p. 959-966, doi : 10.1111/j.1365-2923.2008.03124.x.

Laurier, M., R. Tousignant et D. Morissette (2005). *Les principes de la mesure et de l'évaluation des apprentissages*, 3e éd., Montréal : Gaëtan Morin, éditeur/Chenelière Éducation.

Laveault, D.G. (2002). *Introduction aux théories de tests en psychologie et en sciences de l'éducation*, 2e éd., Bruxelles : De Boeck Université.

Lee, D., M.F. Rushworth, M.E. Walton, M. Watanabe et M. Sakagami (2007). « Functional specialization of the primate frontal cortex during decision making », *Journal of Neuroscience*, *27*(31), p. 8170-8173, doi : 10.1523/JNEUROSCI.1561-07.2007.

Lineberry, M., C.D. Kreiter et G. Bordage (2013). « Threats to validity in the use and interpretation of script concordance test scores », *Medical Education*, *47*(12), p. 1175-1183, doi : 10.1111/medu.12283.

Lineberry, M., C.D. Kreiter et G. Bordage (2014). « Script concordance tests : Strong inferences about examinees require stronger evidence », *Medical Education*, *48*(4), p. 452-453, doi : 10.1111/medu.12417.

Liu, S., H.M. Chow, Y. Xu, M.G. Erkkinen, K.E. Swett, M.W. Eagle et A.R. Braun (2012). « Neural correlates of lyrical improvisation : An FMRI study of freestyle rap », *Sci Rep*, *2*, p. 834, doi : 10.1038/srep00834.

Lubarsky, S., B. Charlin, D.A. Cook, C. Chalk et C.P. Van der Vleuten (2011). « Script concordance testing : A review of published validity evidence », *Medical Education*, *45*(4), p. 329-338, doi : 10.1111/j.1365-2923.2010.03863.x.

Lusk, J.M. et K. Fater (2013). « Postsimulation debriefing to maximize clinical judgment development », *Nurse Educator*, *38*(1), p. 16-19, doi : 10.1097/NNE.0b013e318276df8b.

Messick, S. (1989). « Validity », dans R. Linn (dir.), *Educational measurement*, New York : American Council on Education et Macmillan Publishing, p. 13-104.

Monteiro, S.D., J.D. Sherbino, J.S. Ilgen, K.L. Dore, T.J. Wood, M.E. Young et E. Howey (2015). « Disrupting diagnostic reasoning : Do interruptions, instructions, and experience affect the diagnostic accuracy and response time of residents and emergency physicians ? », *Academic Medicine*, *90*(4), p. 511-517, doi : 10.1097/ACM.0000000000000614.

Muijtjens, A.M., H.V. Mameren, R.J. Hoogenboom, J.L. Evers et C.P. Van der Vleuten (1999). « The effect of a "don't know" option on test scores : Number-right and formula scoring compared », *Medical Education*, *33*(4), p. 267-275.

Nguyen, D.-Q. et J.-G. Blais (2007). « Approche par objectifs ou approche par compétences ? Repères conceptuels et implications pour les activités d'enseignement, d'apprentissage et d'évaluation au cours de la formation clinique », *Pédagogie médicale*, *8*(4), p. 232-251.

Norman, G., M. Young et L. Brooks (2007). « Non-analytical models of clinical reasoning : The role of experience », *Medical Education*, *41*(12), p. 1140-1145, doi : 10.1111/j.1365-2923.2007.02914.x.

Nouh, T., M. Boutros, R. Gagnon, S. Reid, K. Leslie, D. Pace et S.H. Meterissian (2012). « The script concordance test as a measure of clinical reasoning : A national validation study », *The American Journal of Surgery*, *203*(4), p. 530-534, doi : 10.1016/j.amjsurg.2011.11.006.

Page, G. et G. Bordage (1995). « The Medical Council of Canada's key features project : A more valid written examination of clinical decision-making skills », *Academic Medicine*, *70*(2), p. 104-110.

Patel, V.L., G.J. Groen et G.R. Norman (1991). « Effects of conventional and problem-based medical curricula on problem solving », *Academic Medicine*, *66*(7), p. 380-389.

Pelaccia, T., J. Tardif, E. Triby et B. Charlin (2011). « An analysis of clinical reasoning through a recent and comprehensive approach : The dual-process theory », *Medical Education Online*, *16*, doi : 10.3402/meo.v16i0.5890.

Pennaforte, T., A. Moussa, N. Loye, B. Charlin et M.C. Audetat (2016). « Exploring a new simulation approach to improve clinical reasoning teaching and assessment : Randomized trial protocol », *JMIR Research Protocols*, *5*(1), p. e26, doi : 10.2196/resprot.4938.

Piovezan, R.D., O. Custodio, M.S. Cendoroglo, N.A. Batista, S. Lubarsky et Charlin, B. (2012). « Assessment of undergraduate clinical reasoning in geriatric medicine : Application of a script concordance test », *Journal of the American Geriatrics Society*, *60*(10), p. 1946-1950, doi : 10.1111/j.1532-5415.2012.04152.x.

Pottier, P., J.B. Hardouin, B.D. Hodges, M.A. Pistorius, J. Connault, C. Durant, R. Clairand, V. Sebille, J.H. Barrier et B. Planchon (2010). « Exploring how students think : A new method combining think-aloud and concept mapping protocols », *Medical Education*, *44*(9), p. 926-935, doi : 10.1111/j.1365-2923.2010.03748.x.

Raemer, D., M. Anderson, A. Cheng, R. Fanning, V. Nadkarni et G. Savoldelli (2011). « Research regarding debriefing as part of the learning process », *Simulation in Healthcare*, *6*, suppl., p. S52-S57, doi : 10.1097/SIH.0b013e31822724d0.

Ricketts, C. (2009). « A plea for the proper use of criterion-referenced tests in medical assessment », *Medical Education*, *43*(12), p. 1141-1146, doi : 10.1111/j.1365-2923.2009.03541.x.

Roberts, D. (2013). « The clinical viva : An assessment of clinical thinking », *Nurse Educator Today*, *33*(4), p. 402-406, doi : 10.1016/j.nedt.2013.01.014.

Schmidt, H.G. et R.M. Rikers (2007). « How expertise develops in medicine : Knowledge encapsulation and illness script formation », *Medical Education*, *41*(12), p. 1133-1139, doi : 10.1111/j.1365-2923.2007.02915.x.

Schuwirth, L.W. et C.P. Van der Vleuten (2012). « Programmatic assessment and Kane's validity perspective », *Medical Education*, *46*(1), p. 38-48, doi : 10.1111/j.1365-2923.2011.04098.x.

Sibert, L., B. Charlin, J. Corcos, R. Gagnon, P. Grise et C. Van der Vleuten (2002). « Stability of clinical reasoning assessment results with the Script Concordance Test across two different linguistic, cultural and learning environments », *Medical Teacher*, *24*(5), p. 522-527, doi : 10.1080/0142159021000012599.

Swanson, D.B., L.J. Grosso et J. Norcini (1987). « Assessment of clinical competence : Written and computer-based simulations », *Assessment & Evaluation in Higher Education*, *12*, p. 220-246.

Tardif, J. (2006). *L'évaluation des compétences : documenter le parcours de développement*, Montréal : Chenelière Éducation.

Van der Vleuten, C.P., L.W. Schuwirth, F. Scheele, E.W. Driessen et B. Hodges (2010). « The assessment of professional competence : Building blocks for theory development », *Best Practice & Research Clinical Gastroenterology*, *24*(6), p. 703-719, doi : 10.1016/j.bpobgyn.2010.04.001.

Wason, P.C., Evans, J. St. B.T. (1975). « Dual processes in reasoning ? », *Cognition*, *3*, p. 141-154.

CHAPITRE 6

L'utilisation de la formation par concordance comme modalité d'évaluation formative pour entraîner à la prise de décision opératoire

Isabelle Raîche et Bernard Charlin

Les erreurs liées à la prise de décision chirurgicale sont responsables de la vaste majorité des complications survenant durant l'exécution de procédures (Way et al.*, 2003). Cette prise de décision par les chirurgiens experts se fait suivant les principes décrits dans la prise de décision naturaliste. Cette prise de décision se base sur des scripts développés avec la formation, l'expérience et la réflexion et est caractéristique de l'expertise chirurgicale. Actuellement, il est difficile d'enseigner la prise de décision dans le milieu clinique. Dans ce contexte, il est pertinent de rechercher une solution de rechange à l'environnement clinique classique pour enseigner la prise de décision. Par ailleurs, la littérature sur l'évaluation formative peut éclairer la création d'une telle modalité d'enseignement pour les résidents de chirurgie. L'enseignement par concordance, au cours duquel un apprenant peut comparer ses décisions avec celles d'un groupe d'experts et obtenir une rétroaction immédiate, s'inscrit bien dans une perspective d'évaluation formative permettant de développer le jugement clinique dans un contexte de chirurgie. Ce chapitre présente les grandes lignes d'un projet pilote utilisant de courtes vidéos de procédures chirurgicales comme vignettes associées à des problèmes décisionnels authentiques. Ces vignettes ont été créées selon les principes de conception des tests de concordances des scripts. Des chirurgiens experts se sont*

ensuite prononcés sur chaque question. Les réponses des experts ont été intégrées pour former un dispositif d'évaluation formative. Cette modalité d'évaluation formative a été soumise à six résidents et à deux chirurgiens enseignants d'un programme de chirurgie générale et leur impression de ce mode d'évaluation formative a été sondée.

Ce chapitre vise à illustrer comment un outil d'évaluation formative peut être créé en combinant les principes généraux des formations par concordance de script et de l'évaluation formative. Il s'inscrit dans ce collectif en décrivant une modalité d'évaluation formative de compétences complexes.

La première partie du chapitre présente les données de la littérature sur l'utilisation des formations par concordance de script et les principes guidant la création d'instruments permettant une évaluation formative. Ce chapitre se concentre sur la construction d'une formation par concordance comme modalité d'évaluation formative. D'autres modalités auraient pu être choisies, mais les formations par concordance nous ont semblé être un outil intéressant pour répondre aux besoins de nos apprenants et nous avons voulu étudier une adaptation de ce format d'instrument. Nous présenterons donc un exemple issu du contexte clinique de la chirurgie générale pour démontrer le potentiel d'une telle modalité d'enseignement. Cet exemple prend la forme d'un projet pilote mené auprès de chirurgiens enseignants et de résidents d'un programme de chirurgie générale.

1. LA PROBLÉMATIQUE ET UNE REVUE DE LA LITTÉRATURE

La première partie du chapitre présente différents concepts justifiant l'utilisation des formations par concordance de script dans le contexte de la prise de décision opératoire. On explique d'abord le rôle critique de la prise de décision dans l'expertise chirurgicale et la nécessité d'instaurer des modes d'enseignement plus formels que ceux disponibles actuellement. Ensuite, une définition de l'évaluation formative et des principes qui la régissent est donnée. Finalement, un bref résumé des publications portant sur l'utilisation des tests de concordance de script et des formations par concordance dans le domaine médical termine cette section.

1.1. La problématique : le rôle de la prise de décision dans l'expertise chirurgicale et les limites du mode d'enseignement actuel

Selon Way *et al.* (2003), les erreurs liées à la prise de décision sont responsables de plus de 97 % des complications chirurgicales survenant lors de l'ablation de la vésicule biliaire, ce qui montre l'importance de la prise de décision comme caractéristique essentielle de l'expertise chirurgicale. Ces dernières années, la prise de décision au cours de procédures chirurgicales a été étudiée par plusieurs auteurs (DaRosa *et al.*, 2008 ; Flin, Youngson et Yule, 2007 ; Moulton *et al.*, 2007 ; Way *et al.*, 2003 ; Yule et Paterson-Brown, 2012). Dans ce chapitre, le terme « procédure » est utilisé pour décrire une intervention chirurgicale, un geste technique pratiqué en salle d'opération ayant pour but de résoudre un problème clinique. D'après les travaux de ces nombreux auteurs, les chirurgiens utilisent en général un processus de décision appelé *naturaliste* par opposition à la prise de décision *analytique* (Ross, Shafer et Klein, 2006). La prise de décision analytique implique la création d'une série d'hypothèses qui sont ensuite comparées entre elles dans le but de sélectionner la meilleure option possible pour résoudre un problème. Ce type de prise de décision est très exigeant d'un point de vue cognitif et requiert généralement plus de temps que la prise de décision naturaliste. La prise de décision analytique est le type de prise de décision utilisé lorsqu'un clinicien fait face à un patient qui présente une série de symptômes peu communs et non urgents. Le but est alors d'éliminer progressivement et méticuleusement un ensemble de diagnostics en émettant des hypothèses et en les vérifiant une à une. La prise de décision naturaliste, quant à elle, se base sur la reconnaissance d'un scénario et l'application rapide d'une solution adéquate pour améliorer une situation. La réanimation des patients en traumatologie est un bon exemple de décision naturaliste. Le but ici n'est pas de trouver la meilleure réponse possible à un problème, mais de prendre une décision rapide, pertinente et applicable dans un contexte donné. Ce mode de prise de décision requiert généralement moins de ressources cognitives de la part du décideur et permet d'intégrer plus aisément une multitude d'informations (Moulton *et al.*, 2007 ; Yule et Paterson-Brown, 2012). La suite de cette section portera principalement sur la prise de décision naturaliste puisque ce type de prise décision est reconnu comme étant le plus répandu chez les chirurgiens experts.

La prise de décision naturaliste peut être liée à la théorie des scripts (Ross *et al.*, 2006). En effet, les chirurgiens utilisent un éventail varié et complexe de scripts qui leur permettent d'évaluer une situation rapidement en utilisant relativement peu de ressources cognitives et de faire des choix de façon rapide et efficace (Flin *et al.*, 2007 ;

Moulton *et al.*, 2007 ; Sweller, 2003 ; Yule et Paterson-Brown, 2012). En effet, selon cette théorie, avec l'expérience, les experts génèrent des scripts, c'est-à-dire des modèles mentaux, qui leur permettent d'organiser l'information présente dans une certaine situation (Charlin *et al.*, 2007). Les scripts permettent, entre autres, de reconnaître rapidement une certaine présentation clinique comme normale ou anormale et de réagir aux différentes variations de façon appropriée. Les scripts sont le résultat d'une accumulation d'expériences et de connaissances et leur fonction principale est de donner un sens à une situation donnée (Charlin, Tardif et Boshuizen, 2000). Par exemple, ils permettent d'interpréter une grande variété d'informations et d'amalgamer l'ensemble des éléments présents pour en faire un diagnostic. Les scripts permettent aussi aux cliniciens experts de reconnaître rapidement des éléments discordants dans une présentation clinique et d'agir en conséquence. Ils représentent la base de l'expertise clinique.

Malgré les multiples études portant sur le rôle essentiel de la prise de décision dans l'exécution de procédures (Hall, Ellis et Hamdorf, 2003), la prise de décision chirurgicale demeure difficile à enseigner et à évaluer. Traditionnellement, la salle d'opération a été privilégiée comme site d'enseignement et d'évaluation des résidents de spécialités chirurgicales en ce qui concerne leur habileté à la prise de décisions (Hauge, Wanzek et Godellas, 2001). Cependant, les capacités du contexte clinique comme milieu d'enseignement et d'évaluation sont limitées pour plusieurs raisons. D'abord, au cours des procédures, les chirurgiens enseignants font face à de multiples priorités, notamment l'importance d'offrir au patient une procédure optimale et sécuritaire, l'utilisation maximale des ressources hospitalières, du reste souvent limitées, et la présence de plusieurs apprenants au cours d'une même procédure (Moulton *et al.*, 2010 ; Scallon *et al.*, 1992). Dans cet environnement, il est souvent difficile pour les chirurgiens enseignants d'offrir aux apprenants assez de liberté pour pouvoir évaluer de façon directe leur prise de décision (Moulton *et al.*, 2010). De plus, il est difficile de standardiser les procédures dans le contexte clinique pour offrir des conditions d'évaluation sommative ou formative adaptées pour un grand nombre d'apprenants. Enfin, la prise de décision est une activité cognitive souvent difficile à verbaliser, particulièrement pour des experts chez qui les processus de prise de décision sont souvent automatisés (Crandall, Klein et Hoffman, 2006 ; Ross *et al.*, 2006). Par conséquent, il est assez difficile pour des apprenants novices de saisir les subtilités de la prise de décision intra-opératoire durant l'observation de chirurgies (Raîche, 2016). En effet, les novices ont généralement besoin d'instructions pour guider leur attention et leur permettre

d'assimiler les processus cognitifs des experts (Rosen *et al.*, 2010). Combinés, ces facteurs limitent la qualité de l'enseignement offert par les chirurgiens lors de l'exécution de procédures chirurgicales.

En gardant à l'esprit les limites du modèle actuel, il est pertinent de rechercher des solutions de rechange à l'enseignement de la prise de décision chirurgicale en milieu clinique. Traditionnellement, on encourage les résidents à réviser les techniques chirurgicales dans des livres de référence avant les procédures, de manière à se préparer à la performance opératoire. Cependant, une étude menée en urologie a démontré que la lecture de livres de référence techniques n'améliore pas la prise de décision intra-opératoire. Dans cette étude, la lecture comme préparation permet tout au plus aux apprenants novices de mieux reconnaître les étapes d'une chirurgie (Samuelson, Cadeddu et Matsumoto, 2006). D'autres travaux ont exploré l'utilisation de différentes approches à l'extérieur du contexte clinique pour enseigner la prise de décision chirurgicale (DaRosa *et al.*, 2008 ; Guerlain *et al.*, 2004). Ces deux études utilisaient des modalités d'enseignement incluant des rencontres de groupes et des visionnements d'extraits vidéo qui nécessitaient la participation intensive d'experts et semblaient difficiles à implanter à grande échelle en raison du manque de disponibilité des chirurgiens experts pour ce genre d'activités dans le contexte clinique actuel. Compte tenu de l'importance de la prise de décision intra-opératoire dans la compétence chirurgicale, il est pertinent de chercher une façon d'enseigner la prise de décision à l'aide d'une modalité pouvant être utilisée par divers apprenants et menant à une réflexion et à une amélioration des compétences de façon soutenue.

1.2. L'évaluation formative : quelques fondements théoriques

Il existe plusieurs définitions de l'évaluation formative. Ce concept a évolué au cours des dernières décennies. Initialement, l'évaluation formative a été définie comme l'information transmise à l'apprenant, à la suite d'une évaluation, ayant pour but de modifier sa façon de penser ou d'agir et entraînant un apprentissage (Shute, 2008). Cette définition, basée en partie sur les travaux de Bloom en 1968, laisse peu de place à l'apprenant. D'autres auteurs affirment que le rôle de l'apprenant dans l'évaluation formative devrait être davantage reconnu dans la définition que l'on utilise pour créer des modalités d'évaluation formative (Allal et Mottier Lopez, 2005). En effet, ces auteurs décrivent l'évaluation formative davantage comme un dialogue entre l'enseignant et l'apprenant. Au cours d'une série d'interactions, il s'établit un contrat éducationnel où l'apprenant est également chargé de s'autoévaluer et de rechercher des pistes de solutions (Allal et Mottier Lopez, 2005). Selon

ces définitions, on peut penser qu'une approche visant à enrichir l'évaluation formative puisse être utilisée pour améliorer l'enseignement de la prise de décision chirurgicale.

Parmi les éléments clés de l'évaluation formative, on note l'interaction entre l'apprenant et l'enseignant, l'idée de favoriser un changement, la nécessité de fournir de l'information claire et objective, l'importance d'une approche permettant de s'adapter aux besoins individuels des apprenants, et la nécessité d'une évaluation relativement fréquente (Rudolph *et al.*, 2008). La plupart des auteurs suggèrent des modes d'évaluation où les apprenants doivent utiliser leurs connaissances de façon à promouvoir des apprentissages transférables au contexte pratique (Harlen et James, 1997). L'évaluation formative peut prendre plusieurs formes. De manière générale, trois étapes sont reconnues comme essentielles lors d'une évaluation formative : évaluer la performance, offrir de la rétroaction sur la performance et produire un changement chez l'apprenant en utilisant cette rétroaction. Sargeant *et al*. (2009) présentent certaines caractéristiques qui rendent l'évaluation formative plus ou moins susceptible d'être utilisée par l'apprenant. Les commentaires divergents par rapport à l'autoévaluation de l'apprenant demandent en général plus de réflexion avant d'être acceptés ; la crédibilité de la source de rétroaction est importante, de même que l'uniformité des commentaires parmi différentes sources de rétroaction. Enfin, pour introduire des changements à la suite de commentaires, l'apprenant doit sentir qu'il a les moyens de s'améliorer. Harlen et James (1997) suggèrent aussi des critères garantissant que l'évaluation formative mènera à une réflexion et à un changement dans les modèles mentaux des apprenants. Selon eux, une évaluation formative de qualité devrait s'aligner sur les connaissances et aptitudes déjà acquises par les apprenants et leurs connaissances antérieures, établir des liens avec les expériences vécues par les apprenants, être pertinente à leur stade d'apprentissage et être perçue par les apprenants comme utile et riche en potentiel (Harlen et James, 1997). Les travaux de Sargeant *et al*. (2009) utilisent l'intensité de la réflexion suscitée par la rétroaction comme indicateur de l'efficacité de l'évaluation formative. Ces auteurs définissent ce processus de réflexion comme une activité intellectuelle et affective au cours de laquelle un individu explore ses expériences passées pour en dégager une signification et une application concrète. Pour eux, il est difficile de mesurer de façon précise les effets de l'évaluation formative, mais puisque la réflexion est essentielle au changement, on peut penser qu'une évaluation suscitant beaucoup de réflexion est plus susceptible de porter ses fruits (Sargeant *et al.*, 2009).

