

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

L'ÉTABLISSEMENT DE STANDARDS DE PERFORMANCE POUR ÉVALUER
LES COMPÉTENCES PROFESSIONNELLES MINIMALES REQUISES
POUR EXERCER LA PROFESSION D'ACUPUNCTEUR AU QUÉBEC

MÉMOIRE
PRÉSENTÉ
COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN ÉDUCATION

PAR
ISABELLE DION

JUILLET 2021

UNIVERSITÉ DU QUÉBEC À MONTRÉAL
Service des bibliothèques

Avertissement

La diffusion de ce mémoire se fait dans le respect des droits de son auteur, qui a signé le formulaire *Autorisation de reproduire et de diffuser un travail de recherche de cycles supérieurs* (SDU-522 – Rév.10-2015). Cette autorisation stipule que «conformément à l'article 11 du Règlement no 8 des études de cycles supérieurs, [l'auteur] concède à l'Université du Québec à Montréal une licence non exclusive d'utilisation et de publication de la totalité ou d'une partie importante de [son] travail de recherche pour des fins pédagogiques et non commerciales. Plus précisément, [l'auteur] autorise l'Université du Québec à Montréal à reproduire, diffuser, prêter, distribuer ou vendre des copies de [son] travail de recherche à des fins non commerciales sur quelque support que ce soit, y compris l'Internet. Cette licence et cette autorisation n'entraînent pas une renonciation de [la] part [de l'auteur] à [ses] droits moraux ni à [ses] droits de propriété intellectuelle. Sauf entente contraire, [l'auteur] conserve la liberté de diffuser et de commercialiser ou non ce travail dont [il] possède un exemplaire.»

DÉDICACE

À mon partenaire de vie, à mes enfants, à mon directeur de maîtrise, aux personnes œuvrant au sein de l'Ordre des acupuncteurs du Québec ainsi qu'à toutes les personnes qui m'ont soutenue tout au long de l'accomplissement de mon projet de mémoire, ma sincère gratitude et tous mes remerciements pour votre soutien inconditionnel et vos encouragements si précieux pour aller toujours de l'avant, contre vents et marrées.

Merci d'avoir été constamment présent et d'avoir cru en moi et en mes capacités tout au long de cet acte. Grâce à vous, cet accomplissement m'a permis, tout comme Shakespeare le soulignait en 1603, *d'agir, de faire et d'accomplir*.

TABLE DES MATIÈRES

TABLE DES MATIÈRES	iv
LISTE DES FIGURES.....	x
LISTE DES TABLEAUX.....	xii
LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES	xv
RÉSUMÉ	xvii
ABSTRACT	xix
INTRODUCTION	1
1. PROBLÉMATIQUE	5
1.1 Pertinence professionnelle et scientifique de la recherche.....	5
1.2 Problème de recherche	15
2. CADRE THÉORIQUE.....	17
2.1 Introduction	17
2.2 Concepts centraux préalablement définis et contextualisés	18
2.2.1 Approche par compétences.....	18
2.2.2 Concept d'évaluation des compétences.....	20
2.2.3 Concept de standard de performance	23
2.3 Outils produits par l'OAQ.....	29
2.3.1 Polygramme des compétences liées à l'exercice de la profession d'acupuncteur au Québec	29

2.3.2	Cahier d'observation des comportements cliniques destiné aux maîtres de clinique dument accrédités par l'Ordre des acupuncteurs du Québec	37
2.3.2.1	Cadre procédural d'intégration des éléments de compétences	37
2.3.2.2	Grilles d'observation des savoir-faire liés aux comportements cliniques	41
2.4	Établissement des standards de performance	43
2.4.1	Processus d'élaboration des tâches intégratives et des items de l'épreuve clinique	43
2.4.1.1	Modèle d'élaboration des tâches intégratives et des items de l'épreuve clinique	44
2.4.1.2	Procédure d'élaboration et de validation des tâches intégratives et des items de l'épreuve clinique.....	47
2.4.2	Méthodes courantes d'établissement de standards de performance	50
2.4.2.1	Historique des méthodes d'établissement de standards de performance.	50
2.4.2.2	Neuf principes permettant l'établissement des standards de performance	54
2.4.2.1	Classification des méthodes d'établissement de standards de performance	63
2.4.2.2	Analyse des méthodes d'établissement de standards de performance centrées sur le test largement reconnues	66
2.4.2.3	Description sommaire des méthodes d'établissement de standards de performance centrées sur le test ayant servi de levier à la méthode alternative proposée	71

2.5	Méthode alternative d'établissement de standards de performance proposée	74
2.5.1	Aspects liés au contexte de l'épreuve clinique de l'OAQ	74
2.5.2	Méthode alternative inédite permettant d'établir des standards de performance pour l'épreuve clinique de l'OAQ	76
2.6	Objectif spécifique du projet de recherche	83
3.	MÉTHODOLOGIE	85
3.1	Introduction	85
3.2	Outils nécessaires à l'ESP	85
3.2.1	Conception d'un instrument de mesure (l'épreuve clinique)	86
3.2.2	Élaboration des échelles de mesure utiles à l'établissement de standards de performance	89
3.2.3	Préparation de la trousse de formation à l'ESP	90
3.2.4	Développement d'un algorithme de traitements des notations	93
3.3	Sélection et formation des panélistes	95
3.3.1	Sélection des panélistes ayant participé à l'expérimentation	95
3.3.2	Profil des panélistes ayant participé à l'expérimentation	99
3.3.3	Formation des panélistes ayant participé à l'expérimentation	100
3.4	Déroulement de l'expérimentation de la méthode alternative d'établissement de standards de performance	101
3.4.1	Indications préliminaires sur les adaptations apportées à la méthode d'Ebel	102
3.4.2	Déroulement de l'expérimentation en trois étapes	103
3.4.3	Déroulement détaillé de la méthode alternative d'établissement de standards de performance	104
3.4.3.1	L'étape préliminaire	104
3.4.3.2	L'étape expérimentale	105

3.4.3.3 L'étape finale.....	112
3.5 Considérations éthiques et transmissions des résultats aux participants.....	113
4. RÉSULTATS	116
4.1 Validation des items	116
4.2 Répartition des panélistes.....	117
4.3 Résultats obtenus lors de la première phase d'expérimentation.....	117
4.3.1 Résultats des rondes 1 et 2 de la phase I	117
4.3.2 Résultats de la ronde 3 de la phase I.....	129
4.3.3 Résultats de la ronde 4 de la phase I.....	133
4.4 Résultats obtenus lors de la deuxième phase d'expérimentation	135
4.4.1 Résultats des rondes 1, 2, 3 et 4 de la phase II	135
4.4.2 Résultats des rondes 1, 2 et 3 de la phase II	139
4.4.3 Résultat de la ronde 4 de la phase II.....	141
4.5 Résultats obtenus à l'étape finale de l'expérimentation.....	142
5. DISCUSSION	148
5.1 Introduction	148
5.2 Avantages et limites de l'approche alternative.....	148
5.2.1 Rappel des avantages.....	148
5.2.2 Limites et inconvénients.....	151
5.3 Précision et ajustement des échelles de mesure au moyen du modèle multifacette de Rasch	154
5.3.1 Justification.....	154
5.3.2 Échelle de mesure et précision de celle-ci.....	155
5.3.3 Qualité de l'ajustement.....	158
5.3.3.1 Juges	158
5.3.3.2 Items	160

5.3.3.3 Tâches.....	161
5.3.3.4 Critères	164
5.3.3.5 Rondes	166
5.3.3.6 Groupes.....	167
CONCLUSION.....	168
APPENDICES	176
APPENDICE A LETTRE AU PRÉSIDENT DE L'ORDRE DES ACUPUNCTEURS DU QUÉBEC	177
APPENDICE B LETTRE DU PRÉSIDENT DE L'ORDRE DES ACUPUNCTEURS DU QUÉBEC	180
APPENDICE C MÉTHODES D'ÉTABLISSEMENT DE STANDARDS DE PERFORMANCE CENTRÉES SUR LE TEST LARGEMENT RECONNUES	182
Présentation des méthodes courantes d'établissement de standards de performance centrées sur le test.....	183
1 Méthode d'Angoff.....	184
2 Méthode oui/non	188
3 Méthode de Nedelsky	191
4 Méthode d'Ebel	195
5 Méthode de l'appariement au descripteur de l'item.....	199
APPENDICE D CODIFICATION DE LA MODÉLISATION MULTIFACETTE DE RASCH.....	206
APPENDICE E RÉSULTATS DE L'ANALYSE SELON LE MODÈLE MULTIFACETTE DE RASCH	210
APPENDICE F STRUCTURE DE LA GRILLE D'ANALYSE I-SMART	251
APPENDICE G SONDAGE D'OPINION	253
APPENDICE H Lettre d'autorisation d'utilisation et de reproduction du contenu des ouvrages remise par l'OAQ à la chercheuse.....	255

RÉFÉRENCES.....	257
GLOSSAIRE.....	273
INDEX	284

LISTE DES FIGURES

Figure	Page
FIGURE 2.1 – VUE SYNOPTIQUE DU POLYGRAMME DES COMPÉTENCES LIÉES À L’EXERCICE DE LA PROFESSION D’ACUPUNCTEUR AU QUÉBEC	31
FIGURE 2.2 – MATRICE DES COMPÉTENCES DE L’ORDRE DES ACUPUNCTEURS DU QUÉBEC	35
FIGURE 2.3 – CADRE PROCÉDURAL D’INTÉGRATION DES ÉLÉMENTS DE COMPÉTENCES	40
FIGURE 2.4 – CRITÈRES DE SÉLECTION D’UNE MÉTHODE D’ÉTABLISSEMENT DE STANDARDS DE PERFORMANCE	55
FIGURE 2.5 – UTILITÉ DE LA FORMATION DES SUJETS À LA MÉTHODE D’ÉTABLISSEMENT DE STANDARDS DE PERFORMANCE RETENUE	59
FIGURE 2.6 – HISTORIQUE DES MÉTHODES D’ÉTABLISSEMENT DE STANDARDS DE PERFORMANCE	67
FIGURE 3.1 – ILLUSTRATION DES CRITÈRES DE SÉLECTION DES SUJETS	96
FIGURE 3.2 – L’ÉTABLISSEMENT DE STANDARDS DE PERFORMANCE EN TROIS ÉTAPES	103
FIGURE 3.3 – SEIZE CATÉGORIES DE PERFORMANCE ORDONNÉES PAR NIVEAU DE DIFFICULTÉ ET PAR DEGRÉ D’EFFICIENCE	108
FIGURE 4.1 – SCORES BRUTS TOTAUX ISSUS DES PATRONS DE RÉPONSE INDIVIDUELS DE LA PHASE I.....	123
FIGURE 4.2 – ACCORD INTERJUGES DES 2 ÉCHELLES DE MESURE ET LEUR INTERPRÉTATION À LA PHASE I.....	128

FIGURE 5.1 – SOMMAIRE DES PARAMÈTRES (EN LOGIT) POUR L'ENSEMBLE DES SIX FACETTES	157
---	-----

LISTE DES TABLEAUX

Tableau	Page
TABLEAU 2.1 – LES SEPT COMPOSANTES DE LA DIMENSION FONCTIONNELLE	32
TABLEAU 2.2 –LES TROIS COMPOSANTES DE LA DIMENSION PROFESSIONNELLE	33
TABLEAU 2.3 – LA COMPOSANTE DE LA DIMENSION CONTEXTUELLE	33
TABLEAU 2.4 – ILLUSTRATION DE LA PREMIÈRE COMPOSANTE DE LA DIMENSION PROFESSIONNELLE.....	36
TABLEAU 2.5 – ILLUSTRATION DE L'INTÉGRATION DES ÉLÉMENTS DE COMPÉTENCES FONCTIONNELLES À LA DIMENSION PROFESSIONNELLE	42
TABLEAU 2.6 –LE MODÈLE I-SMART	48
TABLEAU 2.7 – CONVENTION DES DESCRIPTIONS DES NIVEAUX DE DIFFICULTÉ	78
TABLEAU 2.8 – CONVENTION DES DESCRIPTIONS DES DEGRÉS D'EFFICIENCE.....	79
TABLEAU 3.1 – PROCESSUS ITÉRATIF DÉTAILLÉ DE LA MÉTHODE ALTERNATIVE D'ESP	106
TABLEAU 3.2 – OUTIL DE NOTATION FOURNI AUX PANÉLISTES LORS DE LA PHASE II.	109
TABLEAU 4.1 – MATRICES D'EBEL DES RONDES DE NOTATION INDIVIDUELLES DE LA PREMIÈRE PHASE	118
TABLEAU 4.2 – STANDARDS PRÉLIMINAIRES INDIVIDUELS ET COLLECTIFS DE LA PHASE I ET LEURS CALCULS.....	125
TABLEAU 4.3 – STANDARDS PRÉLIMINAIRES COLLECTIFS DE LA PHASE I, PROPRES AUX 23 TÂCHES.....	127
TABLEAU 4.4 – ZONES D'INCERTITUDE EXTRAITES DE LA LISTE ORDONNÉE DES ITEMS DE LA PHASE I.....	130

TABLEAU 4.5 – CALCUL DES STANDARDS PRÉLIMINAIRES COLLECTIFS DE LA PHASE I POUR L'ÉPREUVE	134
TABLEAU 4.6 – CATÉGORISATIONS MOYENNES ET CONSENSUELLES DES ITEMS DES ZONES D'INCERTITUDE	136
TABLEAU 4.7 – NOMBRE D'ITEMS PAR FAMILLE DE DIFFICULTÉ ET PAR CATÉGORIE DE PERFORMANCE	138
TABLEAU 4.8 – STANDARDS INTÉRIMAIRES INDIVIDUELS ET COLLECTIFS DE LA PHASE II ET RÉCAPITULATIF	139
TABLEAU 4.9 – STANDARD DE PERFORMANCE CONSENSUEL INTÉRIMAIRE DE LA PHASE II ET SES CALCULS	141
TABLEAU 4.10 – STANDARD DE PERFORMANCE DÉFINITIF ÉTABLI PAR LES PANELS A, B, C ET D ET SES CALCULS	143
TABLEAU 4.11 – STANDARDS DE PERFORMANCE DÉFINITIFS ÉTABLIS AUX 23 TÂCHES ET LEURS CALCULS	144
TABLEAU 5.1 – PRÉCISION ET AJUSTEMENT DE LA FACETTE DES JUGES	159
TABLEAU 5.2 – PRÉCISION ET AJUSTEMENT DE LA FACETTE DES TÂCHES	164
TABLEAU 5.3 – PRÉCISION ET AJUSTEMENT DE LA FACETTE DES CRITÈRES	165
TABLEAU 5.4 – PRÉCISION ET AJUSTEMENT DE LA FACETTE DES RONDES	166
TABLEAU 5.5 – PRÉCISION ET AJUSTEMENT DE LA FACETTE DES GROUPES	167
TABLEAU C.1 – ILLUSTRATION DE L'APPLICATION DE LA MÉTHODE D'ANGOFF AVEC UNE SEULE SESSION D'APPRÉCIATION	187
TABLEAU C.2 – ILLUSTRATION DE L'APPLICATION DE LA MÉTHODE OUI/NON	190
TABLEAU C.3 – NIVEAU DE PASSAGE MINIMAL CALCULÉ EN FONCTION DU NOMBRE DE CHOIX DE RÉPONSE RESTANTS SELON LA MÉTHODE DE NEDELSKY	193
TABLEAU C.4 – ILLUSTRATION DE LA MÉTHODE D'ÉTABLISSEMENT DE STANDARDS DE NEDELSKY	194
TABLEAU C.5 ILLUSTRATION DE LA MÉTHODE D'ÉTABLISSEMENT DE STANDARDS D'EBEL	197

TABLEAU C.6 – ILLUSTRATION DE LA MÉTHODE DE L'APPARIEMENT AU DESCRIPTEUR DE L'ITEM.....	202
--	-----

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

BNQ	Bureau de normalisation du Québec
OAQ	Ordre des acupuncteurs du Québec
OQLF	Office québécois de la langue française
CIQ	Conseil interprofessionnel du Québec
CPMT	Commission des partenaires du marché du travail
DEP	Direction de l'évaluation et de la prospection
DGDP : DEP	Direction générale du développement pédagogique : Département d'éducation et pédagogique
EDSC	Emploi et Développement social Canada
ESP	Établissement de standards de performance
FMMT	Forum des ministres du marché du travail
GDT	Le grand dictionnaire terminologique
IDM	Appariement au descripteur de l'item (<i>Item-Descriptor Matching Method</i>)
IPO	Items à performance observable
IPSFE	Initiative relative aux professionnels de la santé formés à l'étranger

MEQ	Ministère de l'Éducation du Québec
MICC	Ministère de l'Immigration et des Communautés culturelles
MIDI	Ministère de l'Immigration, de la Diversité et de l'Inclusion
MIFI	Ministère de l'Immigration, de la Francisation et de l'Intégration
MTO	Médecine traditionnelle orientale
OIB	Liste ordonnée des items (<i>Ordered Item Booklet</i>)
OPQ	Office des professions du Québec
PCPSS	Programme de contributions pour les politiques en matière de soins de santé
RCP	Reconnaissance des compétences professionnelles
SRHS	Stratégie en matière de ressources humaines en santé

RÉSUMÉ

La reconnaissance des acquis expérientiels des candidats formés à l'étranger représente un enjeu important, notamment au regard du processus d'admission au sein d'un ordre professionnel réglementé. La possibilité de contribuer à l'amélioration de ce processus, qui nous a interpellés, nous a menés à nous questionner : comment établir des standards de performance destinés à évaluer les compétences professionnelles des acupuncteurs formés à l'étranger désirant faire reconnaître leurs acquis expérientiels dans le contexte d'une épreuve clinique afin de favoriser l'accès au permis d'exercice de la profession d'acupuncteur au Québec? C'est cette question qui nous a conduits à réaliser une recherche-action en mesure et évaluation. Pour ce faire, une adaptation du modèle de Doran, l'I-SMART, a d'abord servi à élaborer et à valider les 23 tâches et les 149 items de l'épreuve d'observation clinique. Après avoir recensé les méthodes courantes d'établissement de standards, nous avons développé une méthode alternative à celle d'Ebel. Cette méthode a ensuite été mise à l'essai auprès de panélistes regroupés en deux échantillons intentionnels, auxquels nous avons d'abord proposé des descripteurs élaborés pour deux échelles de mesure renouvelées (difficulté perçue; efficience attendue) et la taxonomie de Bloom a été intégrée à l'échelle de performance (difficulté-efficience). Certains éléments d'autres méthodes ont également été intégrés à la démarche. Quatre opérations itératives avec rétroactions ont ainsi été retenues : la catégorisation, l'assignation, la mise en rang et la délimitation. Lors de la mise à l'essai, les panélistes ont été soumis à quatre rondes de notation par phase; trois individuelles et une consensuelle. Une série de matrices ont par la suite permis, à l'aide l'algorithme d'Ebel, d'établir les standards individuels et consensuels pour chaque panel et un algorithme a été développé pour la collecte et le traitement des données. La deuxième phase s'est limitée aux zones d'incertitude résultant de la liste ordonnée des items de la première phase. L'accord interjuges de la première phase a pu être vérifié par une analyse corrélationnelle linéaire et grâce à l'échelle de Guilford (1965). Au final, une fois la liste ordonnée des items constituée, un double standard a été établi sur la base du pourcentage d'items moyens issus des consensus de l'ensemble des panels.

Comme nous l'avions anticipé, l'expérimentation s'est déroulée dans le temps imparti et a, pour l'épreuve clinique et pour chacune des 23 tâches I-SMART, permis d'établir des standards de performance raisonnables et acceptés par les parties prenantes. Au terme des catégorisations individuelles, les zones d'incertitude avaient été fortement

réduites et un degré d'accord consensuel élevé a été observé pour les items résiduels. Au final, sept points de coupure ont permis de délimiter huit catégories de performance.

Tout comme Cizek (2012) le préconise, cette méthode alternative a l'avantage de fournir des balises claires sur lesquelles appuyer les jugements et, grâce à des algorithmes, d'être économique, fiable et efficace, en plus de fournir des résultats valides et applicables. Dans la présente recherche-action, les standards établis se limitent à leur acceptabilité auprès de l'Ordre des acupuncteurs du Québec. Ils sont applicables dans d'autres contextes. En effet, ces standards reposent sur un procédé d'inférence ayant mené à des conclusions logiques fondées sur un raisonnement déductif et, considérant la petite taille de l'échantillon, à l'acceptation incontestée des données. Les nombreuses catégories de performance représentent toutefois un inconvénient à la méthode alternative proposée et celles-ci devraient être revues à la baisse. Cela étant, outre quelques items ou tâches, les mesures estimées grâce à la modélisation multifacette de Rasch indiquent, en moyenne, un bon degré de précision et de fiabilité, une discrimination acceptable ainsi que des indices de rajustement satisfaisants entre les patrons de réponse observés et ceux prédits par le modèle multifacette. La cohérence interne au sein de la méthode alternative proposée s'avère ainsi satisfaisante, la répartition des items et des tâches se révèle équilibrée et, comme mentionné précédemment, il appert que certaines catégories peu utilisées pourraient être retranchées sans nuire aux résultats.

Mots clés : établissement de standards, standards de performance, méthode d'Ebel, méthode de l'appariement au descripteur de l'item, méthode d'Angoff, évaluation des compétences professionnelles, approche par compétences, reconnaissance des compétences, référentiel de compétences.

*ESTABLISHING PERFORMANCE STANDARDS TO EVALUATE
THE MINIMUM PROFESSIONAL COMPETENCIES
REQUIRED TO PRACTISE AS AN ACUPUNCTURIST IN QUEBEC*

ABSTRACT

Recognition of the prior experience of internationally educated candidates is an important issue, particularly regarding the admission process to a regulated professional order. Being able to contribute to the improvement of this process, which we found appealing, led us to ask ourselves: how can performance standards be established for assessing the professional competencies of foreign-trained acupuncturists who wish to have their experiential learning recognized in the context of a clinical trial in order to facilitate access to the license as an acupuncturist in Quebec? It is this question that led us to carry out an action research project in measurement and evaluation. To do this, an adaptation of Doran's model, the I-SMART, was first used to develop and validate the 23 tasks and 149 items of the clinical observation trial. After identifying common standard setting methods, we developed an alternative method to Ebel's. This method was then tested with panelists grouped into two purposive samples, to whom we first offered descriptors formulated for two renewed measurement scales (perceived difficulty; expected efficiency) and Bloom's taxonomy was incorporated into the performance scale (difficulty-efficiency). Some elements of other methods have also been incorporated into the approach. Four iterative operations with feedback were thus retained: categorization, assignment, ranking, and delimitation. During testing, panelists were subjected to four rounds of scoring per phase; three individual and one consensual. A series of matrices were then used to establish individual and consensual standards for each panel using Ebel's algorithm and an algorithm was developed for data collection and processing. The second phase was limited to the areas of uncertainty resulting from the ordered list of items from the first phase. The interrater agreement of the first phase could be verified via a linear correlational analysis and using the Guilford (1965) scale. Finally, once the ordered list of items was compiled, a double standard was established based on the percentage of average items obtained by a consensus of all panels.

As anticipated, the experiment was completed within the allotted time and, for the clinical trial as well as for each of the 23 I-SMART tasks, established reasonable

performance standards that were accepted by the parties involved. Upon completion of the individual categorizations, the areas of uncertainty were greatly reduced and a high degree of consensus agreement was observed for the residual items. Finally, seven cutting scores were used to identify eight performance categories.

Just as Cizek (2012) indicates, this alternative method has the advantage of providing clear markers on which to base judgments and, through the use of algorithms, is cost-effective, reliable, and efficient, as well as providing valid and applicable results. In this action research, the standards established are limited to their acceptability by the Ordre des acupuncteurs du Québec. They are applicable in other contexts. Indeed, these standards are based on an inference process that has led to logical conclusions based on deductive reasoning and, considering the small sample size, to the unquestioned acceptance of the data. However, the large number of performance categories is a disadvantage of the proposed alternative method and should thus be revised downwards. That being said, except for a few items or tasks, the measures estimated by the multifaceted Rasch model indicate, on average, a good degree of precision and reliability, acceptable discrimination and satisfactory readjustment indexes between the observed response patterns and those predicted by the multifaceted model. The internal consistency within the proposed alternative method thus proves to be satisfactory, the distribution of items and tasks proves to be balanced and, as mentioned above, it appears that certain under-used categories could be removed without affecting the results.

Keywords : *standard setting, performance standards, Ebel method, Item-Descriptor Matching Method, Angoff method, professional skills assessment, Skills-Based Approach, recognition of skills, skills framework.*

INTRODUCTION

Au cours des 20 dernières années, des millions de dollars ont été alloués aux ordres professionnels du Québec pour partager leurs bonnes pratiques, concevoir des outils innovateurs pour reconnaître les acquis expérientiels et pour achever d'élaborer des référentiels de compétences (G. Vaillancourt, MIFI, courriel, 23 juin 2021; Korotkina et Neuwahl, 2012). Or, la nécessité de reconnaître les acquis expérientiels des candidats formés à l'étranger invite à recourir à des instruments et à des procédures valides de haute qualité pour assurer la protection du public et un processus décisionnel juste (CIQ *et al.*, 2010). Tout comme de nombreux intervenants professionnels en santé et en éducation, les acteurs des ordres professionnels québécois se sont mobilisés pour remédier aux différents problèmes rencontrés dans le processus de reconnaissance des compétences et des acquis expérientiels des professionnels formés à l'étranger. Selon une communication récente transmise aux ordres professionnels (G. Vaillancourt, MIFI, courriel, 23 juin 2021), cette mobilisation a entre autres permis de financer la conception d'outils d'évaluation des compétences, d'outils d'autoévaluation, des guides de préparations aux examens et des formations d'appoint, en plus d'accroître considérablement le budget 2021-2022 alloué au programme d'aide à la reconnaissance des compétences (PARC).

Cet enjeu a mené l'Ordre des acupuncteurs du Québec (OAQ) à poursuivre ses efforts en s'engageant dans des démarches, en collaboration avec la chercheuse, pour élaborer une épreuve clinique et pour établir le niveau de maîtrise à atteindre pour permettre de délivrer, en toute confiance, un permis d'exercice de la profession. Notre implication directe au sein de l'OAQ ainsi qu'un intérêt marqué pour cet enjeu ont rendu pertinent

le fait de nous intéresser à l'évaluation des compétences et aux méthodes d'établissement de standards de performance (ESP) pour déterminer des seuils d'entrée dans la profession. Pour établir ces standards, une épreuve clinique devait être préalablement élaborée et validée. De ce fait, un modèle inspiré de Doran et permettant de structurer l'élaboration des tâches et des items de l'épreuve clinique a été proposé à l'OAQ. Une fois l'épreuve clinique élaborée et validée par ce dernier, la recherche pouvait s'y référer pour établir des standards de performance. Il restait ensuite à répondre à la question de recherche de ce mémoire, soit : comment établir des standards de performance destinés à évaluer les compétences professionnelles des acupuncteurs formés à l'étranger désirant faire reconnaître leurs acquis expérientiels dans le contexte d'une épreuve clinique afin de favoriser l'accès au permis d'exercice de la profession d'acupuncteur au Québec?

En plus de contribuer au mandat de protection du public et de rendre le processus de délivrance de permis d'exercice simple, transparent, rigoureux et équitable, cette question contribuera à baliser l'exercice de la profession en plus de favoriser le partage des bonnes pratiques et l'avancement des connaissances scientifiques. Les fondements théoriques apporteront ensuite un éclairage utile à la question de recherche, notamment quant aux concepts centraux de l'approche par compétences et à l'évaluation des compétences et de standard de performance. Une présentation des outils pertinents produits par l'Ordre, notamment le référentiel des compétences et de l'épreuve clinique suivra, et ce, afin de faciliter la compréhension des étapes ayant précédé l'ESP. Cette présentation sera complétée par la description des différents aspects de l'ESP, passant du processus d'élaboration du contenu de l'épreuve clinique aux méthodes courantes d'établissements de standards de performance. Plus précisément, une revue historique des méthodes courantes, des principes et de la classification des méthodes ainsi qu'une analyse des méthodes les plus reconnues et une description sommaire de celles retenues pour élaborer la méthode alternative permettront de contextualiser l'objet de recherche.

Puis, la présentation finale de la méthode alternative mettra d'abord en relief différents aspects liés au contexte de l'épreuve clinique – de nature polytomique et composée d'items à réponses construites élaborées. Elle sera ensuite complétée par la description de la méthode alternative mise en œuvre, y compris des descripteurs des deux échelles de mesure déployées.

L'objectif poursuivi dans le cadre de cette recherche-action consiste, conformément aux besoins de l'OAQ, à établir un double standard de performance : un standard pour chacune des tâches structurées et un pour l'ensemble de l'épreuve clinique. Pour atteindre cet objectif, un plan méthodologique a été établi. Ce plan consiste tout d'abord à assister l'OAQ à concevoir l'épreuve clinique à partir du référentiel de compétences et ultérieurement, l'instrumentation destinée à former les participants et que celle destinée à recueillir et à traiter les données efficacement à l'aide d'un algorithme. Ce plan a également nécessité d'élaborer un processus de sélection et de formation des 20 panélistes sélectionnés pour expérimenter la méthode alternative proposée. Pour les besoins de l'expérimentation, ces panélistes se devaient d'être répartis sur deux phases. En complément, grâce une large recension des écrits portant sur des méthodes centrées sur le test, la recherche a conduit à proposer, en réponse aux besoins de l'OAQ, une méthode alternative à celle d'Ebel. Des éléments de deux autres méthodes, soit celles de Ferrara, Perie et Johnson (2008) et d'Impara et Plake (1997), y ont également été intégrés. Par voie de conséquence, cette méthode innovante nous a mené à expliquer plus à fond le déroulement de l'expérimentation.

Ce mémoire expose au premier chapitre la pertinence professionnelle et scientifique de la recherche à travers divers enjeux et situe la problématique internationale de la reconnaissance des compétences professionnelles acquises à l'étranger. La revue de la littérature vient, au deuxième chapitre, soutenir la pertinence scientifique de la recherche-action portant sur l'établissement des standards de performance au Québec,

celle-ci établissant les fondements théoriques et conceptuels liés qui nous ont guidés vers la question de recherche à laquelle nous nous intéressons. Tout ce qu'il faut savoir sur la mise en œuvre de l'expérimentation est décrit au chapitre portant sur la méthodologie, incluant les considérations éthiques et la transmission des résultats aux participants. Les résultats de la recherche sont ensuite présentés au chapitre suivant, passant de la validation des items aux résultats des différentes étapes de l'expérimentation. Ces résultats sont subséquemment discutés dans le chapitre suivant, où les avantages, les limites et les inconvénients sont exposés, de même que les résultats inférant les précisions et les indices d'ajustements des échelles de mesure au moyen du modèle d'analyse psychométrique multifacette de Rasch. Les facettes d'intérêts analysées ici sont celles des juges, des items, des tâches, des critères de mesure, de même que les rondes de notation et les groupes de panélistes. Ce document se termine avec une conclusion qui rappelle le contenu des chapitres précédents et indique des pistes de recherche et d'actions futures.

1. PROBLÉMATIQUE

1.1 Pertinence professionnelle et scientifique de la recherche

La revue de la littérature effectuée dans le cadre de notre démarche de recherche met en lumière l'existence d'un ensemble de recherches et de travaux issus de divers organismes réglementaires et qui mettent en lumière la nécessité d'évaluer les compétences. Toutefois, il est clairement apparu qu'aucun écrit scientifique ne répond parfaitement aux besoins d'évaluation des compétences professionnelles des organismes réglementés.

En 2012, une revue juridique canadienne précisait dans un article que :

Afin de protéger le public contre des personnes n'ayant pas les compétences requises pour exercer [...] une profession [...], la majorité des pays développés ont mis en place des systèmes réglementaires ayant pour fonction la vérification et la surveillance des compétences des individus autorisés à exercer [...] une profession sur leur territoire. [La] qualification professionnelle ne sera pas automatiquement reconnue par les autorités réglementaires provinciales régissant les ordres professionnels [...]. Elle sera évaluée au cas par cas par l'ordre [...] auquel il souhaite appartenir [...]. L'autorité réglementaire évaluera [notamment] les compétences pratiques des candidats et leur imposera diverses conditions avant de leur délivrer un permis d'exercice de la profession (Houle et Roux, 2012, p. 81, 85, 89).

Le Canada, suivant les discussions ayant eu cours au Forum des ministres du marché du travail (FMMT), reconnaît que le succès économique à long terme repose sur une main-d'œuvre compétente et mobile (FMMT, 2011).

Dans le document intitulé *Évaluation du Programme de contributions pour les politiques en matière de soins de santé (PCPSS)*, il est stipulé que « l'accord de 2003 des premiers ministres sur le renouvellement des soins de santé et le Plan décennal pour consolider les soins de santé (Accord de 2004) comprenaient l'engagement d'un financement additionnel par l'entremise du Transfert canadien en matière de santé et de programmes sociaux, ainsi que d'un financement ciblé pour atteindre certains objectifs pancanadiens [...] dont notamment l'intégration des professionnels de la santé formés à l'étranger » (Santé Canada et l'Agence de la santé publique du Canada, 2013, p. 2). Dans ce sens, le budget fédéral canadien de 2012 faisait état de secteurs prioritaires à prendre en compte dans le cadre du PCPSS, dont celui de « l'amélioration du processus de reconnaissance des titres de compétences étrangers grâce au Cadre pancanadien d'évaluation et de reconnaissance des qualifications professionnelles acquises à l'étranger, qui cible notamment les professionnels de la santé » (*ibid.*, 2013, p. 12). Dans l'optique de soutenir les efforts déployés pour accroître l'intégration à la population active des professionnels formés à l'étranger – aptes à exercer leurs professions dans un contexte canadien et répondant aux besoins des secteurs de la santé jugés prioritaires –, l'une des cinq composantes du programme nommé « Initiative relative aux professionnels de la santé formés à l'étranger (IPSFE) » a notamment permis de financer, en 2005, la création et le maintien de plusieurs projets. Une autre des composantes du programme, qui veille à l'évolution de la Stratégie en matière de ressources humaines en santé (SRHS), a également permis, en 2008, d'en redéfinir les secteurs prioritaires et de les recentrer sur quatre orientations. L'une de ces orientations permet, entre autres, de financer des projets axés sur les professionnels de la santé formés à l'étranger dans l'optique « d'accroître l'accès aux programmes d'évaluation et de formation et aux possibilités d'intégration dans le milieu de travail » pour les professionnels de la santé formés à l'étranger » (*ibid.*, 2013, p. 57). En tout, au cours des huit premières années, le programme a permis de financer « cent soixante-trois projets » (*ibid.*, 2013, p. ii).

Le gouvernement du Québec s'est pour sa part engagé à partager une multitude de rapports déposés à l'Assemblée nationale (FMMT, 2011, p. 22). Grâce à la collaboration de nombreux partenaires, plusieurs projets ont, depuis quelques années, été mis en œuvre (*ibid.*, p. 22). Parmi ces réalisations, plusieurs ordres professionnels ont, entre autres, élaboré des référentiels de compétences et s'affairent maintenant à développer des outils d'évaluation afin de faciliter le processus menant à l'octroi d'un permis d'exercer une profession au Québec (*ibid.*, p. 22). Depuis 2008, le Québec et la France ont fait un pas de géant lorsqu'ils ont conclu, par l'entremise du ministère de l'Immigration et des Communautés culturelles (MICC, devenu MIDI, puis MIFI), une entente « de reconnaissance mutuelle des qualifications professionnelles [donnant ainsi accès] à 68 professions et métiers » (*ibid.*, p. 22). En 2010, par l'entremise de l'Office des professions du Québec (OPQ), « un poste de Commissaire aux plaintes a [vu le jour] afin de recevoir et d'examiner les plaintes [liées aux failles ou aux insatisfactions relatives aux] mécanismes de reconnaissance des compétences professionnelles » (*ibid.*, p. 22).

En 2013, le ministère de l'Emploi et du Développement social du Canada (EDSC, p. 38) précisait que :

l'efficience du marché du travail favorise un meilleur équilibre entre l'offre et la demande de main-d'œuvre à l'échelle nationale [par l'entremise de] la promotion de l'intégration au marché du travail des nouveaux immigrants, du recours aux travailleurs étrangers temporaires, de la mobilité de la main-d'œuvre partout au Canada ainsi que la diffusion d'information sur le marché du travail [afin de] veiller à ce que les Canadiens et les nouveaux arrivants aient accès à des emplois intéressants. [Ce ministère] se penche maintenant sur les pénuries de main-d'œuvre qualifiée et cherche à accroître l'efficience du marché du travail dans des secteurs clés en recueillant des renseignements sur le marché du travail.

Encore aujourd'hui, plusieurs partenaires s'entendent sur l'importance de déployer les efforts pour favoriser et accélérer la reconnaissance des compétences professionnelles

(RCP) développées par l'expérience professionnelle. Le Conseil interprofessionnel du Québec (CIQ) (CIQ, 2009, 2015) ainsi que les ordres professionnels maintiennent leur ferme engagement à améliorer le processus afin d'accélérer l'accès aux professions réglementées des personnes formées à l'étranger. Récemment, le CIQ a recommandé d'intégrer dans le cadre réglementaire « l'approche par compétences et l'évaluation des acquis expérientiels, afin que ce cadre soit le reflet des nouvelles pratiques des ordres » (*ibid.*, 2015, p. 2).

Le Conference Board du Canada (Bureau du commissaire à l'équité, 2013) estime à cet égard que le coût économique annuel lié au manque de main-d'œuvre qualifiée mérite de se pencher sérieusement sur la reconnaissance des compétences des personnes formées à l'étranger de plusieurs groupes de profession et plus spécifiquement dans le secteur de la santé. Il estime que l'absence de reconnaissance des compétences des immigrants hautement qualifiés représente un coût économique de 4 à 6 milliards de dollars par an (*ibid.*, 2013).

En France, la Direction de l'évaluation et de la prospection (DEP) ainsi que plusieurs chercheurs abondent dans le même sens et s'accordent pour dire que l'importance des départs à la retraite jumelée à la relative faiblesse du nombre de travailleurs qualifiés sortant du système éducatif sont des sujets préoccupants (ministère de la Jeunesse, de l'Éducation nationale et de la Recherche, 2003). Cette situation nécessite, selon ces porte-paroles, d'anticiper les besoins de recrutement d'une main-d'œuvre possédant les qualifications requises (*ibid.*, 2003). Il semble y avoir consensus sur le fait que le contexte socioéconomique nécessite de remédier à la rareté d'une main-d'œuvre qualifiée afin de contribuer efficacement à la santé économique de notre société. Sur le plan politique, l'objectif majeur « en matière d'immigration, de diversité et d'inclusion » est actuellement, pour le CIQ, de « parvenir à une meilleure adéquation entre notre système d'immigration et les besoins en emploi de la société québécoise »

(CIQ, 2015, p. 1). La productivité et la compétitivité accrue mondialement requièrent une contribution significative et progressive des travailleurs hautement qualifiés, principalement en raison de l'augmentation des liens et des relations commerciales avec l'Asie, l'Europe et les autres pays partenaires (Böhning et Beijl, 1996; Boudarbat et Grenier, 2014).

Or, il appert que la mobilité de la main-d'œuvre professionnelle et le contexte normatif de la pratique québécoise présentent plusieurs difficultés connues et bien documentées (MICC, 2005; Houle et Roux, 2012). Le groupe de travail sur l'accès aux professions et métiers réglementés a d'ailleurs souligné que 47 mémoires ont permis de préciser les enjeux et les contraintes liés à la reconnaissance des acquis (*ibid.*, 2005). Ainsi, avant d'intégrer le marché du travail, les candidats formés à l'étranger sont souvent confrontés aux problèmes de la RCP. Boudarbat et Grenier (2014) abondent dans le même sens et précisent que ce sont principalement les nouveaux arrivants hautement scolarisés et possédant une expérience acquise dans leur pays d'origine qui éprouvent actuellement des difficultés à les faire reconnaître sur le marché du travail canadien. Ils sont notamment susceptibles de se heurter à un long processus, complexe et onéreux, sans aucune garantie de résultats (CIQ, 2015; Houle et Roux, 2012). Certains d'entre eux se voient ainsi confrontés aux limites du système scolaire du pays hôte, tant pour la durée des formations que pour la capacité des établissements d'enseignement à les accueillir dans le programme d'étude qui leur aura été prescrit (Kustec, 2012). Ils se heurtent également à leurs propres limites et aux limites de leur pays d'origine. D'ailleurs, plusieurs auteurs mettent en évidence la variabilité de la qualité des systèmes éducatifs et des écarts de rendements constatée d'un pays à l'autre et soutiennent que celle-ci en est tributaire, ainsi que du niveau de développement de leur pays d'origine (plus ou moins prospère) (Industrie Canada, 2008, p. 29; Boudarbat et Grenier, 2014). Tant et si bien qu'en Alberta, la RCP conduit, dans la majorité des cas, soit à un permis restrictif ou temporaire avec une prescription à respecter (40 %), soit

à un refus absolu pour non-conformité aux exigences minimales de cet ordre professionnel (20 %) (Houle et Roux, 2012). Ipso facto, cela se traduit par un taux de chômage supérieur pour cette catégorie de travailleurs, et de ce fait, par un coût économique considérable en plus des enjeux juridiques et déontologiques liés à leur intégration et à leur insertion professionnelle (Houle et Roux, 2012). Bien que le parcours menant à l'obtention d'un permis d'exercice risque d'être mis en péril par une « disponibilité insuffisante des stages » prescrits (CIQ, 2015, p. 3), les 47 ordres professionnels (OPQ, 2017) sont juridiquement autorisés, dans le cadre de leur mandat de protection du public, à établir des règles qui limitent l'accès à une profession pourvu que celles-ci soient valables (Houle et Roux, 2012). La littérature révèle à ce sujet une tendance accrue des ordres professionnels à prescrire davantage des stages aux candidats et une tendance à s'éloigner des prescriptions de cours et d'examens, principalement pour les ordres du « secteur de la santé et des relations humaines » (CIQ, 2015, p. 16). Ces tendances reflètent la propension des ordres professionnels à développer des approches par compétences et à favoriser la reconnaissance des acquis expérientiels.

Corolairement, depuis une quinzaine d'années, les ordres professionnels québécois ont connu une augmentation considérable des demandes de RCP, soit près de 5 000 par année, ce qui représente une croissance supérieure à 500 % (CIQ, 2015). « Le CIQ et ses partenaires (CIQ *et al.*, 2010) soulignent, afin de pallier cette situation, que tous les ordres professionnels ont reçu du gouvernement du Québec la commande de mettre en œuvre un processus efficace d'évaluation des compétences et des acquis expérientiels favorisant l'intégration des nouveaux arrivants au marché du travail. En ce sens, les ordres professionnels et plusieurs intervenants provenant des milieux professionnels (par ex. MICC, CIQ), du secteur de la santé (par ex., Ordre des acupuncteurs du Québec, Ordre des inhalothérapeutes du Québec, Ordre des conseillers et conseillères d'orientation et des psychoéducateurs et psychoéducatrices du Québec) et du secteur

de l'éducation (par ex., Cégep Marie-Victorin et Cégep John Abbott College) sont depuis quelques années mobilisés pour remédier aux différents problèmes rencontrés dans le processus de reconnaissance des compétences et des acquis expérientiels des professionnels formés à l'étranger.

Plusieurs s'entendent pour reconnaître que « l'expérience est la bête noire de la reconnaissance des acquis » (Ouellette, 2006, p. 16; Noël, 2010; Shields, 2007). La diversité des origines, des valeurs, des politiques et des règlements ainsi que des pratiques développées à l'étranger accentue la difficile tâche de reconnaître les acquis scolaires et expérientiels des candidats formés à l'étranger. C'est pour cela que la reconnaissance d'une équivalence exige généralement d'un candidat désirant obtenir un permis d'exercice qu'il se conforme aux conditions d'évaluation de ses compétences par l'entremise d'un stage, d'une formation ou d'un enchaînement d'épreuves professionnelles ou d'une seule (Comité d'adaptation de la main-d'œuvre – Personnes immigrantes, 2013).

Il est important de préciser que la « reconnaissance des compétences est possible [seulement] pour un métier, une profession ou une fonction de travail pour lesquels une norme professionnelle est inscrite au Registre de compétences » (*ibid.*, p. 24) et requiert généralement un nombre minimal d'années d'expérience. Présentement, « la Loi n'impose pas une méthode unique pour évaluer les compétences des personnes » désirant faire reconnaître leurs compétences. Toutefois, certaines pratiques courantes sont plus fréquemment observées (*ibid.*, p. 25). Parmi les méthodes d'évaluation les plus communes, l'une consiste en une mise en situation professionnelle, réelle ou simulée [en milieu de travail ou en établissement d'enseignement,] alors qu'une autre vise à « réaliser une production spécifique » telle qu'un travail personnel jumelé à un entretien (*ibid.*, p. 25). La Commission des partenaires du marché du travail (CPMT) précise que lorsqu'une compétence est démontrée ou appuyée d'une preuve de sa

maîtrise, elle peut être officiellement reconnue si elle respecte les normes de la profession (*ibid.*, 2013, p. 25). Toutefois, lors du processus d'évaluation, il est de la responsabilité du candidat de « faire la démonstration [de son degré de maîtrise réel de] la compétence [qu'il] veut faire reconnaître » (*ibid.*, 2013, p. 28).

Ces diversités normatives doivent être prises en compte dans les différents processus d'admission des ordres professionnels afin de déterminer si un permis d'exercice sera délivré avec ou sans restriction ou s'il sera refusé advenant un niveau de compétence inférieur à ce qui est minimalement attendu et qui nécessiterait, par exemple, une formation d'appoint. Quoique divers documents sont destinés à soutenir la réflexion des ordres professionnels, le CIQ (CIQ, 2010) réitère l'importance de pallier la problématique liée à la reconnaissance des compétences développées par l'expérience professionnelle. Plusieurs efforts concertés visent précisément à soutenir les actions entreprises pour développer des mécanismes et des outils capables d'évaluer les compétences professionnelles des nouveaux arrivants formés à l'étranger. En 2012, un mémoire, intitulé « La reconnaissance des qualifications professionnelles comme condition à l'immigration au Québec? Cadre juridique et enjeux politiques d'une réforme de procédure en amont » précise que la protection du public exige que la compétence professionnelle d'un candidat soit évaluée et établie par un expert dans le domaine (Korotkina et Neuwahl, 2012, p. 119). Dans ce sens, un plan d'action, adopté en 2001 par plusieurs ordres professionnels, réunis au Conseil interprofessionnel du Québec, a mené à l'adoption, en 2006, de deux documents soutenant les démarches éthiques et administratives du processus d'admission : les Principes en matière de reconnaissance d'une équivalence de diplôme et de formation acquis hors du Québec (CIQ, 2006) et la Recension des bonnes pratiques en matière de reconnaissance d'une équivalence de diplôme et de formation acquis hors du Québec (Korotkina, 2012). Ces écrits visent notamment à baliser chacune des étapes nécessaires à la mise en œuvre d'un processus de délivrance de permis qui soit simple, rigoureux, transparent et

équitable, ainsi qu'à améliorer les stratégies et les façons de faire en matière de reconnaissance des compétences et des acquis expérientiels.

En 2012, le CIQ invitait les ordres professionnels à collaborer afin de partager leurs bonnes pratiques, tout en poursuivant leurs efforts pour élaborer des référentiels de compétences et concevoir des outils innovateurs pour reconnaître les acquis expérientiels (Korotkina et Neuwahl, 2012). En outre, les ordres professionnels obtenaient, en 2011-2012, un soutien financier gouvernemental de 5 millions de dollars après avoir suscité l'implication « des acteurs du monde de l'éducation » et la participation du ministère de l'Immigration, de la Diversité et de l'Inclusion du Québec (MIDI), ancienneté nommé MICC (CIQ, 2015). L'année suivante, le ministère de l'Immigration et des Communautés culturelles (MICC, 2013) s'est engagé, avec ses partenaires, à investir près de 1,5 million de dollars pour faciliter l'accès à diverses professions principalement dans le secteur de la santé, notamment pour les ergothérapeutes, les orthophonistes, les physiothérapeutes et les travailleurs sociaux. Entre autres, une subvention estimée à plus de 33 000 dollars a été octroyée à l'Ordre des travailleurs sociaux et des thérapeutes conjugaux et familiaux du Québec pour la conception d'outils d'évaluation des compétences destinés à favoriser l'accès à la profession (*ibid.*, 2013).