Parmi les différentes modalités permettant d'offrir une évaluation formative, la formation par concordance a été reconnue par certains comme menant à une réflexion plus élaborée que les modalités classiques, comme les tests à choix multiples (Cobb *et al.*, 2015). La formation par concordance consiste à placer l'apprenant dans une situation d'incertitude décrite à l'aide d'une vignette clinique authentique. L'incertitude proposée dans le scénario est planifiée et dosée par les concepteurs du test pour permettre d'évaluer la prise de décision du participant. Cette incertitude reflète un cas clinique authentique tel qu'il se présenterait normalement pour le chirurgien. En effet, ce dernier doit souvent prendre des décisions en n'ayant pas toutes les informations qu'il souhaiterait avoir afin de prendre une décision parfaitement éclairée. En somme, les vignettes sont conçues pour comparer l'importance accordée à certains éléments d'une présentation clinique entre un apprenant et un groupe d'experts. Les vignettes peuvent inclure, par exemple, une description de cas, des images ou encore des vidéos. Des hypothèses sont ensuite présentées concernant les étapes suivantes de la prise en charge. Un élément supplémentaire est enfin ajouté à la vignette et on demande à l'apprenant de décrire l'effet de cet élément sur la prise en charge (Charlin, 2014 ; Charlin et Fernandez, 2016).

Les tests de concordance de script (TCS), qui représentent une modalité d'évaluation par concordance, ont été utilisés comme instruments de mesure du raisonnement clinique dans de nombreux domaines, y compris dans certaines spécialités chirurgicales (Lubrarsky *et al.*, 2011 ; Meterissian *et al.*, 2007 ; Park *et al.*, 2010). Cependant, les propriétés psychométriques des TCS ont été remises en question par plusieurs auteurs dont Dionne, Grondin et Latreille, au chapitre 3 du présent ouvrage. En général, on reproche un manque de profondeur aux preuves de validité amassées jusqu'ici pour justifier l'utilisation de ces dispositifs comme modalité d'évaluation sommative (Lineberry, Kreiter et Bordage, 2013). Ce projet ne permet pas de juger des propriétés psychométriques de ce test puisque nous utilisons une adaptation des TCS. Par ailleurs, notre projet ayant une visée formative, nous croyons que les principes ayant mené à la création de TCS permettront de créer un dispositif valable pour les participants.

Les tests de concordance de script ont déjà été utilisés dans des contextes d'évaluation formative. En France, on a utilisé des tests de concordance de script dans le domaine de la réanimation pour évaluer des étudiants de médecine et des internes (Gibot et Bollaert, 2008). L'étude de ces auteurs a montré que la formation améliore la performance au test de concordance de script. Les auteurs notent cependant que les étudiants ont été un peu déstabilisés par le format des tests

de concordance et l'absence de réponse tranchée. Gibot et Bollaert soulignent que les étudiants auraient avantage à suivre plus de formations de ce genre, dans la mesure où elles sont adéquates, pour les rendre plus à l'aise avec le format des formations par concordance. Au Mexique, on a eu recours à la formation par concordance comme mode de formation médicale continue chez 1901 médecins spécialistes (Hornos *et al.*, 2013). La formation par concordance a été bien acceptée chez les participants : 70 % ont effectué les 240 vignettes au cours de l'année de formation et 96 % des participants sondés recommanderaient ce mode d'enseignement à un collègue. Rappelons que cette formation était offerte de façon volontaire comme ressource de formation professionnelle continue à un groupe de cliniciens en pratique clinique active.

2. LA CRÉATION D'UN DISPOSITIF D'ÉVALUATION FORMATIVE

Compte tenu des avantages potentiels de la formation par concordance en ce qui concerne la facilité d'administration, la promotion de la réflexion et l'acceptabilité par les participants, on peut penser que cette modalité pourrait servir de cadre pour la création d'une évaluation formative de la prise de décision chirurgicale. Nous présentons ici un projet pilote mené dans un programme de résidence de chirurgie générale. Il s'agit avant tout d'évaluer la faisabilité d'une telle initiative et de relever les impressions d'utilisateurs potentiels par rapport au dispositif créé qui permet de rendre compte d'apprentissages complexes et de compétences. Le but de ce projet était de concevoir un dispositif d'évaluation formative permettant aux résidents de spécialités chirurgicales d'obtenir de la rétroaction sur la justesse de leurs scripts au moment de la prise de décision en cours de procédure. Il s'agit d'une nouvelle approche puisque jusqu'ici, les concordances de script ont été utilisées dans les disciplines chirurgicales davantage dans un contexte de diagnostic clinique, avec des scénarios décrits sous forme de texte, et ce, dans un contexte d'évaluation sommative.

Lors de la création de l'instrument, nous avons choisi d'utiliser le modèle théorique de Downing (2006). Ce modèle décrit 12 étapes nécessaires au développement de tests : planification générale, définition des contenus, création d'un tableau de spécification, création d'items, assemblage du test, production du test, administration du test, établissement des notes, détermination des notes de passage, distribution des résultats, création d'une banque d'items et production d'un rapport technique. Dans le cadre de ce projet, nous avons choisi les étapes qui nous semblaient les plus pertinentes pour commencer la

construction de notre dispositif d'évaluation, sachant qu'il s'agissait de la première phase d'un processus itératif. L'annexe A présente les étapes complétées jusqu'à présent. Au cours des prochaines itérations de ce projet, d'autres étapes seront franchies. Le projet pilote présenté dans ce chapitre contribuera à éclairer les prochaines étapes de l'élaboration.

2.1. La planification générale

2.1.1. *Les orientations générales*

Avant d'amorcer la construction du dispositif d'évaluation formative, nous avons défini les orientations générales de ce projet. Nous avons choisi de créer un dispositif d'évaluation à visée formative plutôt que sommative, le but premier de ce travail étant de tenter d'améliorer l'enseignement aux résidents. Il a également été décidé de concentrer nos efforts sur l'élaboration d'un dispositif d'évaluation pour la prise de décision intra-opératoire en raison de l'absence de dispositifs similaires et des limites de l'enseignement actuel décrites dans les sections précédentes.

La revue de littérature présentée dans les sections précédentes et l'expérience des auteurs ont influencé notre décision de baser le dispositif d'évaluation formative sur les principes des tests de concordance de script. Ce dernier repose sur l'idée que les différents jugements des cliniciens peuvent être comparés à l'aide de questions les forçant à prendre des décisions. Un item dans un test de concordance doit inclure un cas fondé sur une situation clinique réelle. On suggère ensuite une conduite possible et on ajoute un élément supplémentaire à la présentation clinique. On demande au participant de choisir si l'ajout de ce nouvel élément clinique rend la conduite initialement suggérée plus ou moins pertinente en qualifiant leur réponse sur une échelle de Likert (Fournier, Demeester et Charlin, 2008). Dans ce projet, nous tentons effectivement de comparer le jugement d'experts et des participants en les obligeant à prendre une décision. Un cas basé sur une situation réelle est soumis aux participants sous la forme d'un extrait vidéo. On leur demande ensuite de choisir une action à entreprendre après leur avoir expliqué le but de cette étape de la procédure. Au lieu de suggérer une conduite, par exemple « Vous pensiez faire une biopsie », on dit : « Vous cherchez à exposer le pédicule vasculaire », puis on demande de choisir une action. Il n'y a pas d'échelle de Likert pour l'instant puisque ce format semblait initialement difficile à inclure avec les vidéos. On pourrait cependant en ajouter au cours de la deuxième itération du projet. Les items ont été conçus en tentant d'inclure des situations d'incertitude relative reliées

à la prise de décision telles qu'elles sont vécues en salle d'opération. Par exemple, nous avons pris soin de vérifier que l'incertitude n'était pas due à une mauvaise formulation. Nous avons choisi d'inclure des clips où plusieurs conduites pourraient être adéquates, où une décision devait être prise avec des informations limitées et où des indices visuels guidant les décisions étaient présents comme c'est souvent le cas lors de véritables chirurgies. Sachant que nous présentons une adaptation des tests de concordance de script, nous croyons en avoir respecté les principes essentiels.

2.1.2. *Le choix d'une procédure*

Après avoir défini les grandes lignes du projet, nous avons dû choisir une procédure comme pierre angulaire de ce projet. Il nous a semblé qu'il serait plus simple de débuter par une seule procédure plutôt que de ratisser large et de tenter de couvrir des concepts généraux. L'expertise est reconnue comme « locale », en d'autres termes, difficilement transférable d'une situation à une autre, et nous avons pensé qu'il serait plus réaliste de tenter d'améliorer la prise de décision pour une seule procédure et que cette approche nous permettrait d'explorer plus en profondeur les éléments clés de la prise de décision pour cette intervention (Ericsson, 2006).

La chirurgie laparoscopique est de plus en plus utilisée en chirurgie générale (Musselman *et al.*, 2012). Cette approche est associée à moins de douleur postopératoire, une récupération plus rapide des patients et un risque moindre de hernie incisionnelle (Ricca et Lacaine, 2009). Cette approche chirurgicale utilise une caméra et des instruments de petit calibre pour exécuter des procédures chirurgicales plus ou moins complexes. Les chirurgiens doivent baser leurs décisions intra-opératoires majoritairement sur les informations visibles sur l'écran projetant l'image de la procédure en cours. Les sensations tactiles sont d'importance moindre que durant les procédures en chirurgie ouverte. De plus, comme les procédures nécessitent l'utilisation d'une caméra, il est possible de les enregistrer et d'utiliser les enregistrements dans un but de formation (Lee, Seo et Hong, 2015). Les extraits vidéo basés sur des procédures sont de plus en plus utilisés par les résidents devant exécuter des procédures techniques (Koya *et al.*, 2012). À ce stade, la plupart des vidéos produites dans le cadre de l'enseignement chirurgical présentent en quelques minutes les étapes clés, exécutées par des experts, d'une procédure de plusieurs heures. Pour les apprenants, ces vidéos permettent de réviser ces étapes et de créer une image mentale de ce à quoi la chirurgie devrait ressembler. Cependant, ces vidéos sont peu utiles pour expliquer

comment procéder afin d'exposer les tissus et améliorer le déroulement des procédures, et on met rarement l'accent sur la gestion des complications ou leur prévention (Koya *et al.*, 2012).

La colectomie droite laparoscopique a été sélectionnée comme procédure cible pour notre projet. Il s'agit d'une opération relativement commune en chirurgie générale. Elle est pratiquée autant par les chirurgiens généraux en dehors des grands centres que par les chirurgiens surspécialisés en milieu universitaire. Il s'agit d'une procédure complexe, requérant de multiples décisions qui doivent prendre en compte, entre autres, l'état du patient, la progression de la procédure et la réponse des tissus. Cette opération a été choisie, car les résidents s'attendent à être formés de façon adéquate pour mener à bien l'une de ces procédures de façon autonome, mais le volume de ces procédures en milieu universitaire est tel qu'il est possible que les résidents ne parviennent pas au niveau de compétence souhaité avec l'entraînement chirurgical offert à l'heure actuelle.

En appliquant les principes de l'évaluation formative au visionnement de vidéos chirurgicales, on comprend le potentiel des formations par concordance. En effet, ces formations permettent aux apprenants d'interagir avec des experts à leur rythme et selon leurs besoins ; en outre, leur format électronique et portable permet une utilisation fréquente de la part de l'apprenant. Dans le cadre de ce projet pilote, un instrument d'évaluation formative a été créé selon ces principes.

2.2. La définition des contenus à inclure dans le dispositif d'évaluation formative

Suivant le modèle de Downing, nous avons commencé la construction de l'instrument par une démarche d'évaluation des contenus. Puisque l'évaluation formative doit s'appuyer sur les besoins des apprenants, il importe de commencer le processus de conception d'un dispositif d'évaluation formative par l'identification des besoins des apprenants. Pour ce faire, plusieurs modalités existent. En général, le processus d'identification des besoins combine des entrevues avec des experts et des apprenants, une revue de la littérature et des guides de pratique et des observations cliniques (Hornos *et al.*, 2013 ; Madani *et al.*, 2016).

Au moment d'établir les contenus à inclure dans le dispositif d'évaluation, des entrevues ont été menées auprès de deux chirurgiens enseignants ayant plusieurs années d'expérience. Ces entrevues ont été menées selon les principes de l'analyse de tâche cognitive (Crandall *et al.*, 2006). Le but des analyses de tâches cognitives est d'amener les experts à expliquer les processus mentaux qui guident leur prise de

décision, les éléments utilisés dans chaque situation pour guider la prise de décision et les règles, souvent développées avec l'expérience et la réflexion, qui soutiennent leur jugement. En fin de compte, les analyses de tâches cognitives visent à transmettre l'expertise cognitive d'experts à des novices en amenant un ou plusieurs experts à expliciter leurs processus cognitifs. Les analyses de tâches cognitives ont été utilisées avec succès pour plusieurs procédures chirurgicales (Raîche, 2016 ; Smink *et al*., 2012 ; Sullivan *et al*., 2008). De plus, parmi les études portant sur le développement de programmes visant à enseigner la prise de décision chirurgicale, l'analyse de tâche cognitive a parfois servi de base pour déterminer les contenus à enseigner (DaRosa *et al*., 2008). Dans notre exemple, trois entrevues semi-dirigées d'environ une heure chacune ont été réalisées avec chaque expert. Les enregistrements ont été analysés à l'aide d'une grille prédéterminée. Cette grille a été créée lors de travaux antérieurs selon des principes décrits précédemment (Raîche, 2016). Un exemple est présenté en annexe (annexe B). Les données recueillies ont été regroupées d'abord selon les principales étapes de l'intervention chirurgicale, comme la prise de décision durant le contrôle du pédicule vasculaire, la dissection latérale et l'exploration de l'arrière-cavité des épiploons. Pour chaque étape, les réponses des experts ont été divisées en unité de sens. Une unité de sens est définie comme une partie de matériel recueilli ayant un sens propre et spécifique et pouvant être utilisé lors d'un codage (Van der Maren, 1995). Ensuite, pour chaque étape, les unités de sens se rapportant à la prévention de complications, les règles personnelles, les sources d'erreurs communes et les indices visuels et repères anatomiques recherchés par chaque expert ont été relevés. Au cours de ce processus, les experts ont survolé un vaste contenu. De plus, une attention particulière a été portée à l'identification des aspects cognitifs de la procédure semblant les plus difficiles à assimiler pour les apprenants, encore une fois, en essayant de calquer le dispositif d'évaluation sur les besoins des apprenants. Les éléments les plus souvent mentionnés dans la section sur les erreurs communes incluent le rôle de l'instrument de la main non dominante, l'optimisation de la rétraction, la gestion des complications, l'identification dynamique du plan et l'ajustement de l'exposition pour maintenir le plan de dissection.

Ensuite, six apprenants ont également été sondés après avoir visionné un exemple de formation par concordance. On leur a demandé d'indiquer les aspects de la procédure où ils percevaient des lacunes qui pourraient être corrigées par une formation par concordance. Les résidents ont noté la détermination du plan de dissection approprié, le contrôle des saignements intra-opératoires et l'identification du pédicule vasculaire comme étant des éléments pouvant bénéficier

d'un entraînement particulier. Aucune analyse de tâche cognitive n'a été réalisée auprès des résidents. L'analyse de tâche cognitive est une démarche complexe permettant d'établir les processus cognitifs d'un participant dans un but de transmission de connaissances. Dans le cas présent, il aurait été intéressant de se livrer à cet exercice avec les résidents pour connaître avec certitude leurs lacunes personnelles, mais il nous a semblé qu'une approche plus simple d'autoévaluation serait suffisante. Il s'agissait surtout de s'assurer que l'analyse de tâche cognitive menée auprès des experts répertoriait l'ensemble des contenus perçus comme pertinents par les résidents.

Des instruments de mesure de performances chirurgicales ont été consultés pour s'assurer que les aspects reconnus comme étant des composantes de l'expertise chirurgicale étaient présents dans le dispositif d'évaluation formative. L'échelle produite par l'American Society of Colon and Rectal Surgeons pour mesurer la compétence opératoire dans la réalisation de colectomies laparoscopiques a été consultée, de même que des échelles plus générales comme GOALS (Vassiliou *et al.*, 2005) et OSATS (Martin *et al.*, 1997).

Enfin, huit enregistrements de colectomies droites laparoscopiques ont été examinés à plusieurs reprises par un chirurgien. Le but de ces visionnements était de repérer des segments durant lesquels il était possible d'observer un arrêt de progression des procédures ou un point tournant où une décision devait être prise.

À la suite de ce processus de définition des contenus à enseigner, plusieurs aspects requérant une prise de décision durant une colectomie laparoscopique ont été relevés. La reconnaissance du plan de dissection, la prévention et la gestion des complications, l'identification et la dissection des pédicules vasculaires, l'optimisation de l'exposition de même que la progression sécuritaire et efficace de la procédure ont été considérées comme les éléments clés de la prise de décision intra-opératoire pour cette procédure. Il est intéressant de noter que ces catégories sont relativement générales et qu'elles pourraient être applicables à plusieurs procédures.

2.3. L'élaboration d'un tableau de spécification

Downing (2006) décrit les éléments composant un tableau de spécification. Selon lui, le tableau doit inclure le type d'items utilisés pour le test (réponse construite ou réponse sélectionnée), les supports utilisés pour les items (vidéos, photo), le type d'objectifs (ou les compétences) devant être évalués par le dispositif ; par exemple se basant sur

la taxonomie de Bloom, la clé de correction, un plan d'interprétation d'éventuels scores, le nombre d'items pour chaque élément de contenu et le temps alloué pour effectuer l'épreuve.

Pour ce projet, nous avions déjà choisi d'utiliser les principes de la création des tests de concordance de script pour guider notre démarche d'élaboration des items et nous voulions utiliser des items à réponses sélectionnées pour faciliter l'utilisation du dispositif à plus grande échelle. Les modifications expliquées à la section 2.1.1 ont été intégrées pour mieux adapter le dispositif à notre contexte. Rappelons que nous avons choisi d'utiliser des clips vidéo comme support pour les questions de façon à inclure un élément d'évaluation des perceptions visuelles des participants. En utilisant les données recueillies durant l'étape de définition des contenus, nous avons pu déterminer les éléments à évaluer avec le dispositif. En nous basant sur la taxonomie de Bloom, les objectifs cognitifs d'*analyse*, de *synthèse* et d'*évaluation* nous ont semblé les plus représentatifs du but de ce projet : offrir une évaluation formative de la prise de décision intra-opératoire. En effet, nous souhaitions créer un dispositif qui nécessiterait de la part des étudiants d'analyser les clips vidéo pour en faire ressortir les repères indispensables à la prise de décision. Ensuite, nous voulions faire en sorte que les apprenants doivent mettre en relation les différentes options offertes dans chaque item avec les buts exprimés dans la question et les informations cliniques fournies par les clips vidéo dans une démarche d'évaluation. Enfin, nous souhaitions obliger les étudiants à synthétiser l'ensemble des informations pour les opérationnaliser dans le choix d'une réponse.

En ce qui concerne la clé de correction, nous avons choisi de la créer en interrogeant des experts et en regroupant leurs réponses sous forme de diagrammes circulaires (figure 6.3). Nous voulions une représentation visuelle de la variation des réponses entre les experts qui soit facile à consulter par les résidents. Les justifications des réponses ont été soumises aux participants à titre d'information supplémentaire. Les participants au projet n'ont pas soumis leurs réponses et n'ont donc pas reçu de score, le but de ce projet étant de leur fournir du matériel pour nourrir leur réflexion plus que de noter leur performance.

Le nombre d'items nécessaire pour chaque section, de même que la longueur optimale du test, tant pour le nombre d'items total que pour le temps requis, demeurent incertains et devront être clarifiés en fonction des réponses des participants au projet pilote.

2.4. La création d'items

Dans le cadre de ce projet, un item comprend un clip vidéo, une question et un choix de réponses. Une banque d'items a été conçue dans un contexte d'évaluation formative et en suivant les étapes suggérées pour la création des tests de concordance de script (Fournier *et al.*, 2008 ; Rudolph *et al.*, 2008). Comme la concordance de script repose sur des données cliniques authentiques, une bibliothèque de procédures a été créée représentant différentes approches chirurgicales chez différents patients. Ensuite, les vidéos de procédures ont été examinées à la recherche d'extraits illustrant les contenus à enseigner. Ces derniers ont ensuite été isolés, puis des marqueurs ont été ajoutés aux vidéos pour orienter les participants (figure 6.1). Les exemples présentés dans ce chapitre sont en anglais puisque les résidents du programme d'Ottawa sont anglophones.

Figure 6.1
Photo annotée pour orienter les apprenants

Une ou deux questions ont été produites pour chaque extrait. Un choix de réponses a ensuite été préparé incluant soit des flèches indiquant différentes sections sur une image fixe issue de la vidéo, soit une description d'un choix de gestes (figure 6.2).

Les tests de concordance de script présentent habituellement une structure de scores s'apparentant à une échelle de Likert. Dans ce contexte, ces échelles nous ont initialement semblé difficilement applicables lors du visionnement des vidéos de procédures chirurgicales. En effet, il nous a semblé plus facile de produire des items en demandant aux participants quelle serait la prochaine étape dans une situation donnée. Les questions à choix multiples nous ont paru une meilleure modalité à ce stade initial de développement. Il nous a semblé que cette approche permettait de comparer les scripts des apprenants avec ceux

des experts. Les questions et les choix de réponses ont été conçus de manière à offrir plus d'une option valable pour chaque question. Cette approche avait pour but de cerner si les schémas de réponse des résidents seraient les mêmes que ceux des chirurgiens experts dans des conditions d'incertitude contrôlée. Au cours du processus de validation des items, des pistes de solution permettant l'utilisation d'échelles de Likert sont apparues ; des items utilisant des échelles de Likert seront créés sous peu pour que le dispositif d'évaluation formative respecte davantage les principes des évaluations et des formations par concordance. Dans une deuxième itération de ce projet, il sera intéressant d'inclure des items avec des échelles de réponses de Likert pour voir les avantages et les inconvénients de ces échelles dans notre contexte particulier.

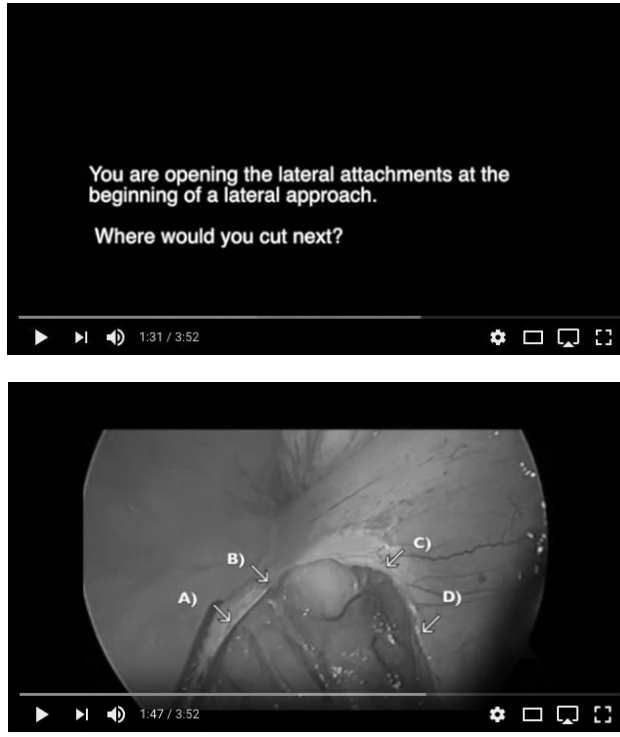


Figure 6.2
Exemple d'item

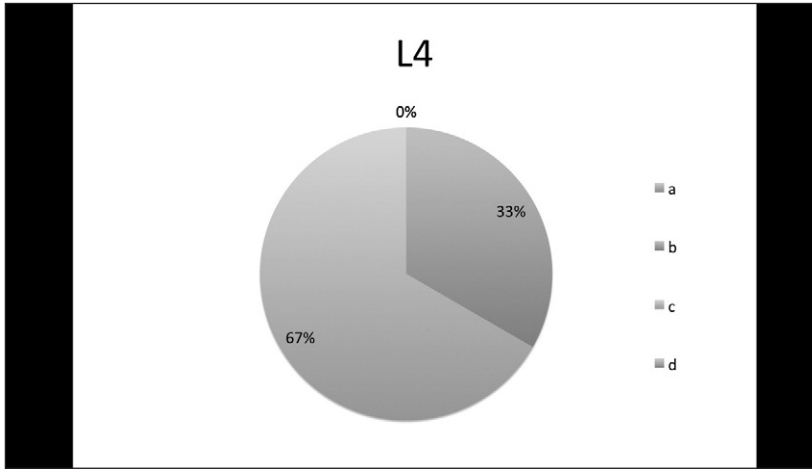
Chaque question a été examinée par deux experts pour en assurer la clarté et vérifier que l'incertitude dans la question était reliée au dilemme clinique et non à la formulation de la question ou à une mauvaise qualité de l'enregistrement vidéo.

Initialement, 81 items ont été mis au point selon le processus décrit ci-dessus. Cette banque initiale d'items a été présentée à 12 experts pour obtenir une clé de correction. Les propriétés des items ont été étudiées dans un exercice de validation de contenus. Onze items ont obtenu des réponses complètement uniformes, se comportant plus comme des questions à choix multiples simples que comme des items présentant des divergences dans les réponses d'experts. Les autres items comportaient un certain niveau de variation dans les réponses d'experts. Dans un deuxième temps, la banque d'items a été passée en revue avec un troisième chirurgien expert afin de relever les items les plus représentatifs de la prise de décision intra-opératoire. Cette entrevue avait pour but de déterminer les critères rendant certains items supérieurs aux autres. Comme entrevue préliminaire, cet entretien cherchait à définir la « supériorité » dans le contexte de ce projet. Cette entrevue tentait de cerner les critères qui permettraient de juger les items et de les classer. On cherchait à relever les items les plus représentatifs de la prise de décision opératoire, ceux ayant la plus grande pertinence par rapport au contexte clinique, ceux qui présentaient des situations cliniques fréquentes et reconnues par le chirurgien enseignant comme sources fréquentes d'erreurs. En général, les items se rapportant aux plans de dissection ont été favorisés, tandis que les questions liées à la reconnaissance des structures anatomiques ont été perçues comme étant de plus faible valeur. Après examen, la divergence dans les réponses d'experts pour certains items semble être liée à la durée du clip et à la difficulté pour les experts participants à s'orienter en visionnant une séquence opératoire aussi courte. Les items ont été corrigés en fonction des suggestions de l'expert.

2.5. La conception et l'assemblage du dispositif d'évaluation formative

Pour ce projet, de la banque initiale de 81 items, neuf ont été présentés à des résidents en tant qu'évaluation formative. Les critères pour le choix des items faisant partie du projet pilote étaient notamment la pertinence des items par rapport au contenu identifié initialement, les items ayant un certain niveau de divergence dans les réponses d'experts et les items jugés les plus clairs par l'expert. On a délibérément choisi un petit nombre d'items pour que ce dispositif d'évaluation ne nécessite pas un trop grand investissement de temps, sachant que les participants potentiels à ce projet devaient inclure cette activité dans leur horaire chargé. De plus, le nombre idéal d'items dans un tel dispositif n'a pas encore été défini ; c'était d'ailleurs l'une des questions que nous voulions éclaircir avec ce projet.

Ces neuf items ont été présentés à trois chirurgiens experts. Les experts ont répondu aux questions et fourni une courte explication avec leur réponse. Les réponses de tous les experts ont ensuite été présentées par question sous forme de diagramme circulaire et les explications regroupées dans des tableaux pour chaque item. La figure 6.3 présente un exemple des réponses fournies aux apprenants. Dans cet exemple, pour la question L4, 33 % des experts ont choisi la réponse B et 67 % ont choisi la réponse C. Le tableau qui suit le diagramme explique le choix des experts.



L4

Expert	Answer	Justification
1	B	c is too far lateral and may get you behind the ureter and gonadal vessels
2	C	You don't want to be too lateral as you may injure retroperitoneal structures
3	C	On the clip, I think they were cutting too lateral, the plan in more medial when looking at the shinny mesentery. C will get them back in a better plan. There is areolar tissue and difference if fat color

Figure 6.3
Exemple de réponses fournies aux apprenants (L4)

Une vidéo regroupant les neuf items a été créée et diffusée sur YouTube pour être accessible aux participants[1].

Les réponses des experts ont été regroupées dans un fichier PowerPoint et envoyées aux résidents et chirurgiens enseignants participant au projet pilote (voir figure 6.3). Pour ce projet, les participants n'ont pas reçu de scores totaux.

3. LE PROJET PILOTE : LE SONDAGE DE LA RÉACTION DES APPRENANTS ET DES ENSEIGNANTS ET LES CONSIDÉRATIONS LOGISTIQUES

3.1. Les réactions des apprenants

3.1.1. *La méthode*

Six résidents du programme de chirurgie générale de l'Université d'Ottawa ont accepté de participer à ce projet pilote afin de discuter du potentiel d'un dispositif d'évaluation formative faisant appel aux principes des formations par concordance. Ces participants ont été choisis parce qu'ils étaient engagés dans un stage de chirurgie colorectale au moment de la mise en œuvre du projet pilote. Trois résidents de troisième année, deux moniteurs cliniques ayant terminé leur résidence et étant enrôlés dans un programme de formation surspécialisée et un résident dans sa dernière année de formation ont participé au projet pilote.

Les résidents ont visionné une vidéo présentant neuf items incluant un court clip issu d'une procédure et une question à choix multiples. Les résidents ont aussi eu accès à un fichier PowerPoint présentant la distribution des réponses des douze experts et les explications des choix de réponses des trois experts participant au projet pilote. Les étapes du projet pilote sont présentées à l'annexe C.

Un questionnaire a été envoyé par courriel aux participants pour sonder leurs réactions. Le but de cette étape était d'obtenir de l'information sur l'impression des apprenants au sujet du dispositif d'évaluation formative lui-même, mais aussi à propos de la faisabilité et de la valeur d'une telle modalité d'évaluation formative. Les questions portaient sur l'utilité potentielle du dispositif d'évaluation, les contenus pouvant être enseignés de cette façon et le niveau de compétence nécessaire pour pouvoir bénéficier de ce genre de stratégie. Quelques

1. <https://www.youtube.com/watch?v=fzdLXWrDgUE&sns=em>.

autres questions portaient sur des aspects logistiques. Le questionnaire comportait sept questions ouvertes. Il a été conçu au moment de la conception du tableau de spécifications, en tenant compte des aspects du dispositif d'évaluation restant incertains et pour appuyer la validité apparente et de contenu. Les réponses ont été analysées en vue de trouver des thèmes communs.

3.1.2. *Les résultats*

Les participants ont souligné la différence entre cet exercice et le visionnement de vidéos chirurgicales disponibles en ligne ou l'observation en salle d'opération. Tous les résidents ont évoqué l'effet positif de ce dispositif d'évaluation en raison de son aspect actif. Le fait de devoir répondre à une question force à réfléchir et les participants ont eu l'impression de s'impliquer davantage dans cet exercice par rapport à une observation sans aide cognitive.

L'accès à des réponses d'experts a permis de mettre en lumière certaines lacunes et les participants ont apprécié le fait de pouvoir obtenir une rétroaction immédiate de qualité. Certains résidents ont indiqué que les réponses d'experts leur donnaient l'impression d'avoir accès aux processus cognitifs utilisés par ces derniers pour prendre des décisions intra-opératoires. En particulier, certaines règles non écrites ont été mentionnées par les experts et des résidents ont observé que ces règles n'avaient jamais été décrites de façon explicite durant une procédure.

Les résidents ont noté que cet exercice permettait de clarifier des sous-étapes nécessaires pour accomplir les étapes généralement décrites dans les livres de référence. Les participants ayant davantage d'expérience ont relevé que ce genre de contenu est difficile à cerner en dehors du contexte clinique. Les moniteurs cliniques qui combinent souvent des tâches d'enseignant et d'apprenant ont mentionné que les erreurs les plus souvent commises par les apprenants concernent l'exécution efficace de sous-étapes au cours d'une procédure et que ce genre d'évaluation formative avait, selon eux, le potentiel d'accélérer la progression des résidents.

Toutefois, le format des réponses d'experts fournies dans le document d'évaluation formative a semblé déstabiliser certains participants ; plusieurs auraient souhaité avoir une réponse claire pour chaque item. Il a été suggéré d'ajouter une séquence vidéo illustrant la meilleure réponse de façon à consolider l'apprentissage. Le groupe d'apprenants auquel le dispositif d'évaluation formative a été présenté n'avait jamais été exposé à une formation par concordance auparavant. Cette gêne relative eu égard au format a été notée par d'autres auteurs

(Gibot et Bollaert, 2008). Il est vrai que les résidents sont plus exposés à des questions à choix multiples classiques où il n'existe qu'une seule bonne réponse. La notion d'incertitude que renferme cette évaluation formative était relativement nouvelle pour les apprenants.

Le format des questions et la longueur des extraits vidéo ont aussi été discutés. Globalement, les résidents ont trouvé qu'il était relativement difficile de s'orienter dans les clips vidéo. Des annotations avaient été ajoutées pour aider les résidents à s'orienter par rapport aux concepts anatomiques, aux étapes de la chirurgie ayant été réalisées et aux buts de l'étape en cours. Cependant, plusieurs participants ont mentionné devoir regarder les clips plusieurs fois avant de bien comprendre les structures présentées à l'écran. Cette observation est intéressante puisque l'une des hypothèses de travail à la base de ce projet est que l'utilisation de vidéos comme supports pour les questions force les apprenants à effectuer une tâche d'interprétation des données visuelles qui ne serait pas présente dans un dispositif d'évaluation utilisant des données textuelles comme question. La capacité des résidents à interpréter les informations visuelles fournies dans les vignettes vidéo fait partie des habiletés devant être évaluées par cet exercice. Malgré le souhait de fournir suffisamment d'information aux apprenants pour leur permettre de prendre une décision, les items ont été expressément créés pour permettre aux résidents de pratiquer leur habileté de perception visuelle. Il n'est pas difficile d'imaginer que l'utilisation d'un même item pour former à l'interprétation visuelle et à la prise de décision puisse être difficile, surtout chez les novices, car on sait qu'ils ont besoin de plus de ressources cognitives que les experts lors de la perception d'une situation (Endsley, 2006). Selon les commentaires des participants, il serait probablement pertinent pour les résidents moins expérimentés d'avoir des items spécialement conçus pour améliorer les habiletés de perception et d'autres comportant davantage d'annotations et des clips vidéo plus longs pour faciliter leur apprentissage et éviter toute confusion. Une autre solution potentielle consisterait à ajouter des commentaires dans le document de réponses sur ce que les experts perçoivent pour faciliter l'apprentissage de la perception.

3.2. Les réactions des enseignants

3.2.1. *La méthode*

Les commentaires de plusieurs experts ont été recueillis durant le processus de création des items, 12 chirurgiens au total. De plus, deux chirurgiens enseignants ont accepté d'effectuer l'exercice comme s'ils

étaient des résidents (annexe C). Leurs commentaires ont été recueillis par courriel à l'aide d'un questionnaire similaire à celui utilisé pour recueillir les commentaires des résidents. Ce questionnaire cherchait à mieux définir la logistique de ce dispositif d'évaluation formative et à sonder la validité apparente et de contenu de ce mode d'évaluation. Les commentaires ont été regroupés et analysés.

3.2.2. *Les résultats*

En général, les chirurgiens concernés ont beaucoup apprécié l'exercice. Plusieurs ont noté le côté novateur de ce dispositif d'évaluation formative. Ils ont mentionné que ce genre d'exercice avait le potentiel de mieux préparer les résidents pour la salle d'opération. En effet, certains ont indiqué que les vignettes vidéo utilisées illustraient des erreurs communes qui nécessitent souvent plusieurs commentaires similaires au cours d'une même procédure et que ce genre d'exercice, en permettant d'enseigner en dehors de la salle d'opération, a la capacité d'aider les résidents à mieux assimiler les commentaires en évitant la surcharge cognitive.

Les chirurgiens ont relevé que ce genre d'exercice met l'accent sur les repères anatomiques, la reconnaissance des plans de dissection et la prévention et la gestion des complications. Ils ont reconnu que ce genre de contenu est difficile à enseigner avec les ressources disponibles à l'heure actuelle, ce qui rend cet exercice d'autant plus pertinent dans le contexte actuel.

Les chirurgiens enseignants sont d'avis que les résidents devraient utiliser cette formation pour se préparer à la réalisation d'interventions, un participant ayant même proposé que ce genre d'exercice devienne obligatoire.

3.3. Les considérations logistiques

3.3.1. *La méthode*

Les participants ayant répondu au sondage courriel ont aussi donné leur opinion sur les considérations logistiques entourant ce dispositif d'évaluation formative.

3.3.2. *Les résultats*

Tout d'abord, s'agissant de la population cible, les apprenants et les enseignants ont indiqué que cet exercice, tel qu'il est construit actuellement, ne serait pas adéquat pour des apprenants n'ayant jamais observé

la procédure. Pour pouvoir répondre aux questions de façon acceptable, il leur a semblé nécessaire d'avoir des notions anatomiques de base, de connaître les étapes de la procédure et d'avoir une image mentale globale de ce à quoi cette procédure doit ressembler. Les participants ont mentionné que les résidents dans la deuxième moitié de leur formation seraient les plus aptes à bénéficier de ce genre d'exercice.

Les huit participants au projet pilote (deux chirurgiens et six résidents) ont mis entre 10 et 30 minutes pour répondre aux questions et lire les réponses d'experts. Les participants ont répondu aux questions seuls, en dehors des heures de travail clinique, soit dans un contexte similaire à celui dans lequel ils préparent leur participation à des chirurgies. Ils ont proposé de limiter le nombre d'items à un maximum de 15 pour tirer le maximum de bénéfices de chacun d'eux.

Les résidents et les enseignants ont souligné que la facilité d'accès à un tel exercice aurait une influence importante sur l'utilisation d'un dispositif d'évaluation formative. Les apprenants ont un horaire chargé et disposent de beaucoup de ressources pour soutenir l'apprentissage des techniques opératoires. Bien que cet exercice soit perçu comme offrant des avantages uniques en leur genre, les participants au projet pilote ont mentionné qu'il serait nécessaire que ce dispositif soit accessible sur des appareils électroniques mobiles, idéalement sur des sites Internet déjà fréquentés par les résidents.

4. LA DISCUSSION

Ce projet pilote avait pour but de créer un dispositif d'évaluation formative et de recueillir des commentaires sur sa pertinence et son applicabilité.

4.1. La synthèse des résultats

Tout d'abord, la méthode choisie pour définir les contenus a mené à une meilleure compréhension des éléments clés de la prise de décision chirurgicale. Plusieurs des concepts discutés par les experts et les apprenants et qui sont présents dans des lignes directrices, de même que certains instruments de mesure de compétence chirurgicale, sont des concepts généraux qui pourront éventuellement servir à la création de dispositifs d'évaluation pour d'autres procédures dans différentes disciplines. En effet, il semble que toutes les sources de contenus utilisées pour créer cet instrument accordent une même importance à l'identification et au respect du plan chirurgical, à l'identification des repères anatomiques et à la prévention et à la gestion des complications. La réaction des apprenants et des enseignants au dispositif d'évaluation

permet de penser que la méthode utilisée pour définir les contenus a permis de créer un instrument pertinent pour les résidents en chirurgie générale. En effet, tous les participants ont convenu que cet exercice offrait une occasion unique pour améliorer leurs habiletés à reconnaître le plan de dissection. Par conséquent, la méthode décrite ici pourrait éventuellement servir dans d'autres contextes connexes.