Bien que de nombreuses réalisations aient été mises en œuvre, dont « la mise en place de politiques, de structures de services [destinées aux] personnes immigrantes et le développement d'outils de reconnaissance », la problématique liée à la reconnaissance des acquis expérientiels demeure un enjeu pour l'implantation des mécanismes de reconnaissance des compétences professionnelles au Québec. Dans ce sens, lors du 4^e colloque des dirigeants des ordres professionnels tenu au Québec en 2011, la reconnaissance des compétences professionnelles est apparue comme un des enjeux fondamentaux auquel il faudra répondre au cours des prochaines années. Ce défi

s'insère directement dans le mandat confié aux ordres professionnels, dont la mission fondamentale vise à assurer la protection du public, notamment par l'entremise d'un contrôle de l'exercice de la pratique de la profession de ses membres ainsi que par la surveillance de la qualité de leurs pratiques professionnelles.

En dépit du fait que les difficultés liées à la reconnaissance des acquis sont largement connues et documentées, l'évaluation des compétences professionnelles et l'ESP dans le cadre du processus de reconnaissance des compétences des personnes sollicitant un droit de pratique dans un ordre professionnel au Québec sont peu documentés. Quoique certains ordres professionnels du Québec soient dotés d'un référentiel de compétences balisant la qualité de la pratique, il semble, à notre connaissance, qu'aucun d'entre eux ne se soit, à ce jour, muni de mécanismes d'ESP. La recension de la littérature tant francophone qu'anglophone nous a, à cet effet, permis de constater qu'aucune étude scientifique n'avait traité de l'établissement des standards de performance pour évaluer les compétences professionnelles et pour reconnaître les acquis expérientiels des candidats formés à l'étranger.

Les études consultées traitant du sujet étant davantage axées sur l'évaluation des compétences en contexte d'apprentissage, il apparaît que la pertinence scientifique de notre recherche découle de la nécessité de transposer les concepts de l'évaluation des compétences afin d'appliquer la rigueur du monde de l'éducation aux exigences du monde professionnel et des organismes réglementaires. Pour l'OAQ, le contexte d'évaluation des compétences s'inscrit en situations réelles non simulées où, en plus de la performance, il est fondamental de considérer la capacité d'un candidat à juger des ressources pertinentes à intégrer à l'intervention clinique et sa capacité de réaliser une variété de tâches de complexités variables dans un temps raisonnable. L'OAQ a

également besoin d'établir un double standard en ayant recours à des échelles polytoniques pour les 23 tâches et à une échelle dichotomique pour l'épreuve clinique.

La démarche que nous proposons vise conséquemment à apporter une contribution à l'avancement de la science et aux démarches entreprises par les ordres professionnels québécois en matière d'ESP et d'évaluation des compétences professionnelles. Plus précisément notre recherche permettra de pallier l'absence de méthode d'établissement de standards répondant spécifiquement aux besoins liés à l'administration de l'épreuve clinique de l'Ordre des acupuncteurs du Québec (OAQ).

Plusieurs auteurs du secteur de la santé des professions réglementées (Cerqueira *et al.*, 2008) précisent, dans ce sens, que l'établissement d'un seuil de réussite/échec pour une épreuve est une tâche complexe; ce que nous comptons relever avec succès. Pour ce faire, nous nous appuierons sur des méthodes reconnues, proposées par certains auteurs (Cizek, 2012; Cizek et Bunch, 2007; Raîche, 2005; Impara et Plake, 1997; Ebel, 1972; Angoff, 1971; Nedelsky, 1954; Jaeger, 1989; Hambleton, 1980) pour établir des standards de performance pouvant être utilisés dans l'actuel processus d'évaluation des compétences professionnelles. À titre d'exemple, nous nous inspirerons de la méthode Ebel et de la méthode de l'appariement au descripteur de l'item.

1.2 Problème de recherche

À l'instar de l'ensemble des ordres professionnels, l'OAQ a la responsabilité d'évaluer les compétences des candidats formés à l'étranger désirant obtenir le permis obligatoire permettant l'exercice de la profession d'acupuncteur au Québec.

Dans cette optique, nous croyons que notre démarche de recherche, visant l'ESP et l'évaluation des compétences professionnelles, contribuera à soutenir les efforts de

l'OAQ dans la réalisation de son mandat. De surcroît, en 2013, le président de cet ordre professionnel, avec l'appui du conseil d'administration, se disait heureux de l'initiative en soulignant qu'outre sa contribution à l'avancement des connaissances, cette recherche sera « des plus utiles dans l'accomplissement de leur mandat de protection du public et par ce biais, bénéficiera à l'ensemble de la communauté québécoise ».

Il est important de souligner que notre recherche s'insère dans le continuum des travaux déjà entrepris par l'OAQ. Plus précisément, elle fait suite à la conception et à la validation de plusieurs documents, dont un référentiel de compétences, un cadre procédural d'intégration des éléments de compétences et des grilles d'observation des savoir-faire liés aux comportements cliniques.

Dans cette suite logique, notre recherche tentera de répondre à la question suivante : comment établir des standards de performance destinés à évaluer les compétences professionnelles des acupuncteurs formés à l'étranger désirant faire reconnaître leurs acquis expérientiels dans le contexte d'une épreuve clinique afin de favoriser l'accès au permis d'exercice de la profession d'acupuncteur au Québec?

Cette question de recherche nécessite de présenter un cadre théorique afin de situer les principaux concepts liés à notre démarche de recherche. Nous présenterons, dans la section suivante, les concepts centraux de notre recherche, les outils produits par l'OAQ, puis les différents aspects liés à l'établissement des standards de performance, ce qui nous conduira à introduire la méthode alternative que nous proposons et finalement, notre objectif de recherche.

2. CADRE THÉORIQUE

2.1 Introduction

Afin de répondre à la question de recherche, nous allons d'abord présenter les concepts de l'approche par compétence, de l'évaluation des compétences et des standards de performance. Nous allons ensuite présenter les différents outils que l'OAQ a produits afin de répondre à sa responsabilité d'évaluer les compétences des candidats formés à l'étranger. Cela permettra, au final, de bien contextualiser notre démarche.

Pour ce faire, nous allons, dans un premier temps, situer le *Polygramme des compétences liées à l'exercice de la profession d'acupuncteurs au Québec*¹ représentant le référentiel de l'ensemble des compétences décrivant l'exercice de la profession. Dans un deuxième temps, au regard des concepts centraux préalablement évoqués, nous allons présenter le *Cahier d'observation, des comportements cliniques, destiné aux maîtres de clinique dument accrédités par l'Ordre des acupuncteurs du Québec*² (OAQ, 2014a). C'est sur la base d'une bonne compréhension de ces deux outils développés par l'OAQ qu'il nous sera possible d'inférer notre démarche d'établissement des standards de performance.

Pour finaliser notre cadre théorique, puisque c'est l'objet principal de ce travail de

¹ Le référentiel des compétences professionnelles de l'Ordre des acupuncteurs du Québec peut être consulté en invoquant l'hyperlien <https://www.o-a-q.org/getmedia/7550f51d-e415-4ab6-a1bb-560c1d1b7da4/Polygramme-des-competences.aspx>

² <http://o-a-q.org/getmedia/2fa38994-cef0-4da3-b674-f875b8e99f99/Cahier-d-observaiton-des-comportements-cliniques.aspx>

recherche, nous allons décrire des méthodes courantes d'établissements de standards de performance ainsi qu'une méthode alternative plus adaptée aux outils de l'OAQ qui conduiront à la justification et à la présentation de l'objectif spécifique de notre projet de recherche.

2.2 Concepts centraux préalablement définis et contextualisés

Afin de bien contextualiser l'objet de notre démarche de recherche, il nous apparaît d'abord essentiel de définir trois concepts centraux liant l'approche par compétences à l'évaluation des compétences et aux standards de performance. Nous nous limiterons toutefois à certains des auteurs les plus reconnus dans le domaine de la mesure et de l'évaluation. Les concepts sous-jacents sont présentés en annexe dans un glossaire modeste constitué à l'aide de la troisième édition du *Dictionnaire actuel de l'éducation*.

2.2.1 Approche par compétences

Pour bien comprendre le concept de compétences, il est important de se rappeler qu'au cours des années 90, le domaine de l'évaluation a effectué un tournant majeur, passant d'une approche par objectifs à une approche par compétences. Ce changement de paradigme a conduit le monde de l'éducation à axer la formation sur le développement des compétences, favorisant ainsi « l'intégration de savoirs, de savoir-faire et d'attitudes nécessaires à l'accomplissement de tâches et de rôles professionnels » (Louis et al., 1996). À cet égard, Louis, Jutras et Hensler (1996, p. 417) soulignent que plusieurs auteurs s'accordent sur la nécessité de proposer des « définitions qui évoquent des savoir-faire complexes » qui s'inscrivent dans une approche cognitive sous-tendant une capacité d'agir complexe et non pas de le faire par une action unique. Pour Le Boterf (1998), les compétences résultent d'un processus découlant d'opérations humaines résultants de la mise en œuvre d'un ensemble de moyens ou de manœuvres stratégiques permettant à un candidat d'atteindre un résultat satisfaisant ou

de mener à bien un objectif précis. Cet auteur précise également que la compétence est le processus qui « indique qu'un individu réalise avec compétence ses activités en combinant et en mobilisant des ressources », alors que les compétences représentent le produit de ce processus. Legendre (2005), en s'appuyant sur Le Boterf (2007, p. 42), précise pour sa part que la compétence doit nécessairement être tributaire de l'agencement des ressources internes et externes que la personne mobilise « dans un contexte particulier et au sein d'une pratique pertinente ». Jonnaert (2002, cité dans Legendre, 2005), souligne de son côté que la personne désirant démontrer sa compétence doit être capable de sélectionner et de coordonner efficacement une combinaison constituée exclusivement des ressources jugées pertinentes dans une situation et dans un contexte donnés.

Il semble de plus généralement convenu chez plusieurs auteurs reconnus (De Ketele, 2001, 2006; Jonnaert, 2002, cité dans Gérard, 2010; Le Boterf, 1998; Legendre, 2005; Perrenoud, 2002; Roegiers, 2000, 2005) qu'une personne compétente doit pouvoir reproduire cette capacité à mobiliser les ressources à l'intérieur d'une famille de situations problèmes. Ce postulat sous-tend qu'une compétence ne peut nullement être inférée à partir d'une seule tâche, mais qu'elle doit pouvoir répondre à une pluralité de situations homogènes dont le niveau de difficulté ou de complexité doit s'accroître graduellement pour chacun des niveaux de compétence. Dès lors, la capacité d'adaptation d'un candidat à mobiliser convenablement les ressources s'avère capitale pour trouver des solutions adaptées à des situations professionnelles comparables. Certains auteurs (Beauchesne et al., 1993, cités dans Louis et al., 1996) soutiennent d'ailleurs que la mobilisation de l'ensemble des ressources peut être traduite en performance.

Malgré le fait que l'approche par compétences continue d'engendrer les débats sémantiques, idéologiques et politiques, il apparaît que les caractéristiques qui émergent

des définitions de la compétence semblent répondre aux aspirations de la professionnalisation (Le Boterf, 2007) ainsi qu'à la recherche de l'efficacité (Carbonneau, 1993; Louis et al., 1996) et de l'efficience en termes de capital humain et financier (Louis et al., 1996) « d'une industrie impatiente d'accueillir une main-d'œuvre adaptée à ses besoins » (Veer, 2002 dans Legendre, 2005, p. 248). L'intégration de ce changement de paradigme au monde professionnel semble s'être progressivement opérée en raison de la nécessité d'une meilleure adéquation entre les exigences d'une économie en mutation et les pratiques évaluatives destinées à répondre aux besoins contemporains issus du marché du travail (Louis et al., 1996).

2.2.2 Concept d'évaluation des compétences

La compétence suppose qu'une personne est reconnue comme étant capable de mobiliser et de combiner des ressources appropriées (savoirs, savoir-faire et savoir-être – attitudes) et d'accomplir des actes concrets afin de résoudre des problèmes complexes liés à une famille de situations (De Ketele *et al.*, 2001; Roegiers, 2000, 2005; Scallon, 2015, 2004a, 2004b; Tardif, 2006). Lors de situations problèmes, le candidat désirant démontrer sa compétence doit être capable de générer une série de comportements observables ou de performances (De Ketele, 2006; Perrenoud, 2002) lui permettant d'atteindre « un niveau de performance prédéterminé en fonction d'attentes fixées et de résultats désirés » (Legendre, 2005, p. 248).

De nombreux auteurs s'entendent sur le sens accordé au terme « évaluation ». Le *Dictionnaire actuel de l'éducation* et *Le grand dictionnaire terminologique* de l'Office québécois de la langue française (OQLF) définissent l'évaluation comme « une démarche permettant de porter un jugement, à partir de normes ou de critères établis, sur la valeur d'une situation, d'un processus ou d'un élément donné, en vue de décisions pédagogiques ou administratives » (ministère de l'Éducation du Québec (MEQ), 2004,

cité dans l'OQLF, 2012a et dans Legendre, 2005, p. 630). Le degré d'acceptation du sens porté à la mesure est cependant moins accueilli (OQLF, 2012a), vraisemblablement en raison du fait que l'évaluation repose fondamentalement sur des jugements de valeur, somme toute subjectifs, et qu'elle implique de juger judicieusement de la pertinence conjointe des descriptions qualitatives et quantitatives.

L'évaluation des compétences, dans le sens de Tardif (2006, p. 104), doit inclure :

un processus de collecte de données à partir de sources multiples et variées [...] qui permet [au candidat] de rendre compte : [de son] niveau de développement des compétences [...]; de son degré de maîtrise des ressources internes et externes qui peuvent être mobilisées et combinées dans leur mise en œuvre; et de l'étendue des situations dans lesquelles il est en mesure de déployer ces niveaux de développement et ces degrés de maîtrise.

À l'instar de Bourguignon (2011) et des auteurs précédents, nous pourrions soutenir que l'évaluation, dans une approche orientée sur la compétence, nécessite le recours à des tâches suffisamment complexes pour obliger l'application de stratégies de mobilisation et de transfert de ressources internes et externes liées aux capacités d'un candidat. Dans ce sens, Ketele (2006, p. 23) affirme que « être compétent suppose bien plus que la maîtrise de connaissances et de savoir-faire de base (ce que l'on tend à appeler les « ressources ») : cela suppose de pouvoir les mobiliser de façon pertinente dans des situations problèmes à résoudre ou dans des tâches complexes à effectuer ». Scallon (2004b, p. 119) signale dans le même sens que « les critères d'évaluation ne portent plus sur les qualités attendues d'une production particulière, mais sur la capacité à mobiliser des ressources, capacité qui correspond à la définition même d'une compétence ». Plusieurs auteurs conviennent que la mobilisation des ressources présuppose, comme Legendre le précise, que les ressources doivent être disponibles, opérationnelles et judicieusement mobilisables dans le temps imparti pour que la compétence soit démontrable.

Legendre (2005), en s'appuyant sur Le Boterf (2002), souligne également que la compétence ne doit pas se limiter à une seule façon de faire dans une situation donnée, mais qu'elle doit permettre l'accomplissement d'un ensemble de tâches. Wiggins (1989) affirme, dans le sens de Scallon (2015, 2004a), qu'évaluer une ou deux fois une performance est insuffisant pour mesurer l'habileté d'un individu à accomplir une diversité de tâches ou à constater les progrès accomplis. Selon Wiggins (1989, traduction libre, p. 705), pour attester de la compétence, il est nécessaire d'évaluer à plusieurs reprises la performance en ayant recours à une multitude et à une variété d'épreuves afin de certifier la qualité d'une réponse et ainsi, identifier la récurrence des réussites et des échecs. Scallon (2015) abonde également dans ce sens lorsqu'il souligne l'importance d'appuyer le jugement sur une famille de situations afin d'en réduire la subjectivité. L'inférence de la compétence découle conséquemment de la récurrence des comportements observés dans différentes situations comprises dans une même famille d'occurrences.

Scallon (2004a) stipule de plus que l'évaluation des compétences n'est pas la sommation de résultats fragmentaires, mais qu'elle commande de porter un jugement sensé et réfléchi. Il précise qu'il est nécessaire, pour porter ce jugement sur la compétence globale, de s'appuyer sur l'observation d'actions accomplies sur une période évolutive en considérant une pluralité de tâches ou de situations assignées à un moment défini. En d'autres mots, juger de la performance ou de la compétence d'une personne exige d'observer avec discernement l'accomplissement d'une multitude de tâches, réalisées à différents moments, pour ensuite les examiner avec objectivité et impartialité afin de porter un jugement juste et le moins subjectif possible.

Certains auteurs précisent de plus que la compétence suppose l'atteinte « d'un degré de performance [répondant] aux exigences minimales du marché du travail » (Direction générale de l'enseignement collégial (DGEC) du MEQ, 1991 dans Leroux,

2010, p. 62 et dans Legendre, 2005, p. 248) et d'une capacité démontrée à s'adapter à une diversité de situations dans le cadre d'un environnement professionnel (Lemoine, 2002 dans Legendre, 2005).

Pour toutes ces raisons, il apparaît important que l'évaluation des compétences soit finalisée et contextualisée de façon à tenir compte de la « qualité que l'on attribue à une personne jugée apte à exercer efficacement une fonction » dans un domaine qui témoigne de sa capacité à assumer une responsabilité (Landry 1987 dans Legendre, 2005, p. 248). Perrenoud (2004) abonde dans le même sens lorsqu'il mentionne que l'évaluation des compétences requiert une évaluation authentique qui repose sur des tâches contextualisées et des problèmes complexes permettant l'utilisation fonctionnelle de connaissances disciplinaires à des fins de résultats préalablement définis. Il souligne à cet effet que le candidat doit, afin d'attester de l'authenticité de l'évaluation prescrite, connaître la tâche et les exigences à satisfaire avant de se retrouver en situation d'évaluation.

2.2.3 Concept de standard de performance

Ainsi que le soulignent certains auteurs, il est nécessaire, en milieu de travail, d'évaluer la performance pour attester de la compétence (Legendre, 2005; Wiggins, 1989). Leroux et Bigras (2003, p. 29) précisent en ce sens que « l'évaluation en situation authentique repose [...] sur la mesure des performances complexes ». L'appréciation de la performance est, selon Scallon (2004b, p. 27), « au cœur de la méthodologie de l'évaluation ». Scallon (2015, p. 34) ajoute que la performance réfère à « la notion d'accomplissement, de résultat obtenu dans l'accomplissement d'une tâche, ou encore d'exploit ou de réussite » dont on reconnaît la qualité ou la valeur comme convenable dans le domaine. Selon Legendre (2005, p. 1029), la performance s'inscrit nécessairement dans la « mise en œuvre d'une activité orientée vers la réalisation d'une tâche comme indicateur d'une compétence ». Scallon (2015, 2004a; 2004, cité dans

Legendre, 2005, p. 1029) soutient, en ce sens, que la performance est la « manifestation concrète d'une compétence ». Le concept de standard de performance réfère à un niveau de performance. Ce niveau de performance, tel que défini dans le GDT, représente « le seuil à partir duquel on reconnaît qu'un objectif d'apprentissage a été atteint ou qu'une compétence a été développée dans un programme d'études donné » (MEQ, 2004, cité dans OQLF, 2012b), c'est-à-dire, le standard.

Le Boterf (1998) souligne pour sa part qu'inférer l'existence de la compétence d'un individu nécessite qu'un certain nombre de critères de performances soient respectés. Legendre (2005, p. 939) précise à cet égard qu'une norme regroupe un « ensemble des règles, des principes et des critères qui expriment les attentes d'un groupe par rapport à ses membres et qui orientent l'action, les décisions et les idées de ceux-ci ». Il fait référence à la notion de standards de performance en termes de norme partagée par une collectivité et qui précise les exigences stipulées auxquelles un individu doit de conformer. Conséquemment, un standard se définit comme ce « qui est conforme à une norme commune, à un archétype, à un modèle uniforme, à des exigences précises prédéterminées » (Legendre, 2005, p. 1256).

Le concept de standard est apparu, dans le monde de l'éducation, au début des années 1980 sous la forme de modalités d'observation ou de systèmes statistiques appliqués à des résultats (Scallon, 2004b). Selon Scallon (2004b), l'avenue de l'approche par compétences a élargi cette perspective en considérant dès lors l'atteinte d'un niveau de performance prédéterminé ou de la satisfaction d'exigences prédéfinies. Un standard de performance représente conséquemment, pour Scallon (2004b, p. 128), « un niveau de performance que les individus doivent démontrer ». L'Academy of Medical Royal Colleges (2015, traduction libre, p. 17) précise dans ce sens que les standards peuvent se définir comme un modèle structuré où des formulations en lien avec différents aspects du domaine médical sont présentées d'une façon à permettre le jugement de la performance d'un candidat dans le respect des conditions de reconnaissances

généralement reconnues dans la profession. Dans le même sens, l'OAQ, en s'appuyant la définition du GDT, fait référence aux standards de performance comme un « ensemble de règles techniques propres à une organisation ou à une industrie ou communes à des organisations et à des industries qui, pour des raisons d'expérience pratique, sont devenues des références pour la production de biens livrables, la prestation de services, la réalisation de processus divers ou la description de savoir-faire » (Bureau de normalisation du Québec (BNQ), 2007, cité dans OQLF, 2012c).

Pour Cizek (2012, traduction libre, p. 4), le concept de standards de performance est celui qui précise le niveau de performance requis de la personne évaluée pour être classée dans une catégorie de performance donnée d'une épreuve. Cet auteur approfondit ce concept en soulignant l'existence d'un lien explicite avec les modalités procédurales qui permettent d'établir différents niveaux de performance. Il précise à cet effet que deux catégories de performance sont minimalement requises pour distinguer le succès de l'échec à une épreuve certificative ou pour reconnaître le niveau de compétence attendu d'un groupe par rapport à ses membres. Norcini (2003) abonde dans ce sens en précisant qu'un standard de performance représente la valeur professionnelle qu'on accorde aux résultats obtenus dans le cadre d'une épreuve pour évaluer la capacité d'un individu sur la base d'une échelle de mesure qui permet de distinguer en différentes catégories ceux qui performant suffisamment bien de ceux qui n'y arrivent pas.

Cizek et Bunch (2007) précisent que chaque catégorie de performance doit être séparée d'une autre par un point de césure afin d'identifier explicitement les limites de chacune de ces catégories. Le domaine médical définit ce point de césure comme le processus d'établissement d'un seuil de réussite (*cut scores*) d'une épreuve où le standard de performance représente le niveau de performance minimal au-delà duquel l'épreuve est considérée comme réussie (Academy of Medical Royal Colleges, 2015, traduction

libre, p. 16). Les seuils de réussites permettront ultimement de jauger et de rendre compte du niveau de performance démontré au regard des objectifs, des résultats ou des attentes précisées (Cizek et Bunch, 2007, traduction libre, p. 13).

Une des définitions de l'ESP les plus utilisée est celle proposée par Cizek en 1993 :

Le bon suivi prescrit un système rationnel de règles ou de procédures résultant de l'attribution d'un nombre pour différencier deux ou plusieurs états ou degrés de performance (*ibid.*, 1993, traduction libre, p. 100).

Suivant cette définition, Cizek (2012) précise que les modalités procédurales doivent reposer sur les définitions traditionnelles du domaine de la mesure et ainsi, être attachées à un système rationnel de règles ou à des procédures permettant de différencier deux ou plusieurs degrés de performance (Cizek et Bunch, 2007, traduction libre, p. 15). L'auteur précise que c'est le niveau de performance attendu pour chacune des catégories de performance d'une épreuve qui permet un classement des candidats. Il faut souligner que certains auteurs s'y réfèrent comme seuil minimal de maîtrise ou comme niveau de performance minimal jugé convenable pour attester de la réussite ou de l'échec d'une épreuve (Cizek, 2012; Kane, 1998).

Au Québec, le seuil de maîtrise et le seuil de réussite sont employés, selon le contexte, comme des synonymes de standard de performance. Ainsi, le seuil de maîtrise est défini, dans la littérature québécoise, comme le point de césure ou de coupure qui « distingue l'atteinte ou non d'un objectif ou la réussite ou l'échec de la réalisation d'une tâche » (Legendre, 2005, p. 1233). Alors que le seuil de réussite s'emploie, selon le Gouvernement du Québec (1985, cité dans Legendre, 2005, p. 1233), pour établir le « niveau de qualité à partir duquel on considère une performance comme réussie », on remarque que la terminologie utilisée dans le cadre de l'établissement des standards de performance a un sens très rapproché. L'essentiel ici est de comprendre que les *cut scores* (points de coupure) fixent, sur une échelle de mesure, les limites de chacun des

niveaux de performance ou des classements des individus soumis à une épreuve (Cizek et Bunch, 2007).

Dans le cadre de notre recherche, l'OAQ mentionne que les standards de performance représentent les normes de la profession et « désignent ce qu'un acupuncteur doit être capable de réaliser ou d'accomplir dans son domaine à divers moments de sa pratique professionnelle » (OAQ, 2010, 2015, p. 10). Selon l'OAQ, ce concept vient soutenir le processus de « reconnaissance des compétences attendues d'un praticien possédant les qualifications essentielles pour assurer une pratique efficace et sécuritaire » (*ibid.*, p. 41).

Selon le Conseil de l'Europe (Noijons *et al.*, 2011) et l'American College of Cardiology Foundation (Cerqueira *et al.*, 2008), il est également crucial de clairement décrire les niveaux de difficulté des différents énoncés (items) de l'épreuve afin de favoriser une compréhension commune. À titre d'exemple, Block (1979) souligne, en présentant la taxonomie de Bloom, la nécessité de tenir compte de quatre niveaux de comportement – le savoir, la compréhension, l'application et l'intégration – pour classer les éléments d'une épreuve du plus simple au plus complexe. Eisner (2000, p. 3) prétend également que la « Taxonomie des objectifs pédagogiques : Vol. I : Domaine cognitif (Bloom *et al.*, 1956) [est] utilisée dans le monde entier pour faciliter l'élaboration des matériels d'évaluation ». De fait, comme les démarches d'évaluation ont déjà eu recours aux taxonomies d'objectifs pour classer les tâches liées à une épreuve, il apparaît pertinent de tenir compte des niveaux de difficulté des items associés aux tâches.

Notre recension de la littérature met de plus en lumière que l'évaluation de la performance porte, selon Santé Canada et l'Agence de la santé publique du Canada (2013), sur la pertinence et le rendement. Le rendement fait notamment référence à

l'efficacité. Or, selon plusieurs auteurs (*ibid.*, 2013; Gouvernement du Québec, 2009; Lamari, 2011; Scholz et Stein, 2011), la performance se mesure dans certains contextes par l'entremise d'une combinaison d'indicateurs tels que l'efficacité, l'efficience et l'économie. En se référant au contexte professionnel, la revue *Télescope* de l'École Nationale d'Administration publique (ENAP) abonde dans le même sens lorsqu'elle précise que :

[la] complexification des mandats confiés aux professionnels du secteur public [exerce une pression qui mène,] avec l'avènement du nouveau management public, les organisations [à adopter] de nouvelles exigences axées sur le [concept managérial] des 3E : économie, efficacité et efficience (Lamari, 2011, p. 8). [Que] les notions d'efficacité et d'efficience sont au cœur de l'action : [qu']il ne s'agit pas [...] de simplement réduire les coûts (efficience), mais de développer les connaissances et les compétences en fonction des objectifs (efficacité). C'est la condition sine qua non pour répondre aux exigences des parties prenantes internes et externes de l'administration publique (Scholz et Stein, 2011, p. 35). [...] Pour les administrations publiques, cela implique qu'elles soient en mesure de vérifier en permanence si leurs connaissances sont encore suffisantes pour produire à la fois efficacité et efficience, désignées par les termes capacités dynamiques (*ibid.*, p. 36).

Caseau (2011) souligne que la performance représente l'atteinte d'objectifs démontrée par l'efficacité en optimisant simultanément les moyens pour y parvenir, c'est-à-dire l'efficience. Le ministère de l'Emploi et du Développement social du Canada (EDSC, 2013) reconnaît l'importance de prendre en compte l'efficience dans les processus d'évaluation des compétences. Dans ce sens, l'échelle de mesure de l'efficience se rapporte à la capacité d'un candidat à intégrer des ressources pour atteindre efficacement les résultats souhaités. Le degré d'efficience est généralement exprimé en valeur, en coûts ou en temps.

L'efficience est, selon notre recension de la littérature consultée, un levier approprié

pour mesurer la performance. L'efficacité vise notamment à conjuguer la capacité d'une personne à discerner les différents aspects d'une situation problème liée à l'environnement de travail afin d'obtenir les meilleurs résultats au moindre coût (efficacité). L'efficacité mesure l'atteinte des objectifs attendus (résultats), en intégrant d'une part, de façon raisonnée et efficace (efficacité), les connaissances et les savoir-faire spécifiques à une profession et d'autre part, en mobilisant les ressources les plus appropriées (moyens) pour répondre le plus profitablement aux difficultés rencontrées en déployant l'énergie avantageusement (comportement efficace).

Comme l'évaluation de la compétence reconnaît qu'il existe plusieurs façons possibles de réaliser une tâche, le degré d'efficacité semble, en définitive, un repère approprié pour juger de l'efficacité des solutions apportées à un problème.

2.3 Outils produits par l'OAQ

Pour observer les performances et évaluer les compétences au regard des normes de la profession, l'OAQ s'est doté d'outils sur lesquels devait s'appuyer notre démarche d'ESP. Il nous apparaît ainsi essentiel, afin de bien situer le contexte de cette recherche, de présenter le référentiel des compétences défini par l'Ordre et le *Cahier d'observation des comportements cliniques* utilisé, lors de l'épreuve clinique, pour évaluer les compétences.

2.3.1 Polygramme des compétences liées à l'exercice de la profession d'acupuncteur au Québec

Le polygramme des compétences liées à l'exercice de la profession d'acupuncteur au Québec est issu d'une réflexion socioéconomique menée par les ordres professionnels sur les enjeux relatifs à la reconnaissance des compétences développées par l'expérience de professionnels formés à l'étranger. Ce polygramme représente l'aboutissement du processus mis en œuvre par l'organisme pour s'acquitter de ses

obligations en matière de reconnaissance des compétences et des acquis expérientiels. Ce document est devenu un référentiel fidèle et représentatif des normes validées en 2010 qui tiennent compte des particularités de l'exercice liées au contexte contemporain de la profession au Québec (OAQ, 2015). Cet exercice de circonscription de la pratique a été, pour l'OAQ, la première étape essentielle à l'évaluation des compétences professionnelles. Il a en effet permis de constituer un descriptif détaillé des éléments d'une bonne pratique attendue comme référentiel de compétences à partir duquel peut être comparée, pour évaluation, la qualité de la pratique d'un candidat. Dans ce sens, le polygramme décrit l'intégralité des compétences liées à l'exercice de la profession d'acupuncteurs au Québec (OAQ, 2015). Afin de bien situer le fondement de notre recherche, il apparaît essentiel de bien décrire ce référentiel des compétences.

Précisons d'entrée de jeu que le polygramme des compétences liées à l'exercice de la profession d'acupuncteur au Québec articule l'aptitude à exercer la profession au cœur d'une interaction entre trois dimensions qui lient l'ensemble des composantes de l'exercice de la profession. L'OAQ s'inspire de la définition de l'Ordre des infirmières et infirmiers du Québec (Leprohon *et al.*, 2009; Leprohon *et al.*, 2001) pour préciser que les compétences représentent la capacité à utiliser des connaissances pouvant provenir de différents champs de savoir (dimension fonctionnelle) afin de répondre aux exigences de la réalité de l'exercice (dimension professionnelle) pouvant apparaître dans l'ensemble des situations cliniques (dimension contextuelle). La figure 2.1 illustre, à la page suivante, une vue synoptique du polygramme avec ces trois dimensions.

Il est important de préciser que l'OAQ a octroyé une autorisation spécifique, disponible à l'appendice H, pour la conception et la reproduction des figures et des tableaux présents dans ce mémoire.

Vue synoptique

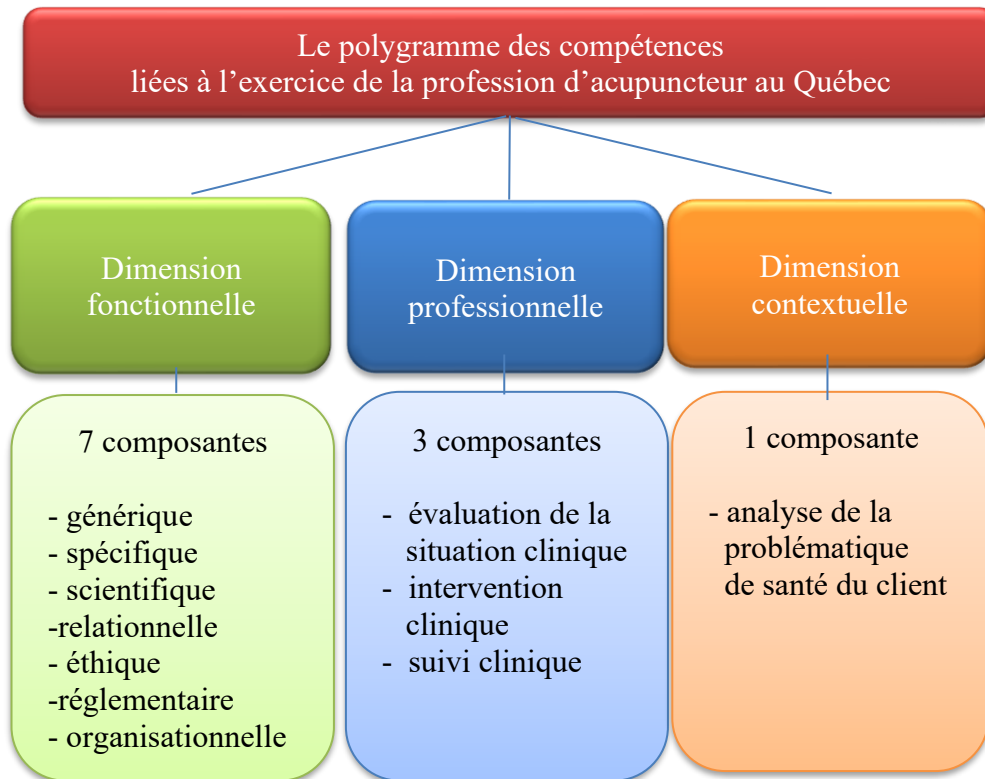


Figure 2.1 – Vue synoptique du polygramme des compétences liées à l'exercice de la profession d'acupuncteur au Québec

Les tableaux 2.1, 2.2 et 2.3 expliquent les 11 composantes relatives aux 3 dimensions (fonctionnelle, professionnelle et contextuelle) du polygramme des compétences liées à l'exercice de la profession d'acupuncteur au Québec adopté en 2010 par le conseil d'administration de l'OAQ, puis publié en 2015 (OAQ, 2015).

La dimension fonctionnelle, présentée comme fondement (tableau 2.1), propose sept composantes qui intègrent différents champs de savoirs pouvant alimenter la réflexion sur divers aspects de l'exercice de la profession.

Tableau 2.1 – Les sept composantes de la dimension fonctionnelle (OAQ, 2015, p. 22)

Dimension fonctionnelle	
Générique	La composante générique qui est constituée d'une compétence qui a trait à la capacité de l'acupuncteur à intégrer, dans son jugement clinique, les outils d'ordre conceptuel qui soutiennent la communication, la documentation et l'intervention.
Spécifique	La composante spécifique qui regroupe trois compétences qui ont trait à la capacité de l'acupuncteur de définir la situation clinique du client, de l'interpréter et d'établir un plan d'intervention fondées et documentées selon les paradigmes qui définissent les cadres théorique et conceptuel de la MTO.
Scientifique	La composante scientifique qui allie quatre compétences qui ont trait à la capacité de l'acupuncteur de comprendre et d'interpréter la situation clinique du client selon les paradigmes qui définissent les cadres théorique et conceptuel des sciences occidentales.
Relationnelle	La composante relationnelle qui comprend une compétence qui a trait à la capacité de l'acupuncteur d'établir une relation thérapeutique constructive avec le client : communiquer avec lui, le conseiller, l'aider à prendre en charge sa santé.
Éthique	La composante éthique qui présente une compétence qui a trait à la capacité de l'acupuncteur à agir en fonction d'un système de valeurs propre au contexte de la pratique.
Réglementaire	La composante réglementaire qui est définie par une compétence qui a trait à la capacité de l'acupuncteur de se conformer aux lois et aux règlements qui encadrent l'exercice.
Organisationnelle	La composante organisationnelle qui comporte trois compétences qui ont trait à la capacité de l'acupuncteur d'organiser ses activités en vue de prodiguer des soins de façon optimale.

La dimension professionnelle, structurée autour de sept compétences regroupées en trois composantes, présente (tableau 2.2) l'ensemble du processus de la relation clinique avec le client.

Tableau 2.2 –Les trois composantes de la dimension professionnelle (OAQ, 2015, p. 23)

Dimension professionnelle	
L'évaluation de la situation clinique	La composante de l'évaluation de la situation clinique regroupe trois compétences fondamentales qui permettent à l'acupuncteur d'effectuer une collecte et une synthèse des données d'un problème ou d'une situation, de les analyser et de les interpréter de façon critique pour en dégager les solutions appropriées afin de planifier une intervention rapidement, efficacement et à bon escient.
L'intervention clinique	La composante de l'intervention clinique regroupe spécifiquement les éléments liés au traitement.
La continuité clinique	La composante de la continuité clinique est constituée de trois compétences qui regroupent les activités permettant à l'acupuncteur de répondre à son obligation de moyen, notamment d'optimiser son intervention, de promouvoir la santé et d'être en relation avec d'autres professionnels liés au dossier.

La dimension contextuelle (tableau 2.3), constituée d'une composante qui décrit une seule compétence, situe l'intervention clinique de l'acupuncteur dans une problématique qui lie l'évolution de l'état de santé du client dans l'ensemble de sa relation au monde.

Tableau 2.3 – La composante de la dimension contextuelle

Dimension contextuelle	
L'analyse de la problématique de santé du client	La composante de l'analyse de la problématique de santé du client est constituée d'une seule compétence qui permet d'évaluer sa capacité à agir en fonction de la problématique de santé du client.

Globalement, le *Polygramme des compétences liés à l'exercice de la profession d'acupuncteur au Québec* regroupe 3 dimensions qui distinguent 11 composantes chapeautant les 22 compétences qui décrivent explicitement le référentiel sur lequel les standards de performances seront établis. Quatorze de ces compétences définissent la dimension fonctionnelle de la profession, 7 compétences cernent la dimension professionnelle et 1 compétence caractérise la dimension contextuelle.

La figure 2.2, présentée à la page suivante, expose dans une matrice la relation (cases grises) entre les 22 compétences et les 11 composantes liées aux 3 dimensions du polygramme des compétences cliniques des acupuncteurs du Québec.

La matrice des compétences de l'Ordre des acupuncteurs du Québec	Fonctionnelle							Professionnelle			Contextuelle
	La composante Générique	La composante Spécifique	La composante Scientifique	La composante Relationnelle	La composante Éthique	La composante relative à la Réglementation	La composante Organisationnelle	La composante de l'Évaluation de la situation clinique	La composante de l'Intervention clinique	La composante de la Continuité clinique	La composante de l'Analyse de la problématique de santé du client
Les compétences associées à la dimension fonctionnelle											
1. Développer une vision critique de la profession	✓										
2. Analyser l'état de santé selon les principes de la MTO		✓									
3. Construire un plan de traitement conforme aux principes de la MTO		✓									
4. Choisir des méthodes d'intervention liées à la MTO		✓									
5. Évaluer la situation clinique sous l'angle des sciences occidentales			✓								
6. Faire des liens entre les problèmes de santé et les divers types d'intervention en sciences occidentales			✓								
7. Situer l'action de l'acupuncture en regard des sciences occidentales			✓								
8. Situer la MTO en regard des sciences occidentales			✓								
9. Établir une relation thérapeutique constructive				✓							
10. Agir en fonction d'un système de valeurs propre au contexte de la pratique					✓						
11. Conformer sa pratique aux lois et aux règlements qui encadrent l'exercice de la profession						✓					
12. Administrer une clinique d'acupuncture							✓				
13. Aménager un environnement clinique confortable et sécuritaire							✓				
14. Optimiser son intervention clinique							✓				
Les compétences associées à la dimension professionnelle											
15. Effectuer la collecte des données								✓			
16. Interpréter les données recueillies								✓			
17. Planifier l'intervention								✓			
18. Appliquer le protocole d'intervention									✓		
19. Faire le suivi de l'intervention auprès du client										✓	
20. Favoriser la responsabilisation du client dans la prise en charge de sa santé										✓	
21. Interagir avec divers intervenants dans un contexte professionnel										✓	
Les compétences associées à la dimension contextuelle											
22. Situer l'état de santé du client dans l'ensemble de sa relation au monde											✓

Figure 2.2 – Matrice des compétences de l'Ordre des acupuncteurs du Québec (reproduction autorisée)

Le polygramme des compétences cliniques des acupuncteurs du Québec propose une opérationnalisation qui décompose chacune des 22 compétences en plusieurs éléments de compétence associés à un certain nombre de critères de performance.

Le tableau 2.4 illustre l'opérationnalisation de la première composante de la dimension professionnelle à partir d'un extrait fragmentaire du document intitulé *Polygramme des compétences liées à l'exercice de la profession d'acupuncteur au Québec* (OAQ, 2015).

Tableau 2.4 – Illustration de la première composante de la dimension professionnelle

4.3.1. La composante de l'évaluation de la situation clinique		
L'évaluation de la situation clinique regroupe trois compétences fondamentales qui permettent à l'acupuncteur d'effectuer une collecte et une synthèse des données d'un problème ou d'une situation, de les analyser et les interpréter de façon critique pour dégager les solutions appropriées afin de planifier une intervention rapidement, efficacement et à bon escient.		
Compétences	Éléments de compétence	Critères de performance
1. Effectuer la collecte des données	1. Identifier les besoins du client	1. Cerner les motifs principaux de consultation
		2. Évaluer les attentes du client
		3. Appliquer une approche centrée sur le client
	2. Recueillir l'information sur l'état de santé du client au moyen des méthodes de la MTO.	1. Réaliser des quatre temps de l'examen
		2. Modifier la stratégie d'examen en fonction de l'information émergente
	3. Obtenir de l'information selon les données diagnostiques biomédicales, les antécédents médicaux et de santé	1. Procéder, selon les circonstances, à l'examen physique au moyen de techniques occidentales
		2. Utiliser les données provenant d'autres intervenants
		3. Évaluer les fonctions cognitives, l'état émotionnel et le comportement

L'encadré prédominant, situé dans la première ligne du tableau 2.4, décrit la composante de l'évaluation de la situation clinique. La colonne de gauche contient le nom de la première compétence de cette composante. La deuxième colonne présente les trois éléments de compétence qui y sont associés. La troisième colonne propose deux ou trois critères de performance, selon l'élément de compétence évalué, qui permettent de préciser les modalités selon lesquelles les éléments de compétence doivent être démontrés. À titre d'exemple, pour effectuer la collecte des données, le candidat doit, entre autres, être capable de recueillir l'information sur l'état de santé du client au moyen des méthodes de la MTO qui incluent de pouvoir réaliser les quatre temps de l'examen.

2.3.2 Cahier d'observation des comportements cliniques destiné aux maîtres de clinique dument accrédités par l'Ordre des acupuncteurs du Québec

L'OAQ, après avoir défini le cadre référentiel des compétences du polygramme des compétences, a entrepris, afin d'objectiver davantage l'évaluation des candidats désirant intégrer la profession, de préciser les critères de performance sur lesquels s'appuie l'observation des comportements attendus d'un praticien possédant les compétences nécessaires pour assurer une pratique efficace et sécuritaire (OAQ, 2014a, p. 1). Pour ce faire, un cadre procédural d'intégration des éléments de compétences a, dans un premier temps, été défini. Ensuite, des grilles d'observation des savoir-faire liés aux comportements cliniques et destinées aux maîtres de clinique dument accrédités par l'OAQ ont également été élaborées.

2.3.2.1 Cadre procédural d'intégration des éléments de compétences

Les dirigeants de l'OAQ, suivant la définition retenue de la notion de compétence et les recommandations des spécialistes du domaine de l'évaluation, ont élaboré un processus d'intégration liant les différents savoir-faire observables associés aux

comportements attendus d'un candidat compétent. Ils ont retenu comme prémisse qu'il était impératif, afin d'observer objectivement les comportements attendus des candidats, d'opérationnaliser chacun des énoncés de compétence.

Dans cette optique, l'OAQ a entrepris, en 2012, une deuxième démarche de validation qui a permis de constituer un cadre procédural d'intégration et suivant laquelle ont été définies, en 2013, les grilles d'observation des savoir-faire liés aux comportements cliniques.

Le cadre procédural d'intégration, présenté à la figure 2.3, défini et validé par l'OAQ en juin 2012, s'appuie sur les définitions de la compétence présentées par Le Boterf (1997), Tremblay (1994) et Leroux. Celles-ci supposent une mise en relation, une mobilisation et la combinaison d'un ensemble de ressources « qui s'intègrent et se combinent les unes aux autres » (Leroux, 2010, p. 64) afin de soutenir un agir complexe.

À partir de ces définitions, l'OAQ a validé un cadre de hiérarchisation de l'ensemble des composantes du polygramme des compétences cliniques des acupuncteurs du Québec. Ce cadre fait ressortir la notion de capacité d'agir complexe, à la base de la reconnaissance de la compétence telle qu'énoncée dans les définitions retenues (OAQ, 2013). Le cadre procédural proposé intègre, en effet, l'ensemble des éléments de compétence du polygramme des compétences liées à l'exercice de la profession d'acupuncteur au Québec dans une structure qui permet d'apprécier la capacité d'agir complexe en situation clinique. La procédure reconnaît la mobilisation potentielle de l'ensemble des éléments des compétences fonctionnelles et contextuelles dans la mise en œuvre d'un savoir-agir professionnel. Le cadre d'intégration identifie, à cet effet, cinq des sept compétences liées à la dimension professionnelle comme étant la pierre angulaire qui regroupe 23 éléments de compétences, qualifiés de primaires, dans la gestion quotidienne de situations cliniques. Les 44 autres éléments, qualifiés de

complémentaires, sont davantage reconnus comme ressources contribuant indirectement à la qualité de la relation clinique que l'acupuncteur a la responsabilité de développer. Un candidat désirant démontrer sa capacité à effectuer une bonne collecte des données pourrait, par exemple, appliquer un des éléments de la compétence relationnelle qui présente les principes d'une relation thérapeutique basée sur le partenariat thérapeute/client. Dans le même sens, il est impossible d'interpréter les données recueillies sans être capable d'utiliser les principales constituantes de la théorie de la médecine traditionnelle orientale (MTO) telle qu'énoncée dans la composante spécifique de la dimension fonctionnelle.

La figure 2.3 présente, à la page suivante, une vue synoptique de l'application de la démarche procédurale d'intégration des cinq compétences associées aux trois composantes de la dimension professionnelle incluses dans l'épreuve d'évaluation des acquis expérientiels.