La méthode utilisée pour concevoir les items démontre le potentiel de vignettes vidéo dans la création de formations par concordance. En effet, il s'agit à notre connaissance d'un premier essai de formation par concordance basée exclusivement sur l'utilisation de vidéos. Le recours à des questions à choix multiples au lieu d'une autre modalité de réponses s'explique par les ressources disponibles au moment de la création des items. L'emploi d'échelles de Likert est recommandé lors de la création de formation par concordance. Au cours des prochaines étapes de ce projet, des échelles de Likert seront incluses. Il aurait aussi été intéressant d'utiliser une méthode de réponses comparable à celle utilisée par Madani *et al.* (2016) ou Schlachta *et al.* (2015) dans laquelle on demande à l'apprenant de tracer une ligne sur une image fixe pour indiquer où se situe le plan ou la meilleure ligne de dissection (Madani *et al.*, 2016 ; Schlachta *et al.*, 2015). Cette méthode de réponses novatrice ne nous était malheureusement pas accessible au moment de créer les items. Nous avons choisi le format des choix de réponses pour faciliter l'utilisation de l'instrument par un grand nombre de participants, en pensant qu'il pourrait être intéressant de comparer les réponses d'apprenants de différents niveaux d'expérience au cours d'un projet futur. En outre, il est possible que le format avec choix de réponses facilite la révision des items par les apprenants. En effet, il est plus facile de constater qu'on a choisi une réponse différente de celles des experts que de comparer des subtilités dans des réponses ouvertes. Cet aspect sera préservé par l'utilisation d'échelles de Likert.

Le format actuel du dispositif d'évaluation formative semble approprié étant donné qu'il s'agit d'un projet pilote. Cependant, ce projet a permis de faire ressortir les besoins des apprenants concernant le format optimal. Les participants au projet pilote sont d'avis que, pour être utilisé de façon régulière, l'instrument devrait être disponible sur des appareils mobiles avec un support Internet. Le format actuel, avec les vignettes vidéo et les questions disponibles sur YouTube et les réponses présentées sur un fichier PowerPoint, rend l'utilisation de ce dispositif d'évaluation plus malaisé. Idéalement, les vignettes vidéo et les questions seraient disponibles sur un même écran suivi des réponses des experts et de leurs explications. Une plateforme a été développée à la suite d'une collaboration entre une firme privée et l'Université de

Montréal ; une prochaine version du dispositif d'évaluation formative bénéficierait de l'utilisation de cette plateforme spécialement créée pour supporter les formations par concordance.

4.2. Les liens avec la littérature sur l'évaluation formative

Les commentaires des participants suggèrent que le dispositif d'évaluation créé respecte les principes de l'évaluation formative. Rudolph *et al.* (2008) listent des éléments clés dans la création de dispositifs d'évaluation formative. Premièrement, par définition, l'évaluation formative suppose une interaction entre un apprenant et un expert enseignant. Dans notre dispositif, cette interaction se réalise au moyen des questions soumises aux participants et aux réponses d'experts incluses dans le dispositif. Ce format permet aux apprenants d'avoir accès aux processus cognitifs d'expert et de les comparer avec les leurs. Ces réponses d'experts ont le mérite d'expliquer leur raisonnement. De plus, comme le dispositif inclut les réponses de plusieurs experts, cela permet aux résidents « d'interagir » avec tous ces experts simultanément et de comparer leur différente façon de penser. Ensuite, Rudolph *et al.* (2008) notent l'importance d'offrir une rétroaction claire et honnête. Le dispositif, en forçant les apprenants à faire un choix avant de leur donner accès aux réponses, leur permet de comparer leur décision avec celles d'un groupe d'expert. Le format des choix de réponses favorise aussi une rétroaction claire où un apprenant peut voir immédiatement si les experts partagent son point de vue et s'il y a de la variance dans les réponses d'experts ; ils ont aussi la possibilité de comparer leur raisonnement avec celui des experts. Le dispositif créé peut s'adapter aux besoins particuliers des apprenants, une autre caractéristique importante de l'évaluation formative de qualité selon Rudolph *et al.* En ayant la possibilité de revoir la vignette pour bien comprendre les explications des experts, les apprenants peuvent décider de passer plus ou moins de temps sur divers aspects de la prise de décision chirurgicale selon leurs besoins. Finalement, Rudolph *et al.* (2008) soulignent l'importance d'une évaluation formative fréquente pour en maximiser les bénéfices. Le dispositif créé, portable et accessible en tout temps, permet aux participants de l'utiliser aussi fréquemment qu'ils le souhaitent. On peut penser que la création de plusieurs dispositifs d'évaluation formative du même genre pourrait former une banque de modules de formation pour chaque procédure offrant aux utilisateurs un accès constant à des évaluations formatives particulières. Gardant en tête l'importance d'offrir une évaluation formative adaptée à chaque résident, on pourrait inclure dans une telle banque de dispositifs d'évaluation des questions dont le degré de difficulté

serait adapté à chaque niveau d'apprenants et des dispositifs contenant des items destinés à travailler certains aspects de la prise de décision intra-opératoire, par exemple la reconnaissance du plan de dissection.

Harlen et James (1997), quant à eux, soulignent l'importance d'une évaluation basée sur le contexte pratique. Le dispositif créé dans le cadre de ce projet, en utilisant des vignettes vidéo provenant de procédures réelles, est définitivement ancré dans le contexte clinique où les habiletés promues par l'évaluation formative seront utilisées. Ces deux auteurs notent aussi comme il est important qu'une évaluation formative soit perçue par les apprenants comme utile et riche en potentiel. Les commentaires des participants au projet pilote indiquent justement que les résidents perçoivent ce genre d'évaluation comme pertinente à leur apprentissage.

Sargeant *et al.* (2009) soulignent l'importance de la crédibilité des sources d'évaluation formative dans l'utilisation de la rétroaction fournie aux apprenants. Dans le cas de notre dispositif d'évaluation, en utilisant des chirurgiens experts pour fournir des explications, on s'assure de la crédibilité de nos sources. Les choix des experts et leur explication étaient présentés de façon anonyme. On peut penser qu'ajouter le nom des experts, et potentiellement une courte biographie, pourrait accroître leur crédibilité auprès des apprenants. De façon intéressante, Sargeant *et al.* (2009) indiquent que des commentaires constants à travers plusieurs sources de rétroaction sont plus susceptibles d'être intégrés par les apprenants. Le concept des formations par concordance, en intégrant la notion d'incertitude planifiée dans les scénarios présentés, rend les consensus improbables. Les commentaires des résidents ayant utilisé le dispositif d'évaluation formative vont dans le même sens que Sargeant *et al.* (2009). Plusieurs se sont dits relativement mal à l'aise avec l'idée qu'il y ait plusieurs bonnes réponses possibles et ont suggéré de soumettre seulement la « meilleure » option pour faciliter leur apprentissage. D'un autre côté, Sargeant *et al.* (2009) et d'autres auteurs (Petty et Cacioppo, 1986) expliquent que des opinions dissonantes fournies par différents enseignants tendent à promouvoir plus de réflexion chez les apprenants à condition que les sources d'opinion contraire aient un niveau de crédibilité équivalent. Dans ce contexte, il est possible, en effet, que le format des formations par concordance, en incluant un élément d'ambiguïté, bien que rendant l'apprenant incertain de la conduite à adopter dans une situation donnée, suscite davantage de réflexion. Rappelons que Sargeant *et al.* soutiennent aussi que la qualité d'une évaluation formative se mesure en fonction de la réflexion qu'elle engendre.

4.3. Les limites du projet

Ce projet pilote a des limites. Le dispositif d'évaluation formative a été offert à un petit échantillon de résidents. Leurs réactions par rapport à l'utilisation de ce dispositif vont dans la même direction, mais il serait nécessaire de recueillir les commentaires d'autres utilisateurs potentiels pour pouvoir prétendre généraliser les trouvailles de ce projet pilote. Par exemple, il n'y avait pas de résidents durant leur première ou deuxième année de formation postdoctorale dans cette étude. Les participants au projet ont indiqué que les bénéfices de ce mode d'évaluation formative sont plus susceptibles d'être ressentis par les résidents qui sont dans la seconde moitié de leur résidence. Il serait intéressant de vérifier si les résidents moins expérimentés ont une opinion différente.

Lors de la conception d'un dispositif d'évaluation formative, on cherche à promouvoir l'amélioration des habiletés des apprenants utilisant ce dispositif. Pour l'instant, il n'a pas été possible de mesurer l'efficacité du dispositif créé. Bien peu d'instruments de mesure de l'acuité de la prise de décision intra-opératoire existent et ils ont en général été utilisés dans un contexte particulier souvent difficile à reproduire (Madani *et al.*, 2016 ; Samuelson *et al.*, 2006). Leur validité dans le contexte de ce projet reste à définir. Idéalement, une évaluation formative de la prise de décision intra-opératoire devrait mener à une amélioration des compétences techniques des apprenants. À ce stade-ci, un tel résultat reste à démontrer.

4.4. Les recherches futures

Ce chapitre se voulant d'abord et avant tout l'illustration d'un projet pilote, il ouvre la porte à plusieurs autres projets. D'abord, pour considérer l'implantation de ce dispositif à plus grande échelle, il sera nécessaire de rendre le format plus accessible aux participants. L'utilisation d'une plateforme permettant une navigation plus facile entre les items sera nécessaire. Cette plateforme devra aussi offrir la possibilité de visionner le même clip vidéo à plusieurs reprises au gré du participant. On pourrait aussi inclure une option pour pouvoir visionner une plus longue partie de la procédure autour de la vignette pour aider les participants à s'orienter.

Éventuellement, il sera pertinent d'évaluer l'incidence de l'utilisation d'une telle modalité d'évaluation formative sur un groupe de résidents plus étendu. À ce moment-là, il sera important d'obtenir l'opinion de tous les acteurs impliqués dans l'entraînement des apprenants, tant chez les acteurs œuvrant dans l'administration du programme de

résidence que chez les chirurgiens enseignants, sans oublier les résidents eux-mêmes. Pour l'instant, il ne semble pas exister de dispositif d'évaluation sommative de la prise de décision intra-opératoire idéal pour étudier les effets de ce dispositif de formation par concordance.

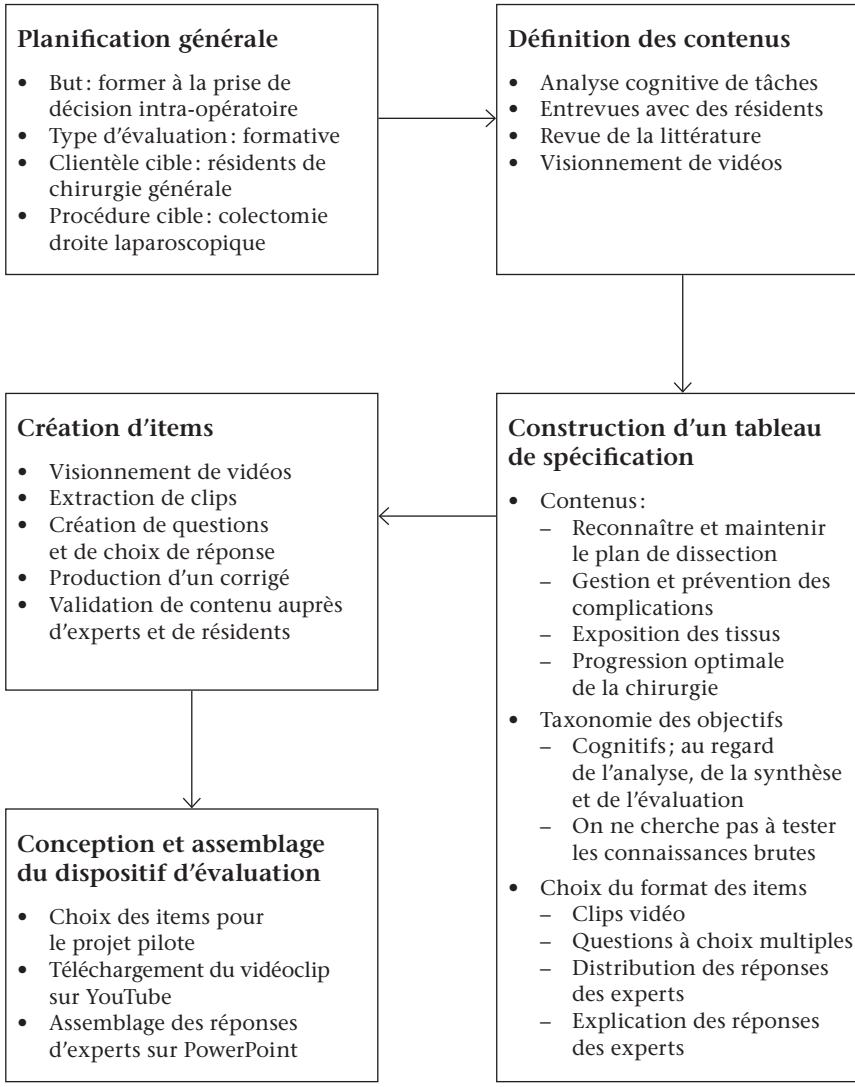
Au cours de ce projet, de nombreux participants ont mentionné que cette modalité d'évaluation formative serait mieux adaptée aux résidents seniors. Il serait intéressant de mener un projet auprès de résidents juniors pour évaluer leurs besoins et construire des items qui leur conviendraient davantage. On peut penser que des items centrés sur l'identification du plan de dissection ou les repères anatomiques seraient plus pertinents pour eux que des items relatifs à la gestion de la procédure. Il serait utile d'avoir une banque d'instruments contenant 10 à 15 items avec un degré de difficulté croissant. À ce stade-ci, par contre, le degré de difficulté de chaque item reste à définir. Un projet explorant les critères rendant un item plus difficile qu'un autre serait pertinent pour permettre de créer des dispositifs comportant des degrés de difficulté adaptés au niveau de formation des résidents.

Finalement, ce projet pilote s'est construit autour de la colectomie droite par laparoscopie, on peut penser qu'il serait utile de développer des modalités d'évaluation formative pour d'autres procédures clés en chirurgie générale. Pour établir l'ordre de priorité, il sera utile de consulter les résidents et leurs enseignants.

CONCLUSION

Le but de ce chapitre était d'expliquer le rôle potentiel des formations par concordance comme modalité d'évaluation formative de la prise de décision intra-opératoire en utilisant l'exemple de la colectomie droite laparoscopique. Dans le cadre de ce collectif, ce chapitre se voulait une fenêtre sur l'évaluation formative. En nous basant sur les principes qui sous-tendent l'évaluation formative, nous avons montré qu'une formation par concordance basée sur l'utilisation de vidéo peut offrir une rétroaction de qualité aux résidents de spécialités chirurgicales. Plusieurs aspects de ce projet restent à explorer, notamment l'effet tangible de cette modalité d'enseignement sur la formation globale des résidents et la place que ce genre d'évaluation formative devrait avoir dans les programmes d'enseignement.

ANNEXE A. ÉTAPES DU DÉVELOPPEMENT DU DISPOSITIF D'ÉVALUATION SUIVANT LE MODÈLE DE DOWNING



ANNEXE B. CONTENUS ESSENTIELS POUR LA PRISE DE DÉCISION DANS LA COLECTOMIE DROITE LAPAROSCOPIQUE

Tâche	Ce que font les experts pour la simplifier	Indices recherchés par les experts	Erreurs courantes et complications possibles
Première incision		• Insérer le premier trocart à mi-section de l'abdomen; au maximum du pneumopéritoine. • À mi-parcours entre le xiphoïde et la symphyse: là où débute la branche droite de l'artère colique moyenne.	• Risque d'interférence avec le positionnement des trocarts. • Les trocarts et la caméra peuvent se heurter. • Risque pour la visualisation (moins important avec l'utilisation d'un embout souple ou d'un endoscope angulé).
Positionnement des trocarts	• Avant l'insertion des autres trocarts, vérifier qu'il n'y a pas de complication liée à l'introduction du premier trocart, puis déterminer si la chirurgie peut être faite en laparoscopie. • Séparer les adhérences près du site des trocarts d'abord. Un trocart supplémentaire peut être nécessaire. • Injecter l'anesthésie locale et utiliser l'aiguille pour orienter l'insertion, connaître la trajectoire du trocart. • Toujours sous vision directe. • Rechercher l'artère épigastrique.	**Trocarts de travail:** • Côté gauche. • Le premier: au moins à 2 doigts de l'épine iliaque antéro-supérieure gauche, au-dessus et médialement. • Ajuster le positionnement des trocarts selon la taille de l'abdomen une fois insufflé; placer les trocarts plus près de l'anatomie cible si le patient a un abdomen très grand. • Latéralement au grand droit. • À une largeur de main d'écart. • Centré autour de la caméra. **Trocarts d'assistance:** • Sus-pubien, quadrant supérieur droit. • Côté supérieur droit: pour plus de confort pour le chirurgien, s'il tient la caméra en même temps.	• Lésions aux autres organes. • Limitation de la mobilité de l'instrument pendant l'intervention. • Chirurgie moins ergonomique. • Lésion de l'artère épigastrique ou erreur d'identification.

ANNEXE C. RÉSUMÉ DU PROJET PILOTE SUR LA VALIDITÉ APPARENTE ET LA FAISABILITÉ DE CETTE FORMATION PAR CONCORDANCE

Participants

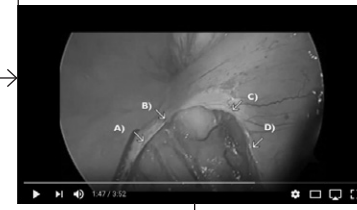
6 résidents
2 chirurgiens enseignants

Intervention

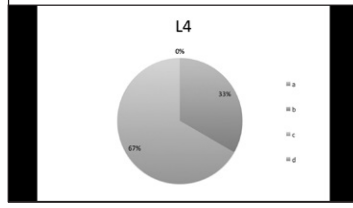
Ont répondu à neuf items présentés sur YouTube et PowerPoint. Un item inclut :

Une question

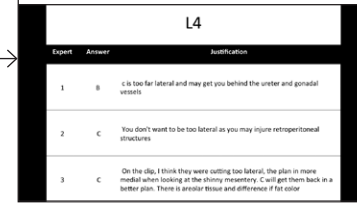
Suivie d'un choix de réponses



Ils ont comparé leurs réponses avec celles de 12 experts



Ils ont pu consulter les explications de 3 experts



Sondage sur les perceptions

Ils ont finalement rempli un questionnaire sondant leur perception

1. Trouvez-vous cet exercice utile ? Pensez-vous qu'il vous aiderait à améliorer vos compétences techniques ?
2. Pour quel niveau de formation conviendrait-il ?
3. Sur quels scénarios ou aspects de l'intervention cet exercice devrait-il porter (p. ex. anatomie, localisation des vaisseaux, maîtrise des saignements, recherche de plans, choix des instruments, rétraction, etc.) ?
4. Combien de temps vous a-t-il fallu pour effectuer cet exercice ?
5. Combien de questions devrait-on inclure dans une séance de formation ?
6. Si cet exercice était disponible, est-ce que vous l'utiliseriez pour vous préparer à la performance en salle d'opération ?
7. Avez-vous d'autres commentaires ?

BIBLIOGRAPHIE

Allal, L. et L. Mottier Lopez (2005). « L'évaluation formative de l'apprentissage : revue de publications en langue française », dans *L'évaluation formative : pour un meilleur apprentissage dans les classes secondaires*, Paris : Organisation de coopération et de développement économiques (OCDE), p. 265-290.

Charlin, B. (2014). « Formation par concordance : raisonner et agir en contexte d'incertitude », *Le Bulletin du CPASS*, (6), p. 4-7.

Charlin, B., H.P. Boshuizen, E.J. Custers et P.J. Feltovich (2007). « Scripts and clinical reasoning », *Medical Education*, *41*(12), p. 1178-1184, doi : 10.1111/j.1365-2923.2007.02924.x.

Charlin, B. et N. Fernandez (2016). « Préparer et animer une formation par concordance », dans T. Pellacia (dir.), *Comment [mieux] enseigner la médecine et les sciences de la santé ?*, Bruxelles : De Boeck Éditeur, p. 325-331.

Charlin, B., J. Tardif et H.P.A. Boshuizen (2000). « Scripts and medical diagnostic knowledge : Theory and applications for clinical reasoning instruction and research », *Academic Medicine*, *75*(2), p. 182-190.

Cobb, K.A., G. Brown, R. Hammond et L.H. Mossop (2015). « Students' perceptions of the Script Concordance Test and its impact on their learning behavior : A mixed methods study », *The Journal of Veterinary Medical Education*, *42*(1), p. 45-52, doi : 10.3138/jvme.0514-057R1.

Crandall, B., G. Klein et R. Hoffman (2006). *Working Minds : A practitioner's Guide to Cognitive Task Analysis*, Cambridge : The MIT Press.

DaRosa, D., D.A. Rogers, R.G. Williams, L.S. Hauge, H. Sherman, K. Murayama *et al*. (2008). « Impact of structured skills laboratory curriculum on surgery residents' intraoperative decision-making and technical skills », *Academic Medicine*, *10*(83), suppl., p. S68-S71.

Downing, S.M. (2006). « Twelve steps for effective test development », dans S.M. Downing et T.M. Haladyna (dir.), *Handbook of Test Development*, Mahwah : Lawrence Erlbaum Associates, p. 3-25.

Endsley, M.R. (2006). « Expertise and Situation Awareness », dans K.A. Ericsson, N. Charness, P.J. Feltovich et R.R. Hoffman (dir.), *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge : Cambridge University Press, p. 633-652.

Ericsson, A. (2006). « The Influence of Experience and Deliberate Practice on the Development of Superior Expert Performance », dans A. Ericsson, N. Charness, P.J. Feltovich et R.R. Hoffman (dir.), *The Cambridge Handbook of Expertise and Expert Performance*, New York : Cambridge University Press, p. 683-704.

Flin, R., G. Youngson et S. Yule (2007). « How do surgeons make intraoperative decisions ? », *Qual Saf Health Care*, *16*, p. 235-239, doi : 10.1136/qshc.2006.020743.

Fournier, J.P., A. Demeester et B. Charlin (2008). « Script concordance tests : Guidelines for construction », *BMC Medical Informatics and Decision Making*, *8*, p. 18, doi : 10.1186/1472-6947-8-18.

Gibot, S. et P.-E. Bollaert (2008). « Le test de concordance de script comme outil d'évaluation formative en réanimation médicale », *Pédagogie médicale*, *9*(1), p. 7-18, doi : 10.1051/pmed :2008037.

Guerlain, S., K.B. Green, M. Lafollette, T.C. Mersch, B.A. Mitchell, G.R. Poole *et al.* (2004). « Improving surgical pattern recognition through repetitive viewing of video clips I », *EE Transaction on Systems, Man and Cybernetics*, *34*(6), p. 699-707.

Hall, J.C., C. Ellis et J. Hamdorf (2003). « Surgeons and cognitive processes », *British Journal of Surgery*, *90*, p. 10-16, doi : 10.1002/bjs.4020.

Harlen, W. et M. James (1997). « Assessment and learning : Differences and relationships between formative and summative assessment », *Assessment in Education*, *4*(3), p. 365-379.

Hauge, L.S., J. Wanzek et C. Godellas (2001). « The reliability of an instrument for identifying and quantifying surgeons' teaching in the operating room », *The American Journal of Surgery*, 181, p. 333-337.

Hornos, E.H., E.M. Pleguezuelos, C.A. Brailovsky, L.D. Harillo, V. Dory et B. Charlin (2013). « The practicum script concordance test : An online continuing professional development format to foster reflection on clinical practice », *The Journal of Continuing Education in the Health Professions*, *33*(1), p. 59-66, doi : 10.1002/chp.21166.

Koya, K.D., K.R. Bhatia, J.T. Hsu et A.C. Bhatia (2012). « YouTube and the expanding role of videos in dermatologic surgery education », *Seminars in Cutaneous Medicine and Surgery*, *31*(3), p. 163-167, doi : 10.1016/j.sder.2012.06.006.

Lee, J.S., H.S. Seo et T.H. Hong (2015). « YouTube as a potential training method for laparoscopic cholecystectomy », *The Annals of Surgical Treatment and Research*, 89(2), p. 92-97, doi : 10.4174/astr.2015.89.2.92.

Linberry, M., C.D. Kreiter et G. Bordage (2013). « Threats to validity in the use and interpretation of script concordance test scores », *Medical Education*, *47*(12), p. 1175-1183.

Lubrarsky, S., B. Charlin, D. Cook, C. Chalk et C. Van Der Vleuten (2011). « Script concordance testing : A review of published validity evidence », *Medical Education*, *45*, p. 329-338.

Madani, A., Y. Watanabe, E. Bilgic, P.H. Pucher, M.C. Vassiliou, R. Aggarwal et L.S. Feldman (2016). « Measuring intra-operative decision-making during laparoscopic cholecystectomy : Validity evidence for a novel interactive Web-based assessment tool », *Surgical Endoscopy*, doi : 10.1007/s00464-016-5091-7.

Martin, J.A., G. Regehr, R. Reznick, H. MacRae, J. Murnaghan, C. Hutchison et M. Brown (1997). « Objective structured assessment of technical skill (OSATS) for surgical residents », *British Journal of Surgery*, *84*, p. 273-278.

Meterissian, S.H., B. Zabolotny, R. Gagnon et B. Charlin (2007). « Is the script concordance test a valid instrument for assessment of intra-operative decision-making skills ? », *The American Journal of Surgery*, *193*, p. 248-251.

Moulton, C.A., G. Regehr, L. Lingard, C. Merritt et H. MacRae (2010). « Operating from the other side of the table : Control dynamics and the surgeon educator », *Journal of the American College of Surgeons*, *210*(1), p. 79-86, doi : 10.1016/j.jamcollsurg.2009.09.043.

Moulton, C.A., G. Regehr, M. Mylopoulos et H.M. MacRae (2007). « Slowing down when you should : A new model of expert judgment », *Academic Medicine*, *82*(10), p. S109-S116.

Musselman, R.P., T. Gomes, B.P. Chan, R.C. Auer, H. Moloo, M. Mamdani, M. Al-Omran, O. Al-Obeed et R.P. Boushey (2012). « Changing trends in rectal cancer surgery in Ontario : 2002-2009 », *Colorectal Disease*, *14*(12), p. 1467-1472, doi : 10.1111/j.1463-1318.2012.03044.x.

Park, A., M. Barber, A.E. Bent, Y. Dooley, C. Dancz, G. Sutkin et E. Jelovsek (2010). « Assessment of intraoperative judgment during gynecologic surgery using the script concordance test », *American Journal of Obstetrics and Gynecology*, *203*, p. 240-246.

Petty, R.E. et J.T. Cacioppo (1986). « The elaboration likelihood model of persuasion », *Advances in Experimental Social Psychology*, *19*, p. 123-192.

Raîche, I. (2016). *Observational Learning of Junior Residents during Surgery : Exploring Promoters and Barriers to Learning*, Master of Arts in Education – Health Professions Education, University of Ottawa.

Ricca, L. et F. Lacaine (2009). « Laparoscopic surgery for colon cancer : A critical reading of the randomized trials of survival », *Journal de chirurgie (Paris)*, *146*(2), p. 136-142, doi : 10.1016/j.jchir.2009.05.018.

Rosen, M.A., E. Salas, D. Pavlas, R. Jensen, D. Fu et D. Lampton (2010). « Demonstration based-training : A review of instructional features », *Human Factors*, *52*(5), p. 596-609.

Ross, K.G., J. Shafer et G. Klein (2006). « Professional Judgments and "Naturalistic Decsion Making" », dans K.A. Ericsson, N. Charners, P.J. Feltovich et R.R. Hoffman (dir.), *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge : Cambridge University Press, p. 403-420.

Rudolph, J.W., R. Simon, D.B. Raemer et W.J. Eppich (2008). « Debriefing as formative assessment : Closing performance gaps in medical education », *Academic Emergency Medicine*, *15*(11), p. 1010-1016, doi : 10.1111/j.1553-2712.2008.00248.x.

Samuelson, M.L., J.A. Cadeddu et E.D. Matsumoto (2006). « Laparoscopic decision making : impact of preoperative reading and laparoscopic experience », *The Journal of Urology*, *176*(4), part. 1, p. 1553-1557, doi : 10.1016/j.juro.2006.06.100.

Sargeant, J.M., K.V. Mann, C.P. Van der Vleuten et J.F. Metsemakers (2009). « Reflection : A link between receiving and using assessment feedback », *Advances in Health Sciences Education. Theory and Practice*, *14*(3), p. 399-410, doi : 10.1007/s10459-008-9124-4.

Scallon, S.E., D.J. Fairholm, D.D. Cochrane et D.C. Taylor (1992). « Evaluation of the operating room as a surgical teaching venue », *Canadian Journal of Surgery*, *35*(2), p. 173-176.

Schlachta, C.M., S. Ali, H. Ahmed et R. Eagleson, R. (2015). « A novel method for assessing visual perception of surgical planes », *Canadian Journal of Surgery*, *58*(2), p. 87-91.

Shute, V.J. (2008). « Focus on formative feedback », *Review of Educational Research*, *78*(1), p. 153-189, doi : 10.3102/0034654307313795.

Smink, D., S.E. Peyre, D.I. Soybel, A. Tavakkolizadeh, A.H. Vernon et D.J. Anastakis (2012). « Utilization of a cognitive task analysis for laparoscopic appendectomy to identify differentiated intraoperative teaching objectives », *The American Journal of Surgery*, *203*, p. 540-545.

Sullivan, M.E., A. Ortega, N. Wasserberg, H. Kaufman, J. Nyquist et R. Clark (2008). « Assessing the teaching of procedural skills : Can cognitive task analysis add to our traditional teaching methods ? », *The American Journal of Surgery*, *195*, p. 20-23, doi : 10.1016/j.amjsurg.2007.08.051.

Sweller, J. (2003). « Evolution of human cognitive architecture », *The Psychology of Learning and Motivation*, *43*, p. 215-266.

Van der Maren, J.-M. (1995). *Méthodes de recherche pour l'éducation*, 2e éd., Montréal : Les Presses de l'Université de Montréal.

Vassiliou, M.C., L.S. Feldman, C.G. Andrew, S. Bergman, K. Leffondre, D. Stanbridge et G.M. Fried (2005). « A global assessment tool for evaluation of intraoperative laparoscopic skills », *The American Journal of Surgery*, *190*(1), p. 107-113, doi : 10.1016/j.amjsurg.2005.04.004.

Way, L.W., L. Stewart, W. Ganter, K. Liu et C.M. Lee (2003). « Cause and prevention of laparoscopic bile duct injuries : Analysis of 252 cases from a human factor and a cognitive psychology perspective », *Annals of Surgery*, *237*(4), p. 460-469.

Yule, S. et S. Paterson-Brown (2012). « Surgeons' non-technical skills », *Surgical Clinics of North America*, *92*(1), p. 37-50, doi : 10.1016/j.suc.2011.11.004.

CHAPITRE 7

Le rôle de l'évaluation de programme dans le domaine de la santé

Maud Mediell et Eric Dionne

Nous proposons un chapitre abordant la complexité reconnue de la pratique et de la recherche dans le domaine de la santé, et ce, particulièrement lorsqu'il s'agit de développer de nouveaux partenariats, notamment avec des disciplines provenant des sciences humaines et sociales telles que l'évaluation de programme. Aujourd'hui, plus que jamais, nous nous trouvons au cœur d'une période où la recherche et la pratique interdisciplinaire en santé permettent de renforcer la validité des connaissances acquises ainsi que le transfert de ces dernières et, par conséquent, l'efficacité des interventions en prévention, promotions et soins de santé. Cependant, un enjeu majeur : le « dialogue de sourds entre les disciplines » (Nadeau, 2005) persiste et menace la mise en œuvre de telles recherches et pratiques interdisciplinaires. De fait, si aujourd'hui, la nécessité et l'utilité du recours à l'évaluation de programme dans le domaine de la santé sont incontestables, l'évaluateur se doit d'inscrire sa pratique au cœur de multiples traditions scientifiques (Brousselle et al.*, 2011) : l'évaluation fondée sur l'épidémiologie, l'évaluation économique en santé, l'évaluation de programme en sciences sociales et la pratique clinique au Canada (approche centrée sur le patient, offre active de soins de santé en milieu francophone minoritaire, etc.). Aussi, verrons-nous que les praticiens de l'évaluation de programme doivent*

s'approprier les concepts clés de ces différentes traditions et déterminer les liens qu'elles entretiennent avec l'évaluation de programme, et ce, afin de comprendre les enjeux relatifs à leurs pratiques dans le monde complexe de la santé.

Au Canada, la santé publique constitue le volet du système de santé chargé de la protection, de la surveillance, de la prévention et de la promotion de la santé. Elle est le fruit d'un ensemble de connaissances scientifiques, d'habiletés et de valeurs qui se traduisent par des actions collectives, et ce, par l'entremise de programmes, de services et d'institutions visant la protection et l'amélioration de la santé de la population canadienne (Last, 2007). En tant que construit complexe, le terme « santé publique » peut décrire un concept, une institution sociale, un ensemble de disciplines scientifiques et professionnelles et de technologies, ou une pratique. De fait, les champs de spécialisation en santé publique ne cessent de croître, de même que les habiletés et les connaissances attendues des praticiens de la santé publique. Ainsi, retrouve-t-on une multitude de sciences de la santé publique, soit des activités scientifiques qui contribuent aux fondements scientifiques de la pratique, des services et des systèmes de la santé publique (Last, 2004).

Aujourd'hui, l'épidémiologie, en tant que science fondatrice de la santé publique, pose les assises de la création, de la mise en œuvre et de l'évaluation d'actions pratiques et appropriées en matière d'intervention en santé. De plus, les interventions largement utilisées dans les services en santé et les pratiques de santé publique affichent les caractéristiques des interventions complexes (Craig *et al.*, 2008). Il s'agit d'interventions mobilisant de multiples composantes à la fois autonomes et interactives, ce qui engendre un grand nombre de problèmes pour les évaluateurs, en plus des difficultés pratiques et méthodologiques que toute évaluation réussie doit surmonter. Aussi, Craig *et al.* (2008) font référence à la difficulté de normaliser la conception et la mise en œuvre des interventions et de leurs évaluations, et ce, du fait de leur sensibilité aux caractéristiques du contexte local, de la difficulté organisationnelle et logistique, des défis dans l'application de méthodes expérimentales dus à la longueur et à la complexité des chaînes causales liant l'intervention aux résultats.

On observe deux tendances dans les pratiques évaluatives actuelles en épidémiologie et en santé publique, toutes deux inscrites dans une tradition scientifique positiviste (voire postpositiviste). La première tendance, l'épidémiologie évaluative, repose sur l'appréciation normative (fonctions de contrôle et de suivi, et procédés de vérification de la conformité des composantes de l'intervention et d'assurance

qualité) et la recherche évaluative (démarche scientifique visant à poser un jugement *a posteriori* sur une intervention [mesure du rendement théorique, de la productivité, des effets et des rendements réels, etc.]). La deuxième tendance, l'évaluation économique en santé, se rapporte à des pratiques de gestion administrative et de mesure de la performance (suivi de gestion et reddition de comptes). Cependant, ce type de pratiques évaluatives en santé publique fait face à certaines limites, notamment au regard du développement, du monitorage et de l'évaluation d'interventions en santé caractérisées par leur complexité, où la définition de l'*evaluand* (l'objet, l'intervention, le programme évalué) demeure instable (Potvin, Bilodeau et Gendron, 2008). Au regard de ce constat, l'évaluation d'interventions implantées par les organisations de santé, interventions souvent compliquées, mais majoritairement complexes (Morrel, 2005), fait de l'évaluation un exercice encore plus difficile et conduit ses praticiens à faire face à des défis conceptuels, méthodologiques et opérationnels majeurs. Cette situation représente un enjeu important que l'on se doit de considérer et de surmonter afin de garantir la production de résultats évaluatifs utiles et utilisables (Potvin, Bilodeau et Gendron, 2011). Aussi, il apparaît important de ne pas uniquement se concentrer sur l'efficience des services de santé, mais également de prêter attention aux aspects développementaux des programmes d'intervention, aux besoins des personnes concernées par les programmes de santé (les parties prenantes primaires, secondaires et tertiaires), à la bonne mise en œuvre des programmes, à l'influence de l'environnement politico-socio-économique et culturel sur les composantes de ces derniers, au contexte dynamique et changeant, etc.

Selon nous, il y a un besoin général d'évaluation en santé publique et une nécessité de multiplier les évaluations de programme, et ce, tout au long de leur existence : du diagnostic de besoin à la mesure de la performance. En effet, il nous semble primordial d'accroître le rôle de l'évaluation, afin d'être en mesure de développer, entre autres, les pratiques de monitorage efficace du développement, de la mise en œuvre et des effets des interventions complexes en santé publique (et dans le domaine de la santé en général). De telles pratiques s'inscrivent d'ailleurs dans des approches évaluatives centrées sur le développement de programme et l'utilisation des résultats évaluatifs. En effet, l'évaluation développementale semble être une approche pertinente dans ces circonstances complexes, et ce, grâce au rôle de l'évaluateur dans le développement, l'implantation et l'évaluation de l'intervention. En outre, l'évaluateur fait partie intégrante de l'équipe de gestionnaires ; il facilite l'apprentissage expérientiel tout en favorisant l'utilisation du processus et des résultats de l'évaluation (Rey *et al.*, 2013). Enfin, l'évaluateur n'est pas présent pour porter un jugement sur

le fonctionnement et le travail de l'équipe de gestionnaires, mais son rôle consiste à les accompagner dans le processus de développement d'interventions contribuant dans le même temps à son succès.

Dans la première partie de ce chapitre, nous abordons les fondements de l'épidémiologie et des pratiques évaluatives en santé publique au Canada. Pour ce faire, nous proposons une description de ce qu'est l'épidémiologie et de ce que l'on entend par intervention complexe en santé, et ce, avant de faire un état des lieux des pratiques évaluatives des interventions en santé publique (épidémiologie évaluative et évaluation économique en santé). Dans la deuxième partie, nous présentons l'évaluation développementale, un modèle évaluatif dans le cadre duquel l'évaluateur fait partie de l'équipe de gestionnaires, responsable de la création et de l'implantation de programmes en santé, programmes qui, par essence, et souci d'efficacité, sont inscrits dans une collaboration interdisciplinaire (entre les sciences de la santé et les sciences humaines et sociales). Nous appuyant sur un exemple concret, nous montrons comment, à titre de coordonnateur, l'évaluateur peut faciliter la relation partenariale, le développement et l'évaluation de programmes de santé innovants, programmes composés d'interventions possédant les caractéristiques des interventions complexes.

1. LES FONDEMENTS DE L'ÉPIDÉMIOLOGIE ET DES PRATIQUES ÉVALUATIVES EN SANTÉ PUBLIQUE AU CANADA

1.1. Qu'est-ce que l'épidémiologie ?

Science fondatrice de la santé publique, l'épidémiologie décrit la relation entre la santé ou la maladie et les facteurs liés à la santé des populations. Si l'on se réfère au dictionnaire de l'épidémiologie, cette dernière est définie comme l'« étude de la distribution et des déterminants de santé ou d'évènements liés à la santé dans des populations données, et l'application de cette étude à la lutte contre les problèmes de santé » (Last, 2004, p. 78). L'épidémiologie est également considérée comme une philosophie et une méthode qui engendre des données essentielles, sur un large éventail de problèmes reliés à la santé, et ce, afin que les professionnels en santé publique soient en mesure de développer, d'implanter et d'évaluer l'efficacité des interventions dans les programmes de prévention et de promotion de la santé (Detels, 2009). En somme, aujourd'hui, l'épidémiologie ne concerne plus uniquement les épidémies de maladies transmissibles, mais également

les évènements touchant à la santé (contrôle des maladies, comportements des individus, mesure de prévention et de promotion de la santé, utilisation des services de santé, etc.).

Dans le cadre de son accompagnement des missions de santé publique, on recense 11 fonctions de l'épidémiologie (Detels, 2009 ; Morris, 2007)[1], cependant dans ce chapitre, nous nous attardons uniquement à la vérification de l'efficacité des stratégies d'intervention et à l'évaluation des programmes de santé publique. La première activité reliée à l'épidémiologie évaluative consiste à tester l'efficacité des stratégies d'intervention : cette fonction est reliée à l'un des objectifs premiers de la santé publique, à savoir, prévenir l'apparition d'une maladie par le biais d'une intervention dans le processus de manifestation de cette dernière (la vaccination, les campagnes antitabac, etc.). Cependant, avant de mettre en place de telles campagnes de prévention, il est nécessaire de prouver leur efficacité, et ce, avant qu'elles ne soient diffusées à plus large échelle. En outre, cette fonction de l'épidémiologie constitue la mise en œuvre de phases de pilotage des interventions (mise en place de dispositif d'essai, recherche de volontaires, coordination du pilotage, évaluation, etc.). La deuxième activité d'épidémiologie évaluative porte sur l'évaluation des programmes de santé publique : les départements de santé publique mettent en œuvre un nombre considérable d'activités afin de promouvoir la santé dans les communautés. De fait, l'évaluation continue de ces activités est primordiale pour garantir l'efficacité de leur rapport coût-bénéfice. Cette épidémiologie évaluative s'inscrit dans la tradition scientifique quantitative de l'épidémiologie, soit le calcul de probabilités, de statistiques, et la mise en œuvre de devis expérimentaux. Enfin, il existe une profusion de termes relatifs aux études évaluatives en santé publique liés aux diverses activités implantées et au cycle de gestion des programmes d'intervention. On parle de suivi de gestion, de surveillance, de reddition de comptes, autant de termes évoquant le recours aux méthodes quantitatives et à la mesure de la performance (Poissant, 2011).