Application de la démarche procédurale d'intégration; le candidat doit être capable de faire :						
1. L'évaluation de la situation clinique			2. L'intervention clinique	3. Le suivi clinique		
1.1 Effectuer la collecte des données (Compétence 16)	1.2 Interpréter les données recueillies (Compétence 17)	1.3 Planifier l'intervention (Compétence 18)	2.1 Appliquer le protocole d'intervention (Compétence 19)	3.1 Faire le suivi de l'intervention (Compétence 20)	Favoriser la responsabilisation	Interagir avec divers intervenants dans un contexte professionnel
Éléments de compétence primaires (Éléments évalués dans le bloc 3 du parcours de mise à jour professionnelle)					Éléments de compétence complémentaires	
1.1.1 Identifier les besoins du client	1.2.1 Sélectionner l'information pertinente à retenir	1.3.1 Déterminer une orientation thérapeutique cohérente	2.1.1 Choisir des méthodes de traitement	3.1.1 Évaluer l'évolution de la situation clinique	Aider le client à améliorer ou à maintenir sa santé, son bien-être et sa qualité de vie	Communiquer des informations à d'autres professionnels de la santé
1.1.2 Recueillir l'information sur l'état de santé du client au moyen des méthodes de la MTO.	1.2.2 Utiliser un processus d'analyse rigoureux	1.3.2 Définir un plan de traitement thérapeutique cohérent	2.1.2 Adapter son intervention à la situation clinique	Éléments évalués dans le bloc 2 du parcours de mise à jour professionnelle - Assurer la mise à jour complète et précise du dossier - Gérer les dossiers des clients - Utiliser la terminologie liée aux concepts de la MTO	Gérer la relation d'aide	Collaborer avec d'autres professionnels de la santé pour coordonner un plan de traitement
1.1.3 Obtenir de l'information selon les données diagnostiques biomédicales, les antécédents médicaux et de santé	1.2.3 Savoir repérer les situations cliniques présentant des signes de gravité, les référer et en assurer le suivi professionnel	- Construire un plan de traitement conforme aux principes de la MTO - Choisir des méthodes d'intervention liées à la MTO	2.1.3 Exécuter le traitement		Reconnaître la continuité du processus liant la santé et la maladie	Identifier les axes de collaborations multidisciplinaires
1.1.4 Adapter sa stratégie de communication selon les différentes situations		1.3.3 Établir une entente avec le client sur le plan d'intervention thérapeutique cohérent	2.1.4 Démontrer un comportement clinique qui assure le contrôle de qualité et l'élimination des risques d'accident		Appliquer les principes d'une relation thérapeutique basée sur le partenariat thérapeute/client	Situer la profession en regard des autres disciplines de la santé vues par la MTO et par les champs de savoir de la science occidentale
1.1.5 Appliquer les principes d'une relation thérapeutique basée sur le partenariat thérapeute/client			2.1.5 Optimiser son intervention clinique			Distinguer les différences entre les conceptions de la santé vues par la MTO et par les champs de savoir de la science occidentale
1.1.6 Gérer la relation d'aide			2.1.6 Intervenir dans des situations d'urgence, de crise ou de violence			Analyser l'action des différents modes thérapeutiques proposés par la MTO en termes scientifiques
1.1.7 Assurer la mise à jour complète et précise du dossier			2.1.7 Repérer et prévenir les dangers d'infection			
			2.1.8 Reconnaître les éléments nécessaires à un environnement de soins de qualité et sécuritaire			
			2.1.9 Poser les gestes qui découlent des valeurs et des principes éthiques de la profession			

Figure 2.3 – Cadre procédural d'intégration des éléments de compétences (OAQ, 2015, p. 4, reproduction autorisée)

2.3.2.2 Grilles d'observation des savoir-faire liés aux comportements cliniques

La démarche de validation (juin 2013) des grilles d'observation des savoir-faire liés aux comportements cliniques avait pour objet de décrire les comportements observables pour chacun des critères de performance associés aux cinq compétences primaires de la procédure d'intégration (OAQ, 2013). Cet exercice a permis de mettre en relation une quantité importante de savoir-faire observables pouvant être intégrés à chacun des critères de performance par un candidat désirant démontrer ses compétences.

Comme le souligne Scallon :

Pour concevoir une tâche ou une situation problème permettant d'inférer une compétence, il faut « interpréter » l'énoncé de cette compétence en termes opérationnels : cela exige de traduire les composantes ou les capacités qui sous-tendent cette compétence en ressources à mobiliser. C'est là un enjeu fondamental de l'évaluation (Scallon, 2004b, p. 120).

L'exemple (tableau 2.5) présenté à la page suivante démontre de quelle façon deux éléments de compétence liés à la dimension fonctionnelle (1.2.1.1 et 1.2.1.2) sont identifiés parmi les savoir-faire observables associés à l'élément de compétence 1.1.7 servant à évaluer la capacité à effectuer la collecte des données (OAQ, 2014a, 2014b).

Tableau 2.5 – Illustration de l'intégration des éléments de compétences fonctionnelles à la dimension professionnelle (OAQ, 2015, p. 15, reproduction autorisée)

Composante 1 de la compétence professionnelle : Le candidat doit démontrer sa capacité de faire « L'évaluation de la situation clinique » Compétence 15 : Effectuer la collecte des données Élément de compétence 1.1.7 : Assurer la mise à jour complète et précise du dossier	
Critères de performance	Savoir-faire observable
2. Inscrire au dossier de façon claire, lisible et concise, en se servant de la terminologie de la MTO, l'opinion de l'acupuncteur quant à l'état énergétique du client	Le dossier présente une vision d'ensemble de la situation clinique détaillée et pertinente qui permet d'apprécier la capacité de l'acupuncteur à : • utiliser la terminologie liée aux concepts de la MTO (1.2.1.1) • utiliser les principales constituantes de la théorie MTO (1.2.1.2) - Identifier le tableau clinique lié à la raison de consultation et au terrain 

Selon l'OAQ (2014b), les grilles d'observation des savoir-faire liés aux comportements cliniques fournissent une liste étoffée de comportements qui accroît l'objectivité des critères de performance. L'OAQ considère de plus que cet outil offre la possibilité d'une observation ouverte et prospective qui permet de repérer les façons de faire satisfaisantes qu'un candidat peut présenter afin de démontrer ses compétences. Celui-ci se voit ainsi offrir la possibilité d'agencer une combinaison de connaissances, d'habiletés et d'attitudes afin de démontrer sa capacité à prendre des décisions judicieuses, à résoudre des problèmes et à agir adéquatement dans un contexte clinique donné (OAQ, 2013). Les savoir-faire observables sont, dans ce sens, vus comme des attributs opérationnels liés aux standards de performance (OAQ, 2013).

2.4 Établissement des standards de performance

Une fois les grilles d'observation des savoir-faire liés aux comportements cliniques constituées, l'objectif subséquent, pour que l'OAQ puisse formaliser son instrument d'évaluation, était d'instaurer des standards de performance. La suite de notre démarche illustre conséquemment les étapes qui ont permis d'établir, à l'aide de méthodes reconnues dans le domaine scientifique, les standards de performance qui servent aujourd'hui à évaluer les compétences professionnelles minimales requises pour exercer la profession d'acupuncteur au Québec.

Suivant la définition de la compétence retenue par l'OAQ, selon laquelle c'est la capacité d'intégrer des ressources internes et externes dans une démarche procédurale qui permet l'inférence des compétences, il appert qu'une étape intermédiaire d'élaboration de tâches intégratives devait être effectuée afin de produire les items de l'épreuve clinique (Cizek *et al.*, 2004) nécessaires à l'établissement des standards.

La suite de notre présentation décrit le processus d'élaboration et de validation des tâches intégratives et des items de l'épreuve clinique avant d'expliquer les méthodes d'établissement de standards.

2.4.1 Processus d'élaboration des tâches intégratives et des items de l'épreuve clinique

Afin d'assurer la rigueur de la mise en œuvre du processus de reconnaissance des compétences des personnes sollicitant un droit de pratique de l'acupuncture au Québec, nous avons entrepris en janvier 2014 une démarche d'élaboration et de validation des tâches intégratives et des items de l'épreuve clinique (Lainey et Dion, 2014). La littérature consultée n'ayant toutefois pas permis de recenser un modèle approprié sur lequel nous appuyer pour normaliser cette opération, nous avons choisi d'adapter un modèle déjà existant afin de pouvoir tenir compte de l'ensemble des impératifs de

l'OAQ. Nous procéderons, conséquemment, à la présentation du cadre qui a permis de structurer l'élaboration des tâches intégratives et des items de l'épreuve clinique.

2.4.1.1 Modèle d'élaboration des tâches intégratives et des items de l'épreuve clinique

La responsabilité de l'OAQ d'assurer une évaluation authentique, valide et objective oblige que le jugement sur l'atteinte du niveau de compétence repose sur des mises en situation cliniques réelles devant objectivement être observables et mesurables.

Aussi, afin de fournir une approche scientifique reconnue sur laquelle appuyer l'élaboration des tâches intégratives, nous avons proposé à l'OAQ une version adaptée du modèle SMART, originalement défini par Doran en 1981 pour élaborer des objectifs. Ce modèle offrait, selon nous, une structure reconnue sur laquelle appuyer notre démarche d'élaboration d'items standardisés. La version adaptée, que nous allons maintenant présenter, qui a été adopté en 2014 par l'OAQ, permettra de situer la structure ayant servi à normaliser l'élaboration des tâches intégratives et des items de l'épreuve clinique.

Le modèle SMART offre selon nous une structure reconnue sur laquelle appuyer la démarche d'élaboration d'items standardisés.

Précisons, pour commencer, que l'acronyme SMART, selon Doran (1981), représente le cadre permettant de structurer un objectif qui doit être :

Spécifique	afin de cibler une amélioration spécifique;
Mesurable	afin de quantifier ou tout au moins proposer un indicateur de progrès;
Assignable	afin de préciser qui fera l'action;
Réaliste	afin de déterminer les résultats qui peuvent être atteints de façon réaliste, étant donné les compétences disponibles;
Temporel	afin de préciser quand les résultats peuvent être obtenus.

Au fil des années, diverses variantes proposées, ont permis à différentes organisations d'adapter le modèle à leurs besoins (Environnement Canada, 2013; Service public fédéral – Personnel et Organisation, 2010). L'Ordre des infirmières et infirmiers de l'Ontario (OIIO) a, par exemple, adopté une variante de l'acronyme qui définit le « S » du SMART comme devant être Spécifique, précis, détaillé, ciblé et clairement rédigé (OIIO, 2014, p. 3). Environnement Canada (2013, p. 23) a, pour sa part, transformé le « A » du SMART pour préciser que l'objectif devait être **atteignable** de façon à être réalisable à l'intérieur des contraintes d'ordre opérationnelles, tout en étant tributaire de la disponibilité des ressources, des connaissances et du calendrier. Ce dernier a adapté la définition du « R » du SMART en spécifiant que l'objectif devait être **réaliste**, lié aux priorités et au mandat du gouvernement et contribuer à l'obtention d'un résultat souhaité dans la société canadienne, l'économie et l'environnement (Environnement Canada, 2013, p. 23). L'Ordre des physiothérapeutes de l'Ontario (2012) a quant à lui adapté le « A » et le « T » pour tenir compte des résultats produits par l'action, dans un temps déterminé en fonction des ressources disponibles. Plusieurs auteurs (Ostermeier *et al.*, 2015) ont récemment proposé, dans la *Revue internationale du travail*, une variante du « A », en précisant que les indicateurs SMART, utilisés pour mesurer les progrès accomplis en matière d'emploi, devraient être **adaptés** à l'action de sorte que « les informations fournies par l'indicateur doivent présenter une utilité

réelle pour ceux qui prennent les décisions » (*ibid.*, 2015, p. 319). Ces auteurs (*ibid.*, 2015) ont, eux aussi, transposé chacun des descripteurs, mais cette fois en y intégrant l'ensemble des dimensions dont tenir compte dans l'évaluation.

À l'instar des adaptations contemporaines du modèle SMART de Doran et afin de tenir compte de la réalité associée au contexte de l'évaluation des compétences cliniques des acupuncteurs, nous avons proposé une variante du SMART. En nous appuyant sur Gérard (2009, p. 26-27), qui souligne que « l'approche par les compétences consiste [...] à orienter les actions [...] vers un objectif d'intégration, c'est-à-dire un objectif qui nécessite la coordination entre différentes ressources, entre différents objectifs spécifiques », nous avons misé sur le fait que le modèle I-SMART que nous proposons pouvait avantageusement être appliqué à la construction de tâches intégratives complexes et authentiques.

L'acronyme I-SMART (OAQ, 2014a) proposé représente de fait la structure qui a permis de normaliser et d'associer des situations cliniques réelles à des tâches authentiques favorisant l'intégration d'indicateurs validés par des experts du domaine :

Intégrateurs	afin d'intégrer des situations complexes authentiques et conceptualisées;
Spécifiques	afin de cibler précisément ce qui doit être accompli;
Mesurables	afin de quantifier des comportements et des faits observables pour chacun des indicateurs;
Authentiques	afin d'être jugé recevable par les experts du domaine;
Réalisables	afin d'être humainement faisable considérant les ressources, les rôles et les responsabilités liées à la profession;

Temporels afin de mesurer le temps d'accomplissement observé considérant le référentiel d'évaluation et les raisons.

L'insertion du « I » à l'acronyme offre donc une valeur ajoutée au modèle proposé par Doran en renforçant la nécessité d'une cohérence intégrative pour la construction des tâches servant à évaluer les compétences. Cette adaptation permettait ainsi d'actualiser le modèle de Doran à la réalité contemporaine et d'harmoniser les descripteurs de l'acronyme aux composantes acceptées de la définition de la compétence.

2.4.1.2 Procédure d'élaboration et de validation des tâches intégratives et des items de l'épreuve clinique

Dans le sens de ce que Gérard (2009) propose, il apparaissait important pour l'OAQ d'établir le pont entre les indicateurs de performances et les éléments de compétences afin d'évaluer avec précision, à partir d'éléments observables et concrets, le niveau de maîtrise de chacun des critères de performance associés. L'élaboration des indicateurs de performance s'est conséquemment appuyée sur la liste énoncée dans les grilles d'observation des savoir-faire liés aux comportements cliniques des critères de performances et des savoir-faire observables. Ces indicateurs ont été constitués de manière à favoriser l'intégration de l'ensemble des critères de performances et des savoir-faire observables liés aux différents éléments de compétences. Le nombre total d'indicateurs a conséquemment été déterminé par la quantité de ressources potentiellement intégrables dans chacun des éléments de compétence et selon le nombre de situations cliniques dont il faut tenir compte.

L'I-SMART sur lequel s'appuie cette opération se définit comme ce qui est présenté, à la page suivante, au tableau 2.6.

Tableau 2.6 –Le modèle I-SMART

I	Les situations cliniques réelles regroupent un certain nombre de tâches qui doivent être suffisamment intégratives pour permettre aux candidats évalués de démontrer, dans le sens de la définition de la compétence retenue, leur capacité à mobiliser des ressources appropriées leur permettant de répondre de façon satisfaisante aux situations problèmes réelles liant l'ensemble des éléments de compétences cliniques de la profession. Dans ce sens, chacun des indicateurs doit pouvoir être intégré à une tâche suffisamment complexe pour répondre aux critères nécessaires à l'évaluation d'une compétence.
S	Les tâches intégratives, qui correspondent aux 23 éléments de compétences, ainsi que les indicateurs qui y sont associés se doivent d'être spécifiques , c'est-à-dire d'être suffisamment explicites pour cibler de façon précise ce qui doit être accompli par les candidats évalués.
M	Les indicateurs associés à chacun des éléments de compétences doivent être définis à partir des savoir-faire observables liés aux critères de performance présentés dans le polygramme des compétences cliniques des acupuncteurs du Québec, de façon à permettre des observations mesurables et quantifiables.
A	Les indicateurs associés à chacune des tâches doivent être authentiques et acceptables du point de vue des experts du domaine.
R	Les tâches ainsi que les indicateurs doivent être réalisables et tenir compte des ressources disponibles, des rôles et des principales responsabilités, ainsi qu'être rattachées au contexte d'exercice de la profession qui s'inscrit dans la vision et la mission de l'Ordre des acupuncteurs du Québec.
T	Chaque tâche et chaque indicateur doit préciser le temps imparti à son exécution afin de permettre l'observation des performances. Ce délai doit être réaliste afin d'aider les candidats à canaliser leurs efforts lors de l'exécution des tâches et d'aider le maître de clinique à observer les performances dans les limites de temps imparties à leur accomplissement.

L'étape d'élaboration des tâches intégratives a permis d'identifier les situations cliniques susceptibles de favoriser l'intégration des ressources associées aux 23 éléments de compétences devant être réussis par un candidat minimalement compétent. Le « I » du I-SMART a servi, pour ce faire, à préciser la notion de situations problèmes suffisamment complexes pour favoriser l'intégration des différentes constituantes des éléments de compétences devant être évalués. La portion SMART a permis de délimiter

la structure d'un certain nombre de ressources devant être intégrées pour démontrer la qualité de ce qui est réussi et de ce qui doit être amélioré. L'élaboration des tâches I-SMART s'est conséquemment appuyée sur les critères de performances et les savoir-faire observables contenus dans l'observation des compétences cliniques validées en 2013 par l'OAQ. Ceci a permis d'objectiver les observations nécessaires à mesurer le degré d'efficience démontré lors de la réalisation des tâches liées aux situations cliniques réelles. Au final, cette démarche nous a permis d'établir des standards de performance pour chacun des 23 éléments de compétences en considérant la quantité et le niveau de difficulté des indicateurs de performance (items) susceptibles d'être intégrés dans la résolution des différentes situations problèmes.

Il est cependant important de noter que cette liste d'indicateurs ne doit pas être présentée comme exhaustive. Tout comme le précise Gérard (2009), l'évaluateur doit éviter de définir à priori une liste d'indicateurs indispensables, car il risque de s'empêcher de tenir compte des savoir-faire auxquels il n'avait pas pensé. Dans ce sens, l'évaluation du niveau de maîtrise de chacun des critères doit, selon Gérard (2009), permettre de s'attarder à l'ajout d'indicateurs novateurs, concrets, observables et contextualisés. Ce même auteur précise que les indicateurs fournissent des informations objectives et documentées, basées sur des règles d'utilisations précises, qui orientent l'observateur sur ce qu'il doit concrètement examiner avec rigueur pour procéder à une inférence déductive admissible du niveau de maîtrise d'un critère par un candidat (Gérard, 2009) lorsqu'elle permet de la justifier. Gagné et Godin (1999) soulignent dans ce sens que les items à réponses construites (ouvertes) permettent d'inclure les options de réponse significatives et de diminuer le risque de négliger certaines options de réponse pertinentes. À l'instar de Gérard, Gagné et Godin, nous croyons que l'amélioration continue des indicateurs permettra de garantir, par le recoupement d'indicateurs qualitatifs et quantitatifs, la collecte de l'ensemble des informations pertinentes, valides et fiables nécessaires à la vérification de ce qui doit

concrètement être vérifié par des faits observables tout en assurant l'atteinte des mêmes résultats par différents observateurs.

L'étape de la validation des tâches et des items a été effectuée, grâce à la grille d'analyse I-SMART présentée en annexe, par un comité constitué de quatre experts occupant une fonction significative dans le processus de reconnaissance des compétences des acupuncteurs exerçant au Québec.

2.4.2 Méthodes courantes d'établissement de standards de performance

La littérature recense une multitude de méthodes pour l'établissement des standards de performance. Afin d'orienter et de canaliser notre recherche vers les méthodes les plus appropriées à une épreuve clinique authentique, il est utile, dans un premier temps, d'explorer brièvement l'historique des méthodes d'ESP. Dans un deuxième temps, une analyse comparative des principes convergents et divergents des méthodes les plus reconnues sera effectuée afin de retenir les principaux éléments appropriés à une épreuve clinique constituée d'items à performance observable (IPO). Cette analyse amènera finalement à présenter la méthode alternative proposée pour établir des standards de performance à une épreuve clinique dans un contexte d'octroi d'un permis d'exercice d'une profession au Québec.

2.4.2.1 Historique des méthodes d'établissement de standards de performance

Quoique les méthodes d'ESP remontent à des milliers d'années, ce n'est qu'au milieu des années 1950 qu'elles ont attiré l'attention des chercheurs (Cizek, 2012, traduction libre, p. 15). À cette époque, les standards de performance étaient simplement établis sur l'acceptation d'un seuil de réussite exprimé en termes de pourcentage de réponses correctes, dans une fourchette généralement prédéterminée et comprise entre 60 % et 75 % (Ebel, 1965, cité dans Cizek, 2012, traduction libre, p. 15). Cette méthode,

autrefois qualifiée d'arbitraire, permettait principalement une comparaison de la productivité d'un candidat par rapport aux autres candidats et une mise en rang croissante des scores de rendement obtenus par ces derniers (Raîche, 2005). Cette conception du standard de performance, dans l'optique de promouvoir la sélection des candidats les plus performants, a donné naissance à l'interprétation normative des résultats scolaires. Les seuils de réussite de cette époque n'en demeurent pas moins imparfaits et critiqués du fait qu'ils ne permettent pas d'attester de la compétence authentique d'un candidat à exercer un métier ou une profession (Raîche, 2005). Pour bon nombre de chercheurs de cette époque, une incertitude subsistait face à l'aspect arbitraire de l'interprétation des résultats d'un test sur la base d'un standard qui ne tient pas compte du niveau de difficulté des items du test. Ce courant d'incertitude semble à l'origine des débats sur la conceptualisation de la notion de compétence, notamment sur la notion de situation complexe et authentique apte à conférer une certaine transférabilité et un réinvestissement dans le monde du travail.

Les méthodes d'ESP contemporaines trouvent leur premier fondement théorique en 1954, alors que la première méthode centrée sur le test a été introduite par Nedelsky. Cet auteur est le premier à avoir conceptualisé la notion d'un individu minimalement compétent en se représentant un étudiant situé à la frontière entre le succès et l'échec à un test constitué d'items à choix de réponses multiples (Cizek, 2012). La tendance à faire passer des tests à choix de réponses multiples, où les seuils de réussite persistaient à être très bas, s'est maintenue jusque dans les années 1970 et 1980. Ce n'est qu'au début des années 1980, pour ajouter au rehaussement du niveau de crédibilité des baccalauréats, qu'a progressivement émergé une nouvelle forme de référence intégrant des critères de mesure qualitatifs dans le processus d'évaluation de la performance. Selon cette classification critériée (absolue), le résultat obtenu au test (la performance) n'est plus comparé aux résultats des autres candidats, mais plutôt à des résultats spécifiquement attendus pour confirmer l'atteinte du standard établi. L'ESP impose,

dans ce contexte, de délimiter une frontière franche entre les deux catégories de performance traditionnelle (succès ou échec) dont les résultats sont avant tout liés à des indicateurs de performance quantitatifs (%). Ici repose le déclenchement de l'établissement des points de coupure.

Au cours des années 1990, de nouvelles méthodes d'ESP ont été introduites pour accroître l'efficacité des programmes postsecondaires et pour répondre aux nouvelles exigences en éducation. Les parties prenantes revendiquaient alors la nécessité d'établir des standards de performance plus rigoureux qui s'appuyaient sur l'utilisation d'items plus difficiles et plus complexes. Paradoxalement, la tendance de cette époque consistait à faire passer des tests sans tenir compte du niveau de difficulté des items, ce qui avait pour conséquence de favoriser des variations dans l'exigence des épreuves qui devaient normalement déterminer le même niveau de compétence. Selon Cizek, (2012, traduction libre, p. 28), cette réalité peut, dans certains cas, mener à deux erreurs de classification, soit : a) un seuil de performance excessivement élevé qui peut conduire à faire échouer un candidat qui devrait normalement réussir l'épreuve ou b) un seuil de performance relativement faible permettant à un candidat incompetent ou quasi compétent de réussir l'épreuve. Ce n'est qu'à la fin des années 1990 que les méthodes d'ESP sont devenues plus claires et explicites sur la procédure utilisée pour établir les standards de performance. Certaines recommandations ont été successivement formulées pour bonifier la procédure. Un jugement sur le niveau de difficulté des items de l'épreuve, une période d'entraînement à l'exercice de jugement avec rétroaction, une expérimentation du format du test et de rétroaction normative ont entre autres été intégrés au processus (Cizek, 2012, traduction libre, p. 16-17). À partir des années 2000, le nombre de catégories s'est également accru, passant de deux catégories de performance (succès ou échec) à trois catégories (par exemple : *Basic*, *Proficient*, *Advanced*) (Cizek, 2012, p. 48). Cette avancée a permis de favoriser la

réduction des écarts observés entre les seuils de réussite des différentes institutions d'enseignement.

Aujourd'hui, toutes les méthodes d'ESP ont invariablement recours au jugement d'experts. Les pouvoirs discrétionnaires qui leur sont attribués nécessitent cependant une procédure permettant d'atténuer au maximum la subjectivité de leur tâche. C'est pourquoi l'aspect procédural des méthodes d'ESP justes et équitables est assujéti à des formalités particulières et à des règles prescrites par des savants largement reconnus et respectés du milieu. Cette procédure, servant de guide pour la prise de décision, s'appuie sur des recommandations judicieuses afin d'établir une norme de rendement minimale d'usage ou un standard de performance raisonnable et accepté par les parties prenantes du domaine. Le panel d'experts constitue, dans ce sens, une disposition qui permet d'encadrer le pouvoir discrétionnaire dont dispose chaque panéliste pour jauger de la capacité d'une personne à atteindre un niveau de rendement ou de performance minimale dans un domaine donné par rapport à une personne standard et statuer à ce propos (Chênevert, 2015). C'est grâce à la procédure ayant conduit les experts à conceptualiser une personne minimalement compétente que les normes de rendement ou de performance guident la prise de décision sur le niveau compétence d'un candidat. Norcini (2003) souligne à cet effet que dans le cadre d'une épreuve professionnelle, l'ESP repose généralement sur une méthode qui permet d'asseoir le jugement de la valeur professionnelle sur une procédure systématique aspirant à parvenir à une décision par consensus. C'est ce qui représentera, au final, le résultat reconnu valable et déclaré recevable pour une épreuve (*ibid.*, 2003).

Ce sont des principes utiles qui servent de guide aux différentes propositions de classification et qui permettent d'orienter efficacement la prise de décision. Cet outil de sélection décrit et explique le processus à suivre pour obtenir des standards de performance acceptables et pour comprendre les raisons qui justifient le choix d'une

méthode plutôt qu'une autre. Il sert également à guider l'élaboration d'une méthode alternative innovante. Pour guider notre choix et orienter notre investigation dans la bonne direction, nous considérons qu'il est utile de présenter les principes fondamentaux communs à l'ensemble des classifications qui ont marqué l'histoire de l'ESP.

2.4.2.2 Neuf principes permettant l'établissement des standards de performance

À l'ère du XXI^e siècle, l'ESP repose, pour être socialement acceptable selon Cizek (2012), sur neuf principes définis suivant le recensement des principaux points de l'ensemble des méthodes. L'auteur souligne que les chercheurs du domaine de la mesure et de l'évaluation recommandent généralement d'appuyer toute démarche d'ESP sur neuf étapes pour en assurer la qualité. Il nous apparaît conséquemment essentiel, afin de donner une assise théorique robuste et de soutenir la modalisation de notre démarche, de présenter lesdites propositions.

Ainsi, selon une traduction libre d'une adaptation moderne de Cizek (2012, p. 52-64), le processus d'ESP propose de :

1. Choisir une méthode d'ESP et planifier la première rencontre avec les experts du panel;
2. Sélectionner un panel large et représentatif et déterminer sa composition;
3. Définir des descripteurs pour les catégories de performance (PLDs);
4. Former les panélistes à l'utilisation de la méthode d'ESP choisie;
5. Colliger les notations;
6. Fournir une rétroaction aux panélistes sur leurs notations et faciliter les discussions;
7. Colliger à nouveau les notations et établir les standards de performance;
8. Sonder l'opinion des panélistes sur le processus d'ESP et les standards de performance recommandés;

9. Rassembler la documentation technique et les preuves de validité de la procédure d'ESP et des résultats obtenus.

Le premier principe consiste à choisir une méthode d'ESP et à planifier la première rencontre du panel, qui se déroule, selon les recommandations de Cizek et Bunch (2007, traduction libre, p. 42-43; Cizek, 2012, p. 53-54), selon six critères de sélection présentés dans la figure 2.4.

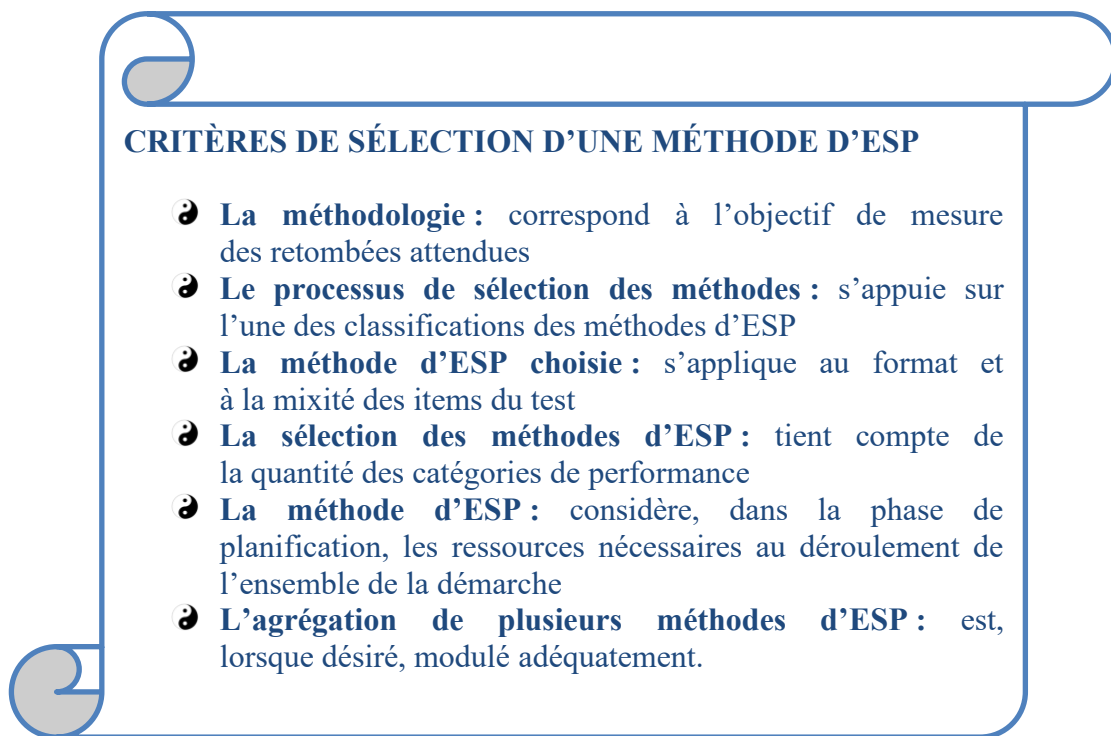


Figure 2.4 – Critères de sélection d'une méthode d'établissement de standards de performance

Le premier critère prescrit que la méthodologie adoptée doit correspondre à l'objectif de mesure des retombées attendues. Par exemple, aux fins de différenciation, une méthode d'évaluation du classement dite certificative serait appropriée pour identifier systématiquement un certain nombre ou une proportion de candidats selon leur

positionnement par rapport à d'autres candidats. Aux fins de prédiction, une méthode critériée serait appropriée pour anticiper la performance d'un candidat par rapport à certains critères de performance (*ibid.*, p. 42). Pour l'évaluation, une méthode d'évaluation de la capacité professionnelle serait finalement appropriée pour mesurer l'évolution de la performance dans le temps.

Le deuxième critère prescrit que le processus de sélection des méthodes doit s'appuyer sur l'une des classifications des méthodes d'ESP. Cela permet une classification selon la nature et l'ampleur des tâches assignées aux panélistes (Cizek, 2012, traduction libre, p. 53) ou, à l'égard des divers niveaux de difficulté des items, une classification selon le niveau de complexité des connaissances, des compétences et des habiletés évaluées par le test (Cizek et Bunch, 2007, traduction libre, p. 42). Ce deuxième critère de sélection semble de plus devoir prendre en compte le type de notation des scores (dichotomique ou polytomique), l'ampleur du test (quantité d'items) et la quantité de niveaux de performance. Nous tenons à préciser qu'une notation dichotomique se limite à deux choix de réponses, alors que dans le cas d'une notation polytomique, le nombre de choix de réponse est supérieur à deux.

Le troisième critère prescrit que la méthode choisie doit s'appliquer au format et à la mixité des items du test. L'outil de mesure se présente habituellement sous la forme de réponses fermées à choisir (items à réponse choisie) ou sous la forme de réponses ouvertes à construire (items à réponse construite). Pour la première catégorie de tests constitués d'items à choix multiples (polytomiques; fermés), c'est la méthode Angoff qui s'applique habituellement (Cizek, 2012, p. 54), les autres formats d'items fermés étant généralement des items d'appariement ou à choix simple (choix dichotomique) entre deux options de réponse (par ex., oui/non). Les formats d'items à réponses ouvertes sont pour leur part inclus dans les tests oraux, les tests à réponses élaborées et les tests à réponses courtes ou lacunaires. Ce format d'items est aussi souvent associé à une échelle descriptive à cotation critériée par niveaux de performance ou par portrait

de comportement.

Le quatrième critère prescrit que la sélection des méthodes doit tenir compte de la quantité des catégories de performance exigée. Chaque catégorie de performance est délimitée par un point de coupure. La quantité de catégories de performance, qui peut varier de deux à plusieurs, détermine conséquemment le nombre de points de coupure (uniques ou multiples). Par exemple, toutes les méthodes s'appliquent pour établir un seuil qui permet de délimiter les catégories réussite et échec.

Le cinquième critère prescrit que la méthode d'ESP doit, dans la phase de planification, tenir compte des ressources nécessaires au déroulement de l'ensemble de la démarche. Ce critère de sélection doit d'abord prendre en considération le temps de production, de préparation du matériel de formation, des invitations et celui consacré à la mise en œuvre de la procédure. Les ressources et les contraintes des parties prenantes doivent également être prises en compte. La programmation doit présenter, de façon détaillée, les principales instructions et opérations auxquelles devront prendre part les membres du panel. Cette étape peut, à titre d'exemple, inclure les différentes phases du déroulement de la méthode (une ou plusieurs rondes), l'échéancier, les coûts estimés et le temps qui, selon plusieurs auteurs, peut s'étaler sur une période allant jusqu'à cinq jours (Hambleton *et al.*, 2000b).

Finalement, le sixième critère prescrit que l'application de plusieurs méthodes d'ESP doit, lorsque désiré, être modulée adéquatement. Le recours simultané à plus d'une méthode doit reposer sur une décision politique qui s'appuie sur des règles préalablement déterminées par les parties prenantes. Puisqu'aucune recherche n'a été produite pour guider une telle procédure, l'objectif consiste à préciser de quelle manière seront utilisés les résultats issus des différentes méthodes pour établir le standard de performance du test (par ex., corroborer ou fusionner les résultats ou en sélectionner un seul, etc.) (Cizek et Bunch, 2007, p. 44).

Ces six critères de sélection d'une méthode d'ESP offrent un guide intéressant pour s'assurer d'une bonne adéquation aux besoins de l'organisme désirant les établir et pour planifier la première rencontre du panel constitué. Le principe suivant sert de guide pour procéder à la sélection dudit panel.

Le deuxième principe implique la constitution d'un panel d'experts qui soit large et représentatif des parties prenantes. L'échantillon, constitué à partir de la population ciblée se doit d'être suffisamment large (taille de l'échantillon), diversifié et représentatif de cette population. Les panélistes peuvent, par exemple, être sélectionnés selon certains critères géographiques ou selon d'autres caractéristiques pertinentes et, si désiré, être répartis en sous-groupe. Selon Cizek (2012, p. 55), puisqu'aucune recommandation spécifique n'est précisée dans la littérature, il est important que la composition d'un panel et le nombre de panélistes soient préalablement déterminés en tenant compte de la réalité des parties prenantes et en fonction des besoins de la démarche.

Le troisième principe, quant à lui, consiste à fournir la définition des descripteurs pour chacune des catégories de performance (PLDs). Les descripteurs des niveaux de performance (échelles descriptives) se doivent d'être formulés méthodologiquement afin de permettre une différenciation et d'assurer une démarche rigoureuse en matière de mesure et d'évaluation. Le nombre de niveaux doit être pertinent, cohérent et utile. Chaque niveau doit préciser la qualité des réponses et du travail attendus de la part du candidat. La dialectique de la formulation de ces critères de performance se doit d'être claire et précise.

Le quatrième principe implique quant à lui la formation des panélistes à l'utilisation de la méthode d'ESP. La durée de la formation, propre à chaque méthode, doit être minutieusement programmée.

La figure 2.5 qui suit présente, selon Cizek (2012, traduction libre, p. 57), les éléments justifiant l'utilité de la formation des sujets.

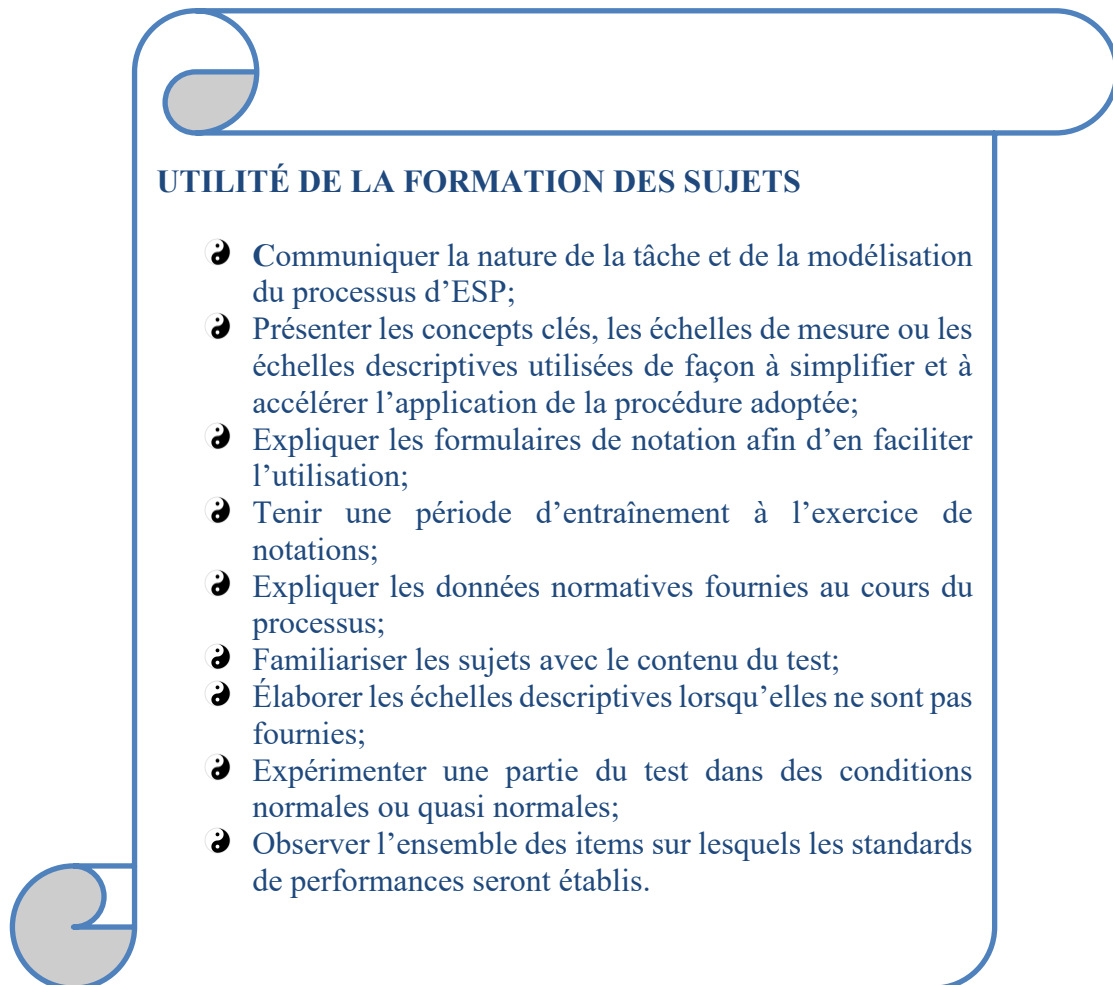


Figure 2.5 – Utilité de la formation des sujets à la méthode d'établissement de standards de performance retenue

Le cinquième principe consiste à colliger les notations. Cette étape du processus d'ESP inclut la collecte des notations et le traitement statistique des données. Il est d'abord important, afin d'éviter des goulots d'étranglement, de choisir une méthode de saisie et de collecte des notations qui soit efficace et efficiente. L'utilisation de l'informatique

par rapport à une procédure manuscrite, favorise des gains optimaux : de temps, de précision et de commodité. Le traitement des données doit ensuite être réalisé à l'aide de statistiques descriptives variées (par ex., l'étendue) (*ibid.*, p. 58).

Le sixième principe concerne la rétroaction envers les panélistes sur leurs notations et la facilitation des discussions. La rétroaction vise à offrir aux participants l'occasion de revisiter, s'ils le jugent à propos, leurs notations préliminaires. Des informations normatives sont alors mises à leur disposition avant d'entamer les rondes de notation subséquentes, et ce, afin de soutenir la prise de décision pouvant mener à un niveau d'accord convenable. Cela permet aux panélistes de réviser, s'il y a lieu, leurs notations à la hausse ou à la baisse et, advenant le cas, d'identifier puis d'éliminer les erreurs, ou encore, de dissiper de fausses interprétations et des mécompréhensions concernant les instructions ou le processus (*ibid.*, p. 58). Cette démarche itérative (rondes subséquentes) est considérée comme valable puisqu'elle permet généralement d'accroître le niveau d'accord entre les panélistes, de même que leur niveau confiance dans les standards établis (*ibid.*, p. 59). Les deux principales formes de rétroaction communément introduites entre les rondes de discussions sont : a) des statistiques descriptives sur les notations des panélistes ou b) la performance moyenne des candidats et la distribution des scores pour les items ou les tâches du test. D'autres formes de rétroaction peuvent également être introduites à cette étape du processus. La fiabilité des notations abordées à l'étape 9 et les arguments des pairs sont des exemples qui peuvent influencer les jugements des panélistes. La cohérence interne permettrait également, afin d'accroître le niveau d'accord, d'identifier les incohérences apparentes parmi les notations. Par exemple, il pourrait être demandé à un panéliste qui a systématiquement attribué des notations élevées pour les items difficiles par rapport aux items faciles d'expliquer sa position (*ibid.*, p. 59).

Le septième principe concerne la nouvelle cueillette des notations et l'ESP. Celui-ci repose couramment sur un algorithme de calcul fondé sur la moyenne et sur la variance des standards d'une performance individuelle des panélistes. C'est l'importance de la variance qui détermine le niveau de confiance (précision) face au standard établi. L'interprétation de la variance donne un indice du niveau de confiance qui se dégage du panel par rapport aux standards de performance établis. Une petite variance traduit un niveau de confiance fort et, à l'inverse, un niveau de confiance faible est inféré par une plus grande variance. Il peut cependant arriver que la médiane soit préférée à la moyenne lorsque l'échantillon est petit ou que la distribution est asymétrique (*ibid.*, p. 59). L'algorithme est généralement réalisé après chacune des rondes de notation. La première ronde est habituellement réalisée sans données normatives et sans intervention des pairs, qui sont susceptibles d'influencer le jugement des panélistes. La deuxième ronde, au contraire, cherche souvent à éclairer le jugement des panélistes et possiblement, à accroître le niveau d'accord entre eux en fournissant des informations normatives et les arguments des pairs. Ces informations et ces discussions sont effectuées, selon les différentes méthodes, avant d'entreprendre la deuxième ronde de notation ou lors du déroulement de celle-ci. Le processus est généralement constitué de deux rondes. Toutefois, certains contextes peuvent nécessiter une ou deux rondes supplémentaires.

Le huitième principe consiste pour sa part à sonder l'opinion des panélistes sur le processus d'ESP et sur les standards de performance recommandés. L'étape du sondage d'opinion sert à vérifier le niveau de confiance accordé par les panélistes au processus choisi pour l'ESP, de sa mise en œuvre et des résultats obtenus (Cizek 2012). Ce sondage doit être élaboré dans le but de vérifier les qualités inhérentes à la méthode d'ESP. Le processus utilisé était-il clair et précis, facile à appliquer par les panélistes? Sa mise en œuvre était-elle adaptée, systématique et rigoureuse? Les résultats obtenus apparaissent-ils comme fiables? Ce sont des exemples de questions pouvant être

incluses dans le sondage afin de contribuer à la validité des standards de performance établis.

Le neuvième principe implique l'assemblage de la documentation technique et des preuves de validité de la procédure d'ESP et des résultats obtenus. Le dernier des neuf principes d'ESP permet d'en étayer les preuves de validité. La preuve de validité prédominante réside, en premier lieu et comme le souligne Kane (2001, cité dans Cizek, 2012), dans la démonstration de la cohérence procédurale. Cette démonstration exige, en conformité avec les huit principes énoncés précédemment, que les documents liés au choix de la procédure, à la sélection des panélistes et aux différentes étapes de réalisation de la démarche, soient présentés de façon détaillée. En second lieu, la preuve de validité de la démarche d'ESP et des résultats obtenus s'appuie sur la démonstration de la légitimité interne et externe de la procédure. La validité interne est principalement liée à la cohérence de la méthode ainsi qu'à la consistance intra et interjuges. La notion de cohérence réfère d'abord au degré de possibilité selon lequel la même méthode, utilisée à nouveau, pourra fournir une estimation équivalente de la moyenne, de la variance et de l'erreur standard de la moyenne. Cette cohérence réfère ensuite, pour une reproduction de la méthode, à l'estimation des ressources en termes de temps et de coût. La consistance intra et interjuges renforce finalement la validité interne par la corrélation statistique entre les notations chez un même juge et entre ses pairs. Cela permet de mesurer le niveau de cohérence entre les jugements et, conséquemment, d'augmenter la fiabilité des conclusions internes. Il apparaît important, lorsque le contexte le permet, de tenir compte de la sensibilité aux changements de notation susceptible d'affecter significativement les résultats.

La validité externe réfère, quant à elle, au potentiel de généralisation des résultats. Elle s'appuie sur la comparaison des résultats obtenus avec des données externes provenant de différentes sources. L'un des principaux moyens de validation externe consiste à démontrer que les résultats obtenus par différentes méthodes, dans les mêmes

conditions et par un panel de même taille, sont sensiblement équivalents. Scallan (1981, p. 9) soutient que « le seuil le plus exigeant » parmi les méthodes Nedelsky, Angoff et Ebel est celui de la méthode d'Ebel. Plusieurs auteurs soulignent que cette méthode génère systématiquement le seuil de performance le plus élevé de ces trois méthodes (Livingston et Zieky, 1983; Poggio *et al.*, 1981). Un autre moyen de validation externe consiste, pour un panel d'experts, à comparer les résultats de leurs jugements à des résultats empiriques déjà colligés ou à des données qualitatives disponibles. Un dernier exemple de moyens qui permettent de renforcer la validité de la méthode est la récurrence de résultats similaires obtenus par l'application successive de la même méthode dans des conditions comparables. La méthode devient, avec le temps, de plus en plus valide.

En conclusion, les neuf principes ont pour rôle de guider les actions à entreprendre et contribuent à démontrer et à rehausser la validité des résultats établis. Nous tenons à préciser ici que ces étapes sont jugées essentielles pour établir des standards de performance recevable. Cizek (2012, traduction libre, p. 53) souligne que ces principes s'appliquent généralement aux méthodes centrées sur les items d'un test (par ex., Angoff, Ebel, Nedelsky) et, de façon fragmentaire ou variable, aux méthodes centrées sur les personnes (par ex., les groupes contrastés). Cet auteur (*ibid.*, p. 53) soutient de plus qu'il est rare que les étapes de la procédure dérogent de ces neuf principes. Dans ces conditions, la première étape consiste, afin de se conformer aux neuf principes, à sélectionner la méthode la plus appropriée. Ce processus de sélection nous amène conséquemment à faire la recension des diverses classifications proposées pour catégoriser les méthodes existantes.

2.4.2.1 Classification des méthodes d'établissement de standards de performance

Considérant l'importance des retombées sociales et du devoir de protection du public des ordres professionnels, il nous apparaît essentiel, dans le cadre de notre démarche

de recherche, d'explorer plus en profondeur les méthodes les plus reconnues de la catégorie qui s'avèrera la plus appropriée au contexte actuel. Au fil des ans, selon Cizek (2012), la quantité considérable de méthodes et la multiplicité des propositions de méthodes alternatives ayant recours à diverses combinaisons d'approches méthodologiques appartenant à des classifications distinctes, a conduit plusieurs auteurs, afin de s'y retrouver plus efficacement, à proposer différentes classifications pour tenter de regrouper l'ensemble des méthodes d'ESP.