1. Les 11 fonctions de l'épidémiologie sont les suivantes : 1) décrire le spectre d'une maladie ; 2) décrire l'histoire naturelle d'une maladie ; 3) réaliser des diagnostics communautaires ; 4) décrire l'image clinique d'une maladie ; 5) relever les facteurs augmentant ou diminuant les risques de contracter une maladie ; 6) indiquer les précurseurs d'une maladie et des syndromes ; 7) tester l'efficacité des stratégies d'intervention ; 8) investiguer les épidémies dont l'étiologie demeure inconnue ; 9) évaluer les programmes de santé publique ; 10) élucider les mécanismes de transmission d'une maladie ; 11) élucider les déterminants moléculaires et génétiques responsables de la progression d'une maladie.

Avant de décrire les pratiques évaluatives en santé, il nous semble important de définir ce que l'on entend par interventions en santé. Aussi, proposons-nous un aperçu des caractéristiques descriptives des interventions en santé, et ce, afin que le lecteur soit en mesure de pleinement saisir les caractéristiques de ces interventions qualifiées de complexes, et les enjeux relatifs à leur évaluation.

1.2. Qu'entend-on par interventions complexes en santé ?

Il existe une pléthore d'interventions en santé, interventions qui peuvent être réparties en fonction de quatre volets de la prévention en santé, soit la prévention clinique, la promotion de la santé, la protection de la santé et la politique publique favorable à la santé (Goldsmith, Hutchison et Hurley, 2004). Dans le cadre de ce chapitre, nous abordons uniquement les trois premiers volets préventifs (et leurs activités respectives) puisqu'ils font souvent partie d'un même modèle d'intervention en santé, concernant une seule et même problématique, chacun mobilisant des angles d'action différents (Goldsmith *et al*., 2004). Le premier volet, la prévention clinique, propose des interventions centrées sur les individus potentiellement à risque ou ceux qui désirent recevoir des soins cliniques (les services de santé offerts peuvent être acceptés ou refusés par le client). Le deuxième volet, la promotion de la santé, propose des interventions souvent sous la forme de campagnes d'information à travers les médias et les différents instituts de santé. La promotion de la santé encourage les comportements individuels censés avoir des effets positifs et freiner les comportements qui entraînent des effets négatifs sur la santé. Cependant, il est important de garder à l'esprit que la décision d'entreprendre ou non une activité bénéfique pour la santé relève de l'individu, l'offre de programme de promotion pour la santé ciblant un groupe ou une population donnée. Le troisième volet, la protection de la santé, propose des modifications dans l'environnement physique et social des individus, et ce, afin de réduire les risques pour leur santé. Dans le cadre de ce type d'intervention préventive, l'individu est passif, il s'agit simplement de respecter les lois et les règlements.

Puisqu'il s'agit d'initiatives s'attaquant à un large éventail de problématiques, bon nombre de ces interventions en santé sont qualifiées d'interventions complexes (Touati et Suarez-Herrera, 2012). Ainsi, ces interventions visent à provoquer des changements, et ce, de manière simultanée, chez les individus, les communautés et les organisations, ces changements étant interreliés. Les interventions sont perçues comme faisant partie intégrante de systèmes organisés d'actions dont le but est de résoudre une ou plusieurs problématiques, dans un contexte et une période donnés (Contandriopoulos *et al*., 2012). Ce système

comprend, dans un environnement donné, des acteurs, une structure (l'ensemble des ressources et des règles), des processus (la relation entre les ressources et les activités) et des objectifs (l'état futur vers lequel le processus d'action est orienté). Enfin, l'intervention est le produit de longues et de multiples chaînes causales qui agissent de manière synchrone, voire asynchrone, et qui mettent en relation l'ensemble des structures, des processus et des résultats (Champagne *et al*., 2011 ; Rogers, 2008). Aussi, voici les caractéristiques des interventions en santé que l'on qualifie de complexes et qu'il importe d'avoir en tête lorsqu'on planifie leur évaluation (Bamberger, Rugh et Mabry, 2012 ; Champagne *et al*., 2011 ; Contandriopoulos *et al*., 2012 ; Rogers, 2008 ; Shiell, Haw et Gold, 2008). Le premier exemple est relatif à la multiplicité des activités mobilisant de nombreux acteurs à la fois interdépendants et fortement autonomes comme les acteurs œuvrant dans un centre hospitalier. Les différents acteurs agissent en fonction de logiques différentes (leurs interactions sont animées de tensions permanentes, entre coopération et compétition) : chaque groupe d'acteurs (les médecins, le personnel soignant, les gestionnaires, etc.), en fonction de leur rôle, conçoit et agit selon leur représentation de l'intervention, de ses composantes, de ses caractéristiques et de ses finalités. Cet exemple illustre toute la difficulté de comprendre le fonctionnement d'un système national de soins de santé et d'agir pour le bonifier ou le rendre plus efficient.

Enfin, l'évaluation d'interventions en santé qualifiées de complexes doit pouvoir tenir compte de chaque fonction associée à ce type d'intervention, à savoir : l'atteinte des objectifs, l'adaptation à l'environnement, la production de services de qualité et le développement de valeurs communes entre les différents acteurs du programme (Contandriopoulos *et al*., 2012).

À présent, nous allons aborder l'épidémiologie évaluative et l'évaluation économique en santé, deux tendances évaluatives observées dans l'évaluation des interventions en santé. De plus, nous allons offrir une description de chacune de ces pratiques, et ce, en discutant de leurs raisons d'être et de leurs limites.

1.3. Deux tendances : l'épidémiologie évaluative et l'évaluation économique des interventions en santé

1.3.1. *L'épidémiologie évaluative*

Les différents volets d'interventions préventives en santé ainsi que l'épidémiologie évaluative s'inscrivent dans l'épidémiologie appliquée, soit « l'application et l'évaluation des découvertes et des méthodes

épidémiologiques aux domaines de la santé publique et des soins de santé » (Last, 2004, p. 79). Par conséquent, cette notion englobe certaines fonctions épidémiologiques citées plus tôt, notamment la définition des priorités en termes de prévention et d'intervention ainsi que l'évaluation des programmes, des politiques et de services de santé. Selon l'Organisation mondiale de la santé (OMS), les interventions en santé s'inscrivent dans la tradition scientifique de l'épidémiologie, et ce, pour justifier la rigueur et l'utilité de leurs procédures. De telles pratiques permettent notamment de réduire les écarts actuels caractérisant l'état de santé, et d'offrir à tous les individus les mêmes ressources et possibilités pour réaliser pleinement leur potentiel de santé (*Charte d'Ottawa*, 1986).

En outre, on remarque une augmentation du recours aux démarches évaluatives et un engouement pour les recherches évaluatives associées aux programmes et aux interventions visant à transformer les comportements et les conditions de vie (Potvin, Bilodeau et Gendron, 2008). En effet, les interventions en santé publique s'inscrivent de plus en plus dans des démarches de recherche, notamment en évaluation, celle-ci représentant un moyen de développer des connaissances pouvant informer l'action et modeler la mise en œuvre de projets d'intervention, etc. Ainsi, si l'épidémiologie permet d'obtenir des informations sur les conditions et l'objet que l'intervention doit transformer, la recherche évaluative est indispensable pour développer des connaissances sur la manière dont une intervention peut agir sur les conditions socio-environnementales (Potvin et Golberg, 2012). L'épidémiologie évaluative en santé semble donc s'inscrire dans une perspective d'appréciation normative et de recherche évaluative (Champagne *et al*., 2011).

Dans le cadre de l'appréciation normative, il s'agit de porter un jugement sur une intervention en comparant sa structure (intrants mobilisés), son processus (extrants et activités) et ses résultats, à des normes et des critères. Par conséquent, il s'agit de vérifier la relation entre le respect des normes et des critères, et les effets observés de l'intervention. Les procédés de vérification de la conformité des composantes de l'intervention s'apparentent aux procédures de vérification de la qualité. L'analyse de la structure et du processus permet de mesurer les écarts (entre les ressources et les normes de ressources, entre les coûts prévus et les coûts réels, entre les extrants et les normes d'extrants, etc.) et les coûts ; l'analyse du processus permet de mesurer la qualité des composantes de l'intervention et la couverture. Enfin, les résultats sont évalués par la prise de mesures directe. La validité de l'appréciation normative repose sur la force causale des liens déterminés par le postulat de départ, entre la structure, le processus et les résultats

de l'intervention (Champagne *et al.*, 2011). Ainsi, l'accent est fortement mis sur la qualité de la mesure et de la stratégie de mesure (procédures de contrôle permettant de garantir la validité et la fiabilité des inférences faites à partir des données obtenues grâce aux instruments de collecte de données).

La recherche évaluative constitue une activité de recherche, inscrite dans une démarche scientifique rigoureuse (tradition scientifique positiviste-postpositiviste), qui vise à analyser et à comprendre les relations de causalité entre les différentes composantes de l'intervention (comprendre le comment et le pourquoi des résultats) et la relation existante entre l'intervention et le contexte dans lequel elle est implantée. Elle permet en outre d'analyser la pertinence, la logique, la productivité, les effets et l'efficience d'une intervention. On retrouve six types d'analyse dans le cadre de la recherche évaluative en santé (Champagne *et al.*, 2011 ; Contandriopoulos *et al.*, 1993 ; Mark, 1990) : 1) l'analyse stratégique, qui porte sur l'analyse de marché, l'analyse des besoins, les méthodes de détermination des priorités, l'analyse de portefeuille, etc. ; 2) l'analyse de l'intervention, qui utilise des méthodes permettant de juger la qualité d'un modèle théorique (sa véracité et sa généralité). Il s'agit d'une généralisation des méthodes visant à vérifier la qualité des instruments de mesure ; 3) l'analyse de la productivité, qui consiste à analyser la manière dont les ressources sont utilisées pour produire les extrants. On parle de productivité physique (mesure en unités physiques) et de productivité économique (mesure en unités monétaires). Les méthodes mobilisées dans ce type d'analyse proviennent des méthodes économiques et de compatibilité analytique ; 4) l'analyse des effets, qui évalue l'influence des interventions sur les états de santé. On retrouve l'efficacité théorique (recherche de laboratoire en environnement contrôlé), l'efficacité d'essai (essais de randomisation), l'efficacité d'utilité (mesure des résultats sur les individus qui ont bénéficié de l'intervention, dans un contexte naturel, et observation du comportement des utilisateurs et des fournisseurs) ; 5) l'analyse de rendement, qui porte sur la relation entre l'analyse des ressources employées et l'analyse des effets obtenus (voir les méthodes de l'évaluation économique à la section 1.3.2) ; et, enfin, 6) l'analyse de l'implantation, qui, d'une part, mesure l'influence de la variation dans le degré d'implantation d'une intervention sur ses effets et, d'autre part, examine l'influence de l'environnement, du contexte dans lequel l'intervention est implantée, sur les effets de l'intervention.

On remarque trois défis relativement à l'évaluation d'interventions en santé (Potvin, Bilodeau et Gendron, 2008). Le premier concerne la complexité, la non-linéarité et l'instabilité dans le processus de définition de ce qui constitue l'objet, le programme, l'intervention en santé à

évaluer (autrement dit l'*evaluand*). En effet, un projet d'intervention de ce genre forme généralement un système dynamique qui évolue dans le temps, le programme devant s'adapter aux conditions d'un environnement constamment changeant, ce qui engendre de perpétuels remaniements dans le programme. Si ces modifications peuvent être minimes *a priori*, à long terme, les parties prenantes sont amenées à changer, créant ainsi de nouvelles collaborations et donc de nouveaux objectifs d'intervention, ce qui peut entraîner des changements considérables pour le programme (Bisset et Potvin, 2007). Il apparaît donc crucial de tenir compte, au moment de la définition de l'*evaluand*, de sa possible évolution, et d'anticiper ses différentes transformations. Le deuxième défi de l'évaluation des interventions en santé fait référence à la rigueur et à l'efficience de la méthode d'investigation. En effet, la pratique évaluative en santé publique s'inscrit dans la tradition scientifique de l'épidémiologie et de ses méthodes quantitatives, ses devis expérimentaux et ses devis scientifiques exigeants. Or, ces exigences se heurtent à la manifestation et à l'influence indéniable d'une dimension sociale, conduisant les chercheurs à emprunter les méthodes propres à la recherche en sciences sociales (Golberg *et al*., 2002). Enfin, le troisième et dernier défi, bien connu du monde de la recherche, est la conséquence ou plutôt la contribution directe de la gestion adéquate des deux premiers défis, à savoir l'apport de connaissances ainsi que les possibilités de généralisation.

Les trois défis qui viennent d'être présentés sont intrinsèquement liés à la confrontation entre les types d'études évaluatives propres à l'épidémiologie évaluative (appréciation normative et recherche évaluative) et les caractéristiques des interventions complexes (multiplicité des composantes, des organisations, des liens causals et des alternatives causales, etc.). En effet, il apparaît primordial que l'évaluation de ce type d'intervention tienne compte de la nature complexe de cette dernière, et de proposer des approches évaluatives adaptées. Aussi, des chercheurs tels que Rickles (2009) sont d'avis qu'il est temps de changer radicalement la manière dont nous investiguons les liens de causalité en ce qui a trait à l'évaluation des interventions complexes. De fait, les devis expérimentaux tels que les ECR (essais contrôlés randomisés) doivent à présent tenir compte du contexte dans lequel l'intervention et l'évaluation se déroulent (Campbell *et al*., 2007). Cependant, cela n'est pas suffisant, il semble nécessaire d'adopter des approches plus configurationnelles et développementales. En effet, évaluer ce type d'intervention signifie être en mesure de saisir la capacité de l'intervention à implanter et à maintenir une tension dynamique entre la réalisation de ces quatre fonctions (l'atteinte des objectifs, l'adaptation à l'environnement, la production de services de qualité et le développement

de valeurs communes entre les différents acteurs du programme). De plus, l'appréciation de la complexité ne repose pas uniquement sur la mesure d'indicateurs de réussite pour chacune de ces quatre fonctions (Contandriopoulos *et al.*, 2012). Elle repose également sur le caractère dynamique des processus qui relient ces quatre fonctions tout en s'appuyant sur la capacité des gestionnaires à coordonner les échanges et les négociations indispensables entre les divers acteurs du système d'action, et ce, afin qu'ils soient en mesure de préserver l'équilibre entre les quatre fonctions associées à ce type d'intervention. Aussi convient-il de recourir à des approches permettant l'accompagnement dans le développement et l'évaluation des interventions complexes, telles que les approches axées sur l'utilisation qui visent à soutenir, en temps réel, le développement d'interventions complexes caractérisées par leurs effets émergents et imprévisibles (Patton, 2008).

1.3.2. *L'évaluation économique*

L'évaluation tient une place importante dans la proposition et le maintien de programme d'intervention en santé publique. En effet, depuis ces vingt dernières années, nous remarquons une augmentation exponentielle des écrits en évaluation économique des services de santé (Husereau *et al.*, 2014). Il n'est plus question de s'assurer uniquement de l'efficacité d'une intervention, mais aussi de sa rentabilité. L'évaluation économique en santé représente une activité analytique importante, car les ressources humaines, financières, temporelles, matérielles, etc., disponibles sont rares, dans un contexte où les besoins sont illimités (Drummond *et al.*, 1993). Ainsi, l'évaluation économique permet, entre autres, aux gestionnaires responsables des ressources d'anticiper l'effet potentiel ou réel de changements apportés aux interventions en santé. De plus, par essence, l'évaluation économique consiste à « comparer les bénéfices et les coûts respectifs de deux interventions (programmes) ou plus. Son objet est de mesurer l'efficience, ou l'utilité de l'argent dépensé pour une intervention par rapport à une autre » (Goldsmith, Hutchison et Hurley, 2004, p. 7, traduction libre). Aussi, dans le contexte canadien actuel, remarque-t-on une accumulation de données probantes produites dans le cadre d'évaluation économique d'interventions en santé (prévention, promotion de la santé, etc.), données qui sont utilisées afin de définir des priorités en matière de santé publique, et de santé en général. En effet, dans le contexte de la recherche sur les services de santé, l'évaluation économique peut aider les chercheurs à démontrer le potentiel économique ou l'incidence économique réelle d'une nouvelle intervention sur le système de santé, encourageant ainsi l'adoption de cette intervention (Husereau *et al.*, 2014). L'évaluation économique se révèle ainsi complémentaire de la

recherche sur les services de santé, domaine de recherche qui s'attarde sur la relation entre les besoins, la demande, l'offre, l'utilisation et les effets des services de santé et porte particulièrement son attention sur les facteurs suivants : qualité, distribution, accessibilité, résultat et efficacité, et ce, sans tenir compte de qui les fournit (Lance, 2004).

L'évaluation économique implique donc toujours une comparaison de différentes possibilités d'actions, à la fois au regard des coûts et des résultats. Une approche pragmatique consiste à différencier quatre modèles méthodologiques selon le traitement des conséquences : études de minimisation des coûts, coût-efficacité, coût-utilité et coût-bénéfice. Chacune de ces méthodes d'analyse économique peut être utilisée dans le cadre de l'évaluation d'interventions préventives. L'un des principes économiques de base est la notion qu'un service de santé devrait offrir le plus grand avantage par coût unitaire. On peut estimer les avantages de différentes façons et celles-ci permettent d'effectuer quatre principaux types d'évaluation économique (Fleurbaey, Luchini et Schokkaert, 2009 ; Lance, 2004 ; Sailly et Lebrun, 1996) : 1) l'analyse coût-efficacité, soit le type d'évaluation économique le plus fréquemment utilisé. Ce type d'étude vise à comparer les coûts et les résultats (ou avantages) obtenus pour différentes stratégies. En considérant un indicateur physique à la fois (le nombre d'années de survie supplémentaires, par exemple), elle permet la classification des stratégies au regard de leurs contraintes budgétaires (ou coût financier) ; 2) l'analyse coût-utilité, soit une « version » enrichie des études de type coût-efficacité, puisqu'elle utilise comme indicateur la pondération de la qualité de vie (combinaison de la quantité et de la qualité de vie gagnée), afin de mesurer l'utilité de l'action évaluée (confrontation du coût de la stratégie au résultat de l'intervention). Ayant recours à des méthodes dites expérimentales, elle mobilise deux types d'indicateurs, les QALYs (*Quality Adjusted Life Years*), qui multiplient chaque année de vie par une valeur pondérée comprise entre 0 et 1 (qui traduisent l'évaluation de l'état de la santé) ; et les HYEs (*Healthy Years Equivalent*), soit le nombre hypothétique d'années de parfaite santé qui pourrait être considéré comme l'équivalent du nombre réel d'années d'un état de santé moins que parfait. L'un des avantages de ce type d'analyse est la comparaison de différentes interventions et de leurs résultats connexes ; 3) l'analyse coût-bénéfice, soit l'évaluation des coûts et des résultats en unités de valeur monétaire (le dollar). Par exemple, si un employé s'absente du travail en raison d'une maladie, la prévention ou la guérison de cette maladie confère un avantage économique direct. On parle d'estimations du coût économique des maladies ; 4) l'analyse de minimisation des coûts, soit la plus aisée parmi les différents

types d'étude économique. On l'utilise lorsque les indicateurs de résultats, soit les avantages de deux interventions, sont égaux. Ainsi, l'intervention à moindre coût sera celle qui sera mise en œuvre.

Comme nous l'avons mentionné précédemment, l'augmentation continuelle de la demande d'accès aux soins et aux services de santé se trouve perpétuellement confrontée au resserrement budgétaire. Aussi, l'évaluation économique des interventions en santé semble-t-elle être un outil incontournable pour la priorisation des services et des interventions de santé, tout en maintenant un système de santé accessible et de qualité. Cependant, il demeure primordial de garder à l'esprit que si les données probantes relatives aux évaluations économiques tiennent une part importante dans le processus décisionnel, les gestionnaires en santé tiennent également compte des travaux des cliniciens ainsi que des épidémiologistes avant de se prononcer sur une intervention (Drummond *et al*., 1993). De plus, si l'évaluation économique a su prouver son utilité, elle doit cependant relever de multiples défis reliés à la prise de décision et à l'utilisation des résultats en santé publique (Lance, 2004). Premièrement, la gestion des contraintes de réalisation n'est pas toujours possible puisqu'en fonction du devis évaluatif, des intrants mobilisés (ou à mobiliser) ainsi que du temps nécessaire, l'étroitesse de la fenêtre décisionnelle eu égard à la vie d'un programme ou d'une intervention ne permet pas la mise en œuvre de l'évaluation ou l'entrave considérablement. Deuxièmement, comme nous avons pu le voir plus tôt, le caractère complexe des modèles d'interventions en santé (et la mobilisation simultanée de trois volets aux angles d'intervention différents) ainsi que les délais très longs de manifestation de changements et la présence d'une multitude d'effets et de personnes concernées (intervenants, bénéficiaires, politiques, etc.) freinent les possibilités d'estimation des conséquences (avantages ou résultats) en prévention. Troisièmement, si l'évaluation économique joue un rôle important dans le processus décisionnel visant l'implantation et le maintien de programmes d'interventions, elle constitue toujours une limite dans le processus d'établissement des priorités parmi les programmes. Enfin, le manque de compréhension des résultats d'évaluations économiques avec la complexité des concepts, des méthodologies évaluatives, des résultats d'analyses économiques, etc., demeure un frein important au processus décisionnel des parties prenantes relativement au maintien d'un programme ou à la priorisation d'une stratégie d'intervention plutôt qu'une autre.

Dans la lignée de la synthèse que nous avons proposée à propos de l'épidémiologie évaluative, les caractéristiques des interventions complexes réduisent l'efficacité et l'utilisation des résultats de

l'évaluation économique en santé. En effet, l'évaluation économique en santé s'inscrit dans la mesure de la performance, une étude évaluative généralement utilisée à la fin du cycle de gestion d'un programme. Or, cinq aspects propres aux interventions complexes représentent des défis pour la mise en œuvre de telles évaluations (Rogers, 2008) :

- La multiplicité des organisations issues de domaines différents : cela accroît la charge de travail au regard de la négociation avec les partenaires des accords en ce qui a trait aux paramètres de l'évaluation et au regard de la réalisation de collectes et d'analyses de données efficaces.
- La multiplicité simultanée des liens de causalité : afin de garantir l'efficacité de leur intervention, les partenaires doivent maximiser l'opérationnalisation de multiples mécanismes causals (et non d'un seul) et l'évaluation doit documenter et soutenir ce processus.
- La multiplicité des mécanismes de causalité agissant dans différents contextes (mécanismes de causalité alternatifs) : les possibilités de réplication d'une intervention dépendent de la compréhension du contexte dans lequel elles sont implantées. Or, l'argument contraire peut se révéler nul quand il existe des solutions de rechange pour obtenir les résultats désirés.
- La non-linéarité et l'apparition d'effets disproportionnés (causalité rétroactive et boucle de renforcement) : à un stade critique, le moindre effet peut faire une grande différence et constituer un tournant décisif dans la vie de l'intervention. Ainsi, un petit effet initial peut engendrer un effet ultime majeur (boucle de renforcement).
- Les effets émergents : certaines mesures ne peuvent pas être élaborées à l'avance, ce qui rend difficile la prise de mesures préintervention et postintervention à des fins de comparaison.

Évaluer les effets des interventions complexes représente un défi majeur pour les évaluateurs, car l'enchaînement des mécanismes causals est si variable qu'il est difficile d'anticiper leurs effets et leur mesure. Cela démontre l'importance de ne pas uniquement se concentrer sur l'efficience et les répercussions des interventions en santé. L'évaluation constitue un construit multidimensionnel qui doit permettre aux parties prenantes de débattre et d'élaborer un jugement sur les caractéristiques essentielles de leur intervention, et ce, en fonction de leurs représentations, de leurs responsabilités, de leurs connaissances, etc. (Contandriopoulos *et al.*, 2012). L'évaluateur se doit d'encourager la participation des parties prenantes dans la planification de l'évaluation et l'élaboration des indicateurs de mesures, des temps d'évaluation, etc.

La complexité de ces interventions est telle qu'il est nécessaire d'inscrire le développement de l'intervention dans une approche évaluative et ainsi multiplier les périodes d'évaluation, et ce, tout au long de la vie de l'intervention : du diagnostic de besoin à la mesure de la performance. De telles pratiques s'inscrivent d'ailleurs dans des approches évaluatives centrées sur le développement du programme et l'utilisation des résultats évaluatifs et permettent, entre autres, d'anticiper les défis que posent les caractéristiques des interventions complexes pour le développement, la mise en œuvre et les évaluations efficaces (et utiles) de ces interventions.

L'évaluation des interventions en santé comporte de nombreux défis, défis qui nous amènent à souligner la nécessité de multiplier les évaluations de ces programmes, et ce, tout au long de leur existence : du diagnostic de besoin à la mesure de la performance. En outre, il nous apparaît nécessaire d'accroître le rôle de l'évaluation, afin de pouvoir développer, entre autres, des pratiques de monitorage efficaces du développement, de la mise en œuvre et des effets des interventions complexes en santé publique (et dans le domaine de la santé en général). Si nous ne remettons pas en cause l'utilité des pratiques évaluatives actuelles en santé publique (épidémiologie évaluative et évaluation économique en santé), nous suggérons de les combiner à des approches en évaluation de programme centrées sur l'utilisation des résultats, telles que l'évaluation développementale. En effet, l'évaluation développementale semble être une approche pertinente dans ces circonstances complexes, notamment grâce au rôle de l'évaluateur dans le développement, l'implantation et l'évaluation de l'intervention.

2. L'ÉVALUATION DÉVELOPPEMENTALE

2.1. Une description

Au fil des dernières décennies, le domaine de l'évaluation de programme a généré un ensemble considérable de connaissances, notamment sur les modèles d'efficience en matière d'implantation et de gestion de programmes (Patton, 2005). En effet, l'une des activités prédominantes dans la pratique des évaluateurs consiste à se prononcer sur la qualité du processus de développement et de l'implantation des programmes évalués. Aussi, sont-ils très souvent amenés à mettre en lumière les imperfections relatives aux activités de planification : évaluation de besoins incomplète, mauvaise répartition des ressources humaines et financières, description insuffisante de la population ciblée, objectifs contradictoires, implantation « bâclée », etc. (Patton, 1994). Ainsi, dans

la continuité du développement de son approche évaluative centrée sur l'utilisation des résultats (de l'évaluation) et du constat selon lequel l'évaluateur peut être un partenaire précieux dans le développement de programme, Patton (1994, 2008, 2016) propose l'évaluation développementale. Dans l'un de ses articles fondateurs sur l'évaluation développementale, Patton (1994) introduit son argumentation en dressant une liste des synonymes du verbe *développer*. Il mentionne, entre autres, les termes suivants : améliorer, évoluer, créer, changer, maturer, etc., soit autant de verbes illustrant les processus de développement et d'implantation de programmes ou d'interventions. Ces termes font d'ailleurs référence au caractère dynamique et itératif des programmes, élément qu'il est nécessaire d'anticiper, et ce, dès la planification. La planification constitue l'une des principales activités dans le développement de programme. Inscrite dans un processus logique et rationnel, elle se traduit par les activités suivantes : évaluer les besoins, proposer des solutions, cibler la population, fixer des objectifs, déterminer et procurer les ressources, implanter le programme et évaluer les résultats. En outre, selon Patton (2008), l'évaluation développementale assiste dans le développement de programme et d'organisations afin d'assurer leur adaptation aux réalités émergentes et dynamiques des systèmes d'action complexes. L'une des principales caractéristiques de cette approche est l'intégration de l'évaluateur dans l'équipe d'administrateurs chargée de la création, de la planification et de l'implantation d'un programme. Ainsi, il participe aux prises de décisions concernant ce programme (de sa création à son évaluation), et ce, en apportant à l'équipe une logique évaluative qui consiste, notamment, à collecter des données (ou rassembler des informations) de manière rigoureuse et systématique, afin de prendre des décisions éclairées (Patton, 1994).

L'évaluation développementale permet de soutenir la création et l'implantation d'interventions complexes, souvent de grande envergure, et ce, en facilitant le processus de cocréation entre les différents collaborateurs et l'évaluateur, la logique évaluative permettant de gérer les incertitudes liées aux multiples exigences de terrain, aux perpétuels changements environnementaux et aux conditions d'implantation (Asher *et al.*, 2016). En effet, une telle perspective développementale permet la mise en œuvre de multiples collectes d'informations, au fil du processus de création, d'implantation et de gestion des programmes, permettant ainsi aux gestionnaires et aux partenaires de prendre des décisions efficaces et avisées (Lam et Shulha, 2015). Enfin, il est important de mentionner que l'évaluation développementale doit être utilisée dans des situations appropriées. McKegg et Wehipeihana (2016) évoquent notamment le recours à ce type d'approche dans le cadre de situations où : 1) le contexte est émergent et « instable »

(environnement très dynamique) ; 2) les variables et les facteurs relatifs à l'implantation sont interdépendants et non-linéaires, ce qui rend les possibilités de planification et de prédiction difficiles ; 3) le contexte est socialement complexe du fait de la collaboration entre des parties prenantes provenant d'organismes, de secteurs ou de « cultures » différentes ; 4) le caractère innovant de la situation requiert un apprentissage constant et un perpétuel aller-retour entre l'*evaluand*, l'environnement et la planification, etc.

2.2. Exemple : le programme international de Médecine et les humanités

Nous travaillons actuellement à la création et à l'implantation d'un programme international de Médecine et les humanités (PMH) visant le partage d'expertise dans le domaine de la médecine et des sciences humaines entre plusieurs universités partout dans le monde. Inscrit dans une volonté de créer un programme d'innovation en pédagogie médicale d'envergure, le partenariat entre l'Université d'Ottawa[2], l'Université Jiao-Tong de Shanghai, l'Université de Lyon et l'Université de la médecine traditionnelle chinoise de Shanghai, doit permettre de : 1) faciliter la création et la diffusion de contenus en médecine et humanités pour améliorer les programmes de formation médicale des partenaires internationaux ; 2) bénéficier de l'expertise de professeurs de différentes disciplines en sciences humaines et sociales pour enrichir les curriculums des programmes d'études de premier cycle et de cycles supérieurs ; 3) favoriser chez leurs étudiants le partage de divers points de vue et perspectives sur les concepts éthiques et philosophiques importants ; et 4) contribuer à l'avancement des connaissances sur l'apport des concepts de sciences humaines en pédagogie médicale.

Compte tenu du caractère innovant et complexe de ce partenariat (soit un partenariat à niveaux multiples, la collaboration de professionnels de différentes disciplines, institutions et cultures, une multitude de projets, une non-linéarité due aux multiples chaînes causales, etc.), nous avons décidé d'inscrire sa conception dans un modèle évaluatif développemental. Ainsi, l'évaluateur développemental du PMH est en mesure de mettre à profit ses connaissances et son expérience en évaluation de programme, pour une planification rigoureuse et efficiente, et

2. L'équipe initiale chargée du développement et de l'implantation du programme international de Médecine et les humanités est constituée des professionnels de l'Université d'Ottawa suivants : le directeur du PMH, Jean Roy, M.D. ; la coordinatrice de l'implantation et de l'évaluation du PMH, Maud Mediell ; le coordinateur du bureau d'internationalisation, Jonathan Gendron-Rossignol ; une experte de contenu en MH, Isabelle Burnier, M.D.

pour limiter ainsi les erreurs de planification, souvent très coûteuses (en temps, ressources humaines et financières, etc.). Ainsi, à ce jour, différents outils de planification de programme tels que le modèle logique et la charte d'implantation (soit deux outils que l'évaluateur désire avoir au moment de l'évaluation et qui sont souvent inexistants) ont pu être développés. Le modèle logique est un outil visuel nous permettant de décrire le programme dans sa logique (ou sa théorie) fondamentale. Ainsi, il s'agit d'un diagramme illustrant le contenu (quoi ?), les destinataires (qui ?) et la raison d'être (pourquoi ?) du programme (Porteous, 2012). Il facilite l'obtention d'une vision commune (entre les différentes parties prenantes) de la logique du programme. Il permet également de s'assurer de la linéarité entre les objectifs, les intrants, les extrants et les effets désirés. Cependant, dans le cadre de programmes ou d'interventions à la fois compliqués (multisites et multiniveaux) et complexes (effets émergents), le modèle logique doit fournir une structure commune permettant d'accommoder les changements et les adaptations sur le plan local (Rogers, 2008). En outre, la représentation de la théorie du programme doit refléter la complexité et la non-linéarité de ce type d'interventions et articuler de manière efficace les multiples mécanismes de causalité impliqués dans la manifestation de changements prévus et émergents (Barnes, Matka et Sullivan, 2003). Ce modèle logique facilite la détermination des différents temps d'évaluation ainsi que l'identification des experts qu'il sera nécessaire d'impliquer dans la création des activités, et ce, afin de développer les indicateurs de mesure qui seront utilisés tout au long de l'évaluation. Enfin, c'est un très bon outil de négociation des prises de décision en ce qui a trait à la structure, au contenu du programme ainsi qu'aux intrants nécessaires pour mener à bien ce projet d'évaluation développementale. De plus, il permet de familiariser les parties prenantes avec les aspects complexes de leur intervention (Rogers, 2008). La charte d'implantation, constitue une version affinée, plus complexe, du modèle logique puisqu'elle permet de décrire la planification et les processus, et d'illustrer les procédures d'opérationnalisation des différents objectifs d'implantation du programme et de leurs activités respectives. Dans la lignée des *SMART Goals*[3], cet outil permet de définir, pour chaque but et chaque objectif : les intrants nécessaires, les personnes responsables, les activités, les résultats attendus, les résultats observés ou la situation d'avancement.

3. *SMART Goals* est un cadre de bonnes pratiques aidant à fixer des objectifs. Un objectif *SMART* doit être spécifique, mesurable, réalisable, réaliste et limité dans le temps. Souvent utilisé pour les examens de rendement, l'acronyme est destiné à aider un gestionnaire ou un autre employé qui est chargé de fixer des objectifs dans le cadre de l'élaboration ou de l'évaluation d'une intervention. Ce type de charte lui permet d'établir exactement ce qui sera nécessaire pour atteindre ses objectifs et de partager cette clarification avec les autres membres du projet.

Par conséquent, une fois le modèle logique et la charte d'implantation finalisés, l'équipe initiale de développement et d'implantation du PMH sera en mesure de rencontrer : 1) les différents partenaires internationaux afin d'obtenir leur rétroaction et d'être en mesure d'apporter les modifications nécessaires au programme ; 2) les experts que nous désirons impliquer dans le développement du contenu des activités afin de définir les indicateurs de mesure évaluative ; et 3) les potentiels organismes de financement extérieurs afin de faciliter la bonne mise en œuvre des diverses évaluations ainsi que le développement de la recherche sur la contribution des concepts de sciences humaines en pédagogie médicale.

2.3. Les défis de l'évaluation développementale

Selon nous, le type de projet que nous venons de présenter constitue un excellent exemple du rôle que peut tenir un évaluateur de programme au sein d'un partenariat interdisciplinaire en santé. En effet, il permet de faire la promotion de l'utilisation de l'évaluation développementale dans le cadre de la création et de l'implantation de projets innovants en santé, tant sur le plan national qu'international. Impliquer un évaluateur de programme dès le stade de la création et de la planification de projet permet de limiter les erreurs de planification, erreurs que les évaluateurs sont habitués à observer lorsqu'ils arrivent trop tard dans la vie de programmes aussi complexes.

Cependant, inscrire un programme ou une intervention complexes en santé dans une approche développementale soulève des défis pour l'évaluateur, mais aussi pour l'équipe de gestionnaires. En effet, le rôle (et les tâches) attribué à l'évaluateur développemental demeure très exigeant (Rey *et al.*, 2013). En effet, il doit à la fois soutenir le développement et l'implantation d'une intervention, faciliter la relation partenariale, soutenir l'utilisation du processus et des résultats d'évaluation et garantir la qualité de l'évaluation et des connaissances qu'elle produit sur les caractéristiques de l'intervention, ses effets, mais aussi sur le développement et la réalisation d'une évaluation développementale. Bien entendu, l'évaluateur fait partie d'une équipe interdisciplinaire de gestionnaires, au sein de laquelle chaque membre détient sa propre expertise et ses propres responsabilités. Cependant, une telle collaboration implique que chaque membre de l'équipe se familiarise avec les pratiques et les rôles de chacun, tout en s'assurant de comprendre ce qui se passe en tout temps. Ce dernier point est crucial, compte tenu de la complexité et de la non-linéarité de telles interventions. La perspective développementale exige une capacité d'anticipation et de rétroaction importante. Il s'agit de planifier les temps d'évaluation (mobilisation

des intrants), les temps de restructuration, le recrutement d'experts de contenu pour le développement des indicateurs de mesure et le développement de méthodologies évaluatives rigoureuses pour chaque composante de l'intervention (évaluation des apprentissages, évaluation économique, etc.), etc. De surcroît, il est primordial de garder à l'esprit que l'évaluateur développemental ne peut pas (et ne doit pas) être chargé de toutes ces activités. En effet, ce dernier fait partie de l'équipe à titre de coordonnateur de l'approche développementale et des différents temps d'évaluation et de restructuration ; il partage son expérience et ses connaissances en développement et en évaluation de programme.

Enfin, inscrire la création et l'implantation d'une intervention complexe dans une approche évaluative développementale se révèle chronophage et très coûteux en ressources humaines, matérielles et financières. Il convient donc de garder à l'esprit ces contraintes, car elles constitueront, entre autres, des points de tension et de négociation au sein des différents groupes d'acteurs (partenaires, gestionnaires, experts, intervenants, etc.).

CONCLUSION

L'actualité canadienne et les enjeux politiques, socio-économiques et culturels en santé publique (et dans le domaine de la santé en général) complexifient la pratique, la recherche et l'évaluation dans le domaine de la santé. Aussi, le caractère complexe des interventions en santé limite-t-il, selon nous, l'efficacité des pratiques évaluatives actuelles en santé publique telles que l'épidémiologie évaluative et l'évaluation économique en santé. Dans le cadre de l'évaluation de ce type d'intervention, les pratiques recensées semblent arriver trop tard dans la vie du programme ou peuvent paraître inadéquates par rapport à la complexité des interventions et de l'environnement dans lequel elles sont implantées. De fait, le temps est propice à l'établissement de nouvelles alliances entre les pratiques évaluatives en santé publique et les approches en évaluation de programme centrées sur l'utilisation des résultats, telles que l'évaluation développementale. En effet, nous venons de voir que le recours à l'approche développementale dans le cadre de la conception, de la mise en œuvre, du monitorage et de l'évaluation de programmes permet de soutenir, en temps réel, la conception des interventions complexes en santé caractérisées par leurs effets émergents et imprévisibles. Dans le cadre de programmes de ce type, l'évaluation (l'évaluateur) permet aux parties prenantes de débattre et d'élaborer un jugement sur les caractéristiques essentielles de leur intervention, et ce, en fonction de leurs représentations, de

leurs responsabilités, de leurs connaissances, etc. De plus, leur participation est fortement encouragée dans la planification de l'évaluation et l'élaboration des indicateurs de mesures, des temps d'évaluation, etc., un élément déterminant en ce qui a trait à l'utilité et à l'utilisation des résultats d'évaluation. Les praticiens de l'évaluation de programme semblent donc en mesure de tenir un rôle déterminant, renforçant, selon nous, l'efficacité des interventions complexes en santé, et ce, en facilitant le développement rigoureux, l'implantation et la gestion de programmes innovants, au service de la santé et du bien-être des individus, des communautés et de la population canadienne.

Enfin, si l'évaluation développementale et les évaluateurs de programme semblent avoir beaucoup à offrir dans le cadre de la conception, de la planification et de l'évaluation des interventions complexes en santé, il apparaît essentiel de documenter de telles pratiques afin de contribuer au domaine de la recherche et de l'évaluation en santé, et de faire la promotion d'approches développementales. Ainsi, multiplier les études empiriques relativement à l'utilisation de l'évaluation développementale permettrait de produire de larges éventails de connaissances, notamment sur la conception de ce type d'approches, les stratégies que les évaluateurs « développementaux » déploient pour s'adapter à la complexité des interventions à évaluer (et leur environnement) et pour garantir l'utilité et l'utilisation du processus et des résultats évaluatifs.

BIBLIOGRAPHIE

Asher, J., N. Foote, J. Radner et T. Warren (2016). « Sciences and How We Care for Needy Young Children : The Frontiers of Innovation Initiative », dans M.Q. Patton, K. McKegg et N. Wehipeihana (dir.), *Developmental Evaluation Exemplars : Principles in Practice*, New York : Guilford Press, p. 103-124.

Bamberger, M., J. Rugh et L. Mabry (2012). « Evaluating Complicated, Complex, Multicomponent Programs », dans M. Bamberger, J. Rugh et L. Mabry (dir.), *Real World Evaluation : Working Under Budget, Time, Data, and Political Constraints*, 2e éd., Thousand Oaks : Sage Publications, p. 393-422.

Barnes, M., E. Matka et H. Sullivan (2003). « Evidence, understanding and complexity : Evaluation in non-linear systems », *Evaluation*, *9*(3), p. 65-84.

Bisset, S.L. et L. Potvin (2007). « Expanding our conceptualization of program implementation : Lessons from the genealogy of a school-based nutrition program », *Health Education Research*, *22*(7), p. 37-48.

Campbell, N.C., E. Murray, J. Darbyshire, J. Emery, A. Farmer, F. Griffiths *et al.* (2007). « Designing and evaluating complex interventions to improve health care », *British Medical Journal*, *334*(7591), p. 455-459.