À la fin des années 1980, la classification binaire classique des méthodes d'ESP se subdivisait en deux catégories : a) les méthodes centrées sur le test ou l'épreuve et b) les méthodes centrées sur le candidat ou l'individu (Jaeger, 1989; Kane 1998; Blais, 2008). La première catégorie propose un regroupement des méthodes dont l'exercice du jugement repose sur les items d'un test ou d'une épreuve, alors que pour la deuxième, cet exercice repose sur les réalisations concrètes des individus (scores au test ou à l'épreuve). Au fil des années, ces classifications ont subi des transformations. La recrudescence des méthodes alternatives et des méthodes mixtes ainsi que la difficulté à classer ces dernières dans les catégories classiques a engendré de nouvelles propositions de classification. Hambleton (1980) a, dans ce sens, proposé une classification reconnue par plusieurs auteurs (Cizek, 2012; Cizek et Bunch, 2007; Haertel et Lorie, 2004; Hambleton et Plake, 2000; Kaftandjieva, 2010; Raîche, 2005). Cette classification propose trois catégories dites appréciative (*judgmental*), empirique (*empirical models*) ou mixte (*combination models*).

Au cours des années 2000, plusieurs auteurs ont proposé une classification des méthodes contemporaines en quatre catégories : a) centrées sur les items ou les tâches d'une épreuve et des échelles de notation (scores ou échelle de mesure), b) centrées sur les personnes (leurs résultats), c) centrées sur les tâches accomplies par des candidats et d) centrées sur les profils de notes (scores ou patrons de réponse) (Hambleton *et al.*,

2000a; Blais, 2008). Les échelles de notation auxquelles réfère la première catégorie permettent de situer la fréquence ou le degré d'occurrence de certains aspects liée à une épreuve. Au cours des années 2010 à 2012, les agences nationales d'accréditation et les États ont privilégié la classification classique (a et c) (Cizek, traduction libre, 2012).

Plus récemment, en 2015, l'Academy of Medical Royal Colleges a proposé une classification renouvelée des méthodes en quatre catégories : a) centrées sur le test ou l'épreuve, b) centrées sur la personne, c) de compromis (une combinaison d'éléments des méthodes de référence et des normes établies) et d) statistiques (fondées sur des résultats réels de candidats). Chacune de ces catégories ont des particularités qui les distinguent quant au rôle que doit exercer chacun des panélistes dans le processus d'ESP. Comme on peut le constater, la classification classique se perpétue et une classification mitoyenne a été ajoutée en réponse à la mixité des méthodes. Les procédés déductifs ont également été retranchés pour former une catégorie distincte, ce qui offre une répartition améliorée permettant de discriminer plus efficacement les méthodes ciblées.

Le choix d'adopter une classification doit reposer sur l'expérimentation éprouvée des méthodes d'ESP répondant aux besoins des parties prenantes. Nous tenons donc à préciser que, selon le contexte lié au développement du référentiel des compétences élaboré par l'OAQ et à l'épreuve clinique appliquée par ce dernier, ce ne sont que les méthodes centrées sur le test ou l'épreuve qui ont été adaptées. Les jugements s'exercent conséquemment sur les items observables d'une épreuve clinique originale qui s'appuient sur la performance attendue d'un candidat hypothétique. Les méthodes centrées sur les personnes, de même que les deux autres catégories présentées plus haut, n'ont donc pas été utilisées dans notre démarche spécifique.

Notre choix s'explique d'abord par le fait que l'épreuve clinique est un nouvel instrument de mesure construit à partir d'un référentiel des compétences contextualisé. Il n'y a conséquemment aucune donnée empirique de disponible pour appuyer les jugements des panélistes. Ensuite, la nature du test s'est avérée déterminante du fait qu'il nous a été impossible d'avoir accès à des candidats qui nous auraient permis d'appliquer une méthode centrée sur les personnes. Le choix de la catégorie regroupant les méthodes centrées sur le test s'avérait donc la seule option adaptée à l'épreuve clinique effectuée en situation réelle ou simulée et réalisée par l'OAQ. L'étape suivante permettra de présenter une analyse des méthodes centrées sur le test étant les plus reconnues dans le but de faire ressortir les différentes caractéristiques applicables au contexte de notre démarche de recherche.

2.4.2.2 Analyse des méthodes d'établissement de standards de performance centrées sur le test largement reconnues

Dans le cadre de cette recherche, les jugements étant essentiellement réalisés sur les items de l'épreuve élaborée à partir du contenu du référentiel des compétences de l'OAQ, il s'est révélé important de sélectionner les méthodes les plus appropriées au type d'épreuve ainsi qu'à son contexte.

La représentativité des connaissances, des habiletés et des compétences minimales essentielles pour exercer une profession dans un domaine réglementé exige que la méthode choisie puisse produire un résultat valable. Nous présenterons conséquemment, les caractéristiques des méthodes centrées sur le test les plus communément répandues, soit celles de Nedelsky, d'Angoff (traditionnelle) et d'Ebel, ainsi que celle proposée par Impara et Plake (oui/non) et celle de l'appariement au descripteur de l'item (*Item-Descriptor Matching Method* : IDM), récemment proposée par Ferrara, Perie et Johnson (2008). Une description élaborée de ces cinq méthodes illustrées par des exemples est disponible à l'appendice C. Afin de situer leurs origines

dans un continuum temporel évolutif, nous présentons, dans la figure 2.6, une vue d'ensemble de la chronologie historique de ces méthodes.

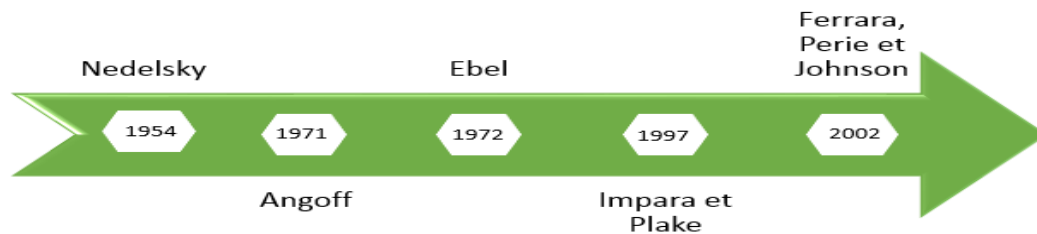


Figure 2.6 – Historique des méthodes d'établissement de standards de performance

D'entrée de jeu, il convient de préciser que toutes les méthodes existantes centrées sur des tests partagent plusieurs caractéristiques. Les nombreuses études statistiques comparatives ont largement permis, pour les quatre méthodes les plus anciennes, soit celles de Nedelsky, d'Angoff (traditionnelle), d'Ebel et d'Impara et Plake (oui/non), de confirmer leur recevabilité et applicabilité. La méthode proposée par Ferrara, Perie et Johnson, quoiqu'ayant à ce jour fait l'objet d'un moins grand nombre d'études, partage néanmoins certains points avec les autres méthodes et apporte quelques éléments de nature qualitative qui permettent de préciser davantage les outils de mesure et l'objet des jugements.

Parmi les caractéristiques communes, soulignons premièrement que toutes les méthodes s'appliquent principalement à des tests académiques destinés à des programmes de certification de grands groupes de personnes. Elles sont également utilisées dans le cadre d'accréditations professionnelles servant à octroyer une licence ou un permis d'exercice ainsi que dans la mesure du rendement en entreprise. Toutes ces méthodes ont de plus recourt aux jugements de panélistes et toutes sont en accord avec les principes, énoncés plus haut, qui recommandent que soit intégrés certains

critères de sélection du panel d'experts. Essentiellement, les panélistes doivent être issus du même champ d'expertise et être représentatifs de l'ensemble de la population ciblée.

L'application de ces méthodes requiert que le panel d'experts estime ou prédise la façon dont un candidat limite devrait minimalement performer à un test (Cizek, 2012). Pour ce faire, toutes les méthodes exigent de conceptualiser ce candidat minimalement compétent. Dans une certaine mesure, cette conceptualisation tend à prétendre à une certaine objectivité tout en demeurant intuitive. Cette capacité d'intuition et l'aptitude des panélistes, ici nommés experts, à saisir le véritable sens du concept permettent de tendre à une apparence d'objectivité, en ce sens qu'elles tiennent compte de la réalité perçue par des sujets pensants, expérimentés et représentatifs d'un domaine auxquels ils appartiennent. En d'autres mots, il s'agit de la capacité à se représenter une personne-standard tout juste qualifiée pour un test ou pour une épreuve particulière ou qui se situe tout juste au-delà du statut de quasi compétent, ou encore, comme le souligne Plake (1998, traduction libre, p. 78), qui satisfait aux conditions requises pour justifier l'obtention d'un permis ou la certification. Pour ces raisons, ces méthodes suggèrent, lors d'une période de familiarisation, de fournir une définition du sens donné à la notion de candidat minimalement compétent. Comme souligné au début du cadre théorique, Nedelsky a été le premier, dans les années 1950, à exiger, afin de contrecarrer la subjectivité individuelle et collective de cet exercice, que les membres du panel s'approprient ce concept avant d'émettre des jugements sur les items d'un test. Cette première définition a été présentée précédemment à la page 54. Ces jugements réfèrent fondamentalement à l'estimation de la probabilité qu'un candidat minimalement compétent puisse réussir chacun des items du test ou à l'estimation de la proportion d'items qu'il serait apte à réussir dans chacun des niveaux de la performance établie.

Pour les quatre premières méthodes (Nedelsky, Angoff, Ebel, Impara et Plake), les seuils des niveaux de performance sont établis sur la base d'un algorithme statistique utilisé pour calculer les tendances centrales individuelles et collectives (médiane ou moyenne). Ces statistiques produisent une estimation du nombre de réponses correctes attendues d'un candidat limite. C'est le nombre de réponses correctes qui représente le seuil de performance de l'épreuve. Quoique les méthodes soient comparables, les seuils de performance qui en résultent peuvent, dans une certaine mesure, différer en raison de la subjectivité des jugements s'inscrivant dans le modèle probabiliste retenu. Comme il est souligné au neuvième principe (à la page 633), malgré le fait que les trois premières méthodes présentent habituellement des seuils similaires, la méthode d'Ebel tend généralement, selon certaines études comparatives, à présenter des seuils légèrement plus élevés. La détermination des seuils de réussite pour chacune de ces quatre méthodes ne requiert aucune donnée normative. Toutefois, il est généralement recommandé, afin d'éclairer l'exercice du jugement, de fournir des données entre les rondes de notation. Seule la méthode de l'IDM donne, comme données normatives, les niveaux de difficulté chiffrés des items extraient de rendements de candidats réels.

Ces quatre premières méthodes proposent aussi des caractéristiques distinctives. La méthode Nedelsky, d'Angoff et du oui/non reposent sur une échelle de mesure simple à une dimension dichotomique. La notation des scores par le panel d'experts consiste conséquemment à classer chacun des items dans l'une ou l'autre des deux catégories prédéfinies (succès ou échec, 1/0, oui/non, compétent/incompétent). La méthode d'Ebel utilise quant à elle une échelle simple à deux dimensions croisées constituée de 12 points, soit 3 niveaux de difficulté (facile, moyen et difficile) et 4 degrés de pertinence (discutable, acceptable, important, essentiel). La notation de scores est, pour cette méthode, polytomique et découle essentiellement des jugements de classification émis par le panel d'experts. Le niveau de difficulté de chacun des items permettant d'appuyer le jugement est pris en compte uniquement dans les

méthodes d'Ebel et de l'IDM. L'ordonnement des items par niveau de difficulté croissant est pris en compte seulement par la méthode de l'IDM, alors que la méthode d'Ebel n'en prescrit aucun.

Les tests constitués d'items à choix de réponse et à réponses construites courtes sont généralement privilégiés pour établir des standards de performance. Les méthodes d'Angoff, d'Ebel et de l'IDM sont généralement appliquées à des items à choix de réponse ou à des items à réponses construites courtes. Dans le cas des items à réponses construites, les jugements des panélistes consistent alors à estimer le score moyen qu'un candidat limite serait apte à obtenir plutôt qu'à estimer, pour chaque item à choix multiples, la probabilité d'obtenir une réponse correcte (Cizek, 2012, traduction libre, p. 23). La méthode de Nedelsky est, quant à elle, la seule à s'appliquer exclusivement à des tests constitués d'items à choix multiples.

Au début des années 2000, Ferrara, Perie et Johnson (2008) ont proposé une méthode qui se distingue des autres. Ces auteurs ont notamment introduit l'usage d'un cahier d'ordonnement des items d'un test réel par niveau de difficulté croissant, soit du plus facile au plus difficile. Ils ont également introduit l'usage de descripteurs. La méthode de l'IDM est la seule à utiliser une échelle qualitative critériée qui présente des descripteurs détaillés des niveaux de performance ainsi que des descripteurs pour chacun des items afin de fournir des repères clairs aux panélistes. Ces descripteurs serviront à l'exercice des jugements d'appariement des items aux descripteurs. Selon Cizek et Bunch (2007), les descripteurs de niveau de performance, spécifiques à chaque niveau, représentent les fondements d'une multitude de méthodes de détermination de points de coupure. Selon eux, ces descripteurs détaillés représentent l'un des éléments déterminant pour appuyer les jugements des panélistes, et ce, sans distinction des préalables exigés par diverses méthodes éprouvées. Ces auteurs font référence à une formulation contextualisée et adaptée aux habiletés attendues pour chacun des niveaux de performance, dont être capable de faire quelque chose dans une situation et dans un

contexte donnés. La tâche des panélistes consiste alors à apparier les items aux descripteurs de performance. Les jugements probabilistes sont ici remplacés par des jugements de classification. L'appariement effectué par les panélistes vise à permettre de cibler les « zones d'incertitude », qui sont déterminées par une étendue d'items servant à établir les seuils de performance.

Il est, à cette étape, primordial de souligner que toutes les méthodes centrées sur le test reposent d'abord sur l'expérience personnelle acquise et sur l'aisance des panélistes à accomplir la tâche selon l'approche privilégiée et ensuite, sur leur intime conviction du bien-fondé de la démarche appliquée. Il est, par exemple, important de s'attarder au niveau de satisfaction des panélistes au déroulement de la démarche ainsi qu'à leur niveau de confiance dans les résultats obtenus.

2.4.2.3 Description sommaire des méthodes d'établissement de standards de performance centrées sur le test ayant servi de levier à la méthode alternative proposée

Pour proposer une méthode alternative fondée sur des méthodes reconnues dans le domaine de la mesure et de l'évaluation, les méthodes postérieures à Nedelsky et à Angoff se sont avérées utiles. Nous présentons conséquemment, et à des fins pratiques, une description sommaire de la méthode d'Ebel, qui s'avère essentielle et fondamentale, et les deux méthodes qui l'ont succédée : la méthode oui/non d'Impara et Plake et la méthode de l'appariement au descripteur de l'item de Ferrara, Perie et Johnson. Nous tenons à rappeler que l'ensemble des méthodes mobilise un ou plusieurs panels d'experts pour porter, de façon indépendante, des jugements probabilistes sur les spécificités de chacun des items d'une épreuve.

La méthode Ebel (1972) mobilise les panels d'experts en vue d'accomplir quatre opérations consistant à :

- Adopter une définition conceptuelle de la notion de minimalement compétent;
- Estimer la difficulté et la pertinence des items afin de les classer dans une matrice selon les deux catégories croisées;
- Estimer, pour chacune des catégories, la proportion d'items qu'un candidat minimalement compétent serait apte à réaliser correctement;
- Définir, pour chaque expert, le seuil de réussite en utilisant les pondérations et le nombre d'items classés dans chacune des catégories de performance.

Comme cette méthode est itérative et réalisée par plusieurs experts, la moyenne des seuils individuels permet d'établir le seuil définitif (standard de performance) de l'épreuve. Cette méthode offre l'avantage de tenir compte des paramètres estimés des items (la difficulté et la pertinence des items).

La méthode oui/non, d'Impara et Plake (1997), requiert d'accomplir trois opérations consistant à :

- Adopter une définition conceptuelle de la notion de minimalement compétent;
- Estimer si un candidat minimalement compétent est oui ou non apte à réussir chacun des items d'une épreuve;
- Définir, pour chaque expert, le seuil de réussite en utilisant le nombre de oui et le nombre total d'items de l'épreuve.

Au final, le seuil de réussite définitif équivaut à la moyenne des seuils établis individuellement. Cette méthode est utile lorsque le niveau de difficulté des items est obtenu empiriquement. Dans le cadre de notre recherche, comme plusieurs standards de performance doivent obligatoirement être atteints, le recours aux variables

dichotomiques oui/non permet de tenir compte du double standard prescrit par l'OAQ. Ultérieurement, lors de l'évaluation des performances, cette méthode sera utile pour identifier un manque d'aptitude à accomplir certaines tâches spécifiques aux standards de qualité établis.

La méthode de l'appariement au descripteur de l'item, de Ferrara, Perie et Johnson (2008), requiert d'accomplir cinq opérations consistant à :

- Apparier chacun des items préalablement ordonnancés empiriquement à un descripteur de performance préétabli;
- Classer les items par ordre de difficulté et par niveau de performance afin d'obtenir une nouvelle liste ordonnée des items;
- Identifier, dans la nouvelle liste ordonnée, les zones d'incertitude à l'aide d'une règle départ-arrêt afin de circonscrire les séquences discontinues (irrégulières), soit là où les jugements des niveaux de performance sont irréguliers et situés entre deux séries de trois jugements consécutifs univoques (réguliers);
- Se repositionner sur l'appariement au descripteur de performance de chacun des items situés dans les zones d'incertitude définies afin de les dissiper graduellement et d'obtenir des points de coupure tranchés entre chacun des niveaux de performance;
- Identifier les points de coupure définitifs délimitant chacun des niveaux de performance en calculant la moyenne des points de coupure obtenus individuellement.

Le standard de performance équivaut conséquemment à la réussite des items se trouvant sous le point de coupure situé entre le descripteur de la performance minimale et le niveau supérieur. L'usage d'une liste ordonnée des items établie sur la base de niveaux de difficulté formels et de descripteurs de performance hiérarchisés s'avère

utile pour classifier les items et appuyer les jugements sur des repères communs précis. L'usage d'une règle identifiant les zones d'incertitude s'avère également utile pour circonscrire avec précision les items présentant des jugements divergents et ainsi, tendre à dégager par la suite des convergences.

2.5 Méthode alternative d'établissement de standards de performance proposée

Suivant la réalité du contexte, il est à noter qu'aucune des méthodes présentées précédemment et en annexe ne s'applique parfaitement à la résolution de notre problème de recherche. Nous allons conséquemment, dans la section qui suit, commencer par circonscrire l'ensemble des aspects liés au contexte de l'épreuve clinique que l'OAQ souhaite administrer. Nous poursuivrons ensuite par l'exposé d'une méthode alternative appliquée pour l'établissement des standards de la performance attendue d'un candidat minimalement compétent.

2.5.1 Aspects liés au contexte de l'épreuve clinique de l'OAQ

Soulignons d'abord que l'épreuve se doit d'être réalisée en situation clinique réelle ou simulée au moyen d'une entrevue semi-dirigée qui s'étale sur une période de quatre demi-journées. Durant cette période, le candidat aura à démontrer qu'il est capable de réaliser avec succès les 23 tâches d'une épreuve clinique qui regroupe au total 149 indicateurs de performance distincts. Concrètement, l'OAQ, après avoir défini son référentiel des compétences et constitué un cahier d'observation des comportements cliniques, avait besoin de réaliser cette étape avant de pouvoir établir un standard de performance (ESP; seuil de passage) attestant de la compétence minimale d'un candidat. Il faut conséquemment déterminer les 23 standards de performance nécessaires pour attester du niveau de performance minimale attendu pour chacune des 23 tâches et un standard de performance complémentaire pour attester de la réussite ou de l'échec de l'ensemble de l'épreuve. Il est finalement important de noter que l'obligation légale, réglementaire et professionnelle de l'OAQ de protéger le public

nécessite que les standards soient établis de façon consensuelle par les experts du domaine afin d'être jugés recevables.

La méthode alternative qui est proposée s'appuie sur la définition de la compétence retenue par l'OAQ et sur les impératifs liés aux 23 éléments de compétence devant être réussis par un candidat minimalement compétent. La capacité d'intégrer des ressources internes et externes dans la démarche procédurale définie par l'OAQ a mené à élaborer 23 tâches, directement liées aux 23 éléments de compétence, auxquelles ont été greffés 149 indicateurs de performance (en moyenne, près de 7 indicateurs par tâche). L'inférence des compétences est ici rattachée aux performances observables lors de la réalisation de ces 149 indicateurs. Dans ce contexte, le candidat a la possibilité de démontrer sa capacité de générer des informations sur un ensemble d'actions représentatives des connaissances, des habiletés et des compétences minimales essentielles pour exercer une profession dans un domaine réglementé. Les 149 indicateurs de l'épreuve clinique de l'OAQ deviennent conséquemment les items sur lesquels doit s'appuyer la démarche d'ESP. Pour établir des standards de performance, comme aucune donnée empirique auprès de personnes évaluées n'est disponible, il est d'abord important de produire une mise en rang de ces 149 items par niveau de difficulté. Afin d'effectuer cet ordonnancement, à l'instar de l'ensemble des méthodes d'ESP, le jugement d'experts du domaine regroupés dans différents panels est sollicité.

Dans ce contexte, les caractéristiques communes aux méthodes centrées sur le test peuvent, comme le souligne Kane (1998), être utilisées sans que l'épreuve ait préalablement eu lieu. Raïche (2005) souligne toutefois que peu de ces nombreuses méthodes ont fait l'objet d'avancées scientifiques adaptées à l'approche par compétences. L'évaluation de la capacité professionnelle au moyen de conditions réelles ou simulées, qui fait appel à la production de réponses complexes colligées lors d'entretiens semi-dirigés, semble très peu documentée. Malgré le fait que les méthodes

centrées sur le test traitent peu des items à réponses construites élaborées ou des items observables, il n'en demeure pas moins qu'elles offrent un certain cadre, des caractéristiques et des principes pouvant répondre aux besoins de l'épreuve de l'OAQ. Cependant, ces méthodes ne sont pas suffisamment appropriées au contexte d'ESP des ordres professionnels. Pour cette raison, une méthode alternative centrée sur les items de l'épreuve clinique est ici proposée.

2.5.2 Méthode alternative inédite permettant d'établir des standards de performance pour l'épreuve clinique de l'OAQ

La complexité de la réalité clinique oblige à recourir à l'application d'une épreuve à réponses construites de notation polytomiques. Plus précisément, il est impossible de situer la faisabilité des tâches de façon dichotomique (succès ou échec) et on est alors contraint de déterminer des indicateurs constitués de plusieurs niveaux de difficulté pour évaluer la performance du candidat.

La méthode d'Ebel (1972), qui est généralement appliquée à des tests de nature dichotomique, propose une avenue pouvant s'appliquer à des items de nature polytomique et à une classification des items par niveau de difficulté. La méthode alternative que nous proposons, qui s'inspire principalement de la procédure d'Ebel, sera appliquée à une épreuve de nature polytomique composée d'items à réponses construites élaborées. Afin de pallier l'absence de données empiriques sur le niveau de difficulté réel des items, le rang du niveau de difficulté de chacun des indicateurs sera estimé subjectivement en fonction de la notation qui leur sera attribuée.

À l'instar de l'ensemble des méthodes, la méthode alternative est centrée sur les items d'un test et elle est liée à la représentativité d'un domaine de connaissances spécifiques qui requièrent le jugement de professionnels du domaine en question. Conséquemment, le panel est ici constitué d'acupuncteurs en exercice aptes à se prononcer sur la

complexité des items et sur le degré de performance attendu d'un candidat. Comme mentionné précédemment (à la page 27), il est crucial, selon plusieurs auteurs (Cerqueira *et al.*, 2008; Noijons *et al.*, 2011), de définir le niveau de difficulté associé à chacun des items. Il apparaît à cet effet, comme le soulignent ces auteurs, que le contenu des items doit préalablement avoir été clairement défini en tenant compte des variations des niveaux de complexité inhérents aux différents contextes dans lesquels ils s'inscrivent. Ces auteurs soutiennent à cet effet que la complexité des items varie grandement en fonction des difficultés liées à la compréhension et aux aptitudes des professionnels formés à l'étranger à fournir des réponses de qualité. Eisner (2000, p. 3; Bloom *et al.* 1956) souligne dans ce sens que l'élaboration d'outils d'évaluation a été, « dans le monde entier », grandement facilitée par l'utilisation de la taxonomie du domaine cognitif énoncée par Bloom. Afin de favoriser une compréhension commune, des descripteurs de chacun des niveaux de difficulté ainsi qu'une explication du concept d'une personne minimalement compétente doivent conséquemment être fournis aux panélistes pour appuyer leur jugement. Afin de rendre les explications claires et le moins équivoques possibles, des descripteurs doivent être fournis aux panélistes. À cet effet, le tableau 2.7 présente à la page suivante l'explication des quatre niveaux de difficulté inspirés de la définition de la compétence retenue par l'OAQ et de la taxonomie de Bloom.

Tableau 2.7 – Convention des descriptions des niveaux de difficulté

Niveau de difficulté	Définition	Niveau taxonomique
Facile	exige simplement de connaître et de comprendre afin d'appliquer des règles, des principes ou des procédés prédéterminés.	○ connaissance ○ compréhension ○ application
Moyen	exige une certaine interprétation de la situation clinique et une discrimination des informations reçues ou des outils à utiliser afin de répondre à des situations cliniques mobilisant un nombre limité de ressources.	○ application - utilisation ○ analyse - interprétation - discrimination
Difficile	exige une capacité d'analyse des situations cliniques et de synthèse nécessaire à la proposition de solutions à des problèmes nécessitant l'intégration d'une multitude de ressources.	○ analyse ○ synthèse
Très difficile	exige un jugement clinique qui permet de discriminer une multitude d'informations pertinentes afin de proposer des solutions appelées à être justifiées.	○ synthèse ○ évaluation - proposition/ recommandation - justification

Le niveau facile exige simplement de connaître et de comprendre, afin d'appliquer des règles, des principes ou des procédés prédéterminés. Le niveau moyennement difficile exige une certaine interprétation de la situation clinique et une discrimination des informations reçues ou des outils à utiliser afin de répondre à des situations cliniques mobilisant un nombre limité de ressources. Le niveau difficile exige une capacité d'analyse des situations cliniques et de synthèse nécessaires à la proposition de solutions à des problèmes exigeant l'intégration d'une multitude de ressources. Le niveau très difficile requiert un jugement clinique qui permet de discriminer une multitude d'informations pertinentes afin de proposer des solutions appelées à être justifiées.

La recension des écrits présentée précédemment a mené à soupeser l'importance

d'inclure les comportements observés chez le candidat afin de les conjuguer aux niveaux de difficulté des items. Dans le contexte de cette recherche, c'est le degré d'efficience qui fait référence à la performance de l'acupuncteur, c'est-à-dire à sa capacité de rendement, ainsi que le rapport entre les résultats obtenus et les moyens mis en œuvre pour y parvenir. Le tableau 2.8 illustre ces degrés d'efficience.

Tableau 2.8 – Convention des descriptions des degrés d'efficience

Degré d'efficience	Rendement	Moyens mis en œuvre	Résultats obtenus
Très fort	remarquablement efficace • effectue le travail aisément dans le temps impartis.	parfaitement appropriés • mobilise systématiquement les ressources nécessaires pour réaliser le travail.	remarquables • produit systématiquement l'effet escompté de façon convenable.
Fort	pleinement efficace • effectue le travail relativement aisément dans le temps impartis.	généralement approprié • mobilise généralement les ressources convenables pour réaliser le travail.	escomptés • produit généralement l'effet escompté de façon convenable.
Moyen	passablement efficace • effectue convenablement le travail en excédant occasionnellement le temps imparti.	modérément appropriés • mobilise occasionnellement des ressources inutiles pour réaliser le travail.	moyens • produit habituellement l'effet escompté de façon convenable.
Faible	partiellement efficace • effectue laborieusement le travail dans un temps excédant fréquemment le temps imparti.	fréquemment inappropriés • mobilise fréquemment plusieurs ressources inutiles pour réaliser le travail.	médiocres • produit rarement l'effet escompté de façon convenable.

Le degré d'efficience très fort requiert que le candidat obtienne un rendement remarquablement efficace par la mise en œuvre de moyens parfaitement appropriés qui se traduisent par des résultats remarquables. Ce premier degré se mesure par l'indice d'un travail accompli aisément dans le temps imparti, où systématiquement, les ressources incontournables sont mobilisées et produisent l'effet escompté. Le degré d'efficience fort requiert un rendement pleinement efficace par la mise en œuvre des moyens généralement appropriés qui se soldent par les résultats escomptés. Ce

deuxième degré se mesure par l'indice d'un travail accompli relativement aisément dans le temps imparti, où généralement, les ressources convenables sont mobilisées de façon appropriée et produisent l'effet escompté. Le degré d'efficacité moyen requiert un rendement passablement efficace par la mise en œuvre des moyens modérément appropriés qui se soldent par des résultats moyens. Ce troisième degré se mesure par l'indice d'un travail accompli convenablement en excédant occasionnellement le temps imparti, où de temps à autre des ressources inutiles sont mobilisées et produisent, à quelques exceptions près, l'effet escompté de façon convenable. Le degré d'efficacité faible requiert un rendement partiellement efficace par la mise en œuvre de moyens fréquemment inappropriés qui se soldent par des résultats médiocres. Ce dernier degré d'efficacité se mesure par l'indice d'un travail accompli laborieusement en excédant fréquemment le temps imparti, où fréquemment, plusieurs ressources inutiles sont mobilisées et produisent rarement l'effet escompté de façon convenable.

Les échelles descriptives des niveaux de difficulté et des degrés d'efficacité visent à appuyer les jugements des panélistes sur le niveau de difficulté de chacun des items et sur le degré d'efficacité attendu d'un candidat minimalement compétent lorsqu'il répond à l'item. La tâche des panélistes consiste donc à estimer le degré d'efficacité attendue qu'un candidat hypothétique minimalement compétent devrait réussir pour chacun des items d'une épreuve en considérant le niveau de difficulté de ceux-ci. Il appert ici qu'une définition du concept de candidat minimalement compétent doit être intégrée à la procédure d'ESP. À cet effet, une adaptation de la définition proposée par Cizek (2012), selon laquelle la notion de candidat minimalement compétent doit être représentée par un candidat situé à la frontière entre le succès et l'échec, a été fournie aux panélistes.

L'ensemble des jugements émis sur les niveaux de difficulté et les degrés d'efficacité de chacun des 149 items résulteront nécessairement en une classification en 16 catégories (4 niveaux de difficulté et 4 degrés d'efficacité). La méthode alternative

s'inspire à cet effet de la méthode d'Ebel, qui propose une grille de classification issue du croisement entre trois niveaux de difficulté et quatre degrés de pertinence. Une fois que ces premiers jugements sont effectués, il est nécessaire de déterminer l'importance relative de chacune des jonctions difficulté/efficience. Il est proposé, toujours suivant la méthode d'Ebel, que les panélistes estiment, pour chacun des croisements, la proportion d'items qu'un candidat minimalement compétent devrait être capable d'accomplir avec succès. Les panélistes doivent conséquemment déterminer cette proportion pour chacune des 16 catégories. Cette pondération, exprimée en pourcentage, est essentielle à la production des statistiques nécessaires à l'ESP. Concomitamment, suivant les recommandations énoncées dans les neuf principes d'ESP, cet exercice permet aux panélistes de se familiariser avec le contenu des 149 items de l'épreuve ainsi qu'avec les échelles descriptives des niveaux de difficulté et des degrés d'efficience. Afin de favoriser une plus grande consistance pour chacune des 16 catégories, la méthode alternative introduit, à l'instar des autres méthodes, un processus itératif. L'intention est ici de permettre aux panélistes de se repositionner, au regard de la moyenne des données, sur les jugements obtenus suite à chacune des rondes. La suite de la procédure exige de mettre en rang les items de chacune des 16 jonctions, du plus facile au plus difficile, de façon à fournir un ordonnancement croissant des 149 items.

La mise en rang des 149 items se solde par une liste ordonnée de ceux-ci. Cette liste comprend les zones d'incertitude représentant une séquence irrégulière de jugement des croisements difficulté-efficience. À cet effet, la règle départ-arrêt proposée par Ferrara Perie et Johnson (2008), qui précise que les zones d'incertitude sont délimitées par deux séries de trois jugements consécutifs univoques, est appliquée. Le but est ici de réduire chacune des zones d'incertitude en points de coupure qui délimitent clairement les différentes catégories. Les panélistes doivent, pour ce faire, se repositionner sur les items situés dans chacune de ces zones. Afin d'obtenir une suite cohérente, ils doivent d'abord se repositionner sur la mise en rang des items pour

ensuite, s'ils le jugent nécessaire, modifier le niveau de difficulté et le degré d'efficacité. La mise en rang finale est effectuée grâce à un exercice consensuel réalisé à partir de la moyenne des jugements. Les pondérations finales sont obtenues grâce à une décision consensuelle sur les pourcentages d'items qu'un candidat minimalement compétent serait apte à réussir dans chacune des 16 catégories difficulté-efficacité. Une succession d'opérations mathématiques, issue du modèle d'Ebel, doit ensuite être effectuée sur les items et les pourcentages attribués à chacune des catégories afin d'établir le standard de performance spécifique à une tâche ou à l'épreuve. Pour chacune des catégories, une multiplication du nombre d'items classés (n) par le pourcentage (p) estimé est d'abord effectuée. Sont ensuite additionnés tous les résultats obtenus. Le total correspond au nombre d'items qu'un candidat minimalement compétent devra réussir pour atteindre le standard de performance établi. Le total obtenu doit ensuite être divisé par le nombre total d'items, d'une tâche ou de l'épreuve, afin de déterminer, en pourcentage, le standard de performance établi par un panéliste. Lorsque plusieurs panélistes réalisent l'exercice, une moyenne pondérée des estimations réalisées par l'ensemble des panélistes est calculée afin de permettre de déterminer les standards de performance.

Il est ici important de rappeler que le candidat doit, pour rencontrer les exigences de l'OAQ, réussir à atteindre à la fois le standard de performance global de l'épreuve et les 23 standards de performance liés aux 23 tâches. Ce double standard permet de situer la performance du candidat au regard des compétences minimales attendues. La performance démontrée par le candidat permet à l'OAQ de lui délivrer ou non un permis d'exercice avec ou sans restriction. Pour ce faire, la performance globale de l'épreuve ainsi que la performance démontrée à chacune des 23 tâches doivent être prises en considération. La réussite du premier standard est déterminée par le nombre d'items devant être minimalement réussis pour l'ensemble de l'épreuve. L'atteinte du deuxième standard, qui oblige la réussite des 23 tâches, est également déterminée par

le nombre d'items devant être minimalement réussis pour chacune des tâches. Les résultats obtenus à l'épreuve et à l'ensemble des tâches sont, au final, convertis en une notation dichotomique (réussi =1 ou échec =0) qui, à l'instar de la méthode oui/non (Impara et Plake, 1997), permet d'attester de la compétence minimale du candidat. Cette conversion permet d'établir le standard de performance final de l'épreuve qui exige, à l'OAQ, que tous les éléments de compétence ciblés soient réussis au-delà du standard établi. L'addition des résultats de l'ensemble des éléments de compétence détermine si un candidat a atteint la compétence minimale attendue.

2.6 Objectif spécifique du projet de recherche

Afin de faciliter l'accès aux professions réglementées et le processus d'intégration des nouveaux arrivants, et considérant l'obligation qu'ont les ordres professionnels de s'assurer que les professionnelles qui obtiennent un permis d'exercice respectent les normes exigées par la profession, force est de constater la nécessité d'établir des standards de performance. Comme constaté, aucune des méthodes centrées sur le test explorées ne convient parfaitement aux besoins spécifiques des ordres professionnels, particulièrement à ceux de l'OAQ. Ce constat nous a conduits à proposer notre méthode alternative d'ESP centrée sur le test, adaptée à une épreuve clinique authentique d'une profession réglementée, qui permet d'effectuer une classification hiérarchique des items et d'établir les standards de performances correspondant aux exigences de cette profession réglementée. Nous nous sommes principalement inspirés de trois méthodes appréciatives reconnues : la méthode d'Ebel, la méthode IDM et la méthode oui/non. La méthode alternative permet ainsi, par le recours au jugement d'experts du domaine, d'établir des standards de performance quantitatifs destinés à évaluer les compétences professionnelles des acupuncteurs formés à l'étranger et désirant faire reconnaître leurs

acquis expérientiels, et ultimement, de favoriser l'accès au permis d'exercice de la profession d'acupuncteur au Québec.

L'objectif spécifique de notre projet de recherche consiste, à l'aide de la méthode alternative proposée, à établir des standards de performance pour chacune des 23 tâches I-SMART, préalablement validés par l'OAQ, ainsi qu'un standard de performance pour l'ensemble de l'épreuve clinique constituée des 23 tâches. Pour atteindre cet objectif, la méthode alternative proposée sera mise à l'essai.

3. MÉTHODOLOGIE

3.1 Introduction

La méthodologie retenue pour la recherche sera décrite dans la présente section. Dans un premier temps, les outils qui ont servi à la collecte des données seront décrits et justifiés en fonction de notre question de recherche. Puisque ce n'est pas l'objet de ce mémoire, le processus de validation du contenu de l'épreuve clinique ne sera pas décrit ici, ni les outils l'ayant permis. Dans un deuxième temps, le processus de sélection et les modalités de la mise en place de la formation des panélistes seront détaillés. Dans un troisième temps, le déroulement de la méthode alternative d'ESP mise à l'essai sera expliqué. Finalement, dans un quatrième temps, les considérations éthiques de la recherche seront présentées, de même que le processus ayant permis la transmission des résultats aux panélistes. En d'autres mots, ce chapitre explique la façon de procéder pour établir des standards de performance destinés à évaluer les compétences professionnelles des acupuncteurs formés à l'étranger et désirant faire reconnaître leurs acquis expérientiels dans le contexte d'une épreuve clinique visant à favoriser l'accès au permis d'exercice de la profession d'acupuncteur au Québec.

3.2 Outils nécessaires à l'ESP

La mise à l'essai de la méthode alternative requiert d'élaborer les outils nécessaires au déroulement de l'ensemble de la démarche d'établissement des standards de performance. L'approche nécessite l'élaboration d'un instrument de mesure (épreuve

clinique), de deux échelles de notation qualitatives (descriptives) des items et d'une trousse de formation. Une fois les standards de performance établis, afin de permettre à l'OAQ de porter un jugement à partir des performances observées et de fournir à l'épreuve clinique aux candidats aux fins d'autoévaluation, il est impératif que l'OAQ élabore une grille d'évaluation (Ordre des acupuncteurs du Québec, 2014a). Cette grille a nécessité le développement d'un algorithme afin d'automatiser le traitement des notations attribuées aux performances observées par un maître clinique.

3.2.1 Conception d'un instrument de mesure (l'épreuve clinique)

Pour établir des standards de performance permettant d'évaluer l'acquisition des cinq compétences professionnelles minimales requises pour exercer la profession d'acupuncteur au Québec, une épreuve clinique a été conçue. Celle-ci permet de mesurer la performance des trois composantes de la dimension professionnelle à évaluer : la situation clinique, l'intervention clinique et la continuité clinique (le suivi). Cinq des sept compétences associées à ces trois composantes ont été retenues. L'épreuve clinique est conséquemment composée de 149 items ventilés entre les 23 éléments de compétences (tâches) associés à ces 5 compétences.

Cet instrument de mesure a nécessité, lors de la phase préliminaire, d'aider l'organisation réglementée à élaborer le contenu à partir du modèle I-SMART adopté par cette dernière. Ce modèle a permis de structurer la transposition du contenu du référentiel des cinq compétences (éléments de compétences, critères de performance, etc.) dans le contexte de l'épreuve clinique. Cette transposition a permis de formuler, de façon standardisée, 23 tâches intégratives complexes et authentiques et 149 indicateurs observables (items) s'inscrivant dans diverses mises en situations cliniques réelles ou simulées.

Ces tâches, de complexités variables, ont été ventilées entre les cinq compétences professionnelles fondamentales à évaluer :

- ☯ effectuer la collecte des données (7 tâches, 69 items, 30 critères);
- ☯ interpréter les données recueillies (3 tâches, 22 items, 12 critères);
- ☯ planifier l'intervention (3 tâches, 14 items, 8 critères);
- ☯ appliquer le protocole d'intervention (9 tâches, 37 items, 37 critères);
- ☯ faire le suivi de l'intervention (1 tâche, 7 items, 6 critères).

L'épreuve clinique situe les trois premières compétences dans l'exercice de l'évaluation de la situation clinique, la compétence suivante dans l'administration du traitement et les deux dernières compétences dans la continuité clinique à assurer. Chacune des compétences correspond donc à un certain nombre de tâches intégratives, que nous allons succinctement décrire, et auxquelles sont rattachés un certain nombre d'items. Afin de soutenir la mise en œuvre des tâches et de permettre l'évaluation des compétences, ces items sont ventilés entre les 93 critères de performance de l'épreuve clinique (Ordre des acupuncteurs du Québec, 2014a).

Sommairement, pour effectuer la collecte des données, trois des sept tâches consistent à identifier de façon récurrente, lors de la première consultation, les besoins du client (trois critères), à réaliser les quatre temps de l'examen (trois critères) et à procéder à la collecte des données sur l'état de santé du client au moyen de techniques occidentales d'examens physiques, de diagnostics biomédicaux et d'antécédents médicaux (trois critères). Les quatre autres tâches consistent à adapter, récurrentement et en tout temps, sa stratégie de communication (trois critères), à appliquer les principes d'une relation thérapeutique basée sur le partenariat thérapeute client (huit critères), à assurer la mise à jour complète et précise de dossiers cliniques variés (quatre critères) et à gérer, de façon opportune, la relation d'aide (six critères). Subséquemment, pour interpréter les données recueillies et en guider l'interprétation, deux des trois tâches

consistent à effectuer récurrentement et dès le début de la relation clinique une analyse rigoureuse (six critères) et à sélectionner, tout au long de la relation clinique, l'information pertinente à retenir (trois critères). La troisième tâche consiste, durant la consultation, à recommander aux intervenants appropriés un client présentant des signes de gravités et à en assurer le suivi (trois critères).

Pour planifier l'intervention, deux des trois tâches consistent à déterminer récurrentement et dès le début de la relation clinique la ou les orientations thérapeutiques cohérentes (trois critères) et à présenter un plan de traitement cohérent (quatre critères). La troisième tâche consiste à obtenir, avant sa mise en œuvre et en tout temps, le consentement libre et éclairé du client au sujet du plan d'intervention thérapeutique présenté (un critère).

Pour appliquer le protocole d'intervention, les neuf tâches consistent à, récurrentement et en tout temps :

- ❷ préconiser une intervention clinique qui respecte la réglementation, les valeurs et les principes éthiques de la profession ainsi que les limites du client et de la relation clinique (quatre critères);
- ❷ adapter son intervention au contexte professionnel (trois critères);
- ❷ exécuter le traitement en adoptant une pratique sécuritaire et respectueuse du client (cinq critères);
- ❷ démontrer un comportement qui assure le contrôle de la qualité et l'élimination des risques d'accident (cinq critères);
- ❷ adopter des attitudes et des comportements professionnels optimisant son intervention clinique (quatre critères);
- ❷ prévenir les dangers d'infections (six critères);
- ❷ organiser l'environnement physique de façon à assurer des soins de qualité et sécuritaires (quatre critères);

- ❷ démontrer des attitudes et des comportements qui découlent des valeurs et des principes éthiques et déontologiques de la profession (un critère);
- ❸ intervenir efficacement dans un contexte d'urgence, de crise ou de violence (cinq critères).

Succinctement, pour faire le suivi de l'intervention, la tâche consiste à ajuster récurremment et avec efficience l'intervention selon l'évolution de situations cliniques multiples et variées (six critères).

Ces 23 tâches sont associées à plusieurs mises en situation favorisant la mobilisation des ressources vraisemblablement intégrables. Ces mises en situation portent sur une série de trois types de problèmes communément rencontrés dans le cadre de l'exercice de la profession : des problèmes simples, des problèmes complexes et des problèmes relatifs au Shen*, notamment des problèmes de communication ou de santé mentale. Pour conclure, l'épreuve est constituée de 149 items à réponses construites élaborées. Ces items ont été formulés conformément à la définition de la compétence et au modèle I-SMART retenus par l'OAQ et destinés à être utilisés par un maître clinique dans le cadre d'entrevues semi-dirigées et d'observations directes.

3.2.2 Élaboration des échelles de mesure utiles à l'établissement de standards de performance

Afin de recueillir de façon ordonnée les appréciations (notations) des panélistes nécessaires à l'ESP, deux échelles ordinales (échelles descriptives) ont été élaborées, l'une se rapportant au niveau de difficulté perçue de l'item et l'autre, au degré d'efficience minimum attendu d'un candidat minimalement compétent.

Afin de favoriser une compréhension commune des échelles de mesure, ces deux échelles de notation qualitatives (d'appréciation) ont servi à élaborer les descripteurs

de chacun des niveaux de difficulté (facile, moyen, difficile, très difficile) et des degrés d'efficacité (faible, moyen, fort, très fort). Puisque ces descripteurs ont précédemment été décrits dans le cadre théorique, ceux-ci ne seront pas présentés ici. Ces descripteurs sont systématiquement utilisés lors de l'expérimentation de la méthode alternative que nous avons mise à l'essai et ils ont été inclus dans la trousse de formation. Une mise en rang des échelles combinées (de 01 à 16) a été déterminée pour guider les panélistes lors de la classification ascendante des items afin de soupeser simultanément le niveau de difficulté et le degré d'efficacité de chacun de ceux-ci. Cette mise en rang combinée a permis de faciliter l'ordonnement des items et d'assurer un ordre logique et cohérent (du plus facile au plus difficile).

3.2.3 Préparation de la trousse de formation à l'ESP

Une trousse rassemblant le matériel, l'équipement et les documents utiles à la formation des panélistes a été constituée pour chacun d'entre eux. Chaque trousse contenait :

- ☛ un formulaire d'information et de consentement;
- ☛ une fiche contenant les deux définitions conceptuelles fondamentales à adopter : la compétence et l'individu minimalement compétent;
- ☛ des fiches contenant les descripteurs de chacune des catégories de performance (quatre degrés d'efficacité), les niveaux de difficulté des items (quatre) et leur ordonnancement combiné (du plus facile au plus difficile);
- ☛ une clé USB, de couleur différente, contenant un chiffrier électronique de notation;
- ☛ un ordinateur portable équipé d'une version spécifique du logiciel Excel;
- ☛ un formulaire de remboursement des dépenses relatives à la participation à la journée d'expérimentation.

Le premier document a permis de recueillir le consentement libre et éclairé de chaque panéliste volontaire. Le chiffrier électronique de notation a servi à consigner individuellement leurs jugements pour chacune des rondes, séparées dans des onglets distincts. Des menus déroulants permettant de sélectionner l'option de notation désirée pour chacun des items de l'épreuve ont été inclus à cet outil. Afin de trier rapidement l'ensemble des items, chacun de ceux-ci y était identifié par un code. Le contenu de ce chiffrier a été protégé tout en permettant la modification des cellules destinées à recueillir les appréciations. Le chiffrier a subséquemment été reproduit pour chacun des panélistes afin de permettre une collecte uniformisée des données et un traitement statistique efficace et rapide entre chacune des séries (rondes) de jugements. Au final, cet outil informatisé a permis de réduire le temps de consignation des jugements.

Une trousse complémentaire rassemblant tout le matériel nécessaire à la formation des panélistes et à l'animation des rencontres a également été constituée. Cette trousse contenait :

- ☛ une liste des participants;
- ☛ les coordonnées du lieu où se déroulait l'expérimentation;
- ☛ une liste du matériel et de l'équipement fourni par l'organisation (imprimante, rallonge électrique, projecteur, etc.);
- ☛ l'échéancier des activités devant être réalisées durant la journée;
- ☛ une clé USB contenant tous les chiffriers d'analyse de notation;
- ☛ un ordinateur portable muni de la même version du logiciel Excel que les panélistes ainsi que tout le matériel de raccordement;
- ☛ un nombre suffisant de sondages et d'exemplaires du référentiel des compétences, ainsi que de l'épreuve clinique;
- ☛ un cahier de bord destiné à noter les observations qui pourraient s'avérer utiles lors de l'analyse des résultats;
- ☛ une trousse de formation des participants.

Afin de produire, en une journée, les statistiques soutenant les jugements entre chacune des rondes de notation de chacune des phases d'expérimentation, une série de chiffriers d'analyse a été conçue. C'est grâce au traitement informatisé des données issues des chiffriers individuels et à la simulation d'une journée complète d'expérimentation pour tester le bon fonctionnement de cette opération qu'elle a pu être réalisée. Un traitement manuel a toutefois été requis tant pour intégrer les données des panélistes que pour transmettre les résultats de leur panel à chacun des panélistes. Diverses fonctions statistiques ont permis, à partir d'onglets individuels, de reproduire fidèlement les données de chacun des panélistes. Cette stratégie a permis d'éviter des erreurs de transcription tout en standardisant le format des données. Ultimement, cette opération a également permis de réduire le temps de traitement des données et de garantir l'intégrité des données. L'OAQ a conséquemment bénéficié d'un gain de temps et d'économies financières non négligeables puisque les statistiques ont été produites au fur et à mesure que les données étaient collectées.

À la fin de l'expérimentation, les données du sondage d'opinion ont été compilées pour confirmer le niveau de confiance face à la procédure d'ESP et aux résultats obtenus. Afin de permettre une analyse quantitative de la confiance accordée à la démarche de recherche, le sondage a été rédigé en s'inspirant essentiellement du formulaire d'évaluation du processus d'ESP proposé par Cizek et Bunch (2007). Une échelle de Likert à quatre points (fortement en accord, en accord, en désaccord, fortement en désaccord) a également été incluse au questionnaire.

Dans le but de répondre à la part de subjectivité individuelle et collective et de favoriser une compréhension commune, il a été nécessaire de présenter les notions conceptuelles fondamentales à adopter, à savoir les définitions relatives à la compétence et à un individu minimalement compétent. Les définitions qui ont été retenues sont celles privilégiées par l'OAQ. Un candidat minimalement compétent se définit ainsi comme une personne standard tout juste qualifiée ou qui se situe tout juste au-delà du niveau

de qualification soit, comme Nedelsky (1954) le souligne, à la frontière entre le succès et l'échec.

3.2.4 Développement d'un algorithme de traitements des notations

Afin de faciliter le processus d'évaluation des performances d'un candidat et de s'assurer que le professionnel qui obtient un permis d'exercice respecte les normes exigées par la profession, conformément au double standard de performance établi, un algorithme a dû être développé.

Lors de l'élaboration de cet outil d'évaluation (algorithme), il a été nécessaire de prendre en considération que l'OAQ a tenu à distinguer la mesure et l'évaluation des compétences en assignant ces responsabilités à des individus distincts, et ce, afin d'assurer une meilleure objectivité du processus d'évaluation et sa transparence. L'algorithme a été conçu en tenant compte de la division des rôles entre le maître clinique et l'OAQ. Le rôle du maître clinique, qui doit être accrédité par l'OAQ, consiste à soumettre le candidat à l'épreuve clinique, à observer les performances, à mesurer l'efficacité observée, à inscrire des commentaires pertinents et à transmettre les résultats à l'OAQ. Le rôle du comité d'admission de l'OAQ consiste pour sa part à édicter une prescription adaptée aux exigences minimales de la profession et conforme aux standards de performance établis et à statuer, sur la base des scores obtenus à l'épreuve clinique, sur l'admissibilité d'un candidat à l'octroi d'un permis d'exercice de la profession au Québec. Pour fonder sa décision, l'OAQ a été amené à se servir des appréciations des comportements observés recueillies auprès d'un maître clinique pour alimenter l'algorithme qui, en fin de compte, produira un rapport d'évaluation des compétences selon une comparaison entre les scores obtenus et les standards de performance établis. Les commentaires recueillis permettront également, à terme, de soutenir la prescription édictée de même que, le cas échéant, l'ensemble des conditions

imposées.

Sommairement, l'algorithme permet la saisie informatisée des appréciations qualitatives attribuées à chacun des 149 items, la conversion de celles-ci en valeurs quantitatives (scores) et un traitement conforme aux pondérations associées à chaque item. L'algorithme permet, au final, de comparer les scores obtenus avec les standards de performances établis, de préciser l'atteinte ou non du standard de l'épreuve clinique et de chacun des 23 standards détaillés ainsi que signaler les forces et les faiblesses observées.

De façon plus détaillée, l'algorithme permet de :

- ❶ calculer le nombre d'items réussis pour chacune des tâches;
- ❷ déterminer, en pourcentage, les scores de réussite de chacune des tâches et celui de l'épreuve;
- ❸ convertir les scores de chacune des tâches en notation dichotomique (réussite = oui, échec = non);
- ❹ additionner l'ensemble des tâches réussies (variable « oui »);
- ❺ identifier les tâches non réussies (variable « non ») en ciblant les performances lacunaires observées pour chacun des items qui y sont associés.

Pour chacune des tâches et pour l'ensemble de l'épreuve, l'algorithme permet donc de rapidement transposer en pourcentage les performances observées afin de déterminer si un candidat possède ou non la compétence minimale exigée pour exercer la profession au regard des standards établis. Ce double standard permet de dresser un diagnostic complet des forces et des faiblesses observées pour chacun des candidats. La compétence minimale est reconnue lorsqu'un candidat atteint simultanément le standard de l'épreuve et les standards des 23 tâches. Lorsque le standard de l'épreuve

est atteint, mais que le standard de chacune des 23 tâches ne l'est pas, l'OAQ peut accorder une reconnaissance partielle des compétences et de ce fait, délivrer un permis d'exercice avec restrictions. Une prescription de formation ciblant les performances lacunaires est alors ordonnée. Lorsque le seuil de réussite de l'épreuve n'est pas atteint, une prescription est édictée sans permis d'exercice. Pour qu'un permis soit délivré, avec ou sans restriction, le seuil de réussite de l'épreuve doit obligatoirement être atteint. Afin de permettre la conversion des performances observées en pourcentage, l'algorithme regroupe un certain nombre de règles et d'opérations qui automatisent la relation entre la liste des items classés par tâche, leur classification par catégorie de performance (difficulté-efficience), les performances observées et les standards de performance établis. Quoique l'outil ait été conçu pour automatiser la transposition des résultats en données chiffrées, il sera important, pour s'assurer que les données sont saisies de façon conforme et d'une lecture appropriée des résultats, de former les décideurs de l'organisme réglementaire à l'utilisation de l'algorithme. Comme cet outil est confidentiel et réservé à l'usage exclusif de l'OAQ, la seule description qui est présentée est celle ci-dessus.

3.3 Sélection et formation des panélistes

Afin de favoriser un essai réussi, nous avons jugé essentiel de prendre en compte le processus de sélection des panélistes avec attention, de dresser le profil de ceux ayant participé à l'expérimentation de la méthode d'ESP proposée et de s'intéresser à leur formation.

3.3.1 Sélection des panélistes ayant participé à l'expérimentation

La sélection des panélistes invités à porter un jugement sur l'ensemble des items retenus nous a d'abord amenés à définir le groupe cible (la population cible) et à déterminer les caractéristiques recherchées. Comme le soulignent plusieurs auteurs, il est important de prioriser un processus de sélection qui fait appel aux pairs (Bullock,

2000; Greenwood et King, 1995; Jensen *et al.*, 2000; Raufaste, Eyrolle et Mariné, 1998; Unsworth, 2001a; Yelder, 2004, cités dans Faucher, 2012). L'établissement des standards de performance exige d'ailleurs que « les panélistes soient représentatifs et possèdent les compétences pour réaliser la tâche demandée » (Blais, 2008, p. 103). Ainsi, puisque la démarche de recherche s'adresse spécifiquement à l'OAQ, les panélistes se devaient manifestement d'en être membres.

La figure 3.1 présente les critères retenus pour sélectionner les sujets qui, de fait, doivent obligatoirement être aptes à comprendre l'objectif de la recherche et les différentes étapes à suivre dans la démarche d'ESP.

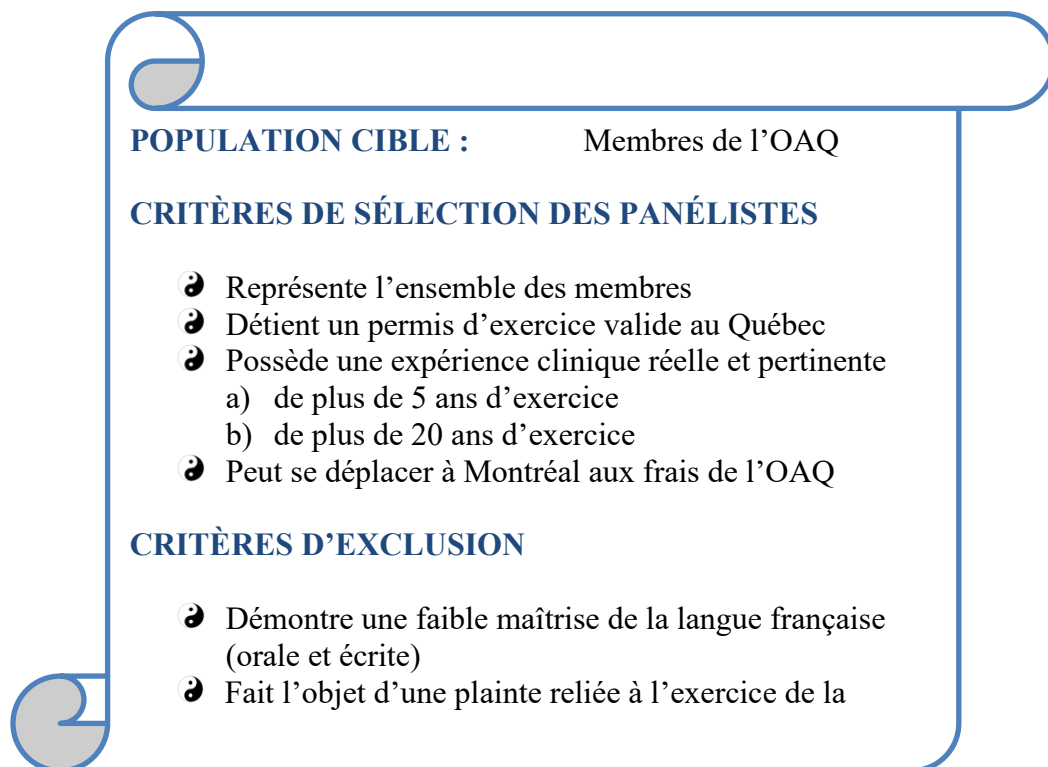


Figure 3.1 – Illustration des critères de sélection des sujets

La sélection des sujets exigeait que les acupuncteurs soient membres de l'OAQ, aptes à représenter l'ensemble des membres de l'organisme, et qu'ils détiennent un permis d'exercice valide au Québec. Ils devaient également posséder une expérience clinique réelle et pertinente et être disposés à se déplacer à Montréal aux frais de l'OAQ. Ce dernier critère a ainsi permis de sélectionner des acupuncteurs exerçants dans les 17 régions administratives du Québec et de prendre en compte les ressources financières et humaines disponibles. Finalement, deux critères d'exclusion ont été pris en compte. Tous les membres de l'OAQ ne maîtrisant pas la langue française orale et écrite et tous ceux faisant l'objet d'une plainte reliée à l'exercice de la profession, que cette plainte ait conduit à une mesure disciplinaire ou non, ont donc été exclus. Précisons, au sujet de la maîtrise de la langue française, que ce critère était important puisqu'une bonne compréhension des diverses étapes de la démarche et de la subtilité inhérente à différents items en était directement tributaire. Quant aux mesures disciplinaires, celles-ci devaient également être prises en compte pour éviter toute remise en question de la compétence du panéliste lorsqu'il s'agit de réaliser les tâches.

Afin d'assurer une meilleure représentativité de la réalité traitée, la démarche appliquée dans le cadre de cette recherche a mené à constituer deux échantillons intentionnels distincts. Le premier échantillon regroupait des membres reflétant le plus fidèlement possible la situation des acupuncteurs peu expérimentés. Cette technique, appelée « l'échantillonnage de cas typiques », est largement répandue pour sélectionner des sujets (Gaudreau, 2011). Le deuxième échantillon a été constitué d'individus inhabituels exerçant une influence notamment appréciable ou étant dans une situation légitime pour l'exercer dans le cadre de la recherche. Cette technique, appelée « l'échantillon de cas politiquement significatifs », permet de recruter de façon intentionnelle (Gaudreau, 2011) des sujets possédant une expertise appréciable dans le domaine ciblé. Dans les deux cas, la sélection de l'échantillon a été réalisée à l'aide d'une liste d'acupuncteurs volontaires préalablement établie par l'OAQ et dont les constituants répondaient à l'ensemble des critères généraux et des critères spécifiques.

Quoique la sollicitation des sujets aurait pu être étendue à l'ensemble des membres de l'OAQ, ces deux types d'échantillonnage, qualifiés de non probabilistes, s'imposaient, comme le souligne Gaudreau (2011), puisque la sélection ne pouvait être aléatoire ni s'avérer coûteuse.

L'échantillonnage de « cas typiques », soit les acupuncteurs possédant un permis récent (valide depuis 5 ans ou moins), donc étant peu expérimentés, s'est alors avéré pertinent au regard, notamment, de la compétence d'entrée dans la profession. Ces participants se veulent représentatifs des caractéristiques les plus répandues ou encore, se situent dans la moyenne des professionnels débutants. Leur rôle consistait à émettre des jugements représentatifs de cette situation particulière en s'appuyant sur des cas et des situations vécues ainsi que sur leur expérience de fraîche date. L'échantillonnage de « cas politiquement significatifs » s'est également avéré pertinent et réalisable. La constitution de ce panel a exigé le recrutement d'acupuncteurs volontaires possédant une expertise considérable. Comme ces acteurs ont été amenés à jouer un rôle important dans la recherche, ils devaient détenir un permis d'exercice valide depuis 20 ans et plus.

Les directives relatives à la constitution des échantillons (panels) ont ainsi d'abord été établies par la présentation des critères de sélection à l'OAQ. Une lettre d'invitation présentant sommairement le déroulement de la démarche d'expérimentation ainsi que les dates retenues a ensuite été acheminée à l'ensemble des membres répondant aux critères de sélection. Les acupuncteurs intéressés étaient alors invités à indiquer leurs disponibilités sur la plateforme en ligne Doodle, à la suite de quoi une liste finale a été établie. Chacune des personnes inscrites sur cette liste a reçu une confirmation l'avisant qu'elle avait été sélectionnée pour faire partie d'un des panels. Quant aux personnes qui n'ont pas été sélectionnées, elles ont toutes reçu un courriel les remerciant de leur intérêt pour la recherche. Les listes des participants à chacun des panels ont finalement servi à consigner leurs signatures et le temps qu'elles allaient allouer à la journée

d'expérimentation. Ces listes ont également permis à l'OAQ de valider la présence de chacun des membres, de procéder au remboursement des frais engendrés par leur participation et d'ajouter à leur dossier de formation continue les heures consacrées à la journée d'expérimentation.

En tout, 10 professionnels acupuncteurs experts (acteurs) ayant 20 ans et plus d'expérience et 10 acupuncteurs novices (participants) possédant 5 ans ou moins d'expérience ont été recrutés. Pour des considérations logistiques, chacun de ces deux échantillons a été subdivisé en deux groupes distincts. Au total, quatre panels (novices A et B; experts C et D) ont ainsi été constitués³. Bien qu'il s'agît d'un échantillon de petite taille, celui-ci permettait malgré tout d'établir une représentativité exportable à l'ensemble des situations comparables. Nous appuyons de fait cette affirmation sur certains auteurs qui mentionnent qu'un échantillon formé de quatre professionnels par niveau est jugé suffisant pour atteindre la saturation des données (Edwards *et al.*, 2004; Embrey *et al.*, 1996; Jensen, Gwyer, Shepard et Hack, 2000, cités dans Faucher, 2012, p. 86).

3.3.2 Profil des panélistes ayant participé à l'expérimentation

Dans le cadre de notre recherche, 85 % des acupuncteurs répondant aux critères de sélection ont pris connaissance de l'invitation à participer à la recherche. Parmi ceux-ci, 20 % ont pris part à la démarche. Les acupuncteurs novices (n=12) avaient entre 1 et 5 années d'expérience clinique et les experts (n=10), entre 20 et 28 ans d'expérience clinique. Un novice s'est toutefois désisté, et un deuxième ne satisfaisait pas aux critères de sélection, ce qui a réduit ce groupe à 10 panélistes (n=10). Au final, le panel

³ Initialement, cinq panélistes par panel étaient prévus. Cependant, au cours de la démarche, certaines disponibilités n'étant plus compatibles, chaque panel a été constitué d'un minimum de quatre panélistes. De plus, lors de l'expérimentation, un expert du groupe C s'est désisté en raison d'une contrainte inattendue. Cet expert a donc été transféré dans le panel D.

de novices était constitué majoritairement d'acupunctrices (90 %), alors que le panel d'experts était majoritairement composé d'acupuncteurs (70 %).

La répartition des novices s'est avérée statistiquement représentative de toutes les tranches d'années d'expérience : quatre panélistes possédaient 5 ans d'expérience (40 %), deux possédaient 4 ans d'expérience (20 %), un possédait 3 ans d'expérience (10 %), deux possédaient 2 ans d'expérience (20 %) et un avait 1 an d'expérience (10 %). La répartition des 10 experts s'est également avérée représentative à l'intérieur d'une fourchette de 20 ans à 28 ans d'expérience acquise au Québec : cinq panélistes possédaient moins de 25 ans d'expérience (50 %) et cinq possédaient 25 ans et plus d'expérience (50 %). L'ensemble des panélistes provenait de six régions administratives différentes : la Montérégie (45 %), Montréal (30 %), Lanaudière (10 %), Laval (5 %), Chaudière-Appalaches (5 %) et la Capitale-Nationale (5 %).

3.3.3 Formation des panélistes ayant participé à l'expérimentation

Une fois les panels constitués, conformément au quatrième principe d'ESP, chaque panéliste a été soumis à une période de formation et de familiarisation à la méthode proposée. Préliminairement, les postes informatiques ont été installés et les trousseaux de formation ont été distribués. La formation a commencé, pour chaque panel, par une présentation sommaire des buts de l'expérimentation, de la nature de la tâche et de la modélisation du processus mis à l'essai. S'en est suivi une courte période de questions-réponses sur le consentement libre et volontaire devant être signé avant le début de l'expérimentation. Les panélistes ont d'abord été invités à prendre connaissance du contenu de l'épreuve clinique (items), des descripteurs, des échelles de notation (niveaux de difficulté et degrés d'efficacité) et de leur mise en rang ainsi que du concept de candidat minimalement compétent. Une période de présentation a permis d'expliquer les concepts clés et d'en assurer une interprétation commune afin d'éviter des erreurs d'interprétation (Egan *et al.*, 2009, traduction libre, p. 562). Dans la même

visée, les échelles de notation et leurs descripteurs ont été expliqués, de même que l'outil permettant la saisie informatisée des appréciations pour chacun des items de l'épreuve clinique. Cela a donné lieu à une période d'échanges entre les panélistes, ce qui a favorisé une compréhension commune de la terminologie utilisée durant l'expérimentation. Une démonstration a ensuite été effectuée afin d'assurer que les panélistes comprenaient bien la façon d'enregistrer les données (notation) dans le chiffrier électronique qui leur a été fourni. Un temps de familiarisation avec cet outil informatisé de notation des jugements a également précédé le début de l'expérimentation. En vue d'assurer le bon fonctionnement de la mise à l'essai de la méthode alternative d'ESP proposée, les panélistes bénéficiaient, tout au long de l'expérimentation et sur demande, d'un soutien individuel.

3.4 Déroulement de l'expérimentation de la méthode alternative d'établissement de standards de performance

Comme décrit précédemment dans le contexte théorique, quatre opérations (tâches) ont été réalisées. La première opération consistait en une catégorisation de chaque item par niveau de difficulté perçu (facile, moyen, difficile ou très difficile) et par degré (très fort, fort, moyen ou faible) d'efficience minimal préconisé pour que l'on considère qu'un candidat est minimalement compétent au regard de cet item. Seize (4*4) catégories de performance regroupant un certain nombre d'items ont ainsi été créées. La seconde opération, selon la matrice de classification d'Ebel, consistait à assigner une pondération à chacune des 16 catégories représentant le pourcentage d'items que devrait raisonnablement réussir un candidat minimalement compétent, et ce, selon le nombre d'items dans chacune des catégories. C'est cette dernière opération qui permettra de calculer chacun des seuils de réussite (standard de performance). La troisième consistait à effectuer une mise en rang à l'aide de la catégorisation afin d'obtenir une liste ordonnée des items. Enfin, la quatrième opération consistait, suite au calcul de la moyenne des rangs, à délimiter les zones d'incertitude calculées.

L'ensemble de ces opérations peut globalement être présenté comme un processus itératif effectué en deux phases constituées chacune de quatre rondes de notation. Afin de permettre une fine compréhension de la mise à l'essai de notre méthode alternative, la suite de notre présentation expose en détail chaque aspect du déroulement de l'expérimentation et fournit les indications préliminaires sur les adaptations apportées à la méthode d'Ebel.

3.4.1 **Indications préliminaires sur les adaptations apportées à la méthode d'Ebel**

Nous nous contentons ici d'évoquer que la méthode alternative propose de modifier les deux échelles de mesure de la méthode d'Ebel. Afin d'augmenter le potentiel de discrimination des items devant être classifiés, un quatrième niveau (très difficile) a été ajouté à l'échelle de mesure de la difficulté (facile, moyen, difficile). L'échelle de mesure de la pertinence proposée par Ebel a aussi été remplacée par une échelle mesurant le degré d'effcience (faible, moyen, fort, très fort) afin de qualifier l'effort devant être déployé et les moyens mis en œuvre pour atteindre les résultats attendus. Afin d'en arriver ultimement à déterminer des standards de performance, notre modèle alternatif réfère également à la méthode de l'appariement au descripteur de l'item (*The Item-Descriptor Matching Method*) pour la mise en rang ascendante des items par niveau de difficulté-effcience ainsi que pour l'usage de la règle départ-arrêt, qui sert à identifier les zones d'incertitude entre les niveaux de difficulté. Finalement, afin de répondre à la règle de l'OAQ, qui oblige à ce que les 23 tâches liées aux 23 éléments de compétence de l'épreuve soient toutes réussies, nous avons ajouté à notre méthode alternative l'utilisation des variables dichotomiques (oui/non), tirées de la méthode d'Impara et Plake.

3.4.2 Déroulement de l'expérimentation en trois étapes

La méthode alternative que nous proposons se définit en trois étapes, soit une étape préliminaire de planification des opérations et d'organisation de l'expérimentation, une étape intermédiaire d'expérimentation et une étape finale d'établissement des standards de performance et développement d'un algorithme (voir la figure 3.2).

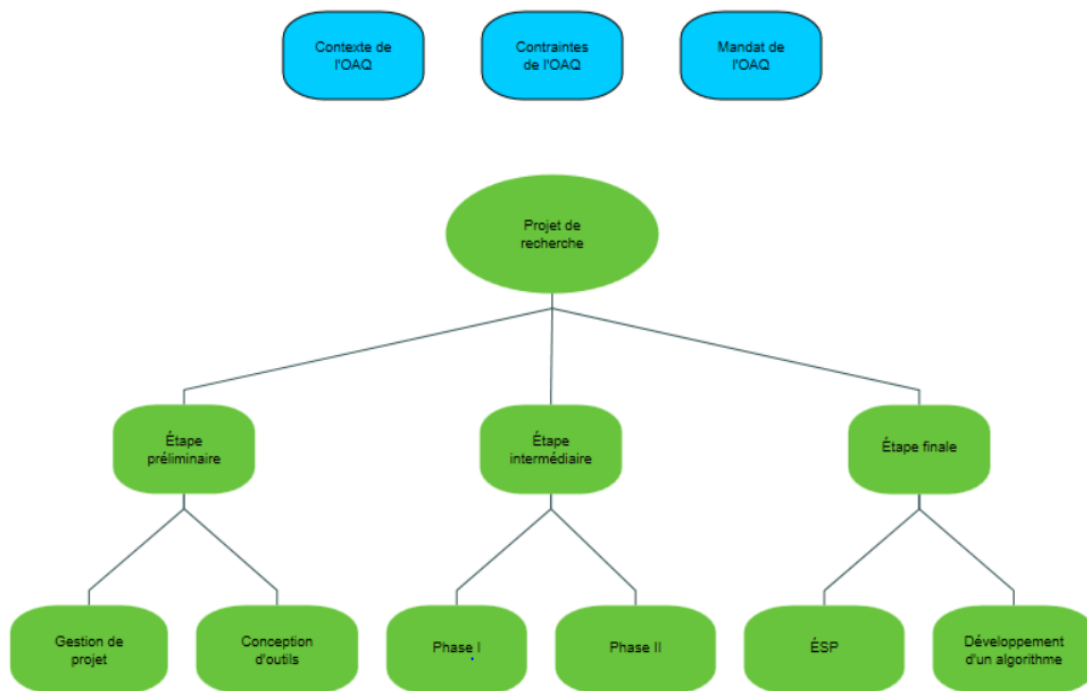


Figure 3.2 – L'établissement de standards de performance en trois étapes

L'étape préliminaire correspond globalement à la gestion du projet de recherche et à la conception des outils nécessaires à son bon déroulement. Elle consiste, plus spécifiquement, à coordonner et à mobiliser l'ensemble des ressources humaines, matérielles et financières nécessaires à la planification et à la mise en œuvre de la procédure permettant d'établir les standards de performance.

L'étape intermédiaire, qui représente spécifiquement notre méthode alternative, peut globalement être présentée comme un processus itératif qui s'appuie sur la récurrence de quatre tâches appliquées à deux phases de quatre rondes chacune. Trois de ces tâches sont effectuées par les panélistes qui prennent part à l'expérimentation, alors que la quatrième tâche est effectuée par la chercheuse.

L'étape finale de la méthode alternative consiste, quant à elle, à établir les standards de performance et à développer un algorithme informatique permettant à la fois de situer les performances au regard des standards de performance de la profession et d'identifier les forces et les faiblesses observées dans le cadre de l'évaluation des compétences professionnelles.

3.4.3 Déroulement détaillé de la méthode alternative d'établissement de standards de performance

3.4.3.1 L'étape préliminaire

Rappelons brièvement que l'étape préliminaire à l'expérimentation de la méthode alternative d'ESP consiste, pour la chercheuse, à rassembler les outils nécessaires à sa mise en œuvre (épreuve clinique, échelles descriptives, trousse de formation), à mener à bien le processus de sélection, à assurer la formation de l'ensemble des panélistes retenus ainsi que l'accompagnement individuel nécessaire. Puisque ces sujets ont été évoqués plus haut, nous nous contenterons d'indiquer que les outils ont été préalablement testés avec succès et que la formation à l'utilisation de cette méthode a été programmée conformément au quatrième principe d'établissement des standards de performance. La nature de la tâche et la modélisation du processus d'ESP ont été communiquées. De plus, les concepts clés et les échelles descriptives ainsi que la structure du chiffrier de notation ont été expliqués. Ensuite, les panélistes ont été familiarisés au contenu de l'épreuve clinique (items), puis formés à l'utilisation des outils et à l'enregistrement des notations dans le chiffrier électronique.

3.4.3.2 L'étape expérimentale

La première phase d'expérimentation de notre méthode alternative d'ESP consistait essentiellement à procéder à la collecte et au tri des données recueillies auprès des panélistes novices (≤ 5 ans d'expérience) et des experts (≥ 20 ans d'expérience) et, afin d'appuyer leurs jugements subséquents, à produire les statistiques descriptives sur la moyenne des jugements entre chacune des rondes de notation. La deuxième phase suivait essentiellement la même séquence que la première, mais spécifiquement pour les zones d'incertitude. Elle a permis, en ayant cette fois recours uniquement à des panélistes experts, de parachever la mise en rang et d'établir les points de coupure définitifs délimitant les frontières de chacune des catégories de performance (difficulté-efficience). Cette deuxième phase a également permis d'accroître le nombre de jugements émis lors de la première phase sur les pondérations de la matrice.

Les quatre rondes de notation de chacune des deux phases de la méthode se sont déroulées en une journée dans les locaux de l'OAQ. Chaque ronde a été entrecoupée d'une pause, nécessaire à la production des statistiques. Les trois premières rondes exigeaient que les panélistes qu'ils se prononcent individuellement, alors que la quatrième ronde visait à parvenir à un consensus. Chaque panéliste a ainsi été amené à effectuer plusieurs séries de jugement en s'appuyant sur son expérience personnelle. Le tableau 3.1 à la page suivante illustre la modulation des quatre tâches (catégorisation, assignation, mise en rang et délimitation) suivant la succession des quatre rondes du processus itératif de notre méthode d'ESP.

Tableau 3.1 – Processus itératif détaillé de la méthode alternative d’ESP

Processus itératif d’établissement de standards de performance									
Phase I									
		Ronde 1		Ronde 2		Ronde 3		Ronde 4	
Consentement	Formation	Catégorisation Assignment	Statistique	Catégorisation Assignment	Statistique	Mise en rang		Assignment	Sondage
Phase II									
		Ronde 1		Ronde 2		Ronde 3		Ronde 4	
Consentement	Formation	Catégorisation Mise en rang Assignment	Statistique Délimitation	Catégorisation Mise en rang Assignment	Statistique Délimitation	Catégorisation Mise en rang	Statistique Délimitation	Catégorisation Mise en rang Assignment	Sondage

Ce processus itératif d’ESP permet d’effectuer une collecte des données issues de plusieurs séries de jugements visant à atteindre un haut niveau d’accord entre les panélistes.

La première phase d’expérimentation a été successivement réalisée auprès d’un groupe de 10 panélistes novices, subdivisé en deux groupes distincts (panels A et B), et d’un groupe de trois panélistes experts (panel C). La première ronde de notation visait à alimenter la matrice et à établir les standards de performance conformément à l’algorithme proposé par Ebel et aux deux échelles de mesure élaborées. Cette ronde a d’abord amené les panélistes à catégoriser chacun des 149 items de l’épreuve en sélectionnant, à l’aide de deux listes déroulantes concomitantes, le niveau de difficulté devant y être attribué et le degré d’efficacité estimé. Ensuite, chaque panéliste a été amené à assigner les pondérations correspondantes à chacune des 16 catégories de performance de la matrice proposée. Il leur a alors été demandé d’indiquer sous forme de pourcentage la proportion (p) d’items qu’un candidat minimalement compétent devrait raisonnablement être en mesure de réussir.

Lors de la deuxième ronde de notation, les panélistes ont émis les mêmes jugements individuels en consultant les résultats moyens de la première ronde.

Les résultats moyens de la deuxième ronde ont ensuite permis, lors de la troisième ronde, d'effectuer une première mise en rang des items, du plus facile au plus difficile. Les panélistes devaient s'efforcer de s'assurer d'une progression logique et cohérente du degré d'efficience attendu d'un candidat minimalement compétent. Pour simplifier cette tâche longue et complexe, une stratégie a été suggérée aux panélistes, c'est-à-dire de filtrer et de trier les items par catégorie. Ainsi, les panélistes ont effectué 16 séries de mise en rang distinctes, soit une série par catégorie de performance (difficulté-efficience).

Suivant cette ronde, 15 points de coupure ont automatiquement été déterminés par le dernier item de chacune des catégories de performance (difficulté-efficience). Aucun jugement n'a, par conséquent, été nécessaire pour identifier ces points de coupure puisque c'est la moyenne de la mise en rang par l'ensemble des panels qui a servi à établir un nouvel ordonnancement des items et à identifier les zones d'incertitude devant servir, lors de la deuxième phase d'expérimentation, à situer les points de coupure entre chacune des 16 catégories de performance (difficulté-efficience). Seuls les trois points de coupure (moyen, difficile, très difficile) délimitant les frontières des quatre niveaux de difficulté (facile, moyen, difficile, très difficile) se sont avérés pertinents lors de la première phase d'expérimentation.

La quatrième et dernière ronde de notation a amené les panélistes à parvenir à un consensus sur les résultats de la tâche d'assignation des pondérations.

À la fin de la journée, pour vérifier les qualités inhérentes à la méthode alternative d'ESP mise à l'essai et ainsi, contribuer à la validité des standards de performance, les

panélistes ont répondu à un sondage d'opinion. Comme Cizek (2012) le souligne, ce sondage permet également de vérifier le niveau de confiance des répondants au regard du processus, de sa mise en œuvre et des résultats obtenus.

Une fois la première phase terminée, une série d'opérations a dû être effectuée afin de produire des statistiques et concevoir les outils de notation ainsi que ceux pour le traitement et l'analyse automatisés des résultats en vue de la deuxième phase d'expérimentation. La production des statistiques descriptives à l'aide des chiffriers électroniques, qui ont permis le traitement automatisé des données recueillies, a d'abord exigé de constituer une base de données cumulative des notations issues de l'ensemble des panels, puis de calculer les moyennes pour toutes les variables (niveaux de difficulté, degrés d'efficacité, mises en rangs et pondérations) associées aux items. À partir des moyennes obtenues, une liste des items ordonnés du plus facile au plus difficile a été produite et les zones d'incertitude, délimitées suivant la règle de départ-arrêt proposée par Ferrara, Perie et Johnson (2008), ont été encadrées par des bordures. Chacune de ces zones d'incertitude était précédée et suivie de trois items, dont la séquence des jugements difficulté-efficacité (catégorisation) s'est avérée univoque. Ainsi, 16 catégories de performances ordonnées (figure 3.3) ont été créées grâce au fusionnement des deux échelles de notation (niveaux de difficulté et degrés d'efficacité). Ces catégories ont été ordonnées numériquement par niveau de difficulté (croissant), puis subdivisées par degré d'efficacité (décroissant).

01.Facile/Très fort	05.Moyen/Très fort	09.Difficile/Très fort	13.Très difficile/Très fort
02. Facile/Fort	06. Moyen/Fort	10. Difficile/Fort	14. Très difficile /Fort
03. Facile/Moyen	07. Moyen/Moyen	11. Difficile/Moyen	15. Très difficile /Moyen
04. Facile/Faible	08. Moyen/Faible	12. Difficile/Faible	16. Très difficile /Faible

Figure 3.3 – Seize catégories de performance ordonnées par niveau de difficulté et par degré d'efficacité

La liste ordonnée des items (de 1 à 149) a permis d'identifier ceux ne respectant pas l'ordre des échelles fusionnées (de 1 à 16). Chaque zone d'incertitude a été délimitée par le premier item dont la catégorisation interrompait la séquence ordonnée et par le dernier item précédant une série de trois items dont la catégorisation s'est avérée ordonnée.

Une fois ces zones d'incertitude délimitées par deux séries de trois jugements consécutifs univoques, la modélisation d'un outil servant à deuxième phase et destiné aux panélistes a été réalisée. Le tableau 3.2 illustre sommairement cet outil (données fictives).

Tableau 3.2 – Outil de notation fourni aux panélistes lors de la phase II

No.Item Référence	Phase I		Phase II	
	Liste ordonnée des items	Catégorisation D/É moyen	Mise en rang	Catégorisation D/É
I-14	1,0	01. Facile/Très fort	1	01. Facile/Très fort
I-1	2,0	02. Facile/Fort	2	02. Facile/Fort
I-7	3,0	02. Facile/Fort	3	02. Facile/Fort
I-11	4,0	02. Facile/Fort	4	02. Facile/Fort
I-4	5,0	04. Facile/Faible		
I-8	6,0	02. Facile/Fort		
I-17	7,0	02. Facile/Fort		
I-15	8,0	06. Moyen/Fort	8	06. Moyen/Fort
I-2	9,0	06. Moyen/Fort	9	06. Moyen/Fort
I-12	11,3	06. Moyen/Fort	10	06. Moyen/Fort
I-18	11,3	05. Moyen/Très fort		
I-9	11,7	06. Moyen/Fort		
I-20	11,7	11. Difficile/Moyen		
I-5	14,0	12. Difficile/Faible		
I-6	15,0	11. Difficile/Moyen		
I-3	16,0	12. Difficile/Faible	16	12. Difficile/Faible
I-10	17,0	12. Difficile/Faible	17	12. Difficile/Faible
I-13	18,7	12. Difficile/Faible	18	12. Difficile/Faible
I-19	18,7	15. Très difficile/Moyen	19	15. Très difficile/Moyen
I-16	19,7	15. Très difficile/Moyen	20	15. Très difficile/Moyen

Pour toutes les variables fusionnées, une liste déroulante ainsi que deux fonctions (filtre et tri) du logiciel Excel ont été intégrées au chiffrier. Cela a permis, d'une part, de faciliter le traitement des données par catégorie de performance et, d'autre part, de faciliter la tâche de mise en rang des items par catégorie. Il faut préciser que des onglets

différents ont été utilisés pour collecter les notations de façon standardisée et afin de segmenter les tâches des panélistes. Une fois terminé, le modèle du nouvel outil de notation a été dupliqué pour chacun des panélistes de la deuxième phase d'expérimentation. À l'instar de la phase préliminaire, une nouvelle série de chiffriers électroniques a également été conçue afin de traiter les résultats des notations entre chacune des rondes de la deuxième phase. Le bon fonctionnement de l'ensemble des chiffriers a été mis à l'essai par la simulation d'une journée complète de cette deuxième phase d'expérimentation.

La deuxième phase de notre méthode alternative consistait à émettre, suivant le processus itératif de quatre rondes, une série de jugements pour les items situés uniquement dans les zones d'incertitude délimitées par les moyennes des résultats obtenus lors de la première phase. La séquence du processus itératif des quatre rondes de la deuxième phase a été réalisée auprès d'un seul groupe d'experts (panel D) et s'est déroulée en une journée. À l'instar des premières journées de la première phase, cette nouvelle journée d'expérimentation a d'abord nécessité l'obtention du consentement des participants et la formation d'usage. À la fin de la journée, les participants ont également répondu à un sondage d'opinion.

La première ronde de notation reprenait essentiellement les quatre tâches individuelles de l'expérimentation : la catégorisation, la mise en rang, l'assignation et la délimitation. À partir des résultats obtenus, chaque panéliste a simultanément réalisé les tâches de catégorisation et de mise en rang des items situés dans les zones d'incertitude. Enfin, sans avoir accès aux résultats finaux moyens de la première phase, chaque panéliste a effectué la tâche d'assignation. Les moyennes ont ensuite été calculées. Une nouvelle liste ordonnée des items a été produite à partir des moyennes et de nouvelles zones d'incertitude ont été délimitées.

Afin de favoriser une réduction des zones d'incertitude, lors des deuxième et troisième rondes, les panélistes ont été invités à reprendre individuellement les tâches de catégorisation et de mise en rang en tenant compte de la moyenne des résultats obtenus lors de la ronde précédente. Une fois la mise en rang réactualisée et la catégorisation renouvelée, de nouvelles zones d'incertitude ont été délimitées et les points de coupure intérimaires s'en sont dégagés naturellement. Au final, uniquement lors de la deuxième ronde, chaque panéliste s'est repositionné sur la tâche d'assignation à partir des moyennes obtenues à la première ronde. Cette dernière tâche a ainsi permis d'enrichir le nombre de notations relatives aux pondérations de chacune des catégories de performance (difficulté-efficience).

La quatrième ronde de la première phase invitait chaque panel (A, B et C) à parvenir à un consensus sur l'assignation des pondérations aux 16 catégories de performance. La quatrième ronde de la deuxième phase reprenait ainsi intrinsèquement les mêmes tâches de la deuxième ronde de notations (catégorisation, mise en rang, assignation, délimitation), mais cette fois-ci, par l'entremise d'une démarche consensuelle. Pour ce faire, le dernier panel d'experts (D), après avoir consulté les statistiques de la troisième ronde, a été amené à parvenir à un accord tant sur la catégorisation et la mise en rang finale des items que sur l'assignation des pondérations aux 16 catégories de performance. Les résultats de cette toute dernière ronde de notation visaient à permettre de produire la liste ordonnée définitive des items et à identifier les points de coupure définitifs délimitant chacune des catégories de performance, qui permettront, au final d'établir les standards de performance de l'épreuve clinique. Les pondérations intérimaires ont ultérieurement été intégrées à celles obtenues par l'ensemble des panels de la première phase d'expérimentation afin de déterminer les valeurs définitives à intégrer à la matrice adaptée d'Ebel.

Une fois le consensus dégagé, la succession d'opérations mathématiques proposée par

Ebel a été appliquée afin d'établir, pour chaque panel, le standard de performance intérimaire de l'épreuve. Cette deuxième phase terminait l'expérimentation de la méthode alternative. Les résultats issus des deux phases d'expérimentation ont permis, lors de l'étape finale, d'établir les standards de performance définitifs.

3.4.3.3 L'étape finale

L'étape finale de la méthode alternative consistait à traiter et à analyser les données de l'ensemble de l'expérimentation afin d'établir un standard de performance pour chacun des éléments de compétence et un standard de performance pour l'ensemble de l'épreuve clinique. L'établissement des standards de performance (ESP) exige, comme le suggère la méthode d'Ebel (1972), d'effectuer une série d'opérations mathématiques sur l'ensemble des notations de tous les panels. Il est important de rappeler que les standards de performance sont établis à partir de la catégorisation définitive des items obtenue par consensus lors de la deuxième phase d'expérimentation et des pondérations moyennes des 16 catégories de performance obtenues lors de la dernière ronde de notation de l'ensemble des panels.

Finalement, la matrice d'Ebel obtenue à la fin de la quatrième ronde a permis de calculer le standard de performance pour l'ensemble de l'épreuve. Il a ensuite été possible, sans avoir encore recours aux panélistes, de produire une matrice d'Ebel pour chacun des 23 éléments de compétence essentiels et d'établir, à l'aide des calculs proposés par Ebel, 23 autres standards de performance servant à baliser l'exercice de la profession. Chacune de ces matrices a préalablement été constituée des items associés à ces éléments de compétence. À titre de rappel, pour réussir, un candidat se devra d'atteindre à la fois le standard de performance pour l'ensemble de l'épreuve et pour chacun des 23 éléments de compétence.

En conclusion, nous évoquons ici que pour établir les standards de performance, nous

avons adapté la procédure d'Ebel et proposé une matrice qui met en lien a) le nombre d'items (n) classé dans chacune des 16 catégories de performance et b) le pourcentage d'items (p) moyen issu des consensus de l'ensemble des panels. Ce sont les opérations mathématiques réalisées sur le contenu de la matrice qui, au final, ont déterminé les standards de performance définitifs. À titre de rappel, un tableau synthèse de la démarche d'établissement des standards a été présenté au tableau 3.1 à la page 107.

3.5 Considérations éthiques et transmissions des résultats aux participants

Dans le cadre de notre projet de recherche, il appert, selon les membres du comité éthique consultés, qu'aucun certificat éthique n'est requis. Cette réalité s'explique par le fait que les deux échantillons étaient issus d'une population spécifique restreinte aux membres de l'OAQ, qui chapeaute le projet, et que la sélection des participants, en conformité avec les caractéristiques retenues, a été faite par l'entremise et à la demande de ses dirigeants. Pour l'OAQ, il s'agit donc d'une activité de nature administrative qui est supervisée par la chercheuse. L'absence de certificat éthique s'explique également par le fait que les discussions n'ont pas été enregistrées et que les tâches demandées n'exigeaient pas des panélistes qu'ils aient à justifier leurs jugements. Le comité pour l'évaluation des projets étudiants impliquant de la recherche avec des êtres humains (CERPE) de l'UQAM a ainsi autorisé que cette recherche soit réalisée au sein même du bureau de l'OAQ par une approche consensuelle non enregistrée. Un formulaire de consentement a tout de même été suggéré pour appuyer la démarche.

Dans cette optique, l'OAQ a acheminé une invitation à participer à la recherche à l'ensemble des membres qui correspondaient aux caractéristiques de nos deux échantillonnages. À cet effet, une invitation a été rédigée en s'assurant de préciser l'objectif de la recherche, les tâches que les panélistes seraient amenées à accomplir et les modalités de participation. Il est important de préciser que l'OAQ a compensé financièrement certaines des dépenses encourues par les panélistes, notamment les frais

de transport et les frais de repas du midi. Le temps consenti a également été comptabilisé dans la banque d'heures de formation continue obligatoire et un certificat a été remis à chacun des panélistes.

La plateforme Doodle a été utilisée par l'OAQ afin de recueillir le consentement libre et volontaire des membres souhaitant participer à la recherche ainsi que leurs disponibilités respectives. La sélection des participants a finalement été effectuée en fonction du nombre de membres disponibles aux dates retenues.

Lors des rencontres, nous avons demandé à chacun des panélistes de signer un formulaire d'information et de consentement (participant majeur) les informant que leur participation à la recherche était libre et volontaire et qu'ils pouvaient se retirer en tout temps sans aucune formalité et sans aucun préjudice. Ce formulaire stipulait que leur participation serait traitée en toute confidentialité et qu'eux-mêmes s'engageaient à ne pas diffuser les informations concernant l'identité des panélistes et les résultats obtenus à la suite de leur participation. Une option a permis aux panélistes qui le souhaitaient de remédier à leur anonymat advenant une éventuelle publication ou diffusion des résultats de l'étude. Ce formulaire se devait d'être suffisamment compréhensible et vulgarisé pour permettre à chacun des participants de décider s'il désirait participer ou non au projet de recherche.

Ce document incluait, conformément aux normes établies par l'UQAM :

- ❷ le titre du projet de recherche;
- ❷ les informations nominatives de la chercheuse responsable du projet et du directeur de recherche dirigeant cette dernière;
- ❷ le titre et le but général du projet;
- ❷ la procédure ou les tâches demandées au participant (panéliste);

- ☛ les avantages et les risques associés à sa participation;
- ☛ les informations relatives à l'anonymat et à la confidentialité assurée et à la participation volontaire;
- ☛ la compensation financière accordée;
- ☛ les coordonnées de la chercheuse pour des questions relatives au projet de recherche ou sur les droits du participant;
- ☛ les tenants et aboutissants liés à la signature du formulaire de consentement.

Afin d'assurer l'anonymat lors de la collecte des données, un code de couleur a été attribué à chaque panéliste. Seule l'équipe de recherche a eu accès aux outils de planification contenant les informations nominatives. Les données normatives seront conservées sous clé au domicile de la chercheuse ou au sein de l'OAQ pour une durée d'au moins cinq années suivant la complétude du mémoire de maîtrise.

Afin d'assurer la confidentialité des notations individuelles, la collecte des données et la transmission des résultats aux participants ont été effectuées par l'entremise de clés USB. Durant les périodes de pauses, les données ont été recueillies individuellement puis dupliquées dans le chiffrier d'analyse statistique approprié. Les résultats ont, à divers moments, été dupliqués dans le chiffrier électronique de chacun des panélistes et chaque clé USB a été redistribuée au bon panéliste. Au final, l'ensemble des clés USB ont été récupérées et la transmission des résultats finaux d'ESP à l'organisme réglementé s'est effectuée en main propre. Lorsque le mémoire de maîtrise sera approuvé par le jury et que les corrections mineures seront approuvées par le Directeur de recherche, une version PDF sera transmise par courriel à l'Ordre des acupuncteurs du Québec. Les participants seront informés postérieurement des résultats par l'entremise d'un communiqué émis par l'Ordre.

4. RÉSULTATS

Ce chapitre présente les résultats quantitatifs obtenus lors de l'expérimentation de la méthode alternative d'ESP. À certains moments, des discussions sont amorcées afin de poser les fondements de l'interprétation statistique et des conclusions qui suivront dans le prochain chapitre.

Ce chapitre expose d'abord le processus de validation des 149 items de l'épreuve, puis la répartition des panélistes selon les phases d'expérimentation suivie des résultats obtenus lors de la première et de la deuxième phase. À la suite de quoi les résultats de l'étape finale de l'expérimentation, soit le standard de performance de l'épreuve et les standards de performance des 23 tâches la constituant, sont présentés.