Champagne, F., A. Brousselle, Z. Hartz et J.-L. Denis (2011). « L'évaluation dans le domaine de la santé : concepts et méthodes », dans A. Brousselle, F. Champagne, A.-P. Contandriopoulos et Z. Hartz (dir.), *L'évaluation : concepts et méthodes*, 2e éd., Montréal : Les Presses de l'Université de Montréal, p. 34-56.

Charte d'Ottawa pour la promotion de la santé (1986). WHO/HPR/HEP/95.1, OMS, Genève.

Contandriopoulos, A.-P., F. Champagne, J.L. Denis et R. Pineault (1993). « L'évaluation dans le domaine de la santé : concepts et méthodes », *Bulletin*, *33*(1), p. 12-17. Version révisée de l'article des mêmes auteurs dans T. Lebrun, J.C. Sailly et M. Amourette (dir.), *Actes du colloque* L'évaluation en matière de santé : des concepts à la pratique, GREGE, 1992.

Contandriopoulos, A.-P., L. Rey, A. Brousselle et F. Champagne (2012). « Évaluer une intervention complexe : enjeux conceptuels, méthodologiques et opérationnels », *Revue canadienne d'évaluation de programme*, *26*(3), p. 1-16.

Craig, P., P. Dieppe, S. Macintyre, S. Michie, I. Nazareth et M. Petticrew (2008). *Developing and Evaluating Complex Interventions : The New Medical Research Council Guidance*, Londres : Medical Research Council.

Detels, R. (2009). « Epidemiology : The foundation of public health », dans R. Detels, M. Gulliford, Q.A. Karim et C.C. Tan, *Oxford Textbook of Global Public Health, Volume 2*, 6e éd., New York : Oxford University Press, section 5.1.

Drummond, M., A. Bradnt, B. Luce et J. Rovira (1993). « Standardizing methodologies for economic evaluation in health care : Practice, problems, and potential », *International Journal of Technology Assessment in Health Care*, *9*(1), p. 26-36.

Fleurbaey, M., S. Luchini et E. Schokkaert (2009). « Évaluation économique en santé : qui a peur de l'étalon monétaire ? », *Revue de philosophie économique*, *1*(10), p. 19-34.

Goldberg, M., M. Melchior, A. Leclerc et F. Lert (2002). « Les déterminants sociaux de la santé : apports récents de l'épidémiologie sociale et des sciences sociales de la santé », *Sciences sociales et santé*, *20*(4), p. 72-128.

Goldsmith, L., B. Hutchison et J. Hurley (2004). « Economic evaluation across the four faces of prevention : A Canadian perspective, a discussion paper », Centre for Health Economics and Policy Analysis, McMaster University, Ontario, Canada, <http://evaluationcanada.ca/distribution/200405_goldsmith_laurie_hutchison_brian_hurley_jeremiah.pdf>, consulté le 25 avril 2017.

Husereau, D., P. Jacobs, B. Manns, T. Hoomans, D. Marshall et R. Tamblyn (2014). « Economic evaluation of complex health system interventions : A discussion paper », Institute of Health Economics, Edmonton, AB, <http://www.ihe.ca/publications/economic-evaluation-of-complex-health-system-interventions-a-discussion-paper>, consulté le 25 avril 2017.

Lam, C.Y. et L.M. Shulha (2015). « Insight on using developmental evaluation for innovating : A case study on the co-creation of an innovative program », *American Journal of Evaluation*, *36*(3), p. 358-374.

Lance, J.-M.R. (2004). *Les approches d'évaluation économique utiles au champ de la santé publique*, Communication préparée dans le cadre des Journées annuelles de santé publique (JASP), novembre, Montréal, Québec.

Last, J.M. (2007). *A Dictionary of Public Health*, New York : Oxford University Press.

Last, J.M. (2009). *Dictionnaire d'épidémiologie, enrichi d'un lexique anglais-français*, Acton Valle, Québec : Edisem/Maloine.

Mark, M. (1990). « From program theory to tests of program theory », *New Directions for Evaluation*, *47*, p. 37-52.

McKegg, K. et N. Wehipeihana (2016). « Developmental evaluation in synthesis. Practitioners' perspectives », dans M.Q. Patton, K. McKegg et N. Wehipeihana (dir.), *Developmental Evaluation Exemplars : Principles in Practice*, New York : Guildford Press, p. 271-288.

Morrel, J.A. (2005). « Complex Adaptative Systems », dans S. Mathison (dir.), *Encyclopedia of Evaluation*, Thousand Oaks : Sage Publications, p. 71-72.

Morris, J.N. (2007). « Uses of epidemiology », *International Journal of Epidemiology*, *36*(6), p. 1165-1172.

Patton, M.Q. (1994). « Developmental evaluation », *American Journal of Evaluation*, *15*(3), p. 311-319.

Patton, M.Q. (2005). « Developmental Evaluation », dans S. Matthison (dir.), *Encyclopedia of Evaluation*, Thousand Oaks : Sage Publications, p. 117.

Patton, M.Q. (2008). « Evaluation Focus Options : Developmental Evaluation and Other Alternatives », dans *Utilization-Focused Evaluation*, 4e éd., Thousand Oaks : Sage Publications, p. 271-307.

Patton, M.Q. (2016). « State of the Art and Practice of Developmental Evaluation », dans M.Q. Patton, K. McKegg et N. Wehipeihana (dir.), *Developmental Evaluation Exemplars : Principles in Practice*, New York : Guilford Press, p. 1-24.

Poissant, C. (2011). *Une démarche participative et négociée pour l'exercice de l'évaluation : Cadre de référence à la Direction de santé publique et d'évaluation de l'Agence de la santé et des services sociaux de Lanaudière. Service de surveillance, recherche et évaluation*, Québec, Agence de la santé et des services sociaux de Lanaudière.

Porteous, N.L. (2012). « La construction du modèle logique d'un programme », dans V. Ridde et C. Dagenais (dir.), *Approches et pratiques en évaluation de programme*, 2e éd., Montréal : Les Presses de l'Université de Montréal, p. 89-108.

Potvin, L., A. Bilodeau et S. Gendron (2008). « Trois défis pour l'évaluation en promotion de la santé. Promotion de la santé : besoins de recherches francophones et perspectives », *International Journal of Health Promotion and Education*, Hors-série, p. 17-21.

Potvin, L., A. Bilodeau et S. Gendron (2011). « Trois conceptions de la nature des programmes : implications pour l'évaluation de programmes complexes en santé publique », *Revue canadienne d'évaluation de programme*, *26*(3), p. 91-104.

Potvin, L. et C. Goldberg (2012). « Two Roles of Evaluation in Transforming Health Promotion Practice », dans I. Rootman, S. Dupérée, A. Pederson et M. O'Neill (dir.), *Health Promotion in Canada : Critical Perspectives on Practice*, 3e éd., Toronto, Canada : Canadian Scholars' Press Inc., p. 254-265.

Rey, L., A. Brousselle, N. Dedobbeleer et M.-C. Tremblay (2013). « Les défis de l'évaluation développementale en recherche : une analyse d'implantation d'un projet "Hôpital promoteur de santé" », *Revue canadienne d'évaluation de programme*, *28*(1), p. 1-26.

Rickles, D. (2009). « Causality in complex interventions », *Medicine Health Care and Philosophy*, *12*(1), p. 77-90.

Rogers, P.J. (2008). « Using programme theory to evaluate complicated and complex aspects of interventions », *Evaluation*, *14*(1), p. 29-48.

Sailly, J.-C. et T. Lebrun (1996). « L'évaluation économique des actions de santé », *Actualité et dossier en santé publique*, *17*, p. 26-31.

Shiell, A., P. Hawe et L. Gold (2008). « Complex interventions or complex systems ? Implication for health economic evaluation », *British Medical Journal*, *336*(7656), p. 1281-1283.

Touati, N. et J.C. Suarez-Herrera (2012). « L'évaluation des interventions complexes : quelle peut être la contribution des approches configurationnelles ? », *Revue canadienne d'évaluation de programme*, *25*(3), p. 17-35.

CONCLUSION

Eric Dionne et Isabelle Raîche

Cet ouvrage collectif visait à présenter des réflexions à caractère théorique et aussi des résultats de recherche sur les plus récentes avancées autant en mesure qu'en évaluation des apprentissages en éducation médicale. Les trois premiers chapitres ont permis de constater que la modélisation de la mesure qui s'appuie sur les traits latents offre des avantages importants au regard, entre autres, de la qualité de la mesure. On constate d'ailleurs que de plus en plus de chercheurs en éducation médicale, intéressés par la mesure, ont de plus en plus recours à ce type de modélisation. Compte tenu des enjeux souvent élevés en éducation médicale, il n'est pas étonnant de constater l'intérêt grandissant pour les aspects théoriques et empiriques associés à l'acte de mesurer. On peut facilement projeter que les prochaines années seront fastes en développement pour tout ce qui touche à la mesure.

En ce qui concerne l'évaluation, les quatre derniers chapitres nous ont montré que des développements intéressants touchent également la démarche visant à porter un jugement sur les apprentissages. Le concept de validité, loin d'être figé, est constamment remis en question et le terrain de l'éducation médicale offre un terreau fertile pour une telle réflexion. Les travaux qui concernent la correction automatisée annoncent également des avancées importantes. En effet, l'approche par compétences exerce une pression sur les dispositifs d'évaluation afin de les rendre plus

cohérents avec les nouvelles perspectives en éducation médicale. L'instrumentation n'est pas en reste puisque les tests de concordance de script sont de plus en plus utilisés comme modalité d'évaluation formative de la prise de décision clinique, qui est une compétence très étudiée en éducation médicale sous divers angles. Enfin, l'évaluation de programme sera de plus en plus mise à profit pour mieux comprendre les programmes de formation en éducation médicale. En effet, les systèmes éducatifs doivent, de plus en plus, faire la démonstration de leur efficacité afin de justifier les budgets considérables qui leur sont consentis.

NOTICES BIOGRAPHIQUES

ANDRÉ-PHILIPPE BOULAIS a commencé sa carrière dans le domaine de l'évaluation au début des années 1980 au Service d'évaluation de l'Association des infirmières et infirmiers du Canada et y a occupé le poste d'agent d'élaboration des examens pendant 12 ans. Il a par la suite poursuivi sa carrière en mesure et évaluation auprès du Collège royal des médecins et chirurgiens du Canada où il a occupé le poste d'associé à la recherche pendant deux ans. Il s'est ensuite joint au Conseil médical du Canada à titre de gestionnaire au Bureau d'évaluation où, s'appuyant sur plus de 30 ans d'expérience en évaluation, il a géré avec succès l'Examen d'aptitude, partie I. Monsieur Boulais a par la suite concentré ses efforts et intérêts dans la recherche et occupe actuellement le poste de chercheur-psychométricien principal au sein du Département de psychométrie et services docimologiques. Il détient une maîtrise en éducation de l'Université d'Ottawa et un baccalauréat spécialisé en psychologie de la même université.

LYNN CASIMIRO occupe le poste de vice-présidente à l'enseignement et à la réussite scolaire (VPE) à La Cité. Outre d'assumer ce poste, elle continue à mener des travaux de recherche à titre de chercheuse à l'Institut du savoir Montfort. Elle s'intéresse particulièrement à l'offre active de services sociaux et de santé en français en contexte linguistique minoritaire.

BERNARD CHARLIN est professeur au Département de chirurgie de l'Université de Montréal. Il a été formé à Montpellier, en France, comme chirurgien cervico-facial. Il détient une maitrise en éducation de l'Université Harvard et un Ph.D. en éducation de l'Université de Maastricht. Son champ de recherche est le raisonnement en situation d'incertitude (théorie, acquisition, évaluation). Il est le créateur des concepts de concordance de script et de formation par concordance. Il a écrit ou coécrit plus de 130 articles dans la littérature avec comité de pairs. Le Collège royal des médecins et chirurgiens du Canada lui a décerné en 2015 le prix Duncan-Graham pour souligner sa contribution remarquable à l'éducation médicale.

ANDRÉ F. DE CHAMPLAIN, Ph. D., occupe le poste de directeur du Département de psychométrie et des services docimologiques au Conseil médical du Canada (CMC). Il participe à un certain nombre de nouvelles initiatives du Conseil, y compris à l'analyse des méthodes actuelles de notation et d'établissement de normes pour les examens du CMC ainsi qu'à plusieurs études visant à mieux étayer et appuyer les politiques et les projets actuels.

ERIC DIONNE est professeur agrégé à la Faculté d'éducation et à la Faculté de médecine de l'Université d'Ottawa au Département d'innovation en éducation médicale (DIEM) et au Centre d'appui pédagogique en santé pour la francophonie (CAPSAF). Directeur de la revue *Mesure et évaluation en éducation*, il est également directeur du Groupe de recherche interuniversitaire en mesure et évaluation des apprentissages en santé (GRIMEAS). Il est aussi chercheur à l'Institut du savoir de l'Hôpital Montfort (ISM) et membre de l'Observatoire interuniversitaire sur les pratiques innovantes d'évaluation des apprentissages (OPIÉVA). Il se spécialise dans la construction et l'étude des propriétés métriques d'instruments permettant l'évaluation de construits complexes dans différents domaines disciplinaires dont celui de la santé.

JULIE GRONDIN détient un baccalauréat en mathématiques, un certificat en informatique et une maîtrise en mesure et évaluation de l'Université de Montréal. Son entreprise DociMétrique collabore à différents projets de développement d'épreuves et d'analyse de données et de qualité psychométriques des tests, ainsi qu'à différents projets de recherche mettant à profit la modélisation de Rasch. Elle œuvre dans différents champs de pratique comme l'éducation, les sciences sociales ou les sciences de la santé. Elle est l'auteure de plusieurs écrits et communications scientifiques.

MARIE-EVE LATREILLE a obtenu un baccalauréat en sciences infirmières en 2006 ainsi qu'une maîtrise *ès arts* en éducation en 2012 de l'Université d'Ottawa. Elle a effectué et publié une recherche sur la mesure et l'évaluation du raisonnement clinique d'étudiant en sciences infirmières à l'aide d'un test de concordance de script. Elle possède plusieurs années d'expérience en soins infirmiers pédiatriques dans une variété d'unités, y compris aux soins intensifs, à l'urgence et aux soins néonataux ainsi que dans les unités de chirurgie, médecine et d'oncologie. Elle occupe le poste de facilitatrice de laboratoire de simulation clinique à l'Université d'Ottawa depuis 2009.

NATHALIE LOYE est professeure spécialisée en mesure-évaluation à l'Université de Montréal. Ses travaux portent principalement sur l'évaluation diagnostique, tant en ce qui a trait aux modalités permettant de la réaliser qu'aux modèles psychométriques visant à analyser les données issues de tests diagnostiques. Elle s'intéresse de manière générale aux modèles de mesure tels que les modèles de classification diagnostique (MCD), le modèle de Rasch ou les différents modèles issus de la théorie de réponse à l'item (TRI). La validité est au cœur de ses préoccupations, tant en ce qui concerne sa définition que les démarches visant à l'attester.

MAUD MEDIELL est doctorante à la Faculté d'éducation de l'Université d'Ottawa. Elle détient un Master 2 (professionnel et recherche) en psychologie sociale de la santé ainsi qu'un diplôme d'études supérieures en évaluation de programme. Elle est consultante en évaluation de programme et travaille actuellement en tant qu'associée de recherche en sciences humaines au Centre d'appui pédagogique en santé pour la francophonie (CAPSAF) de la Faculté de médecine de l'Université d'Ottawa.

MAXIM MORIN est doctorant en mesure et évaluation à l'Université de Montréal et chercheur-psychométricien au Conseil médical du Canada. Il s'intéresse particulièrement à la modélisation des données ainsi qu'aux applications des techniques de traitement du langage naturel en évaluation des apprentissages.

THOMAS PENNAFORTE est pédiatre-néonatologiste formé en France. Il a réalisé un postdoctorat en néonatalogie au Québec et exerce actuellement à l'Hôpital du Sacré-Cœur de Montréal. Son axe de recherche concerne la simulation et l'évaluation du raisonnement clinique, au travers d'une maitrise en éducation médicale puis d'un passage accéléré au doctorat, qu'il poursuit actuellement à la Faculté des sciences de

l'éducation de l'Université de Montréal. Sa recherche a d'ores et déjà été diffusée dans plusieurs congrès internationaux et par le biais de publications. Enfin, le docteur Thomas Pennaforte a bénéficié des prestigieuses bourses Vanier et FRQSC (Fonds de Recherche du Québec – Société et culture).

ISABELLE RAÎCHE est professeure adjointe au Département de chirurgie à la Faculté de médecine de l'Université d'Ottawa. En plus de sa pratique professionnelle à titre de chirurgienne, elle mène des recherches sur l'éducation médicale et plus particulièrement sur la salle d'opération en tant qu'environnement pouvant favoriser l'enseignement, l'apprentissage et l'évaluation des résidents.

JEAN-SÉBASTIEN RENAUD est professeur agrégé au Département de médecine familiale et de médecine d'urgence à la Faculté de médecine et directeur du secteur Évaluation du Vice-décanat à la pédagogie et au développement professionnel continu (VDPDPC). Il détient un doctorat et une maîtrise en mesure et évaluation de l'Université Laval, ainsi qu'un baccalauréat en sciences de la consommation et un certificat en administration de la même institution. Ses travaux de recherche visent à faire avancer la sélection des candidats et l'évaluation des compétences, ainsi qu'à faire progresser la qualité de la formation pour ultimement former de meilleurs professionnels de la santé. Il s'intéresse plus précisément au développement et à la validité des méthodes d'évaluation en pédagogie des sciences de la santé.

JACINTHE SAVARD est professeure agrégée et directrice du programme d'ergothérapie de l'École des sciences de la réadaptation de l'Université d'Ottawa, membre du Groupe de recherche sur la formation des professionnels en santé et service social en contexte francophone minoritaire (GReFoPS) et de l'Institut du savoir Montfort-recherche. Sa recherche se concentre sur l'accès aux services sociaux et de santé en français pour les francophones en situation minoritaire. Elle a élaboré une mesure des comportements d'offre active et un cadre d'analyse des leviers d'action pour la prestation de services intégrés et linguistiquement adaptés dans les communautés de langue officielle en situation minoritaire.

Depuis une centaine d'années, des chercheurs ont développé des théories, des modèles et des techniques permettant de mieux mesurer ou observer les apprentissages des étudiants. Le domaine de la pédagogie des sciences de la santé, pour sa part, est le terrain de nombreuses avancées en mesure et évaluation depuis les années 1990. Cependant, le secteur des construits complexes reste à développer, ce qui suscite de nombreux questionnements. Comment mesurer le jugement clinique ? Sur quelles théories édumétriques doit-on s'appuyer pour étudier les instruments de mesure ? Comment automatiser la notation ? Comment s'assurer de la validité des inférences lors de la mesure ou de l'observation du jugement clinique ? Comment mieux comprendre les mécanismes qui permettent de rendre compte du succès d'un programme d'intervention ? C'est à ces questions que le présent ouvrage veut répondre.

Mesure et évaluation des compétences en éducation médicale est le fruit de la 37e conférence de l'Association pour le développement des méthodologies d'évaluation en éducation (ADMÉÉ-Canada), qui s'est tenue à Gatineau (Québec) en novembre 2015. Il fait état des plus récents travaux de chercheurs qui étudient la mesure et l'évaluation dans le domaine de l'éducation médicale et des pédagogies des sciences de la santé, des échanges qui ont découlé de leur présentation et des défis posés par les problématiques évoquées. Il s'adresse autant aux professionnels qu'aux chercheurs et étudiants qui s'intéressent à ces questions.

ERIC DIONNE est professeur agrégé à la Faculté d'éducation et à la Faculté de médecine de l'Université d'Ottawa. Directeur de la revue Mesure et évaluation en éducation *et du Groupe de recherche interuniversitaire en mesure et évaluation des apprentissages en santé (GRIMÉAS), il est aussi chercheur à l'Institut du savoir de l'hôpital Montfort (ISM-recherche) et membre de l'Observatoire interuniversitaire sur les pratiques innovantes d'évaluation des apprentissages (OPIÉVA).*

ISABELLE RAÎCHE est professeure adjointe au Département de chirurgie à la Faculté de médecine de l'Université d'Ottawa. En plus de sa pratique professionnelle à titre de chirurgienne, elle mène des recherches touchant à l'éducation médicale et plus particulièrement à l'étude de la salle d'opération comme environnement pouvant favoriser l'enseignement, l'apprentissage et l'évaluation des résidents.

ONT COLLABORÉ À CET OUVRAGE

André-Philippe Boulais
Lynn Casimiro
Bernard Charlin
André F. De Champlain
Eric Dionne
Julie Grondin
Marie-Eve Latreille
Nathalie Loye
Maud Mediell
Maxim Morin
Thomas Pennaforte
Isabelle Raîche
Jean-Sébastien Renaud
Jacinthe Savard