4.1 Validation des items

La validation des 149 items I-SMART a préalablement été effectuée par un comité constitué de 4 experts exerçant une position significative dans le cadre du processus de reconnaissance des compétences des acupuncteurs exerçant au Québec. Tous se sont avérés unanimement « en accord » avec la formulation et la pertinence des 23 tâches qui ont été utilisés pour établir les 23 standards de performance, et « fortement en accord » avec la clarté de leur formulation I-SMART ainsi qu'avec les 3 mises en situation professionnelles qui ont été associées aux observables de chacune des tâches.

4.2 Répartition des panélistes

Les 20 panélistes ayant participé à l'expérimentation de la méthode alternative d'ESP ont été répartis, en fonction de leur expérience et de leur disponibilité, en 4 panels (panel A, B, C et D) et en 2 phases. En 2014, 10 novices (panel A et B) et 3 experts (panel C) ont participé à la première phase d'expérimentation. En 2015, sept experts (panel D) ont participé à la deuxième phase. La répartition des novices s'est avérée, comme souhaité, quasi équivalente (panel A, $n=6$; panel B, $n=4$).

4.3 Résultats obtenus lors de la première phase d'expérimentation

La présentation des résultats nous amène a priori à nous remémorer que l'objectif poursuivi lors des deux premières rondes de notation de la première phase d'expérimentation de la méthode alternative d'ESP consistait à alimenter la matrice permettant d'établir les standards de performance conformément à l'algorithme proposé par Ebel et aux deux échelles de mesure présentées. Ensuite, nous souhaitons rappeler que la troisième ronde avait pour objectif de produire la liste ordonnée des items selon la mise en rang moyenne, et à déterminer les zones d'incertitude utiles à la deuxième phase d'expérimentation. Au final, rappelons que la quatrième ronde avait pour objectif d'établir les standards de performance préliminaires consensuels utiles à l'établissement des standards définitifs. Les résultats du processus itératif d'ESP mis en œuvre dans le cadre de la recherche sont présentés dans le tableau 3.1, au chapitre précédent.

4.3.1 Résultats des rondes 1 et 2 de la phase I

Les résultats obtenus aux tâches récurrentes de catégorisation et d'assignation des deux premières rondes de notation individuelles de la première phase sont présentés aux pages suivantes, dans une série de matrices d'Ebel (tableau 4.1). Ces matrices tracent

le portrait des jugements individuels émis par les 10 panélistes novices (panels A et B) et les 3 panélistes experts (panel C) ayant participé à l'expérimentation. Rappelons qu'à chacun des croisements des différentes cotes attribuées aux deux échelles de mesure, le nombre d'items catégorisés (n_i) et le pourcentage de réussite assigné (p_i) par chaque panéliste y est spécifié pour chacune des deux rondes de notation. Chaque matrice indique leurs notations respectives et le seuil de réussite préliminaire de l'épreuve calculé conformément à l'algorithme d'EBEL ($\sum(n_i \times p_i)/N$).

Tableau 4.1 – Matrices d'Ebel des rondes de notation individuelles de la première phase

;Panélistes 1

Niveau de Difficulté		1		2		3		Total n	Total p
Degré d'Efficiency		n	p	n	p	n	p		
1		104	100	42	84	2	70	148	95
1		45	100	1	90			46	100
2		59	100	31	85			90	95
3				10	80	2	70	12	78
2		87	89	57	74	5	65	149	83
1		41	100	2	80			43	99
2		45	80	37	75			82	78
3		1	75	18	70	5	65	24	69
Total général		191	95	99	78	7	66	297	89

;Panélistes 2

Niveau de Difficulté										Total n	Total p
Degré d'Effice	1	2		3		4					
	n	p	n	p	n	p	n	p			
1	54	95	49	87	40	79	5	73	148	87	
1	30	100	7	95	5	90			42	98	
2	9	95	26	90	11	85			46	90	
3	14	85	16	80	21	75	4	75	55	79	
4	1	75			3	65	1	65	5	67	
2	71	98	71	86	6	75			148	92	
1	46	100							46	100	
2	23	95	46	90					69	92	
3	2	85	25	80	6	75			33	79	
Total général	125	97	120	87	46	78	5	73	296	89	

;Panélistes 3

		Niveau de Difficulté 								Total n	Total p
		1		2		3		4			
Degré d'Efficience 		n	p	n	p	n	p	n	p		
	1	84	93	51	74	12	64	2	60	149	84
	1	67	95	22	80	1	75			90	91
	2	17	85	27	70	6	65			50	75
	3			2	65	5	60	2	60	9	61
	2	78	86	60	74	11	62			149	80
	1	50	90	2	85					52	90
	2	28	80	50	75	2	70			80	77
	3			8	65	9	60			17	62
Total général		162	90	111	74	23	63	2	60	298	82

;Panélistes 4

Niveau de Difficulté 										Total n	Total p
		1		2		3		4			
Degré d'Effcience 	n	p	n	p	n	p	n	p			
 1	73	98	34	59	26	33	13	10	146	70	
1	65	100	7	80	3	60			75	97	
2	8	80	19	60	6	40			33	61	
3			8	40	11	30	6	15	25	30	
4					6	20	7	5	13	12	
 2	70	88	59	47	20	25			149	63	
1	58	90	1	60					59	89	
2	12	80	37	50	1	30			50	57	
3			21	40	17	25			38	33	
4					2	20			2	20	
Total général	143	93	93	51	46	30	13	10	295	66	

;Panélistes 5

		Niveau de Difficulté									
		1		2		3		4		Total n	Total p
Degré d'Effcience		n	p	n	p	n	p	n	p		
1		65	100	48	80	25	76	9	54	147	87
1		19	100					1	50	20	98
2		29	100	3	85	3	65	1	60	36	95
3		17	100	45	80	19	80	3	60	84	83
4						3	65	4	50	7	56
2		69	94	65	64	14	86			148	80
1		33	100	1	70					34	99
2		32	90	47	65	5	80			84	75
3		4	80	17	60	9	90			30	72
Total général		134	97	113	71	39	80	9	54	295	83

;Panélistes 6

		Niveau de Difficulté									
		1		2		3		4		Total n	Total p
Degré d'Effcience		n	p	n	p	n	p	n	p		
1		97	99	30	84	22	75			149	92
1		38	100	5	90	8	80			51	96
2		42	100	11	85	4	75			57	95
3		17	95	14	80	10	70			41	84
2		80	100	53	84	14	71	1	75	148	92
1		42	100	11	90			1	75	54	98
2		37	100	25	85	4	75			66	93
3		1	95	17	80	10	70			28	77
Total général		177	99	83	84	36	73	1	75	297	92

;Panélistes 7

		Niveau de Difficulté									
		1		2		3		4		Total n	Total p
Degré d'Effcience		n	p	n	p	n	p	n	p		
1		47	88	70	66	24	29	8	11	149	64
1		3	100							3	100
2		38	90	28	80					66	86
3		6	70	39	60	15	40	2	30	62	55
4				3	20	9	10	6	5	18	10
2		48	93	76	70	18	39	6	17	148	72
1		16	100	1	90					17	99
2		31	90	43	80					74	84
3		1	70	28	60	11	45	2	30	42	55
4				4	35	7	30	4	10	15	26
Total général		95	91	146	68	42	33	14	14	297	68

;Panélistes 8

		Niveau de Difficulté									
		1		2		3		4		Total n	Total p
Degré d'Effcience		n	p	n	p	n	p	n	p		
1		53	96	74	83	19	70	3	60	149	85
1		29	100	9	90					38	98
2		16	95	24	85	3	75			43	88
3		8	85	41	80	14	70	3	60	66	78
4						2	60			2	60
2		37	90	80	49	24	26	8	19	149	53
1		24	95	9	60					33	85
2		13	80	50	50	2	35			65	56
3				21	40	21	25	7	20	49	31
4						1	20	1	15	2	18
Total général		90	94	154	65	43	45	11	30	298	69

;Panélistes 9

		Niveau de Difficulté									
		1		2		3		4		Total n	Total p
Degré d'Effcience		n	p	n	p	n	p	n	p		
1		64	79	58	52	20	18	6	10	148	57
1		27	95							27	95
2		26	70	18	60					44	66
3		11	60	38	50	12	20			61	46
4				2	20	8	15	6	10	16	14
(vide)											
2		44	83	78	46	26	19	1	15	149	52
1		22	95	2	60					24	92
2		22	70	39	50					61	57
3				37	40	21	20	1	15	59	32
4						5	15			5	15
Total général		108	80	136	48	46	19	7	11	297	54

;Panélistes 10

		Niveau de Difficulté									
		1		2		3		4		Total n	Total p
Degré d'Effcience		n	p	n	p	n	p	n	p		
1		46	100	65	80	29	61	9	10	149	78
1		42	100	9	85					51	97
2		4	95	46	80	9	65			59	79
3				10	75	18	60			28	65
4						2	50	9	10	11	17
2		43	92	71	54	25	35	10	14	149	59
1		31	95	2	70					33	93
2		11	85	45	60	1	40			57	64
3		1	75	24	40	23	35	4	20	52	37
4						1	25	6	10	7	12
Total général		89	96	136	66	54	49	19	12	298	68

;Panélistes 11

		Niveau de Difficulté									
		1		2		3		4		Total n	Total p
Degré d'Efficiency		n	p	n	p	n	p	n	p		
1		40	96	32	87	60	80	17	74	149	85
1		19	100	5	100	7	100			31	100
2		17	95	12	90	20	80	12	75	61	85
3		4	85	15	80	33	75	5	70	57	77
2		43	95	66	89	36	83	4	78	149	89
1		29	95	7	95					36	95
2		14	95	44	90	21	85	2	80	81	89
3				15	85	15	80	2	75	32	82
Total général		83	96	98	89	96	81	21	74	298	87

;Panélistes 12

		Niveau de Difficulté									
		1		2		3		4		Total n	Total p
Degré d'Efficiency		n	p	n	p	n	p	n	p		
1		60	100	52	85	33	74	4	30	149	87
1		41	100	1	90					42	100
2		14	100	26	85	4	75			44	89
3		5	95	20	85	25	75	2	40	52	79
4				5	80	4	70	2	20	11	65
2		54	98	55	85	37	75	3	40	149	86
1		34	100	2	90					36	99
2		19	95	35	85	7	75			61	87
3		1	95	17	85	29	75	3	40	50	77
4				1	80	1	70			2	75
Total général		114	99	107	85	70	75	7	34	298	87

;Panélistes 13

		Niveau de Difficulté							
		1		2		3		Total n	Total p
Degré d'Efficiency		n	p	n	p	n	p		
1		72	96	65	82	12	76	149	88
1		40	100					40	100
2		32	90	52	85	2	80	86	87
3				13	70	10	75	23	72
2		50	94	67	84	32	78	149	86
1		37	95	5	90			42	94
2		13	90	48	85	11	85	72	86
3				14	80	21	75	35	77
Total général		122	95	132	83	44	78	298	87

L'analyse des résultats nous permet de constater que les items couvrent une variété de niveaux de difficulté. La grande majorité des items étant perçus, par les 13 panélistes, comme facile (cote 1) ou relativement facile (cote 2 : moyen) à réussir par un candidat minimalement compétent formé à l'étranger, comparativement à une minorité d'items perçus comme difficiles (cote 3) et à une très faible minorité d'items perçus comme très difficiles (cote 4). Globalement, les pondérations associées aux degrés d'efficacité suivent une décroissance. Néanmoins, lors des deux premières rondes de notation des items, nous remarquons que le panéliste 05 a assigné quelques pondérations désordonnées aux croisements des niveaux de difficulté et des degrés d'efficacité supérieurs (à la jonction des cotes 3 et 4). Il semble être le seul à avoir éprouvé des difficultés à assigner des pourcentages de réussite (p_i) cohérents pour les niveaux de

difficulté supérieurs (difficile et très difficile). À ce stade, il nous apparaît que cette erreur d'assignation s'avère mineure puisqu'une ronde consensuelle succède.

Par la suite, lorsque nous analysons les seuils de réussite (standards de performance) préliminaires établis, à la suite de l'application de l'algorithme de calcul d'Ebel à l'ensemble des notations individuelles, nous remarquons un très faible taux d'omission de notation d'items (0,4 % à la première ronde et 0,2 % à la deuxième ronde) et une variation modérée des seuils individuels et collectifs. Nous remarquons qu'à la ronde de familiarisation à la notation des items, le panéliste 01 a omis d'émettre un jugement sur l'un des items de l'épreuve clinique ($P^{01} i_{134}$) et quatre autres en ont omis quelques-uns ($P^{02} i_{44}$; $P^{04} i_{47}, i_{99}, i_{147}$; $P^{05} i_{30}, i_{109}$; $P^{09} i_9$). Les résultats de cette première ronde indiquent que le panéliste 01 (P^{01}) a classifié 104 items comme étant faciles à réussir, 42 comme étant moyennement facile et 2 comme plutôt difficile à réussir, auxquels il a postérieurement assigné des pondérations d'efficience (de 100 %, 90 %, 85 %, 80 % et 70 %) dont le seuil de réussite est de 95 % (141/148 items). À la deuxième ronde de notation, ce panéliste (P^{01}) n'a omis aucune notation d'item et a ajusté le seuil de réussite à 83 %, signifiant ainsi qu'il s'attend à ce qu'un candidat minimalement compétent formé à l'étranger soit apte à réussir 123 items sur 149 items, dont 43 à une très forte efficience (cote 1), 82 à une forte efficience et 24 à une efficience modérée (43 items à 99 % de réussite; 82 items à 78 % de réussite; 24 items à 69 % de réussite). Nous constatons parallèlement qu'à cette deuxième ronde, trois panélistes (P^{02}, P^{05} et P^{06}) ont synchroniquement omis d'émettre un jugement sur l'item 44 (i_{44}) et qu'un autre (P^{07}), alors qu'il n'en avait omis aucun lors de la première ronde, en a omis un distinct (i_{55}). Nous remarquons, de façon concomitante, qu'un autre panéliste (P^{02}) a continûment omis d'émettre un jugement sur l'item 44, qui se définit comme suit : « explique, en tout temps, dans une terminologie claire et accessible, les divers éléments liés aux problèmes de santé complexes ». Quoique notre étude ne nous permette pas de vérifier les causes des omissions, cette double déficience

pourrait s'expliquer par un mode de raisonnement consistant à omettre d'emblée d'émettre un jugement plutôt que de fournir une réponse hasardeuse. Nous remarquons néanmoins qu'en calculant la moyenne pondérée des seuils de réussite des différents panels, ces omissions ont peu d'incidence sur les résultats de l'épreuve clinique. Globalement, nous observons une baisse du seuil moyen chez les deux premiers panels ($p_{seuil\ gr.A}$ 86 % vs 81 %; $p_{seuil\ gr.B}$ 71 % vs 59 %) et une stabilité chez le troisième panel ($p_{seuil\ gr.C}$ 87 %). Bref, une majorité des panels obtient un seuil moyen supérieur à 80 % ($p_{seuil\ gr.A}$ 84 %; $p_{seuil\ gr.C}$ 87 %) alors qu'une minorité obtient un seuil moyen inférieur ($p_{seuil\ gr.B}$ 65 %).

En observant de plus près les scores bruts totaux ($\sum(n_i \times p_i)$) préalables à l'établissement des standards, nous remarquons qu'ils offrent des indications supplémentaires relatives à la prédisposition des panélistes à adopter certains comportements, dont une tendance marquée à porter des jugements d'intensité sévère, modérée ou clément. La figure 4.1 illustre, à la page suivante, l'évolution de la distribution des scores bruts totaux issus des patrons de réponse individuels des panélistes au sein de leur panel respectif et de l'ensemble des panels.

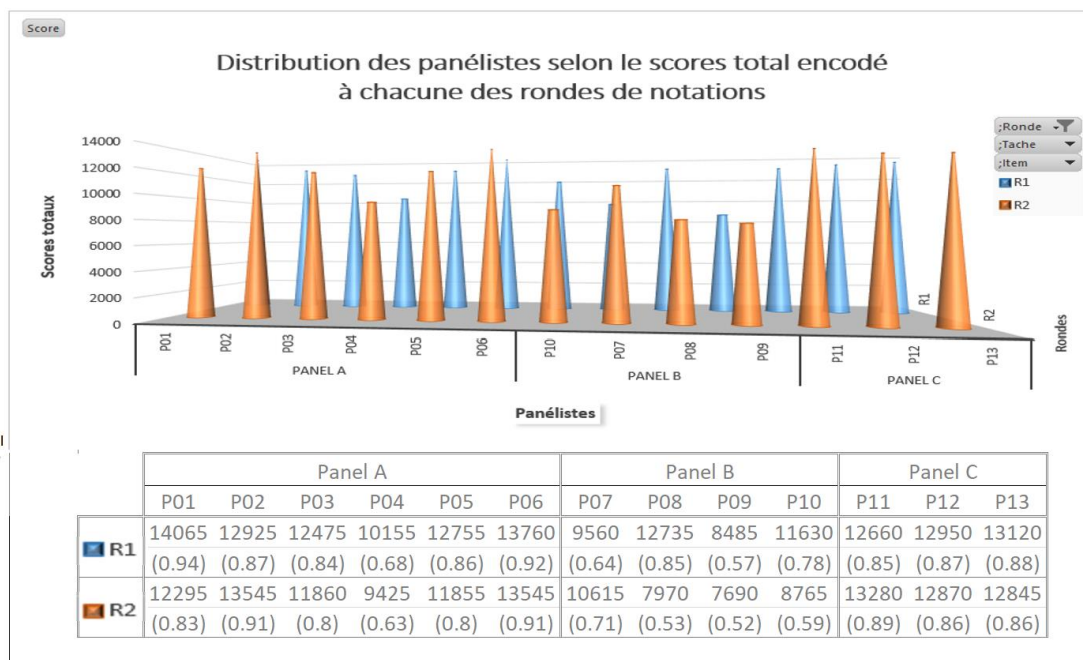


Figure 4.1 – Scores bruts totaux issus des patrons de réponse individuels de la phase I

En observant les variations des scores bruts totaux résultant des différents patrons de réponse individuels des panélistes, nous constatons que les scores moyens de l'épreuve clinique fluctuent entre 8 485 (57%) et 14 065 (94%) à la première ronde et entre 7 690 (52%) et 13 545 (91%) à la deuxième ronde, et que le score total moyen s'atténue d'une ronde à l'autre, passant de 12 098 à 11 274 (équivalent à un pourcentage de réussite de 81 % et de 76 %). Nous remarquons a priori que les scores bruts totaux des panélistes P⁰⁶ et P¹² demeurent relativement stables d'une ronde à l'autre, que ceux de la majorité des panélistes suivent une tendance décroissante et que ceux des panélistes P⁰², P¹¹ et P⁰⁷ suivent une tendance inverse, donc croissante. À la première ronde, une grande majorité des panélistes, soit plus des trois quarts d'entre eux (10 sur 13), affiche un score brut total supérieur à 11 600, donc équivalent à un seuil de 78 %. À la deuxième ronde de notation, nous remarquons que deux d'entre eux (P⁰⁸ et P¹⁰) adoptent un comportement contraire qui abaisse leur score brut total sous la valeur de 9 000 (60 %).

Nous remarquons corolairement d'importantes variations à la baisse au sein du panel B, constitué des panélistes P⁰⁷, P⁰⁸, P⁰⁹ et P¹⁰, comparativement à de faibles variations à la baisse au sein du panel A, constitué des six premiers panélistes. Nous observons que le panéliste P⁰⁷ est le seul à avoir rehaussé son score brut total au-dessus de 9 000, celui-ci passant de 9 560 à 10 615, ce qui représente la plus élevée des variations croissantes de l'ensemble des panels. Soulignons que des variations inverses, mais légèrement moins marquées, sont observées auprès du panéliste P⁰⁴, pour qui le score brut total passe de 10 155 à 9 425. Au sein du panel C, des variations mineures variées sont observées. Globalement, le score brut total le moins élevé de tous les panels est celui du panéliste P⁰⁹ (panel B), suivi par celui du panéliste P⁰⁴ (panel A), ce qui révèle les patrons de réponse les moins exigeants de tous les panélistes et qui illustre une tendance marquée à la clémence. Soulignons que le score brut total du panéliste P⁰⁴ est également le moins élevé de son propre panel et que le score brut total du panéliste P⁰⁷ est le moins élevé du panel B. Toutes rondes confondues, les variations les moins élevées observées par rapport au score total moyen sont celles des panélistes P⁰³ et P⁰⁵, qui démontrent ainsi une tendance marquée à la modération et une constance dans leurs jugements.

À la première ronde, nous observons que les scores totaux les plus exigeants qui illustrent une tendance à juger avec une certaine sévérité – se révèlent être ceux des panélistes P⁰¹ (14 065), P⁰⁶ (13 760) et P¹³ (13 120), alors qu'à la deuxième ronde, ce sont ceux des panélistes P⁰⁶ (13 545), P⁰² (13 545) et P¹¹ (13 280). Toutes rondes confondues, les scores moyens de l'épreuve les plus exigeants de l'ensemble des panels sont ceux des panélistes P⁰⁶ (13 653), P⁰² (13 235) et P⁰¹ (13 180), suivis par ceux des panélistes P¹³ (12 983), P¹¹ (12 970) et P¹² (12 910).

Lorsque plusieurs experts réalisent l'exercice d'établissement de standards au moyen de la méthode alternative expérimentée, une moyenne des estimations est calculée.

Celle-ci permet de déterminer le seuil de réussite de l'épreuve. Rappelons ici que la somme des produits des pondérations multipliées par le nombre d'items de chaque ligne de la table de données d'un panéliste correspond au score total escompté à l'épreuve (nombre total d'items à réussir) et que la division de ce score par le nombre total d'items de l'épreuve (149) correspond au standard de performance préliminaire (seuil de réussite) de l'épreuve, qui est défini individuellement à chacune des rondes de notation. Le standard de performance moyen le plus exigeant se calcule comme suit : $13\,653/149 = 91,6\% (P^{06})$. L'ensemble des résultats de l'algorithme et les détails des calculs des standards de performance préliminaires de l'épreuve clinique, établis individuellement, sont présentés dans le tableau 4.2. Il est à noter que les pourcentages observés au tableau 4.2 sont les mêmes que ceux présentés au tableau 4.1.

Tableau 4.2 – Standards préliminaires individuels et collectifs de la phase I et leurs calculs

Standards préliminaires Seuils (en %)				Algorithme d'EBEL (Nb. d'items à réussir / Nb.items total)			
Panélistes	Rondes R1	R2	Moyenne	R1	R2	Moyenne	
Panel A	86	81	84				
P01	▲ 95	▲ 83	▲ 89	141 / 148 items	123 / 149 items	132	/ 148.5 items
P02	▲ 87	▲ 92	▲ 89	129 / 148	135 / 148	132	/ 148
P03	▲ 84	▲ 80	▲ 82	125 / 149	119 / 149	122	/ 149
P04	▬ 70	▼ 63	▼ 66	102 / 146	94 / 149	98	/ 147.5
P05	▲ 87	▲ 80	▲ 83	128 / 147	119 / 148	123	/ 147.5
P06	▲ 92	▲ 92	▲ 92	138 / 149	135 / 148	137	/ 148.5
Panel B	71	59	65				
P07	▼ 64	▬ 72	▼ 68	96 / 149 items	106 / 148 items	101	/ 148.5 items
P08	▲ 85	▼ 53	▼ 69	127 / 149	80 / 149	104	/ 149
P09	▼ 57	▼ 52	▼ 55	86 / 149	77 / 149	81	/ 149
P10	▬ 78	▼ 59	▼ 68	116 / 149	88 / 149	102	/ 149
Panel C	87	87	87				
P11	▲ 85	▲ 89	▲ 87	127 / 149 items	133 / 149 items	130	/ 149 items
P12	▲ 87	▲ 86	▲ 87	130 / 149	129 / 149	129	/ 149
P13	▲ 88	▲ 86	▲ 87	131 / 149	128 / 149	130	/ 149
Seuil moyen global							
Panels A, B, C	▲ 81	▬ 76	▬ 79	121 / 149 items	113 / 149 items	117	/ 149 items
Panels A, B	▬ 79	▬ 70	▬ 74	117 / 149	105 / 149	111	/ 149
Panels A, C	▲ 86	▲ 84	▲ 85	128 / 149	126 / 149	127	/ 149
Panels B, C	▬ 79	▬ 73	▬ 76	118 / 149	109 / 149	113	/ 149

Afin de faire ressortir les contrastes, les seuils les plus élevés de chacun des panels sont identifiés par un triangle vert et les moins élevés, par un triangle rouge. Globalement, les seuils de réussite de l'épreuve (standards de performance), établis individuellement, fluctuent en moyenne entre 55 % et 92 %. Le seuil moyen le plus élevé observé se révèle être celui du panel C, avec 87 % de réussite attendue, alors que le seuil le moins élevé se révèle être celui du panel B, avec 65 %. Toutes rondes confondues, les seuils moyens des novices (panels A et B combinés, 74 %) se révèlent moins exigeants que ceux établis par les experts (panel C, 87 %). Nous observons que les seuils moyens de l'épreuve de l'ensemble des panels (panels A, B et C) se révèlent corolairement atténués par les seuils inférieurs du panel B. Nous constatons que le seuil moyen du panéliste P⁰⁶ (panel A) se révèle le plus exigeant des panels et que celui du panéliste P⁰⁹ (panel B) se révèle le moins exigeant.

En complément des seuils préliminaires établis lors des deux premières rondes, le tableau 4.3 expose les standards préliminaires des différentes tâches constituant l'épreuve clinique. Rappelons ici que le score total de chacune des tâches divisées par le nombre total d'items associés à cette tâche correspond au standard de performance préliminaire (seuil de réussite) de ladite tâche. Notons ici que le nombre d'items associés à chacune des tâches est précisé dans la partie supérieure du tableau et que la même interprétation des formes géométriques s'applique.

Tableau 4.3 – Standards préliminaires collectifs de la phase I, propres aux 23 tâches

	Tâches																								Total
	T01	T02	T03	T04	T05	T06	T07	T08	T09	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	général	
Nombre d'items	10	19	8	11	13	3	5	8	12	2	3	8	3	1	2	5	5	8	5	6	4	1	7	149	

Seuil moyen	Rondes (panels)																								
	R1																								
	Panel A	Panel B	Panel C	Total R1	Panel A	Panel B	Panel C	Total R2	Total général																
Tâches (Items)																									
T22	97,50	96,25	93,33	96,15	96,67	96,25	91,67	95,38	95,77																
T20	97,86	90,21	99,72	95,91	96,11	89,38	96,67	94,17	95,03																
T21	95,00	94,69	98,33	95,67	91,46	94,06	96,67	93,46	94,57																
T14	95,00	96,25	91,67	94,62	96,67	96,25	86,67	94,23	94,42																
T16	98,97	86,75	94,67	94,14	96,17	79,75	94,33	90,69	92,40																
T15	99,17	82,50	94,17	92,88	96,25	79,38	91,67	90,00	91,44																
T17	94,67	84,25	97,67	92,15	88,67	77,00	94,67	86,46	89,31																
T05	89,94	80,19	90,64	87,10	86,67	73,24	90,26	83,42	85,27																
T18	91,98	82,66	88,75	88,37	84,79	70,63	90,21	81,68	85,02																
T13	90,28	79,58	90,56	87,05	86,94	69,58	89,44	82,18	84,62																
T07	86,67	80,25	92,00	85,92	84,33	68,25	90,67	80,85	83,38																
T12	87,66	77,97	88,13	84,76	84,38	62,97	85,21	77,98	81,35																
T10	80,00	83,75	91,67	83,85	75,83	68,75	90,83	77,12	80,48																
T19	82,33	74,25	99,33	83,77	74,00	63,50	96,67	76,00	79,88																
T11	89,72	74,17	80,00	82,69	85,00	55,83	85,56	76,15	79,42																
T23	85,61	70,36	82,14	80,06	75,36	54,64	83,81	70,93	75,47																
T04	84,61	67,27	78,94	77,87	81,27	53,18	82,88	72,82	75,36																
T06	79,44	70,83	95,56	80,51	71,39	50,42	90,56	69,36	74,94																
T03	81,17	61,72	81,67	75,24	76,98	43,28	86,04	68,70	71,96																
T09	78,47	62,60	80,69	74,10	73,54	46,15	83,33	67,37	70,74																
T01	79,17	54,88	84,33	72,88	77,50	41,13	85,83	68,23	70,56																
T02	81,71	57,43	78,95	73,60	77,15	42,57	80,70	67,33	70,47																
T08	76,04	53,59	80,42	70,14	70,21	37,34	81,88	62,79	66,47																
Total général	85,81	71,28	86,64	81,52	81,40	58,89	87,24	75,82	78,67																

Nous observons que le standard de performance préliminaire établi pour la tâche 22, lors de la première ronde (R1), est de 97,5 % pour le panel A, de 96,25 % pour le panel B et de 93,33 % pour le panel C, alors qu'il est, à la deuxième ronde, respectivement établi à 96,67 %, 96,25 % et 91,67 %. Les tâches sont ordonnées du seuil moyen global le plus élevé au moins élevé, passant de la tâche T22 (95,77 %) à la tâche T08 (66,47 %). Les seuils moyens préliminaires les plus exigeants circonscrivent ici les tâches T14, T15, T16, T17 T20, T21 et T22 alors que les seuils les moins exigeants sont ceux des tâches T1, T2, T3, T8 et T9. Pour les autres tâches (rectangles jaunes), nous observons une séquence interrompue de seuils qui révèle une variabilité entre les seuils des différents panels. Rappelons qu'au terme de la deuxième ronde, la moyenne

des catégorisations servira à produire une liste ordonnée des items utiles à la troisième ronde de notation.

Afin de vérifier la constance et la précision des jugements émis par les panélistes d'un même panel, le coefficient de corrélation linéaire a été appliqué aux deux échelles de mesure (niveau de difficulté, degré d'efficacité) pour juger de l'accord interjuges. Ce choix s'inscrit dans le respect du neuvième principe d'ÉSP présenté au cadre théorique. La figure 4.2 présente ce niveau d'accord pour chacune des échelles de mesure et l'interprétation selon l'échelle proposée par Guilford en 1965.

R1D					
Coefficient de corrélation	Coefficient de corrélation	A n=6	B n=4	C n=3	Intensité de la corrélation
$ R =1$	1	0.3	0.4	0.5	Parfaite
$0,9 \leq R < 1$	0.9				Très forte
$0,7 \leq R < 0,9$	0.7				Forte
$0,4 \leq R < 0,7$	0.4	0.4	0.6	0.5	Modérée
$0,2 \leq R < 0,4$	0.2	0.3			Faible
$0,2 > R < 0,2$	0				Presque négligeable
Total		1	1	1	
* (Échelle de GUILFORD, J.P., 1965)					

R1E					
Coefficient de corrélation	Coefficient de corrélation	A n=6	B n=4	C n=3	Intensité de la corrélation
$ R =1$	1	0.3	0.4	0.5	Parfaite
$0,9 \leq R < 1$	0.9				Très forte
$0,7 \leq R < 0,9$	0.7				Forte
$0,4 \leq R < 0,7$	0.4		0.2	0.5	Modérée
$0,2 \leq R < 0,4$	0.2	0.0	0.1		Faible
$0,2 > R < 0,2$	0	0.7	0.3		Presque négligeable
Total		1	1	1	
* (Échelle de GUILFORD, J.P., 1965)					

R2D					
Coefficient de corrélation	Coefficient de corrélation	A n=6	B n=4	C n=3	Intensité de la corrélation
$ R =1$	1	0.3	0.4	0.5	Parfaite
$0,9 \leq R < 1$	0.9				Très forte
$0,7 \leq R < 0,9$	0.7	0.3	0.6	0.5	Forte
$0,4 \leq R < 0,7$	0.4	0.4			Modérée
$0,2 \leq R < 0,4$	0.2				Faible
$0,2 > R < 0,2$	0				Presque négligeable
Total		1	1	1	
* (Échelle de GUILFORD, J.P., 1965)					

R2E					
Coefficient de corrélation	Coefficient de corrélation	A n=6	B n=4	C n=3	Intensité de la corrélation
$ R =1$	1	0.3	0.4	0.5	Parfaite
$0,9 \leq R < 1$	0.9	0.1			Très forte
$0,7 \leq R < 0,9$	0.7	0.1	0.3	0.5	Forte
$0,4 \leq R < 0,7$	0.4	0.0			Modérée
$0,2 \leq R < 0,4$	0.2				Faible
$0,2 > R < 0,2$	-	0.4	0.3		Presque négligeable
Total		1	1	1	
* (Échelle de GUILFORD, J.P., 1965)					

Figure 4.2 – Accord interjuges des 2 échelles de mesure et leur interprétation à la phase I

La figure 4.2 présente dans les colonnes centrales les coefficients de corrélation au sein des panels A, B et C lors des deux premières rondes de notation. Ces panels sont respectivement représentés par l'échantillonnage $n=6$, $n=4$ et $n=3$. La variation des valeurs des coefficients de corrélation est présentée dans la colonne à l'extrémité

gauche et son interprétation à l'extrémité droite. La figure 4.2 est par ailleurs divisée en quatre parties de manière à distinguer, pour chacune des échelles de mesure (difficulté D, efficacité E), le degré d'accord découlant de chaque ronde (R1, R2). Lorsque nous observons l'échelle d'interprétation des coefficients de corrélation, nous remarquons une amélioration de l'intensité de l'accord interjuges chez tous les panels lors de la deuxième ronde, sauf pour quelques corrélations qualifiées de presque négligeables au regard du degré d'efficacité (en bas, à droite). Ainsi, l'intensité de l'accord interjuges de la seconde ronde de catégorisation des items par niveau de difficulté est comprise entre 0,46 et 1,0 pour le panel A (n=6, 21 paires possibles), entre 0,64 et 1,0 pour le panel B (n=4, 10 paires possibles) et entre 0,83 et 1,0 pour le panel C (n=3, 6 paires possibles). L'intensité de la corrélation fluctue entre modérée, forte et parfaite pour le panel A, alors que pour les panels B et C, elle est soit forte, soit parfaite. Quoique l'intensité de l'accord interjuges relative à la catégorisation des items par degré d'efficacité présente également des améliorations, certaines faiblesses subsistent après la deuxième ronde. L'intensité est ici comprise entre 0,12 et 1,0 pour le panel A et le panel B, et entre 0,75 et 1,0 pour le panel C. Globalement, 57 % de la catégorisation des items par degré d'efficacité pour les panélistes du panel A correspond à un degré d'accord interjuges fluctuant de modéré à parfait par rapport à 43 % pour presque négligeable. L'intensité de la corrélation interjuges du panel B est pour sa part à 40 % parfaite, 30 % forte et 30 % presque négligeable et ex aequo pour le panel C avec 50 % parfaite et 50 % forte.

4.3.2 Résultats de la ronde 3 de la phase I

La liste ordonnée des items obtenue par la moyenne de la mise en rang de l'ensemble des panels lors de la troisième ronde de notation a permis de délimiter, au moyen de la règle départ-arrêt de Ferrara, Perie et Johnson (2008), les zones d'incertitude balisant les séquences interrompues d'items. Comme évoqué dans la méthodologie, les balises des zones d'incertitude, à la base de la modélisation de l'outil de notation destiné aux

panélistes de la deuxième phase, sont délimitées par deux séries de trois jugements consécutifs univoques. La liste ordonnée des items a conséquemment été constituée à partir de la moyenne résultant de l'ensemble des notations individuelles de la deuxième ronde, abstraction faite de leur groupe d'appartenance respectif. La liste préalablement triée de l'item le plus facile au plus difficile, est présentée partiellement (tableau 4.4) afin d'illustrer exclusivement les balises des quatre zones d'incertitude déterminées à partir de la colonne « catégorie de performance », chaque zone étant circonscrite par une bordure rose. À titre de rappel, la figure 3.3 et le texte qui la précède présentent les catégories de performance et expliquent comment celles-ci sont ordonnées par niveau de difficulté et par degré d'efficience. Les degrés d'efficience sont présentés au tableau 2.8.

Tableau 4.4 – Zones d'incertitude extraites de la liste ordonnée des items de la phase I

No.Item Référence	Liste ordonnée des items (rang moyen)	Énoncé de l'item	Catégorie de performance	Points de coupure
53	1		1	Amorce au seuil minimal
[...]	[...]	[...]	[...]	[...]
54	23		1	
106	24		1	
123	25		1	
49	26	démontre, tout au long de la relation clinique, de la bienveillance et de l'empathie envers les clients.	2	
48	27	démontre, en tout temps, des attitudes et des comportements qui respectent l'individualité des clients.	2	
112	28	applique, en tout temps, une méthode respectueuse qui allie une communication des intentions et une absence d'ambiguïté.	1	
133	29	nomme les maladies à déclaration obligatoire devant systématiquement être notées au dossier.	1	
58	30	respecte, en tout temps, le droit des clients à opter pour les soins de santé de leur choix et, le cas échéant, les encourage à consulter d'autres thérapeutes.	1	
134	31	énonce les risques d'infection liés aux situations cliniques.	1	
52	32		2	
47	33		2	
41	34		2	
[...]	[...]	[...]	[...]	[...]

No. Item Référence	Liste ordonnée des items (rang moyen)	Énoncé de l'item	Catégorie de performance	Points de coupure
[...]	[...]	[...]	[...]	[...]
103	38		2	
65	39		2	
69	40		2	
111	41		6	
101	42		6	
122	43		6	
7	44		2	
107	45		2	
11	46		2	
121	47		2	
30	48		2	
140	49		2	
27	50		6	
24	51		2	
8	52		6	
127	53		5	
67	54		2	
117	55		5	
4	56		2	
87	57		1	
25	58		2	
143	59		6	
94	60		2	
39	61		6	
32	62		2	
51	63		6	
43	64		6	
92	65		6	
17	66		6	
119	67		2	
97	68		2	
34	69		6	
128	70		6	
91	71		2	
28	72		6	
130	73		6	
129	74		6	
[...]	[...]	[...]	[...]	[...]

No. Item Référence	Liste ordonnée des items (rang moyen)	Énoncé de l'item	Catégorie de performance	Points de coupure
[...]	[...]	[...]	[...]	[...]
126	105		6	
62	106		6	
70	107		6	
9	108		7	
20	109		7	
31	110		7	
29	111		7	
125	112		6	
33	113		7	
12	114		6	
61	115		7	
21	116		7	
46	117		7	
147	118		6	
18	119		6	
98	120		6	
44	121		7	
93	122		7	
36	123		7	
73	124		7	
84	125		7	
79	126		7	
71	127		7	
5	128		7	
104	129		10	
76	130		11	
96	131		7	
74	132		11	
77	133		11	
22	134		7	
80	135		11	
86	136		7	
75	137		11	
45	138		11	
89	139		11	
[...]	[...]	[...]	[...]	[...]
23	149		11	Arrêt au seuil maximal

La liste ordonnée des items révèle que seuls les trois premiers niveaux de difficulté (facile, moyen, difficile) et les trois degrés d'effcience supérieurs ont été utilisés (moyen, fort, très fort). Près de 61 % de la catégorisation des items est cohérente et logique selon la mise en rang moyenne résultant de la troisième ronde de notation, soit près des deux tiers des items. Corolairement, près du tiers (39 %) des items (58/149) sont circonscrits dans quatre zones d'incertitude. En tout, 58 items ont nécessité de nouveaux jugements lors de la deuxième phase d'expérimentation. La plus grande proportion d'items (19 %) s'inscrit dans la réussite de la tâche T02, avec 11 items sur 58. Nous observons par ailleurs que les items liés aux tâches 6 et 14 sont exclus des zones d'incertitude.

La première zone d'incertitude délimite la première séquence interrompue, là où la mise en rang des six items les range du 26^e au 31^e rang (49, 48, 112, 133, 58, 134). Nous remarquons que la catégorisation de cette première zone oscille entre les deux premières catégories de performance (01, item facile requérant un degré d'effcience très fort; 02, item facile requérant un degré d'effcience fort), alors que la deuxième zone d'incertitude circonscrit la plus grande séquence d'items, entre le 41^e et le 71^e rang, avec près de 21 % des items (31/149). Ici, la catégorisation dénote la plus grande variabilité, oscillant majoritairement entre la catégorie de performance 02 (15 items) et 06 (13 items) et minoritairement entre la catégorie de performance 01 (1 item) et 05 (2 items). La troisième zone d'incertitude, circonscrite entre le 108^e et 120^e rang, délimite pour sa part près de 9 % (13/149 items) des items dont la séquence interrompue oscille entre la catégorie 06 (5 items) et 07 (8 items). La dernière zone circonscrite contient quant à elle 5 % des items (8/149 items), avec une catégorisation oscillant entre la catégorie 07 (3 items), 10 (1 item) et 11 (4 items). Bref, la variabilité de la catégorisation des 58 items situés dans les zones d'incertitude déroge quelque peu de l'ordre des échelles fusionnées (de 01 à 16) et la liste ordonnée des items recèle 7 catégories de performance (01, 02, 05, 06, 07, 10 et 11) plutôt que 16.

Au terme de cette troisième ronde, six points de coupure ont conséquemment permis de délimiter ces sept catégories de performance. La détermination des points de coupure a été réalisée suivant la règle départ-arrêt à partir de la mise en rang des items et de leur classement dans les catégories juxtaposées. Il en a été de même pour les points de coupure délimitant les frontières des trois niveaux de difficulté pertinents (facile = 56 items, moyen = 75 items et difficile = 18 items). Nonobstant ce constat, rappelons que les points de coupure définitifs ont été déterminés lors de la deuxième phase d'expérimentation.

4.3.3 Résultats de la ronde 4 de la phase I

La quatrième et dernière ronde de notation de cette première phase d'expérimentation a finalement permis de parvenir à un consensus au sein de chaque panel sur la pondération à assigner à chacune des catégories de performance. Abstraction faite de la mise en rang des items, la dernière tâche d'assignation a permis, au moyen des pondérations consensuelles, de calculer, pour chacun des panels (A, B et C), le standard de performance préliminaire de l'épreuve au moyen de l'algorithme d'Ebel. La catégorisation moyenne issue de la deuxième ronde a été utilisée à cette fin pour chacun des panels. Le tableau 4.5 expose, pour chacun des panels, le calcul du standard préliminaire de l'épreuve résultant de la première phase d'expérimentation.

Tableau 4.5 – Calcul des standards préliminaires collectifs de la phase I pour l'épreuve

Famille de difficulté	Catégories de performance (niveau de difficulté- degré d'efficacité)	Panel A			Panel B			Panel C		
		n	p	n*p	n	p	n*p	n	p	n*p
Facile	01. Facile-Très fort	38	100%	38	23	100%	23	32	100%	32
	02. Facile-Fort	30	95%	28.5	15	90%	13.5	16	95%	15.2
	03. Facile-Moyen									
	04. Facile-Faible									
Moyen	05. Moyen-Très fort	3	90%	2.7	2	75%	1.5	3	90%	2.7
	06. Moyen-Fort	53	85%	45.05	52	70%	36.4	45	90%	40.5
	07. Moyen-Moyen	12	75%	9	24	65%	15.6	20	85%	17
	08. Moyen-Faible									
Difficile	09. Difficile-Très fort									
	10. Difficile-Fort	1	75%	0.75				9	80%	7.2
	11. Difficile-Moyen	12	65%	7.8	22	45%	9.9	23	80%	18.4
	12. Difficile-Faible				3	40%	1.2			
Très difficile	13. Très difficile-Très fort									
	14. Très difficile-Fort									
	15. Très difficile-Moyen				3	20%	0.6	1	50%	0.5
	16. Très difficile-Faible				5	10%	0.5			
Résultats de l'algorithme		149	88%	131.8	149	69%	102.2	149	90%	133.5

Chacune des catégories de performance dénombre un certain nombre d'items, qui une fois multipliés par la pondération consensuelle, résultent en un nombre d'items devant être réussis à chaque niveau de performance. Nous observons que la très grande majorité des items sont catégorisés dans les trois premiers niveaux de difficulté. Les résultats du panel A indiquent qu'un candidat formé à l'étranger doit être apte à réussir 131,8 des 149 items, soit un standard préliminaire équivalent à 88 %, alors que pour le panel B, le standard préliminaire équivaut à 69 % (102,2 des 149 items) et à 90 % (133,5 des 149 items) pour le panel C.

Rappelons ici que ce n'est qu'au terme de l'expérimentation que la liste ordonnée des items devient définitive, de même que les catégorisations. Seules la liste définitive résultant de la deuxième phase et les pondérations consensuelles résultant de la quatrième ronde de notation de chaque panel serviront, au final, à l'établissement des standards de performance définitifs.

4.4 Résultats obtenus lors de la deuxième phase d'expérimentation

Rappelons que l'objectif de la deuxième phase, tributaire du parachèvement de l'expérimentation de la méthode alternative, vise analogiquement, en ayant recours à des panélistes chevronnés (panel D), à alimenter la matrice permettant d'établir les standards de performance consensuels au moyen de l'algorithme d'Ebel. Soulignons que cette procédure, en complément des assignations, s'est limitée à la catégorisation des items situés dans les zones d'incertitude, les autres items ayant été catégorisés sur la base de leur moyenne respective. Au final, la mise en rang a permis d'atténuer les zones d'incertitude et de fixer les points de coupure définitifs délimitant les frontières des catégories de performance utilisées (difficulté-efficience).

4.4.1 Résultats des rondes 1, 2, 3 et 4 de la phase II

Lors de la deuxième phase d'expérimentation, nous observons un degré élevé d'accord quant aux catégorisations des items circonscrits dans les zones d'incertitude. Un peu plus des deux cinquièmes des catégorisations des 58 items circonscrits dans les zones d'incertitude ont conservé la catégorie de performance de la première phase, alors que près des trois cinquièmes ont été ajustées. Des deux cinquièmes, 67 % des catégorisations moyennes sont demeurées inchangées et 33 % sont demeurées pratiquement inchangées, soit à l'exception d'une (29 %) ou deux rondes (4 %). Des trois cinquièmes, 13 items indiquent une variation de plus ou moins une classe par rapport à la catégorisation de la première phase, 5 items, de plus ou moins deux classes, 5 items de plus ou moins trois classes et 11 items de plus ou moins quatre classes. Le tableau 4.6 présente, à la page suivante, les items de chacune des zones d'incertitude et la catégorisation moyenne de la première phase ainsi que les catégorisations moyennes et consensuelles résultant de la deuxième phase.

Tableau 4.6 – Catégorisations moyennes et consensuelles des items des zones d'incertitude

Zones	# item	PHASE I	PHASE II	Cat.perf. ($\bar{x}$ R2)	Cat.perf. ($\bar{x}$ R3)	Cat.perf. ($\bar{x}$ globale)	Cat.perf. (consensus)
		Cat.perf. ($\bar{x}$ R1)	Cat.perf. ($\bar{x}$ R1)				
Zone 1	48	2	1	1	1	1	1
Zone 1	49	2	1	1	1	1	1
Zone 1	58	1	2	1	1	1	1
Zone 1	112	1	1	1	1	1	1
Zone 1	133	1	1	1	1	1	1
Zone 1	134	1	1	1	1	1	1
Zone 2	4	2	2	6	2	2	6
Zone 2	7	2	2	2	2	2	2
Zone 2	8	6	6	3	3	3	3
Zone 2	11	2	2	2	2	2	2
Zone 2	17	6	6	6	6	6	6
Zone 2	24	2	2	3	2	2	2
Zone 2	25	2	2	2	2	2	2
Zone 2	27	6	6	2	2	2	2
Zone 2	30	2	2	2	3	2	3
Zone 2	32	2	2	2	6	2	2
Zone 2	34	6	6	6	6	6	6
Zone 2	39	6	6	6	2	6	3
Zone 2	43	6	6	6	6	6	6
Zone 2	51	6	6	3	6	6	3
Zone 2	67	2	2	2	2	2	2
Zone 2	87	1	6	3	3	3	3
Zone 2	91	2	2	6	6	6	6
Zone 2	92	6	6	6	6	6	6
Zone 2	94	2	2	2	5	2	6
Zone 2	97	2	2	6	6	6	6
Zone 2	101	6	6	2	2	2	2
Zone 2	107	2	2	6	2	2	2
Zone 2	111	6	6	2	2	2	2
Zone 2	117	5	6	3	3	2	3
Zone 2	119	2	2	6	6	6	6
Zone 2	121	2	2	3	2	2	2
Zone 2	122	6	6	2	2	2	2
Zone 2	127	5	6	6	3	2	3
Zone 2	128	6	6	6	6	6	6
Zone 2	140	2	2	2	2	2	2
Zone 2	143	6	6	6	2	6	6
Zone 3	9	7	7	6	6	6	6
Zone 3	12	6	6	7	7	7	7
Zone 3	18	6	6	7	7	7	7
Zone 3	20	7	7	6	6	6	6
Zone 3	21	7	7	7	7	7	7
Zone 3	29	7	7	7	6	6	7
Zone 3	31	7	7	6	6	6	6
Zone 3	33	7	7	6	7	7	6
Zone 3	46	7	7	6	7	7	7
Zone 3	61	7	7	7	7	7	7
Zone 3	98	6	6	7	7	7	7
Zone 3	125	6	7	7	7	7	7
Zone 3	147	6	6	7	7	7	7

Zones	# item	PHASE I	PHASE II		Cat.perf. ($\bar{x}$ R3)	Cat.perf. ($\bar{x}$ globale)	Cat.perf. (consensus)
		Cat.perf.	Cat.perf. ($\bar{x}$ R1)	Cat.perf. ($\bar{x}$ R2)			
Zone 4	22	7	7	10	11	11	11
Zone 4	74	11	11	11	10	10	9
Zone 4	76	11	11	11	9	10	9
Zone 4	77	11	11	10	10	10	10
Zone 4	80	11	10	11	11	11	11
Zone 4	86	7	7	11	11	11	11
Zone 4	96	7	11	10	9	10	10
Zone 4	104	10	10	7	7	7	7

Globalement, près des trois quarts des catégorisations consensuelles se sont avérées équivalentes à la catégorisation moyenne de la troisième ronde et à la moyenne globale de la deuxième phase d'expérimentation. Ainsi, au terme de la troisième ronde de notation, les zones d'incertitude ont été fortement réduites. Nous recensons au total 14 variations. De ces 14 variations, la catégorisation de 3 items (i^{29} , i^{32} , i^{143}) s'avère équivalente à la moyenne des catégorisations de la grande majorité des résultats obtenus au fil des rondes, celle de 8 items (i^4 , i^{30} , i^{33} , i^{51} , i^{76} , i^{96} , i^{117} , i^{127}) s'avère égale à la catégorisation moyenne de la deuxième ou de la troisième ronde, alors qu'elle s'avère distincte pour 3 autres items (i^{39} , i^{74} , i^{94}).

Toutes rondes confondues, les panels s'accordent tous pour conclure que les items i^{122} , i^{133} , i^{134} de la première zone sont faciles à réaliser et requièrent un très fort degré d'efficience (catégorie de performance 1) et ils adoptent la même catégorie de performance pour les trois autres items (i^{48} , i^{49} et i^{58}). Dans la deuxième zone, tous les panels s'accordent pour conclure, toutes rondes confondues, que les items i^7 , i^{11} , i^{25} , i^{67} , et i^{140} , catégorisés comme faciles, et les items i^{17} , i^{34} , i^{43} , i^{92} et i^{128} , catégorisés de difficulté moyenne, requièrent un fort degré d'efficience (catégories de performance 2 et 6). Les 21 autres items de cette deuxième zone ont été catégorisés dans l'une des 3 catégories de performance suivantes : 2, 3 ou 6. Dans la troisième zone, tous les panels s'accordent pour conclure, toutes rondes confondues, que les items i^{21} et i^{61} , de difficulté moyenne, requièrent un degré d'efficience moyen

(catégorie de performance 7). Sept autres items ont été catégorisés dans la catégorie 7 et quatre autres, dans la catégorie 6. Dans la quatrième et dernière zone d'incertitude, un item (i^{104}) a été catégorisé dans la catégorie 7, deux items (i^{74} , i^{76}), dans la catégorie 9, deux items (i^{77} , i^{96}) dans la catégorie 10 et trois items (i^{22} , i^{80} , i^{86}) dans la catégorie 11.

Au terme de la quatrième et dernière ronde de notation, ces zones d'incertitude ayant été résolues consensuellement se soldent en une catégorisation définitive des items nécessaire à l'algorithme d'Ebel. Le tableau 4.7 présente le nombre total d'items assignés à chacune des catégories de performance préservées.

Tableau 4.7 – Nombre d'items par famille de difficulté et par catégorie de performance

Famille de difficulté	Catégorie de performance	<u>n</u>
Facile	01. Facile/Très forte	31
	02. Facile/Forte	22
	03. Facile/Moyenne	7
Moyen	06. Moyen/Forte	51
	07. Moyen/Moyenne	18
Difficile	09. Difficile/Très forte	2
	10. Difficile/Forte	2
	11. Difficile/Moyenne	16
	Nombre total d'items	149

Nous observons que la quasi-totalité des items sont assignés aux familles de difficulté « facile » (40,3 %) et « moyen » (46,3 %) et que très peu d'items sont assignés à la famille « difficile » (13,4 %). En tout, nous remarquons que sept points de coupure ont été marqués.

4.4.2 Résultats des rondes 1, 2 et 3 de la phase II

Observons d’abord les standards de performance individuels, établis sur la base des catégorisations individuelles, lors des trois premières rondes de notation. Le tableau 4.8 présente les résultats de l’application de l’algorithme d’Ebel et les détails des calculs des standards de performance intérimaires de l’épreuve clinique, établis lors de ces trois premières rondes. La partie inférieure du tableau présente différentes combinaisons avec les seuils préliminaires établis lors de la première phase d’expérimentation.

Tableau 4.8 – Standards intérimaires individuels et collectifs de la phase II et récapitulatif

Seuils (en %)	Rondes						
Panélistes	R1	R2	R3	Moyen			
Panel D							
P14	85	87	87	▲ 86	Panel D	78%	
P18	78	79	80	▬ 79			
P20	82	77	77	▬ 79			
P17	77	79	79	▬ 79			
P15		78	77	▼ 77			
P16	71	76	76	▼ 74			
P19	66	77	77	▼ 73			
Seuil moyen global							
Panels A, B, C, D	80%	77%		78%	117	items / 149	
Panels A, B, C	81%	76%		78%	117	items / 149	
Panels A, B	80%	72%		76%	113	items / 149	
Panels C, D	80%	81%		81%	120	items / 149	
Panel A	85%	81%		83%	124	items / 149	
Panel B	71%	59%		65%	97	items / 149	
Panel C	87%	87%		87%	130	items / 149	
Panel D	77%	79%	79%	78%	116	items / 149	

Le seuil moyen de l’épreuve clinique, qui résulte des trois rondes de notation individuelles du panel d’experts (D), est respectivement de 77 %, 79 % et 79 %, équivalent à un seuil moyen de 78 % (116 sur 149 items). À la première ronde de notation, nous constatons que le panéliste P¹⁵ n’a pas de seuil établi. Ce constat découle du fait que l’algorithme d’Ebel n’a pas pu être calculé puisque ce panéliste a omis d’enregistrer ses notations à la tâche d’assignation.

Tout comme précédemment, les formes illustrent les contrastes entre les seuils les plus élevés, identifiés par un triangle vert, et les moins élevés, identifiés par un triangle rouge. Toutes rondes confondues, nous remarquons que le panéliste P¹⁴ est en moyenne le plus exigeant ($85 \% + 87 \% + 87 \% = 86 \%$) et qu'à l'inverse, le panéliste P¹⁹ est le moins exigeant du panel D. Globalement, les seuils de réussite de l'épreuve (standards de performance), établis individuellement, fluctuent en moyenne entre 73 % et 86 %, avec un seuil moyen de 78 %. Les seuils individuels sont stables ou varient très peu d'une ronde à l'autre, particulièrement de la deuxième (R2) à la troisième ronde (R3) ($\pm 1 \%$). En comparant le seuil moyen global du panel D avec les différents seuils combinés, nous observons, dans la partie inférieure du tableau, qu'il est identique au seuil moyen global de l'ensemble des panels (A, B, C) de la première phase d'expérimentation et au seuil de tous les panels combinés (A, B, C et D). Nous remarquons que le seuil moyen global des panels de novices (A et B) est pour sa part inférieur de 2 % (76 %), alors que celui des panels d'experts (C et D) est supérieur de 3 % (81 %). Toutes combinaisons de tous les panels confondus, le seuil moyen global de la première ronde est remarquablement stable [80 % : 81 %] comparativement à celui de la deuxième ronde [72 % : 81 %]. Nous observons que les panels de novices (A et B) ont considérablement abaissé le seuil moyen global (-8 %) de la première à la deuxième ronde, alors que les panels d'experts (C et D) l'ont légèrement haussé (+ 1 %). Somme toute, le seuil moyen des panels d'experts est globalement stable d'une ronde à l'autre, alors que celui des panels de novices varie faiblement à la baisse.

4.4.3 Résultat de la ronde 4 de la phase II

Observons maintenant, les standards de performance issus de la quatrième ronde de notation de la deuxième phase d'expérimentation. Cette dernière ronde a permis de parvenir à un consensus au sein du panel d'experts (D) sur l'ensemble des tâches (catégorisation, mise en rang, assignation) constituant la méthode alternative d'ESP et de générer une liste ordonnée des items et une catégorisation définitive. Les catégorisations jumelées aux pondérations consensuelles ont permis, à l'aide de l'algorithme d'Ebel, d'établir les standards de performance intérimaires nécessaires à l'établissement ultérieur des standards définitifs. Les détails des calculs des standards de performance intérimaires de l'épreuve clinique, établis consensuellement par le panel D, sont présentés dans le tableau 4.9.

Tableau 4.9 – Standard de performance consensuel intérimaire de la phase II et ses calculs

Algorithme d'EBEL		Panel D		
Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	01. Facile/Très forte	31	90%	28
	02. Facile/Forte	22	88%	19
	03. Facile/Moyenne	7	86%	6
	04. Facile/Faible			
Moyen	05. Moyen/Très forte			
	06. Moyen/Forte	51	76%	39
	07. Moyen/Moyenne	18	72%	13
	08. Moyen/Faible			
Difficile	09. Difficile/Très forte	2	65%	1
	10. Difficile/Forte	2	60%	1
	11. Difficile/Moyenne	16	55%	9
	12. Difficile/Facile			
Très difficile	13. Très difficile/Très forte			
	14. Très difficile/Forte			
	15. Très difficile/Moyenne			
	16. Très difficile/Facile			
	Nombre total d'items	149		116
	Standard intérimaire		78%	

Les pondérations assignées aux trois catégories de performance pertinentes (k_1 , k_2 et k_3) de la famille de difficulté « facile » varient entre 86 % et 90 %. Celles

de la famille « moyen » des deux catégories pertinentes (k_6 et k_7) varient entre 72 % et 76 % et celles de la famille « difficile » des trois catégories pertinentes (k_9 , k_{10} et k_{11}) varient entre 55 % et 65 %. Le nombre d'items (n) issu de la liste ordonnée des items, associé à chacune des huit catégories pertinentes de l'épreuve et à leur pondération (p) respective, résultent en un standard de performance intérimaire de 78 % (116 sur 149 items). Nous remarquons que ce standard intérimaire est équivalent à la moyenne issue des notations individuelles du panel D des trois premières rondes de la deuxième phase, de même qu'au seuil moyen global des panels (A, B et C) de la première phase. Plus précisément, il s'avère légèrement supérieur à celui des panels de novices (A et B : + 2 %) et légèrement inférieur à celui des panels d'experts combinés (C et D : -3 %).

4.5 Résultats obtenus à l'étape finale de l'expérimentation

Subséquentement, les standards de performance définitifs établis à l'aide de la matrice et de l'algorithme d'Ebel résultent a) du nombre d'items (n) classés dans chacune des huit catégories de performance pertinentes et b) du pourcentage d'items (p) moyen issu des consensus de l'ensemble des panels. Le tableau 4.10 présente à la page suivante les données issues des deux phases d'expérimentation, nécessaires à l'établissement du standard de performance définitif de l'épreuve, ainsi que les seuils consensuels harmonisés aux catégorisations définitives de la liste ordonnée des items.

Tableau 4.10 – Standard de performance définitif établi par les panels A, B, C et D et ses calculs

Algorithme d'EBEL			Panel A		Panel B		Panel C		Panel D		Standard	
Famille de difficulté	Catégorie de performance	n	p	n*p	p	n*p	p	n*p	p	n*p	P moyen	N°P moyen
Facile	01. Facile/Très forte	31	100%	31	100%	31	100%	31	90%	28	98%	30
	02. Facile/Forte	22	95%	21	90%	20	95%	21	88%	19	92%	20
	03. Facile/Moyenne	7	85%	6	85%	6	90%	6	86%	6	87%	6
	04. Facile/Faible											
Moyen	05. Moyen/Très forte											
	06. Moyen/Forte	51	85%	43	70%	36	90%	46	76%	39	80%	41
	07. Moyen/Moyenne	18	75%	14	65%	12	85%	15	72%	13	74%	13
	08. Moyen/Faible											
Difficile	09. Difficile/Très forte	2	80%	2	55%	1	80%	2	65%	1	70%	1
	10. Difficile/Forte	2	75%	2	50%	1	80%	2	60%	1	66%	1
	11. Difficile/Moyenne	16	65%	10	45%	7	80%	13	55%	9	61%	10
	12. Difficile/Faible											
Très difficile	13. Très difficile/Très forte											
	14. Très difficile/Forte											
	15. Très difficile/Moyenne											
	16. Très difficile/Faible											
	Nombre total d'items	149		128		113		135		116		123
	Standard intérimaire		86%		76%		91%		78%			
	Standard définitif										83%	

Nous observons que le nombre total d'items (n) de chacune des catégories est commun à l'ensemble des pondérations consensuelles. Les pondérations assignées à la famille de difficulté « facile » varient entre 85 % et 100 %, celles assignées à la famille « moyen », entre 65 % et 90 %, et celles assignées à la famille « difficile », entre 45 % et 80 %. Les standards intérimaires harmonisés calculés par l'algorithme varient entre 76 % et 91 %. Les pondérations moyennes définitives de chacune des catégories de performance pertinentes varient entre 61 % et 98 %. Celles des catégories k_1, k_2 et k_3 , de la famille de difficulté « facile », varient entre 87 % et 98 %, celles des catégories k_6 et k_7 , de la famille « moyen », entre 74 % et 80 % et entre 61 % et 70 % pour les catégories k_9, k_{10} et k_{11} , de la famille « difficile ». Ces résultats marquent l'aboutissement du premier segment de l'objectif de l'expérimentation de la méthode

alternative d'ESP, avec le standard de performance définitif à l'ensemble de l'épreuve clinique équivalent à un seuil de réussite de 83 % (123 sur 149 items).

Rappelons ici que le deuxième segment de l'objectif de l'expérimentation de la méthode alternative d'ESP consiste à établir des standards de performance pour chacune des 23 tâches I-SMART de l'épreuve clinique.

Le tableau 4.11 présente les 23 matrices contenant les données nécessaires à l'algorithme d'Ebel (les catégorisations et les pondérations définitives) et le standard définitif de chacune des tâches.

Tableau 4.11 – Standards de performance définitifs établis aux 23 tâches et leurs calculs

COMPÉTENCE 15 Effectuer la collecte des données											
TÂCHE 1	Famille de difficulté	Catégorie de performance	n	p	n*p	TÂCHE 4	Famille de difficulté	Catégorie de performance	n	p	n*p
	Facile	02. Facile/Forte	2	0,92	1,84		Facile	01. Facile/Très forte	1	1	0,98
		03. Facile/Moyenne	1	0,87	0,87			02. Facile/Forte	2	0,9	1,84
	Moyen	06. Moyen/Forte	3	0,8	2,4			03. Facile/Moyenne	1	0,9	0,87
		07. Moyen/Moyenne	1	0,74	0,74		Moyen	06. Moyen/Forte	3	0,8	2,4
	Difficile	11. Difficile/Moyenne	3	0,61	1,83			07. Moyen/Moyenne	2	0,7	1,48
	Nombre total d'items		10		7,68		Difficile	11. Difficile/Moyenne	2	0,6	1,22
TÂCHE 2	Standard définitif				0,77	TÂCHE 5	Nombre total d'items		11		8,79
							Standard définitif				0,80
	Famille de difficulté	Catégorie de performance	n	p	n*p		Famille de difficulté	Catégorie de performance	n	p	n*p
	Facile	02. Facile/Forte	5	0,92	4,6		Facile	01. Facile/Très forte	6	1	5,88
	Moyen	06. Moyen/Forte	5	0,8	4			02. Facile/Forte	1	0,9	0,92
		07. Moyen/Moyenne	4	0,74	2,96			03. Facile/Moyenne	1	0,9	0,87
	Difficile	11. Difficile/Moyenne	5	0,61	3,05		Moyen	06. Moyen/Forte	4	0,8	3,2
TÂCHE 3	Nombre total d'items		19		14,6	TÂCHE 6	Nombre total d'items		13		11,61
	Standard définitif				0,77		Standard définitif				0,89
	Famille de difficulté	Catégorie de performance	n	p	n*p		Famille de difficulté	Catégorie de performance	n	p	n*p
	Facile	02. Facile/Forte	1	0,92	0,92		Moyen	06. Moyen/Forte	3	0,8	2,4
		03. Facile/Moyenne	1	0,87	0,87		Nombre total d'items		3		2,4
	Moyen	06. Moyen/Forte	4	0,8	3,2		Standard définitif				0,80
		07. Moyen/Moyenne	1	0,74	0,74	TÂCHE 7	Famille de difficulté	Catégorie de performance	n	p	n*p
TÂCHE 3	Difficile	11. Difficile/Moyenne	1	0,61	0,61		Facile	02. Facile/Forte	3	0,9	2,76
	Nombre total d'items		8		6,34		Moyen	06. Moyen/Forte	2	0,8	1,6
	Standard définitif				0,79		Nombre total d'items		5		4,36
							Standard définitif				0,87

COMPÉTENCE 16 Interpréter les données recueillies

TÂCHE 8

Famille de difficulté	Catégorie de performance	n	p	n*p
Moyen	06. Moyen/Forte	2	0,8	1,6
	07. Moyen/Moyenne	2	0,74	1,48
Difficile	09. Difficile/Très forte	2	0,7	1,4
	10. Difficile/Forte	1	0,66	0,66
	11. Difficile/Moyenne	1	0,61	0,61
Nombre total d'items		8		5,75
Standard définitif				0,72

TÂCHE 9

Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	03. Facile/Moyenne	1	0,9	0,87
Moyen	06. Moyen/Forte	5	0,8	4
	07. Moyen/Moyenne	2	0,7	1,48
Difficile	11. Difficile/Moyenne	4	0,6	2,44
Nombre total d'items		12		8,79
Standard définitif				0,73

TÂCHE 10

Famille de difficulté	Catégorie de performance	n	p	n*p
Moyen	06. Moyen/Forte	2	0,8	1,6
Nombre total d'items		2		1,6
Standard définitif				0,80

COMPÉTENCE 17 Planifier l'intervention

TÂCHE 11

Famille de difficulté	Catégorie de performance	n	p	n*p
Moyen	06. Moyen/Forte	2	0,8	1,6
	07. Moyen/Moyenne	1	0,74	0,74
Nombre total d'items		3		2,34
Standard définitif				0,78

TÂCHE 12

Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	01. Facile/Très forte	1	0,98	0,98
	02. Facile/Forte	2	0,92	1,84
Moyen	06. Moyen/Forte	3	0,8	2,4
	07. Moyen/Moyenne	1	0,74	0,74
Difficile	10. Difficile/Forte	1	0,66	0,66
Nombre total d'items		8		6,62
Standard définitif				0,83

TÂCHE 13

Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	01. Facile/Très forte	1	1	0,98
	02. Facile/Forte	1	0,9	0,92
Moyen	07. Moyen/Moyenne	1	0,7	0,74
Nombre total d'items		3		2,64
Standard définitif				0,88

COMPÉTENCE 18 Appliquer le protocole d'intervention

TÂCHE 14

Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	01. Facile/Très forte	1	0,98	0,98
Nombre total d'items		1		0,98
Standard définitif				0,98

TÂCHE 15

Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	01. Facile/Très forte	1	0,98	0,98
	02. Facile/Forte	1	0,92	0,92
Nombre total d'items		2		1,9
Standard définitif				0,95

TÂCHE 16

Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	01. Facile/Très forte	4	0,98	3,92
	02. Facile/Forte	1	0,92	0,92
Nombre total d'items		5		4,84
Standard définitif				0,97

TÂCHE 17

Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	01. Facile/Très forte	3	0,98	2,94
	03. Facile/Moyenne	1	0,87	0,87
Moyen	06. Moyen/Forte	1	0,8	0,8
Nombre total d'items		5		4,61
Standard définitif				0,92

TÂCHE 18

Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	01. Facile/Très forte	3	1	2,94
	02. Facile/Forte	2	0,9	1,84
Moyen	06. Moyen/Forte	2	0,8	1,6
	07. Moyen/Moyenne	1	0,7	0,74
Nombre total d'items		8		7,12
Standard définitif				0,89

TÂCHE 19

Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	03. Facile/Moyenne	1	0,9	0,87
Moyen	06. Moyen/Forte	4	0,8	3,2
Nombre total d'items		5		4,07
Standard définitif				0,81

TÂCHE 20

Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	01. Facile/Très forte	6	1	5,88
Nombre total d'items		6		5,88
Standard définitif				0,98

TÂCHE 21

Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	01. Facile/Très forte	3	1	2,94
	02. Facile/Forte	1	0,9	0,92
Nombre total d'items		4		3,86
Standard définitif				0,97

TÂCHE 22

Famille de difficulté	Catégorie de performance	n	p	n*p
Facile	01. Facile/Très forte	1	1	0,98
Nombre total d'items		1		0,98
Standard définitif				0,98

COMPÉTENCE 19 Faire le suivi de l'intervention auprès du client					
TÂCHE 23	Famille de difficulté	Catégorie de performance	n	p	n*p
	Moyen	06. Moyen/Forte	6	0,8	4,8
		07. Moyen/Moyenne	1	0,74	0,74
	Nombre total d'items		7		5,54
	Standard définitif				0,79

En complément du standard de performance de l'épreuve, dans le cadre de l'évaluation de la performance, ces 23 standards permettent d'attester de la compétence d'un candidat à exercer la profession d'acupuncteur au Québec. Nous remarquons que les standards de performance varient d'une tâche à l'autre dans une fourchette de 72 % à 98 %. Plus précisément, pour attester de la compétence « effectuer la collecte des données », les standards établis aux sept premières tâches (t_1 à t_7) sont respectivement de 77 %, 77 %, 79 %, 80 %, 89 %, 80 % et 87 %, alors que pour attester de la compétence « interpréter les données recueillies », les standards établis sont de 72 %, 73 % et 80 % pour les tâches t_8 , t_9 et t_{10} . Ceux permettant d'attester de la compétence « planifier l'intervention » sont de 78 %, 83 % et 88 % pour les tâches t_{11} , t_{12} et t_{13} . Nous observons des standards supérieurs pour la plupart des tâches t_{14} à t_{22} , qui permettent d'attester de la compétence « appliquer le protocole d'intervention », soit des seuils de 98 %, 95 %, 97 %, 92 %, 89 %, 81 %, 98 %, 97 % et 98 %. Nous remarquons corolairement que la grande majorité des items permettant d'appliquer le protocole d'intervention sont inscrits dans la famille de difficulté des items « facile », où les seuils de réussite attendus sont généralement plus élevés. Finalement, pour attester de la compétence « faire le suivi de l'intervention auprès du client », le standard établi est, pour la tâche t_{23} , de 79 %. Dans le domaine professionnel de l'OAQ, nous observons que l'ensemble des panels s'accordent sur un seuil de réussite minimal équivalent à près des trois quarts des items d'une tâche. La manifestation concrète d'une compétence passe donc, sur la base de leurs perceptions, par un accomplissement de grande qualité.

Ces résultats marquent l'aboutissement de la démarche de recherche, soit l'atteinte de l'objectif spécifique consistant à établir des standards de performance pour chacune des 23 tâches I-SMART de l'épreuve et un standard de performance pour l'ensemble de l'épreuve clinique à l'aide de la méthode alternative d'ESP mise en œuvre. En conclusion, pour réussir l'épreuve clinique, un candidat devra concomitamment atteindre un standard de performance de 83 % à l'épreuve et le standard de performance spécifique établi pour chacune des tâches de l'épreuve.

5. DISCUSSION

5.1 Introduction

Pour aborder la discussion des résultats, nous présentons d'abord un rappel des avantages que procure la méthode alternative d'ESP mise en œuvre et nous traitons de leur authenticité ou inauthenticité, ainsi que des limites. Par la suite, nous abordons la précision et l'adéquation des résultats obtenus à la suite de l'application de la méthode alternative au moyen du modèle multifacette de Rasch.

5.2 Avantages et limites de l'approche alternative

5.2.1 Rappel des avantages

De nombreux avantages anticipés ont été évoqués dans les chapitres précédents, mais qu'en est-il, en réalité? La méthode alternative d'ESP a l'avantage, par sa nature centrée sur le test, d'être utilisable avec une épreuve clinique non encore administrée et à notations polytomiques. Cette méthode offre la particularité d'être réalisable en quatre jours de quatre rondes, soit une journée par panel. Cela équivaut à la durée par ronde décrite pour certaines méthodes éprouvées définies par Cizek (2012). En l'absence de données empiriques, cette méthode alternative présente aussi l'avantage de tenir compte du paramètre de difficulté estimé des items pour estimer la performance attendue (l'efficacité). Le processus itératif par rondes successives et la mise en rang des items permettent de corriger les incohérences dans les jugements émis aux rondes précédentes. Les catégorisations résultant de la matrice d'Ebel à deux dimensions permettent de leur côté de juger avec une relative précision le niveau de difficulté de

l'épreuve clinique. Le nombre d'items par niveau est relativement équilibré, avec une majorité d'entre eux qui sont perçus comme faciles et moyennement faciles et une minorité étant perçus comme difficiles. Ces catégorisations contribuent à évaluer si la répartition des items est équilibrée et adaptée au contexte, soit dans le cas présent, celui de l'OAQ, et la capacité du candidat minimalement compétent formé à l'étranger. L'atout majeur de la première ronde est qu'elle fournit des jugements individuels indépendants. Les rondes subséquentes et les rétroactions (moyennes calculées) offrent pour leur part de multiples occasions de se repositionner au regard des moyennes et contribuent à la transparence et à la rigueur de la méthode ainsi qu'à une meilleure cohérence et convergence des résultats. Sans conteste, une part de subjectivité humaine et de biais subsiste. Pour tendre à minimiser celle-ci, les décisions consensuelles ont l'avantage d'accroître, auprès des ordres professionnels, la recevabilité des standards de performance établis par l'exercice du jugement. La mise en rang, jumelée à la règle départ-arrêt (Ferrara Perie et Johnson, 2008), accroît cette recevabilité par son efficacité à circonscrire les items présentant des jugements divergents, à réduire les zones d'incertitude, à délimiter clairement les catégories de performance (points de coupure) et à produire une liste ordonnée des items et des catégorisations cohérentes.

L'usage du modèle I-SMART a facilité l'élaboration des items de façon structurée et a significativement contribué à leur compréhension par des formulations claires. À l'unanimité, les experts se sont prononcés fortement en accord avec l'I-SMART. La définition du concept de personne minimalement compétente et les descripteurs des échelles de mesure qualitatives employés (difficulté perçue de l'item, degré d'efficacité attendu) ont grandement contribué, avec un niveau d'accord atteignant 95 %, à rendre les explications claires et précises et ont favorisé une compréhension partagée au sein des panels. Ces échelles, par leurs propriétés ordinales descriptives, ont contribué efficacement à recueillir les appréciations de façon ordonnée. De plus, l'échelle d'efficacité tient compte de l'efficacité personnelle dans l'estimation de la

performance attendue d'un professionnel en considérant non seulement la capacité de rendement, mais également, le rapport entre les résultats escomptés et les moyens mis en œuvre pour y parvenir, ce qui représente un atout majeur.

Le principe d'efficience s'est toutefois avéré complexe pour certains, mais la formation et l'exercice pratique ont permis de pallier cette complexité et ont contribué à la compréhension et à l'utilisation des descripteurs. Seul un novice ayant déclaré avoir éprouvé des difficultés à suivre les instructions et à compléter avec précision ses notations s'est prononcé en désaccord avec l'utilité et la clarté des descripteurs, en précisant qu'il a eu peine à comprendre le principe d'efficience. Lors de la deuxième phase, les catégories de performance (01 à 16) issues de la fusion des deux échelles et leurs descripteurs bonifiés par la taxonomie de Bloom ont singulièrement facilité la classification et l'ordonnement logique et cohérent des items tout en fournissant un cadre simplifié sur lequel appuyer les jugements. Un expert a corolairement souligné que les verbes d'action de la taxonomie et la fiche synthèse contenant la numérotation des catégories de performance et les niveaux de difficulté intégrés ont été d'une grande utilité pour réaliser la mise en rang des items. Pour la plupart, cet ordonnancement s'est avéré simple, mais a constitué la tâche la plus fastidieuse du fait qu'elle a exigé une grande concentration. Les cadres de référence communs ci-dessus ont permis de minimiser les risques de distorsion cognitive et, comme le souligne Cizek (2012), ont défini des balises sur lesquelles appuyer les jugements individuels.

De plus, les outils informatisés ont significativement contribué 1) à réduire le temps de consignation des jugements et de traitement des données, 2) à éviter des erreurs de transcriptions et 3) à assurer l'intégrité des données et un traitement standardisé. Corolairement, le traitement statistique entre chacune des rondes s'est avéré efficace et

rapide. Bref, l'informatisation a permis un gain de temps considérable et des économies pécuniaires non négligeables.

5.2.2 Limites et inconvénients

Comme toute méthode d'établissement de standards, la méthode alternative comporte des limites méthodologiques. Nous reconnaissons ici que le recours au jugement comporte une part de subjectivité et que la petite taille de l'échantillon d'évaluateurs issu de deux techniques non probabilistes invite à la prudence quant à la généralisation des résultats à l'ensemble de la population. Les résultats (non nominatifs) de la recherche se limitent ici à un procédé d'inférence menant à des conclusions logiques fondées sur un raisonnement déductif et sous-tendant l'acceptation inconditionnelle des données disponibles. Les techniques d'échantillonnage non probabilistes sont largement répandues pour sélectionner des sujets (Gaudreau, 2011) et tendent tout de même vers une certaine représentativité de la population étudiée. Ainsi, les standards établis reposent sur leur acceptabilité auprès de l'OAQ. Quoiqu'aucune étude antérieure ne permet de vérifier si les standards établis au moyen de la méthode alternative seraient sensiblement les mêmes avec un plus large échantillonnage ou une méthode distincte d'établissement de standards, la littérature scientifique portant sur des méthodes ayant recours au jugement témoigne toutefois d'une ouverture en spécifiant qu'aucune méthode, du fait de sa nature subjective, ne génère de standards strictement identiques. Afin d'atteindre une forte représentativité de la population et de minimiser le risque d'erreurs de mesure, le recours à un plus grand échantillon d'évaluateurs serait une avenue intéressante à explorer.

Une autre limite réside dans la validité externe restreinte, étant donné qu'aucune méthode existante ne répond aux enjeux de l'OAQ. Le potentiel de généralisation des résultats est par conséquent limité puisqu'aucune comparaison n'est possible avec des résultats empiriques issus de multiples sources. Une étude de généralisation des

résultats avec des méthodes distinctes (simultanée ou non), dans les mêmes conditions, avec un échantillon de même taille et répondant aux mêmes critères de sélection, apporterait vraisemblablement un éclairage intéressant, notamment pour vérifier si des résultats seraient sensiblement équivalents à ceux obtenus.

Une limite non négligeable réside finalement dans le fait que la méthode alternative mise en œuvre en est à sa première expérimentation. Les descripteurs des degrés d'efficacité ont semblé complexes pour certains panélistes lors de la ronde de familiarisation. Les catégorisations issues de la liste ordonnée des items ont mis en lumière que le degré d'efficacité faible est non pertinent dans le contexte de l'épreuve clinique de l'OAQ et du mandat de protection du public des ordres professionnels. Considérant le fait que le nombre de niveaux doit être pertinent, cohérent et utile, nous sommes d'avis que le nombre de catégories d'efficacité aurait pu se limiter à trois plutôt qu'à quatre. Cette adaptation permettrait de réduire la quantité de catégories de performance (difficulté-efficacité) à 12 plutôt qu'à 16. Cela aurait pour avantage de simplifier les tâches cognitives des panélistes. Dans les futures expérimentations, nous recommandons d'éliminer la catégorie « faible » de l'échelle de mesure de l'efficacité et de s'assurer que les descripteurs sont bien compris de tous dès la phase de familiarisation. Notons ici que la phase de familiarisation suppose des limites humaines dans un contexte de primauté de participation à la première expérimentation de la méthode alternative. Les données obtenues sont tributaires de la vigilance et de l'attention qu'apportent les panélistes à chacune des étapes à réaliser. L'omission d'une action peut engendrer la perte de données utiles à l'établissement des standards, d'où l'importance d'encadrer individuellement les panélistes lors de cette phase.

Comme toutes méthodes reconnues, la méthode alternative apporte son lot de désagréments et d'inconvénients. Un désagrément soulevé relève du grand degré de concentration requis pour réaliser la mise en rang. Cette tâche a été ressentie comme

longue et fastidieuse, mais simple à réaliser. L'absence de temps de repos pour la chercheuse est un désagrément mineur considérant que ce temps a été consacré au traitement des données entre les rondes, ce qui a permis de réduire considérablement la durée du travail des panélistes.

Outre ces désagréments mineurs, la méthode alternative comporte certains inconvénients qui pourraient éventuellement s'avérer plus importants. À titre indicatif, une erreur humaine accidentelle a généré des données manquantes. Cette erreur provient d'un panéliste qui, lors de la ronde de familiarisation (ronde 1), n'a pas enregistré ses notations à la tâche d'assignation. Cette erreur d'inattention, survenue au moment de l'annonce de la pause, a empêché le calcul du seuil établi par ce panéliste puisque seules les classifications des items étaient reportées dans l'algorithme d'Ebel. Un suivi individualisé a permis de résoudre le problème. Il s'agissait d'un incident isolé occasionné par un moment de distraction. De fait, comme les données manquantes se sont limitées à des jugements isolés, cette erreur constitue un inconvénient mineur. Quoique notre étude ne nous permette pas de vérifier la cause des erreurs pouvant survenir en cours d'expérimentation, celles-ci pourraient s'expliquer par un mode de raisonnement consistant à omettre d'emblée d'émettre un jugement plutôt que de fournir une réponse hasardeuse. Lorsque nous calculons la moyenne pondérée des seuils de réussite (de l'épreuve clinique) des différents panels, nous remarquons toutefois que ces omissions ont peu d'incidence sur les résultats globaux. Une difficulté technique pourrait également expliquer certaines omissions, notamment la minceur des bordures des cellules délimitant les items et leur catégorisation, ou encore, l'utilisation de la fonction tri ou filtre du chiffrier électronique. L'ajout d'une troisième ronde individuelle lors de la deuxième phase a généré très peu de variations, soit une valeur ajoutée négligeable, ce qui nous porte à croire que cette ronde pourrait être éliminée de la procédure de la méthode alternative.

Quoique l'expérimentation de la méthode alternative comporte certaines limites, la présente recherche contribue malgré tout à l'avancement des connaissances spécifiques à l'établissement des standards de performance pour les ordres professionnels et des connaissances scientifiques. Elle fournit en effet des résultats valides et applicables pour l'OAQ.

5.3 Précision et ajustement des échelles de mesure au moyen du modèle multifacette de Rasch

Dans cette section, nous justifions d'abord l'utilité du modèle multifacette de Rasch pour ensuite l'appliquer à l'établissement de la précision de l'échelle de mesure ainsi qu'au jugement de la qualité de l'ajustement des différentes facettes au modèle. Seuls les patrons de réponse de la première phase d'expérimentation ont ici été analysés.

5.3.1 Justification

Il a été jugé pertinent de présenter à cette discussion une analyse complémentaire de la qualité de la démarche retenue. Il a été décidé de traiter de cette analyse dans ce chapitre plutôt que dans celui sur les résultats, car il ne s'agit pas d'un des objectifs de ce dernier, où nous voulions spécifiquement nous concentrer sur l'établissement des standards, et non sur une modélisation de la qualité de la démarche mise à l'essai. À cet effet, une analyse par modélisation à facettes multiples (*A Many-Facet Rasch Analysis*) (Linacre, 1994) a été jugée appropriée. Ce type d'analyse permet d'étudier simultanément les propriétés métriques des scores observés selon diverses facettes, ici au nombre de six. Un des avantages de la modélisation à facettes est que toutes les facettes partagent une même échelle de mesure unidimensionnelle et que les paramètres estimés de chacune de celles-ci peuvent ainsi être plus facilement comparés. Les facettes d'intérêts retenues ici sont celles des tâches, des items, des juges, des rondes, des groupes et des critères des échelles de mesure déployées (S.1 pondération, S.2 difficulté et S.3

efficience). Chaque facette sera analysée séparément. L'étendue des paramètres associés à chacune de ces six facettes seront analysés, ainsi que la précision de ceux-ci et leur ajustement (*fit*). L'estimation de la valeur des paramètres est ici effectuée par la procédure du maximum de vraisemblance conjointe (*Joint Maximum Likelihood Estimation*, JMLE) à l'aide du logiciel Facets. Le détail du code source utilisé est disponible à l'appendice D.

5.3.2 Échelle de mesure et précision de celle-ci

La mesure de chacun des paramètres associés à chacune des facettes est calculée sur une échelle de logit. La valeur du logit est égale à $\log_e(p/1 - p)$, où p correspond à la probabilité d'une catégorie spécifique d'une facette : un logit compare donc la probabilité d'apparition de cette catégorie avec toutes les autres catégories de la facette. Un logit peut prendre des valeurs entre moins l'infini et plus l'infini. Cette unité de mesure permet de différencier une réponse d'une autre sur la base de l'écart observé entre la valeur mesurée d'un attribut propre à la personne (β_n : sévérité ou clémence) et le niveau de difficulté (δ_i) que cette personne perçoit de l'item. Ainsi, la probabilité de réussite attendue pour un item i dépend de la localisation de la personne n le long du continuum de la facette mesurée (trait latent β_n). Soulignons toutefois qu'en raison de la présence de l'échelle d'efficience dans les données, la notation des items a dû être inversée de manière à ce qu'un score élevé accordé à un item corresponde à un item facile, et non à un item difficile. La notation a aussi été modifiée pour le degré d'efficience. Dans ce contexte, pour le niveau de difficulté perçue « facile » ainsi que pour le degré d'efficience « très fort » attendu, la cote 1 a été convertie en cote 4.

Dans le cadre de l'établissement de standards, une mesure élevée (+ logit) aux facettes juges, items et tâches s'interprètent comme un juge qui se montre plus exigeant (sévère) que les autres (mesures inférieures), et que l'item ou la tâche sont perçus comme plus faciles à réussir, etc. Pour chaque mesure logit, une estimation de la précision est

donnée, soit son erreur type (S. E., *Standard Error*). Nous espérons que l'erreur type du paramètre estimé soit la plus petite possible : traditionnellement, nous souhaitons que la valeur soit inférieure à 0,20. Une valeur inférieure à 0,30 est toutefois raisonnable (S. E. < 0,30) (Linacre, 2019).

L'échelle de logit illustre, à la figure 5.1, les paramètres de chacune des facettes dans une colonne de sorte à les aligner (*Measr*) sur le continuum de la règle verticale, graduée de $\pm 2,6$. Aucun item n'est trop facile ou trop difficile. Les multiples facettes couvrent une étendue raisonnable du continuum de l'échelle unidimensionnelle des logits (*Measr*). Les éléments des deux premières facettes (tâches, items) sont ordonnés selon la difficulté perçue, du plus difficile (logit négatif) au plus facile (logit positif), où la tâche la plus difficile de l'épreuve est la tâche t_8 tandis que la tâche t_{20} est la plus facile. Au regard des items, l'item i_{104} est le plus difficile (-2,13 logit) et l'item i_{55} est le plus facile (2,62 logit). Ainsi, la répartition des items et des tâches est bien équilibrée sur le plan de la difficulté perçue. La troisième facette ordonne les juges selon leur degré d'exigence prédit, du moins sévère (j_9 plus clément) au plus sévère (j_6). Les critères utilisés pour établir les standards de performance sont localisés près de la moyenne de la mesure (zéro logit), symbolisée par l'astérisque (*). Le critère de difficulté perçue se situe ici à l'opposé du critère d'efficacité attendue et du critère de pondération de réussite attendue. Les catégories de réponse de chacun des critères (S.1, S.2 et S.3) sont ordonnées de la catégorie 1 à k .

Les estimations des mesures des multiples facettes affichent un bon degré de précision, puisque de faibles erreurs types (S. E. $\leq 0,30$) sont observées aux tableaux (5.1 à 5.5). Quelques items (i_{55} , i_{105} , i_{108} , i_{109} , i_{114} , i_{124} , i_{132} , i_{135} , i_{136} , i_{137} , i_{139} et i_{141}) affichent des erreurs types légèrement supérieures, mais demeurent relativement faibles (S. E. < 0,42). La capacité des juges a été estimée avec une relativement bonne

précision (S. E.~0,05). La précision des mesures obtenues aux différentes facettes à l'aide du modèle est donc acceptable.

Vertical = (1N,2N,3N,4N,5A,6A,SL) Yardstick (columns lines low high extreme)= 0,15,-2.6,2.6,End

Measr	Taches	Items	Juges	Rondes	Groupes	Criteres	S.1	S.2	S.3
		55					(6) 91-100 severit	(4) Facile	(4) Tres fort
		100 105	6 1						---
		124	13 2 3					---	
2	20								
	21	59	12				---		
	22		11						
		87 109	5						
	16	41							
	15	1 11 108 114							3 Forte
	14	48 120	4						
		24							
1	17	14 47 116	8 10				5 81-90		
		7 58 99						3 Modere	
		27							
		4 25							
		8 72 115 123 127	7						
		30 32 65 85 137 141							
	13	17 54							
		113 136							
		83 94				Difficulte	---		
		69							
		34 78 139	9						
		39 70							
	5	28 53 63 81 91 92 101 135							
	18	26 49 97 138 143 149							
		35 132 148							---
		43 128					4 71-80		
0		15 38 52 64 73 106 142 145	1 2	A B C	Efficiency				
		2 21 88 110 144							
		12 20 71 76 130							
	7	9 18 29 31 42 66 67 103 146							
	12	62 77 90 121 129 147							
		33 51 74 75 86							
	19	5 36 102 133							
	10	79 84				Ponderation	3 61-70		
		119 131							
		22 44 46							
	11	80 82 89 93 112 122							
		68							
		95							
-1	4 6								
	23	6 50					2 51-60	2 Difficile	
		37 57 98							
	3	3 10 117							
	1	13 40 56 60 126							2 Moyenne
	2 9	16 19 45 107							
		23 111 140							
		134							
	8	96 125							
		61							
		118							
-2									
		104							
							(1) 0-50	(1) Tres difficile	(1) Faible
Measr	Taches	Items	Juges	Rondes	Groupes	Criteres	S.1	S.2	S.3

S.1: Model = 2,2,2,2,2,1,PONDERATION ; Criteres: Ponderation
S.2: Model = 2,2,2,2,2,2,DIFFICULTE ; Criteres: Difficulte
S.3: Model = 2,2,2,2,2,3,EFFICIENCE ; Criteres: Efficiency

Figure 5.1 – Sommaire des paramètres (en logit) pour l'ensemble des six facettes

5.3.3 Qualité de l'ajustement

Deux statistiques sont utilisées pour juger de la qualité de l'ajustement de chacun des paramètres des facettes. La première est le coefficient de corrélation point-bisériel (rp_{bis}) (*PtMea* aux tableaux suivants) entre le score observé et la mesure prédite (avec le score total) de la facette spécifique : en théorie classique des tests, il est identifié comme le coefficient de discrimination. Nous nous attendons à une corrélation rp_{bis} positive. Plus le coefficient est élevé, plus l'élément de la facette est jugé discriminant. Une valeur négative signifierait que plus la réponse à cet élément de la facette est élevée, moins le score total à la facette le sera (ce qui indiquerait un problème avec cet élément). Nous signalons qu'une corrélation inférieure à 0,10 s'avère relativement faible et que l'élément de la facette est alors jugé peu discriminant : nous espérons des valeurs supérieures à 0,80. Un coefficient supérieur à 0,40 est toutefois considéré comme hautement discriminant (McGahee et Ball, 2009; Kelly *et al.*, 2002). La seconde statistique, fréquemment utilisée avec les modélisations de Rasch, est l'indice d'ajustement des patrons de réponse au modèle unidimensionnel de mesure, nommé « infit » (*Information-Weighted Fit Statistic*). Cet indice a pour principale fonction de permettre de vérifier (quantifier) le degré de prévisibilité des mesures (logit) par rapport aux données réelles observées et s'interprète sur la base du carré moyen (*MnSq*, *Mean Square*). L'indice d'ajustement d'usage équivaut à 1 et est considéré comme infructueux ou irrecevable pour une épreuve d'observation clinique si sa valeur est supérieure à 1,70 (*misfit*) (Bond et Fox, 2012, 2015; Linacre, 2012; Penta et al. 2005). L'indice d'ajustement *outfit* (*Outlier-Sensitive Fit Statistic*), qui permet de détecter les patrons de réponse éloignés de ceux attendus, n'est cependant pas utilisé, car il a été démontré que son interprétation pose des problèmes sérieux (Béland, 2015).

5.3.3.1 Juges

Au tableau 5.1, au regard de la facette des juges, nous notons que les coefficients de corrélation rp_{bis} entre chaque score de juge et le score total de l'épreuve s'avèrent tous

positifs, mais s'approchent rarement de 0,80 lors de la première phase. La corrélation positive la plus forte s'avère celle du juge j_9 ($rp_{bis} = 0,79$); la plus faible, celle du juge j_2 ($rp_{bis} = 0,46$). En moyenne, les mesures associées aux juges de la première phase d'expérimentation s'avèrent donc moyennement discriminantes ($rp_{bis} = 0,64$).

L'indice d'ajustement *infit* de chaque juge indique une différence acceptable entre les patrons de réponse observés et ceux prédits par le modèle. Le patron de réponse du juge j_{13} se révèle être le mieux ajusté (*infit* = 0,66), tandis que celui du juge j_8 est le moins bien ajusté (*infit* = 1,69). Aucun indice n'excède la valeur critique maximale admise (*infit* < 1,70). L'indice moyen indique un bon ajustement au modèle Rasch (*infit_{moy}* = 1,05).

Tableau 5.1 – Précision et ajustement de la facette des juges

ANALYSE DE LA FACETTE		+	MODÈLE	AJUSTEMENT INFIT	CORRÉLATION	
	JUGES	MESURE (Logit)	Erreur type	Carré moyen	rp _{bis}	
	8	1.03	.04	☆	1.69	.69
	4	1.18	.04	☆	1.52	.71
	2	2.08	.05	☆	1.06	.46
	11	1.73	.04	☆	1.21	.48
	3	2.06	.05	☆	.84	.65
	9	.39	.04	☆	1.13	.79
	7	.79	.04	☆	.92	.71
	5	1.60	.04	☆	1.00	.64
	10	1.02	.04	☆	1.08	.78
	6	2.48	.05	☆	.83	.53
	1	2.42	.05	☆	.92	.58
	12	1.87	.04	☆	.81	.64
	13	2.11	.05	☆	.66	.62
Moyenne (Dénombre : 13)		1.60	.04	☆	1.05	.64
Écart-Type (Population)		0.63	.00	★	.28	.10
Écart-Type (Échantillon)		0.66	.00	★	.29	.10

5.3.3.2 Items

En ce qui a trait aux items (tableau disponible à l'appendice E), 95 % affichent un coefficient de corrélation positif entre chacun d'entre eux et le score total de l'épreuve. Quelques-uns ($i_{51}, i_{116}, i_{137}, i_{139}$ et i_{141}) présentent toutefois une corrélation excessivement proche de zéro et ainsi, peu discriminante ($rp_{bis} < 0,10$). En plus de ces items, 5 % de ceux dont la corrélation est négative ($rp_{bis} < 0$) posent problème ($i_{41}, i_{58}, i_{59}, i_{106}, i_{124}, i_{125}, i_{140}$ et i_{142}). Ces derniers ne contribuent donc pas suffisamment à la mesure. Différentes raisons peuvent expliquer cette situation. Ces items peuvent être mal formulés, présenter des ambiguïtés ou être en contradiction avec la direction de la mesure modélisée : il y aurait lieu soit de les réviser, soit de les retirer.

L'indice d'ajustement *infit* des items de l'épreuve d'observation clinique se situe entre 0,58 et 2,39. Globalement, la qualité des libellés des items est bonne. La grande majorité (92 %) des items s'ajustent de façon adéquate au modèle de mesure, tandis qu'une minorité sont mal ajustés (*misfit* : $i_{87}, i_{100}, i_{106}, i_{112}, i_{115}, i_{116}, i_{127}, i_{128}, i_{134}, i_{135}, i_{136}$ et i_{140}). Presque tous les items mal ajustés relèvent de la compétence c_{18} , « appliquer le protocole d'intervention », tandis qu'une minorité d'entre eux (i_{100} et i_{87}) relèvent des compétences c_{17} , « planifier l'intervention », et c_{16} , « interpréter les données recueillies ». Les trois items présentant une valeur d'*infit* plus élevée que 1,70 sont aussi ceux ayant un indice de discrimination faible (i_{116}) ou négatif (i_{106} et i_{140}). Une analyse plus approfondie est requise pour tenter de connaître la cause des valeurs problématiques observées pour ces derniers items. Ceux-ci s'avéraient mal adaptés, instables et imprévisibles lors des rondes individuelles de la première phase, mais qu'en est-il à la fin de l'expérimentation? Pour répondre à cette question, nous avons jugé pertinent d'investiguer les zones d'incertitude résultant de la première phase d'expérimentation et, afin de vérifier si leur niveau de difficulté était extrême, la mise en rang finale de la liste ordonnée des items définitive. Les zones d'incertitude utiles à la deuxième phase d'expérimentation sont en effet composées de

la moitié des items mal ajustés ($z_1 : i_{112}, i_{134}; z_2 : i_{87}, i_{127}, i_{128}$ et i_{140}). La liste ordonnée des items résultant de la deuxième phase permet de constater que les items mal adaptés ont un paramètre de difficulté extrême (facile), excepté pour l’item i_{128} (modéré). En approfondissant l’analyse, nous remarquons que la mise en rang des items permet d’expliquer partiellement la raison sous-jacente aux items mal ajustés. Le rang des items de la première zone ($28^{\text{e rang}} = i_{112}, 31^{\text{e rang}} = i_{134}$; très forte efficacité attendue) est situé à la limite de la frontière du point de coupure ($31^{\text{e rang}}$) entre les catégories de performance (difficulté-efficacité) k_1 et k_2 , là où une efficacité extrême est attendue (très forte ou forte). Près de la moitié des items de la deuxième zone ($i_{127}, i_{87}; 56^{\text{e}}, 58^{\text{e rang}}$; moyenne efficacité attendue) sont situés à la frontière du point de coupure ($60^{\text{e rang}}$) entre les catégories k_3 et k_6 , là où une efficacité modérée est attendue aux items faciles, et où une forte efficacité est attendue aux items de difficulté modérée.

En conclusion, la méthode d’ESP mise à l’essai permet, sans la modélisation de Rasch, d’identifier la plupart des mauvais ajustements de sorte à y remédier lors de la deuxième phase. Les mauvais ajustements des items émanent principalement du paramètre de difficulté inférieur (items faciles) à la limite des points de coupure entre deux catégories. Les mauvais ajustements de ces items indiquent des perceptions individuelles divergentes et sont probablement difficiles à répondre pour les juges. Rappelons ici que quoique certains items soient mal ajustés lors de la première phase d’expérimentation, tous les items ont été jugés pertinents dans le cadre de l’évaluation des compétences requises pour exercer la profession, donc ils ont été conservés. En moyenne, les items s’ajustent suffisamment au modèle de Rasch ($infit_{moy}=1,16$).

5.3.3.3 Tâches

Lorsque nous analysons les tâches (tableau 5.2), nous constatons qu’en moyenne, les mesures qui y sont associées s’avèrent moyennement discriminantes (rp_{bis} moyen =

0,49). Les mesures modélisées de plus de 90 % des tâches s'avèrent relativement fiables considérant le coefficient de corrélation rp_{bis} toujours positif, soit variant entre 0,27 et 0,74, mais jamais supérieur à 0,80 lors de la première phase. Les tâches t_1 , t_2 et t_{13} y contribuent davantage ($rp_{bis} \approx 0,70$) alors que les tâches t_{10} et t_{21} y contribuent plus faiblement ($rp_{bis} \approx 0,30$). Deux tâches avec une très faible corrélation, négative de surcroît ($rp_{bis_{t_{14}}} = -0,11$ et $rp_{bis_{t_{22}}} = -0,14$), posent problème. En effet, la probabilité d'atteindre la performance attendue à ces tâches décroît à mesure que la capacité s'accroît, ce qui révèle une distorsion aberrante qui nécessite d'être étudiée. En approfondissant l'analyse, nous constatons que ces tâches se rapportent à la même compétence « appliquer le protocole d'intervention » (c_{18}), qu'elles sont les seules de l'épreuve à être composées d'un unique item (i_{106} et i_{142}) et à porter sur la réglementation, l'éthique et la déontologie. Nous en avons discuté avec les représentants de l'OAQ. Ces derniers nous ont indiqué que ces deux tâches découlent d'une formalité juridique qui s'inscrit dans la formation obligatoire pour l'obtention d'un permis d'exercice de la profession d'acupuncteur au Québec. Ceux-ci reconnaissent que ces items gagneraient à être scindés afin d'optimiser la valeur de chacune de leurs composantes. Lorsque l'on se réfère à la liste ordonnée des items, celle-ci indique que ces tâches (ces items) s'avèrent plutôt faciles à réussir et qu'une très forte efficience est attendue, ce qui est cohérent avec la mission fondamentale des ordres professionnels, soit celle d'assurer la protection du public, et avec la nécessité de suivre la formation obligatoire pour exercer cette profession au Québec. Toutefois, il se peut que certains juges aient perçu ces tâches comme étant plus difficiles pour des candidats formés à l'étranger, mais ayant acquis une grande expérience. Nous constatons que parmi les items qui affichaient de faibles valeurs du coefficient de corrélation rp_{bis} ($rp_{bis} < 0,10$), la majorité d'entre eux s'inscrivaient d'ailleurs parmi les tâches les moins discriminantes de l'épreuve. L'insertion d'items plus discriminants à ces tâches augmenterait la valeur des coefficients de corrélation. Nous sommes ici d'avis qu'une tâche ne devrait pas être constituée d'un seul item.

L'indice d'ajustement *infit* des tâches se situe entre 0,81 et 1,81. Les données de près de 90 % des tâches s'ajustent adéquatement au modèle de mesure. Leurs patrons de réponse sont donc proches de ceux attendus. Toutefois, les patrons de réponse des tâches t_{14} , « préconiser une intervention clinique qui respecte la réglementation, les valeurs et les principes éthiques de la profession ainsi que les limites du client et de la relation clinique », t_{20} , « prévenir les dangers d'infection » et t_{21} , « organiser l'environnement physique de sorte à assurer des soins de qualités et sécuritaires » posent problème ($infit_{MnSq} > 1,70$). Comme soulevé précédemment, la tâche t_{14} est constituée d'un seul item (i_{106}) mal ajusté (*misfit*) et insuffisamment discriminant. Une investigation nous a permis de constater que le groupe d'experts gr_C perçoit l'item i_{106} comme plus difficile que ne l'estiment les groupes de novices (gr_A, gr_B ; item perçu de niveau de difficulté facile, très forte efficience attendue) et qu'il exige une efficience inférieure à celle attendue par une large majorité d'entre eux. La tâche t_{20} intègre un des items peu discriminants (i_{137}), ce qui est potentiellement explicable par des catégorisations d'efficience divergentes de la part de quelques juges. La tâche t_{21} est constitué de deux items peu discriminants (i_{139} et i_{141}) et d'un insuffisamment discriminant (i_{140}). Une investigation plus approfondie serait à envisager pour connaître les raisons expliquant pourquoi les patrons de réponse de ces deux tâches diffèrent de ceux prédits par le modèle. L'indice moyen indique que les tâches s'ajustent suffisamment au modèle ($infit_{moy} = 1,24$).

Tableau 5.2 – Précision et ajustement de la facette des tâches

ANALYSE DE LA FACETTE	TÂCHES	+ MODÈLE		AJUSTEMENT INFIT CORRÉLATION	
		MESURE (Logit)	Erreur type	Carré moyen	rp_{bis}
	14	1.22	.22	★ 1.81	-.11
	22	1.67	.26	☆ 1.41	-.14
	17	.97	.08	☆ 1.51	.60
	20	1.98	.11	★ 1.78	.37
	21	1.85	.13	★ 1.71	.27
	16	1.36	.10	☆ 1.56	.51
	19	-.42	.06	☆ 1.56	.42
	15	1.30	.14	☆ 1.43	.56
	10	-.49	.10	☆ 1.32	.30
	6	-.92	.08	☆ 1.30	.45
	18	.14	.06	☆ 1.22	.59
	11	-.66	.08	☆ 1.19	.55
	7	-.21	.07	☆ 1.13	.54
	5	.20	.04	☆ 1.13	.58
	1	-1.21	.04	☆ 1.07	.71
	4	-.90	.04	☆ .99	.64
	12	-.28	.05	☆ .97	.66
	3	-1.13	.05	☆ 1.01	.63
	2	-1.26	.03	☆ 1.00	.71
	9	-1.27	.04	☆ .94	.62
	8	-1.52	.05	☆ .83	.56
	13	.59	.10	☆ .82	.74
	23	-1.00	.05	☆ .81	.52
Moyenne (Dénombre : 23)		.00	.09	☆ 1.24	.49
Écart-Type (Population)		1.11	.06	★ .30	.22
Écart-Type (Échantillon)		1.14	.06	★ .31	.23

5.3.3.4 Critères

En ce qui concerne les critères, au tableau 5.3, nous observons une forte corrélation rp_{bis} critères-total moyenne et positive pour chacun d’eux : la valeur n’atteint toutefois jamais 0,80 lors de la première phase. La corrélation du critère de pondération et d’efficacité ($rp_{bis} \simeq 0,70$) s’avère légèrement plus forte que celle du critère de difficulté ($rp_{bis} = 0,65$). Cela infère que les critères sont valablement mis en œuvre pour évaluer les paramètres de l’épreuve clinique. En moyenne, les mesures associées aux critères s’avèrent moyennement discriminantes ($rp_{bis} = 0,68$).

L’indice d’ajustement *infit* des critères s’inscrit dans un intervalle acceptable, entre 0,90 à 1,20. Nous remarquons que les mesures des critères d’efficacité attendue et des difficultés perçues se révèlent près de l’ajustement d’usage ($infit \simeq 1$) et légèrement

mieux ajustées que celles du critère de pondération de réussite attendue ($infit = 1,20$). En moyenne, les critères s'ajustent bien au modèle de Rasch ($infit_{moy} = 1$).

Tableau 5.3 – Précision et ajustement de la facette des critères

ANALYSE DE LA FACETTE	CRITÈRES	+ MODÈLE		AJUSTEMENT INFIT CORRÉLATION	
		MESURE (Logit)	Erreur type	Carré moyen	rp_{bis}
	1 Ponderation	-.49	.02	☆ 1.20	.71
	3 Efficience	.02	.03	☆ .91	.69
	2 Difficulté	.47	.03	☆ .90	.65
Moyenne (Dénombre : 3)		.00	.02	☆ 1.00	.68
Écart-Type (Population)		.39	.00	☆ .14	.02
Écart-Type (Échantillon)		.48	.01	☆ .17	.03

Nous remarquons de plus que les catégories extrêmes inférieures (k_1) des critères, localisées à une mesure supérieure à $-2,0$ logit, sont très peu utilisées et devraient être abandonnées en raison de leur faible endossement (voir l'appendice E). Seules les catégories les plus probables devraient être conservées. Le cas échéant, la catégorie très difficile (k_1) de l'échelle de difficulté, la catégorie faible (k_1) de l'échelle d'efficience et la catégorie de pondération inférieure à 50 % sont celles à abandonner ou à regrouper. Ce remaniement aurait pour effet de réduire les catégories de performance (difficulté-efficience) à 9, plutôt qu'à 16, ce qui simplifierait l'exercice du jugement et qui irait dans le sens de nos recommandations. Contrairement à ce que nous croyons, le niveau de difficulté « très difficile » ajouté n'a pas permis d'augmenter le potentiel discriminant des items. De plus, les pondérations inférieures à 50 % semblent superflues. Les catégories les plus probables (k_4, k_5 et k_6) sont celles dont les pondérations excèdent 70 %. Dans une étude ultérieure, nous croyons qu'il serait approprié, pour un ordre professionnel, de limiter les catégories de réponse aux pondérations supérieures à 50 %.

5.3.3.5 Rondes

Nous constatons, au tableau 5.4, que la valeur de la corrélation rp_{bis} est positive et moyenne pour chacune des deux rondes individuelles de la première phase de l'ESP. La corrélation la plus forte est celle de la deuxième ronde ($rp_{bis} = 0,74$); elle donc discrimine mieux, ce qui est logique. Ainsi, les patrons de réponse obtenus avec des rétroactions et à la suite d'une période d'apprentissage sont de qualité supérieure à ceux obtenus lors de la période de formation. Il y a donc un effet d'apprentissage bénéfique. La formation, les rétroactions et les itérations successives contribuent vraisemblablement à améliorer la précision des mesures. En moyenne, les mesures associées aux rondes s'avèrent moyennement discriminantes ($rp_{bis} = 0,68$).

L'indice d'ajustement *infit* de chaque ronde indique une différence acceptable entre les mesures observées et celles prédites par le modèle. Les mesures de la deuxième ronde se révèlent corolairement être mieux ajustées ($infit_{r2} = 1,01$) que celles de la première ronde ($infit_{r1} = 1,14$). En moyenne, l'indice d'ajustement des rondes est très près de l'ajustement d'usage.

Tableau 5.4 – Précision et ajustement de la facette des rondes

ANALYSE DE LA FACETTE	RONDES	+	MESURE (Logit)	MODÈLE Erreur type	AJUSTEMENT INFIT Carré moyen	CORRÉLATION rp_{bis}
	ronde 1		.00	.02	☆ 1.14	.63
	ronde 2		.00	.02	☆ 1.01	.74
Moyenne (Dénombre : 2)			.00	.02	☆ 1.07	.68
Écart-Type (Population)			.00	.00	★ .06	.05
Écart-Type (Échantillon)			.00	.00	★ .09	.08

5.3.3.6 Groupes

Nous constatons, au tableau 5.5, que la valeur de la corrélation rp_{bis} est positive et moyenne pour tous les groupes. Les corrélations du groupe B ($rp_{bis} = 0,75$) sont toutefois plus fortes que celles des groupes A et C ($rp_{bis} \simeq 0,60$). Le groupe composé des juges les plus cléments (gr_B) discrimine donc plus que les groupes les plus sévères, possiblement en raison de la plus petite taille de l'échantillon. Nous n'avons pas constaté de distinction marquée entre les scores moyens observés des groupes de novices et des scores des groupes d'experts. Toutefois, celui du groupe B s'est révélé légèrement inférieur à ceux des groupes A et C. En moyenne, les mesures associées aux groupes s'avèrent moyennement discriminantes ($rp_{bis} = 0,65$). Ainsi, la cohérence interne des groupes est satisfaisante lors de la première phase d'expérimentation.

L'indice d'ajustement *infit* de chaque groupe indique une différence acceptable entre les mesures observées et celles prédites par le modèle. Les meilleurs ajustements sont ceux des groupes A (*infit* = 1,05) et C (*infit* = 0,90). En moyenne, l'indice d'ajustement des groupes se révèle être très près de l'ajustement d'usage (*infit_{moy}* = 1,05).

Tableau 5.5 – Précision et ajustement de la facette des groupes

ANALYSE DE LA FACETTE	GROUPES	+ MODÈLE		AJUSTEMENT INFIT		CORRÉLATION rp_{bis}
		MESURE (Logit)	Erreur type	Carré moyen		
2	B	.00	.02	☆	1.20	.75
1	A	.00	.02	☆	1.05	.62
3	C	.00	.03	☆	.90	.58
Mean (Count: 3)		.00	.02	☆	1.05	.65
S.D. (Population)		.00	.00	☆	.12	.07
S.D. (Sample)		.00	.00	☆	.15	.09

CONCLUSION

Au cours des cinq dernières années, l'OAQ, avec notre collaboration, a mis en œuvre un processus d'ESP à une épreuve d'observation clinique. L'objectif de notre recherche consistait à établir des standards de performance pour cette épreuve et pour chacune des 23 tâches I-SMART qui y sont associées à l'aide d'une méthode alternative. Une adaptation de la méthode d'Ebel a été proposée. Les échelles d'appréciation d'origine de la méthode d'Ebel ont été modifiées, la règle départ-arrêt de la méthode d'appariement au descripteur de l'item y a été intégrée, ainsi que, au final, la méthode oui-non. Un processus itératif scindé en deux phases de quatre rondes de jugements a été retenu. Ce processus regroupait quatre opérations : la catégorisation des items par niveau de difficulté et par degré d'efficacité, l'assignation de pondérations aux différentes catégories de performance, la mise en rang ascendante des items selon leurs catégorisations respectives et, afin de raffiner les jugements, la délimitation de zones d'incertitude des items dont l'ordonnancement se serait avéré irrégulier.

L'expérimentation de la méthode alternative a permis d'établir une liste ordonnée des items et le double standard prescrit par l'organisme réglementé. Au chapitre des résultats, les items et les tâches I-SMART ont été validés en primauté par plusieurs experts pour s'assurer de la qualité de l'instrument de mesure. Les formulations se sont avérées claires et pertinentes. Dix panélistes novices et dix experts ont participé à

l'expérimentation, dont trois groupes à la première phase et un à la deuxième phase. À la première phase d'expérimentation, l'épreuve s'est avérée réalisable et bien équilibrée en ce qui concerne les niveaux de difficulté. Les niveaux de difficulté ont principalement été perçus comme faciles ou modérés. Globalement, les pondérations ont été assignées suivant une progression logique et cohérente. Un panéliste semble avoir éprouvé des difficultés à assigner des pondérations cohérentes aux items perçus comme difficiles lors de la ronde de familiarisation. Le taux d'omissions s'est avéré négligeable. Une variation modérée du seuil préliminaire de l'épreuve a été observée lors des rondes individuelles. À la suite d'une rétroaction, une tendance à l'abaissement des seuils a été observée, de même que l'accroissement de l'accord interjuges, toujours supérieur ou égal à modéré.

Lors de la première phase, près des deux tiers des catégorisations des items se sont avérées cohérentes et logiques et trois catégories sur quatre des échelles de mesure des critères de difficulté perçue et d'efficience attendue se sont avérées pertinentes, ainsi que près de la moitié des catégories de performance. Les résultats de la recherche suggèrent que la matrice d'Ebel trois par trois est suffisante et que les trois niveaux de difficulté originels sont suffisants et pertinents. Les seuils moyens préliminaires individuels se sont avérés supérieurs à 55 % et supérieurs à 70 % lors du consensus des panels, suggérant que les panélistes perçoivent qu'un candidat minimalement compétent formé à l'étranger se doit d'être minimalement apte à réussir plus de la moitié des items de l'épreuve clinique pour exercer la profession d'acupuncteur au Québec.

Lors de la deuxième phase d'expérimentation, la majorité des catégorisations des items circonscrits dans les zones d'incertitude ont été ajustées et une minorité a conservé la catégorisation résultant de la première phase. La mise en rang des items a contribué à réduire les zones d'incertitude. Près des trois quarts des catégorisations consensuelles

se sont avérées équivalentes à la moyenne résultant des jugements individuels du plus grand panel d'experts, inférant un degré d'accord interne satisfaisant. L'épreuve clinique s'est avérée réalisable, avec une liste ordonnée des items constituée d'une majorité d'items faciles et modérément faciles à réussir. Les seuils moyens intérimaires individuels se sont avérés supérieurs aux deux tiers de réussite attendue, se révélant relativement stables d'une ronde à l'autre et ayant particulièrement très peu varié entre la deuxième et la troisième ronde. Cela nous a amenés à nous questionner sur la valeur ajoutée de cette troisième ronde individuelle. Les standards intérimaires consensuels se sont avérés près de la moyenne issue des notations individuelles et de la moyenne des standards préliminaires, soit près de 80 %. Ultiment, le plus petit panel de novices s'est avéré le plus clément de tous, suivi par le plus grand panel d'experts. Au final, huit catégories de performance (difficulté-efficience) se sont avérées pertinentes, mettant ainsi à l'index les catégories k_4 (facile-faible), k_5 (moyen-très fort), k_8 (moyen-faible) et k_{12} (difficile-faible).

Lors de l'étape finale, la méthode alternative a permis d'établir le double standard de performance à partir de l'algorithme d'Ebel, des catégorisations de la liste ordonnée des items définitive et de la moyenne des coefficients des pondérations consensuelles de réussite attendue de l'ensemble des panels. Les plus hautes performances attendues pour exercer la profession relèvent de la compétence « appliquer le protocole d'intervention », qui regroupe les tâches t_{14} à t_{22} , perçues comme les plus faciles à réussir. Au final, les standards de performance des 23 tâches I-SMART de l'épreuve et le standard de performance de l'épreuve clinique établissent qu'un candidat minimalement compétent formé à l'étranger doit à tout le moins être en mesure de réussir près des trois quarts des items (ou plus).

La méthode alternative est adaptée à l'approche par compétence et utilisable avec une épreuve d'observation clinique non administrée intégrant des réponses construites

élaborées. Elle a l'avantage d'être itérative, de recourir aux jugements d'experts, d'être réalisable dans un temps conforme aux méthodes éprouvées décrites par Cizek (2012) et elle permet de produire des standards de performance jugés recevables par l'OAQ. L'adaptation de la matrice d'Ebel à deux dimensions permet de tenir compte du paramètre de difficulté estimé des items pour estimer la performance attendue sur le plan de l'efficacité et permet de générer des catégorisations hiérarchiques équilibrées. La procédure a l'avantage de recourir à des jugements indépendants, sans et avec rétroactions, et à des décisions consensuelles, favorisant ainsi la transparence, la rigueur et la cohérence des résultats. La mise en rang jointe à la règle départ-arrêt s'avère simple à réaliser et permet efficacement de pallier les catégorisations incohérentes et l'absence de données empiriques tout en contribuant à accroître la recevabilité des standards par l'exercice du jugement auprès de l'Ordre, et en atténuant la part de subjectivité et de biais. Suivant Cizek (2012), les divers cadres de référence communs minimisent les risques de distorsions cognitives et fournissent des balises claires sur lesquelles appuyer les jugements. L'intégration de la taxonomie de Bloom aux catégories de performance de la deuxième phase fournit un cadre simplifié sur lequel appuyer les jugements et s'inscrit dans le sens des recommandations de Skakun et Kling (1980). L'informatisation assure l'intégrité des données et un traitement statistique standardisé efficace et rapide, en plus de permettre de prévoir avec précision le temps de réalisation de l'expérimentation et d'éviter certains biais, notamment les erreurs de transcription. Au final, cette méthode représente un gain de temps considérable et des économies pécuniaires non négligeables. Elle permet aussi d'inférer que l'épreuve est perçue comme réalisable pour un candidat formé à l'étranger minimalement compétent et efficace.

La méthode alternative constitue une première expérimentation qui comporte certaines limites, notamment qu'elle ne permet pas d'exclure les profils atypiques ni de prendre en compte le degré de sévérité (clémence) caractérisant les juges. Pour certains

individus, la primauté des descripteurs de l'échelle d'efficience a engendré, lors de la ronde de familiarisation, une interprétation divergente qui a nécessité une assistance individuelle. De plus, le recours à une catégorisation à quatre niveaux plutôt qu'à trois s'est avéré non pertinent pour les échelles de mesure; la catégorie très difficile et la catégorie faible efficience seraient à abandonner. La principale limite aux données manquantes, engendrée par une erreur humaine, porte l'attention et la vigilance sur celle des individus et sur celle liée à l'opération de mise en rang qui, quoique simple, est considérée comme longue et fastidieuse. Ces limites n'ont toutefois pas nui aux résultats globaux. Malgré les différentes limites de la méthode alternative, la recherche fournit des résultats valides et applicables pour l'OAQ et contribue à l'avancement des connaissances spécifiques à l'établissement des standards de performance des ordres professionnels et des connaissances scientifiques.

L'analyse des données de la première phase au moyen du modèle multifacette de Rasch s'est avérée utile, en ce sens que les propriétés psychométriques évaluées démontrent des preuves de validité préliminaires de bon augure. Les propriétés métriques ont permis d'estimer valablement le niveau de sévérité latent des juges considérant le niveau de difficulté perçue et d'inférer que les patrons de réponse individuels sont raisonnablement convergents aux prédictions du modèle. Les estimations des mesures affichent un bon degré de précision et de fiabilité. Les items, les tâches et les juges, ainsi que les éléments des autres facettes sont localisés à l'intérieur d'une étendue convenable (± 3 logit) sur l'échelle de mesure unidimensionnelle (Linacre, 2010; Andrich et Styles, 2004). Leur dispersion est bien équilibrée, leur ordonnancement est cohérent et il permet des comparaisons et des interprétations crédibles et cohérentes avec le trait latent mesuré. L'épreuve clinique est corolairement reconnue comme bien équilibrée, avec des items de difficulté variés, ni trop faciles, ni trop difficiles, inférant une preuve de validité de sa structure et de son contenu. D'infimes problèmes d'ajustement des données aux prescriptions du modèle et de pouvoir de discrimination

déficient ont été décelés aux facettes items et tâches lors de la première phase. Ces mauvais ajustements sont principalement issus de la compétence « appliquer le protocole d'intervention » et d'items perçus comme faciles à réussir. Tous les autres éléments des multifacettes (13 juges, 137 items, 20 tâches, 3 critères, 2 rondes, 3 groupes) ne présentent pas de problèmes majeurs au regard de l'ensemble des patrons de réponse (contribuent valablement à la variable mesurée) et s'ajustent bien aux prescriptions du modèle de Rasch.

Ainsi, la grande majorité des scores observés aux éléments des multifacettes contribuent valablement à la variable mesurée et sont corolairement convenablement localisés sur l'échelle de mesure unidimensionnelle. Les zones d'incertitude ont permis de circonscrire la moitié des items mal ajustés (6/12), inférant que la méthode alternative d'ESP permet en bonne partie de déceler et de pallier ces mauvais ajustements lors de la deuxième phase. Dans le même sens, les éléments des multiples facettes se sont avérés positivement corrélés avec un pouvoir discriminant et une cohérence interne raisonnable, excepté pour quelques items et tâches dont le pouvoir discriminant est faible ou déficient. Dans tous les cas, la corrélation moyenne entre les scores observés et estimés et l'ajustement moyen au modèle de Rasch se sont avérés acceptables pour l'ensemble des éléments des multifacettes, inférant qu'en moyenne, ceux-ci discriminent relativement bien et que les mesures prédites s'avèrent de relativement bons prédicteurs des mesures des scores observés. Le tout démontre des qualités psychométriques satisfaisantes permettant d'inférer que l'instrument est un outil de mesure raisonnablement fiable et précis des données observées lors de la première phase d'expérimentation. Les items mesurent correctement les construits, outre quelques-uns étant mal adaptés ou difficiles à juger. Ces derniers sont cependant négligeables considérant que nous ne pouvions ni les perfectionner ni les retirer de l'analyse et que la modélisation de Rasch a été réalisée postérieurement à l'expérimentation de la méthode alternative. Les catégories de l'échelle de difficulté et

de l'échelle d'efficience fonctionnent conformément aux attentes et le positionnement des seuils délimitant chacune des catégories affiche une séquence ordonnée valide. De plus, outre l'échelle de pondération, ils suivent une progression monotone d'une catégorie à l'autre (Penta *et al.*, 2005). Les catégories « très difficile » et « faible » efficience et de pondération inférieure à 50 % n'émergent pas, leur utilisation étant négligeable (respectivement 3 %, 3 %, 14 %) et leurs probabilités d'endossement n'étant jamais les plus élevées. Cela vient appuyer le constat selon lequel trois catégories seraient suffisantes pour favoriser un pouvoir discriminant optimal et dont la catégorie de pondération inférieure à 50 % devrait être éliminée, et ce, considérant qu'elle est peu utile pour un ordre professionnel. De surcroît, les seuils situés entre les six catégories de pondération définies, a posteriori, ne suivent pas l'ordre postulé. L'évaluation de la précision de l'outil de mesure et la signification des résultats obtenus représentent un apport intéressant qui gagnerait à être intégré à la méthode alternative pour améliorer le contrôle de la qualité de l'instrument et la cohérence interne.

En explorant les pistes d'améliorations envisageables, nous avons d'abord retenu le questionnaire suivant : aurait-on obtenu des standards similaires avec un plus grand échantillon de juges ou avec d'autres approches d'établissement de standards? En vue de vérifier si nous arrivons sensiblement aux mêmes valeurs de seuils, il serait pertinent de reproduire l'expérimentation de la méthode alternative avec un plus grand échantillon de juges, composé cette fois d'étudiants et d'acupuncteurs. En effet, le recours à des étudiants en voie d'obtenir d'un permis d'exercer la profession au Québec combiné à des acupuncteurs formés au Québec et à l'étranger et ayant obtenu un permis d'exercer au Québec permettrait de vérifier l'hypothèse d'équivalence des seuils. Il serait également pertinent de comparer les résultats obtenus à l'approche alternative éprouvée avec d'autres approches d'établissement de standards (par exemple, la méthode d'Ebel d'origine), d'appariement au descripteur de l'item ou avec une autre méthode, comme celle du signet. Ensuite, nous avons retenu le questionnaire

suivant : comment intégrer l'analyse de Rasch à la méthode alternative? Il serait intéressant d'élaborer des indices utiles pour porter un jugement sur la qualité des seuils obtenus par la méthode alternative. Une analyse multifacette comme celle de Rasch pourrait être réalisée pour les résultats de chaque groupe de la première phase de la méthode alternative afin de s'assurer d'intégrer l'intégralité des mauvais ajustements dans les zones d'incertitude. Une telle analyse serait surtout utile pour guider le travail des juges à chacune des itérations. Il serait ainsi possible de donner une rétroaction sur le niveau de difficulté de chacun des items et sur la sévérité de chacun des juges, tous deux estimés selon la modélisation multifacette, à chaque itération de façon à ce que les appréciations puissent être réajustées par la suite.

APPENDICES

APPENDICE A

**LETTRE AU PRÉSIDENT DE L'ORDRE DES ACUPUNCTEURS DU
QUÉBEC**

Salaberry-de-Valleyfield, le 12 août 2013

Monsieur Raymond Bourret
Président de l'Ordre des acupuncteurs du Québec
Ordre des acupuncteurs du Québec
505, boul. René-Lévesque Ouest, bureau 1106
Montréal (Québec) H2Z 1Y7

Objet :

Projet de mémoire
L'établissement des standards de performance destinés à évaluer les
compétences professionnelles donnant accès au permis d'exercice de la
profession d'acupuncteur au Québec

Monsieur le Président,

C'est avec enthousiasme que je vous sou mets mon projet de maîtrise en vue d'obtenir un mandat officiel de l'Ordre des acupuncteurs du Québec.

Comme discuté lors de notre entretien du 28 juin dernier, le sujet de ma démarche est lié à l'établissement des standards nécessaires à l'évaluation des compétences professionnelles.

L'objet de ma recherche est de définir, à l'aide de méthodes quantitatives reconnues dans le domaine scientifique, les standards de performance permettant l'accès à la profession. Cette recherche, tel qu'on peut le constater, s'insère dans le continuum des démarches déjà entreprises par l'Ordre des acupuncteurs du Québec. Il apparaît pertinent, suite à la validation en juin dernier des indicateurs de performance, d'établir les standards sur lesquels l'Ordre pourra s'appuyer pour octroyer un permis d'exercice de l'acupuncture au Québec. Ainsi mon projet permettra d'outiller l'Ordre afin de soutenir les efforts consentis depuis plusieurs années pour faire face à la problématique de l'évaluation professionnelle des compétences des candidats étrangers.

La démarche prévue est constituée de plusieurs étapes. La première étape permettra de définir les niveaux de difficulté de l'ensemble des observables de la dimension professionnelle (les indicateurs de performance validés en juin 2013) afin de pouvoir

les hiérarchiser. Cette étape sollicitera la contribution de trois groupes distincts, soit : des acupuncteurs experts sélectionnés par l'Ordre, des enseignants du Collège de Rosemont et des acupuncteurs comptant moins de 5 ans de pratiques. Cette étape sera réalisée en étroite collaboration avec le comité d'admission responsable de l'administration du polygramme des compétences.

Pour la deuxième étape, les experts et les enseignants seront amenés à identifier, en se basant sur une méthode scientifiquement reconnue (Angoff, Bookmark, etc.), le niveau de compétence minimal qu'un candidat doit démontrer pour assurer la sécurité du public.

Lors de la troisième étape, les résultats seront présentés en plénière afin de favoriser des discussions pouvant mener au meilleur consensus possible et ajuster, s'il y a lieu, les « scores » obtenus.

L'étape suivante servira à recueillir, grâce à la collaboration des enseignants du département d'acupuncture, des données quantitatives auprès de la clientèle étudiante (~ 100 étudiants) du collège de Rosemont.

L'étape finale consistera à analyser les données recueillies et à en tirer des conclusions afin de présenter les standards établis à un comité d'expert aux fins de validation finale. Le projet de mémoire sera réalisé sous la supervision de monsieur Gilles Raïche, directeur de l'unité de programme de cycles supérieurs et professeur au Département d'éducation et pédagogie à l'Université du Québec à Montréal ainsi que directeur du Collectif pour le développement et les applications en mesure et évaluation (Cdame).

Espérant obtenir votre soutien ainsi que celui du conseil d'administration, je vous prie d'accepter, Monsieur Bourret, mes cordiales salutations.

Isabelle Dion, B.A.A.
Étudiante à la maîtrise en Éducation
Université du Québec à Montréal

APPENDICE B

LETTRE DU PRÉSIDENT DE L'ORDRE DES ACUPUNCTEURS DU QUÉBEC



Ordre des acupuncteurs du Québec

Le 20 septembre 2013

Madame Isabelle Dion, B.A.A.
Étudiante à la maîtrise en Éducation
Université du Québec à Montréal

Objet : votre projet de mémoire de maîtrise

Madame Dion

J'ai le plaisir de vous informer que, lors de sa réunion régulière du 23 août 2013, le conseil d'administration de l'Ordre des acupuncteurs du Québec a donné son aval au projet que vous nous avez présenté visant l'établissement de standards de performance destinées à évaluer les compétences professionnelles donnant accès au permis de pratique de la profession d'acupuncteur au Québec.

Nous sommes des plus heureux de cette initiative qui, outre sa contribution à l'avancement des connaissances, nous sera des plus utile dans l'accomplissement de notre mandat de protection du public et par ce biais bénéficiera à l'ensemble de la communauté québécoise.

Je serai heureux de vous rencontrer pour préciser les divers aspects de la mise en œuvre de ce projet.

Veuillez agréer, madame Dion, l'expression de mes sentiments les meilleurs.

Le président

Raymond Bourret, Ac., M.A.

c.c. Me François Houle, directeur général, secrétaire général et conseiller juridique.

505, boul. René-Lévesque Ouest, bureau 1106, Montréal (Québec) H2Z 1Y7

☎ 514.523.2882 / 1.800.474.5914 ☎ 514.523.9669

info@o-a-q.org www.o-a-q.org

APPENDICE C

MÉTHODES D'ÉTABLISSEMENT DE STANDARDS DE PERFORMANCE CENTRÉES SUR LE TEST LARGEMENT RECONNUES

Présentation des méthodes courantes d'établissement de standards de performance centrées sur le test

La présente annexe propose les descriptions des principales méthodes centrées sur le test reconnues desquelles nous nous sommes inspirés pour proposer notre méthode alternative. Ces descriptions viendront préciser : la nature du test utilisé (notation dichotomique ou polytomique), le type d'item auquel la méthode s'applique (à réponses choisies ou à réponses construites), si des données empiriques sont fournies ou non aux experts et finalement, le nombre de rondes de discussions généralement nécessaires. La procédure associée à chacune des méthodes d'établissement de standards sera également décrite et appuyée d'un exemple d'application. La première des méthodes décrites sera la méthode d'Angoff qui, selon plusieurs auteurs, est l'une des plus fréquemment utilisées pour établir des standards de performance (Kane *et al.*, 1999; Mills et Melican, 1988 : cités dans Cizek, 2012; Office of Professional Examination Services, 2016). Ensuite seront présentées la méthode oui/non, la méthode de Nedelsky, celle d'Ebel et celle de l'appariement au descripteur de l'item, proposée par Ferrara, Perie et Johnson.

1 Méthode d'Angoff

La méthode d'Angoff (1971) est un procédé rationnel et systématique qui permet au chercheur d'appliquer un algorithme de calcul simple afin de déterminer le seuil de réussite à une épreuve. Cette démarche permet d'intégrer le jugement d'un certain nombre d'experts d'un domaine déterminé sur la probabilité qu'un candidat dont le niveau de compétence est jugé limite puisse réussir chacun des items de l'épreuve. Toutes les méthodes alternatives subséquentes ont également eu recours au jugement centré sur les items d'un test.

La méthode d'Angoff utilise des tests dont la notation des items est de nature dichotomique, c'est-à-dire avec deux choix de réponse (oui/non ou 0/1) qui ne sont pas associés à des données empiriques pour appuyer le jugement ou pour établir des standards de performance. Les données obtenues suite à cette démarche sont ensuite traitées à l'aide d'une méthode statistique descriptive.

Entre 1971 et 2001, plusieurs auteurs ont proposé des méthodes alternatives ayant recours à cette même formule (Berk, 1984a, 1984b; Impara et Plake, 1997; Plake et Giraud, 1998 et Taube, 1997). Seuls Impara et Plake (1997) ont fourni des données empiriques concrètes aux experts pour appuyer à la fois le jugement et l'établissement des standards de performance. Berk (1984a, 1984b) a, pour sa part, proposé d'en fournir seulement pour appuyer le jugement, alors que Taube (1997) a proposé d'en fournir uniquement pour établir les standards de performance à l'aide de statistiques descriptives et d'items auxquels les réponses sont théoriques. En 1995, Hambleton et Plake ont quant à eux utilisé des tests de nature polytomique (à plusieurs choix de réponse), alors que Loomis et Bourque (2001 dans Cizek et Bunch, 2007) ont utilisé des tests de nature dichotomique et polytomique (entre 1994 et 1998) (Kaftandjieva, 2010). Initialement, et dans les années qui ont suivi, la méthode d'Angoff ne précisait

aucun type d'item spécifique. Seuls Hambleton et Plake (1995) ont proposé une méthode alternative, nommée « la méthode d'Angoff étendue » en spécifiant qu'elle pourrait s'appliquer à des items à réponses construites et ainsi, à une notation polytomique.

Différents auteurs (Loomis et Bourque, 2001 dans Cizek et Bunch, 2007; Plake et Giraud, 1998) ont adopté, à l'instar d'Angoff, une seule session (*round*) d'appréciation par les experts, alors que d'autres en ont appliqué jusqu'à quatre. Cizek *et al.* (2004, traduction libre, p. 40) soutiennent, comme Impara et Plake (1997), que généralement deux ou plusieurs sessions d'appréciation devraient être déployées et que peu importe le nombre de sessions, le standard de la performance recommandée se calcule généralement sur les données obtenues à la fin du processus. D'autres auteurs (Buckendahl *et al.*, 2002, traduction libre, p. 253) soulignaient que les méthodes alternatives les plus populaires ont généralement recours à deux sessions de jugement séparées par des données sur les performances concrètes de candidats.

Selon Raïche (2005) et plusieurs autres auteurs (Angoff, 1984; Cizek, 2012; Cizek et Bunch, 2007; Loomis *et al.*, 2000), ce procédé consiste d'abord à demander à des experts de décrire ce qu'est une personne minimalement compétente (niveau limite). Il doit y avoir unanimité entre les experts sur une convention qui détermine l'écart de notation maximale acceptable (fourchette de notation) exprimée en pourcentage. Notons bien qu'il est recommandé, pour établir des seuils de performance efficacement, d'offrir une période d'entraînement aux panélistes. Ceci est encore plus important lorsque le contexte implique des productions complexes et l'observation de performances (Scallan, 1981). Ensuite, l'expert doit estimer le pourcentage de candidats minimalement compétents qui pourraient réussir chacun des items du test. Cet expert doit par la suite inscrire, dans chacune des cases d'un tableau, le pourcentage associé à chacun des items.

L'opération suivante consiste à rassembler tous les formulaires et à compiler tous les résultats des experts dans un tableau synthèse. Par la suite, le chercheur doit additionner les pourcentages attribués à chaque item par chacun des experts et faire la moyenne à chaque item : ce qui donne le pourcentage moyen de réussite par item. Ensuite, on calcule la moyenne des moyennes obtenues à l'étape 4, ce qui donne le résultat attendu, soit le seuil de passage au test, le standard, pour une personne minimalement compétente. Finalement, en plénière, on présente le tableau synthèse et le seuil de passage calculé en identifiant les écarts observés entre les experts. Les experts ayant fourni des pourcentages forts différents de ceux des autres experts doivent expliquer le raisonnement qui les a conduits à leur appréciation. Suite à une discussion, tous les experts pourront décider de modifier ou non leurs pourcentages.

Lorsqu'il y a plusieurs sessions d'appréciation, au terme de l'étape finale, on répète les étapes 3 à 6 lorsque les discussions ont amené les experts à modifier leurs notations. La nouvelle moyenne devient alors le seuil de réussite qu'un candidat limite devrait minimalement atteindre pour obtenir un permis d'exercice. Ce pourcentage représentera alors la norme minimale acceptable pour exercer la profession.

Un exemple d'application de la méthode d'Angoff est présenté, à la page suivante, au tableau C.1. Celui-ci illustre la détermination du standard de passage à un test constitué de cinq items devant être évalués par trois experts qui ont déterminé que leur fourchette d'appréciation ne devrait pas excéder 25 %. Une seule session d'appréciation est prise en compte. Le tableau C.1 présente les résultats de cet exemple fictif.

Tableau C.1 – Illustration de l'application de la méthode d'Angoff avec une seule session d'appréciation

ÉVALUATION DES EXPERTS de la probabilité d'obtenir une réponse correcte à un item				SEUIL DE RÉUSSITE PAR ITEM	
No. de l'item	Expert #1	Expert #2	Expert #3	Moyenne des probabilités par item	
1	65 %	65 %	70 %	Moyennes des experts	67 %
2	85 %	70 %	80 %	Moyennes des experts	78 %
3	70 %	65 %	75 %	Moyennes des experts	70 %
4	75 %	60 %	70 %	Moyennes des experts	68 %
5	80 %	65 %	50 %	Moyennes des experts	65 %
Nombre d'items total : 5	Somme des 5 items / 5	Somme des 5 items / 5	Somme des 5 items / 5	Moyenne des moyennes des experts	
Moyenne	75 %	65 %	69 %	70 %	
Nombre total d'experts : 3				Seuil de réussite	

Chacun des trois experts (#1, #2, #3) inscrit dans les cases centrales son appréciation en pourcentage d'obtention d'une réponse correcte à chacun des cinq items. Ici, l'expert #1 a estimé à 65 % le pourcentage de candidats minimalement compétents qui obtiendraient une réponse correcte à l'item numéro 1. En bas de chaque colonne, on observe la moyenne calculée des valeurs attribuées par un expert pour l'ensemble des items. Le premier expert a donc estimé le seuil de réussite de l'ensemble des items à 75 %, alors que le deuxième et le troisième ont respectivement estimé le seuil de réussite à 65 % et à 69 %. Dans la dernière colonne de chacune des lignes est indiquée la moyenne attribuée à un item par l'ensemble des experts. Nous pouvons donc observer que le seuil de réussite au premier item est de 67 % et que celui des items subséquents est de 78 %, 70 %, 68 % et 65 %. Au coin inférieur droit, le seuil de réussite à l'ensemble de l'épreuve est égal à la moyenne de ces cinq valeurs, soit 70 %.

Imaginons, pour les besoins de notre exemple, que leur fourchette d'appréciation ait été établie à 25 %. Nous constatons que l'item 5 déroge à la convention par un écart d'appréciation séparant les experts #1 et #3. Ces deux valeurs extrêmes présentent un écart supérieur à la fourchette convenue de 25 % ($80\% - 50\% > 25\%$), ce qui mène les experts à une discussion sur les raisons ayant conduit à ces appréciations respectives. Advenant la modification de l'appréciation par l'un ou l'autre des experts, les moyennes seront recalculées. Supposons que l'expert #1 réduise sa notation à 75 %, son nouveau seuil de réussite serait de 74 % ($((65\% + 85\% + 70\% + 75\% + 75\%) / 5 \text{ items})$), celui de l'item 5 deviendrait 63 % ($((75\% + 65\% + 50\%) / 3 \text{ experts})$) et le seuil de réussite de tous les experts passerait de 70 % à 69 % ($((74\% + 65\% + 69\%) / 3 \text{ experts})$), soit 1 % de moins que lors de l'évaluation préliminaire.

La méthode d'Angoff peut aussi être appliquée exactement de la même façon avec des items à réponses polytomiques ordonnées, sauf que l'appréciation donnée par les experts n'est plus un pourcentage, mais plutôt un score qui prend une valeur entre le minimum et le maximum attribuable à chacun des items du test. Toutefois, globalement la méthode d'Angoff est assez exigeante pour des experts, principalement lorsque le test est composé de nombreux items. C'est pourquoi la méthode du oui/non lui est préférée par certains.

2 Méthode oui/non

La méthode oui/non, telle que proposée en 1997 par Impara et Plake, constitue une variante à la méthode d'Angoff d'origine (1971) présentée plus haut. Cette variante, comme nous allons le présenter plus bas, reprend un certain nombre de suggestions formulées par différents auteurs afin de faciliter l'établissement des standards de performance.

La méthode oui/non, tout comme celle d'Angoff, s'appuie sur le jugement centré sur les items d'un test de nature dichotomique. Elle peut toutefois s'appliquer à tous les types d'items, qu'ils soient à réponses choisies ou à réponses construites. Impala et Plake (1997) soutiennent que deux sessions d'appréciations sont nécessaires pour favoriser des résultats plus représentatifs de la réalité.

L'approche proposée par Impara et Plake (1997) tient également compte des recommandations visant à simplifier la tâche des experts. Ils ajoutent même que cette méthode pourrait s'avérer plus valide que la méthode d'Angoff originale. Cizek et Bunch (2007), Impara et Plake ont contribué à l'amélioration du processus de mise en œuvre de la méthode d'Angoff originale, principalement par la réduction du niveau de difficulté d'exécution de la tâche exigée des experts et la facilité de compréhension et d'application de la procédure.

La procédure a été simplifiée. D'abord, on remplace le jugement exprimé en termes de pourcentage par un jugement exprimé en termes de réussite (oui) ou d'insuccès (non) pour chacun des items. Ensuite, en considérant un seul répondant minimalement compétent plutôt qu'un groupe hypothétique de 100 répondants. Cette méthode a pour effet de réduire la complexité et le temps de réalisation de la tâche.

Nous présentons la procédure telle que décrite par Impara et Plake, ainsi que Subhiyah, Featherman et Hawley (2002, cité dans Cizek et Bunch, 2007; Downing *et al.*, 2003). Cette procédure très simple comporte uniquement 2 étapes. La première consiste à demander aux experts de déterminer si un candidat hypothétique minimalement compétent peut ou non réussir chacun des items d'une épreuve. Cette étape permet aux experts d'indiquer « oui » lorsqu'ils estiment que ce candidat pourrait tout juste réussir l'item et « non » dans le cas contraire (Cizek et Bunch, 2007; Downing *et al.*, 2003; Plake, 1998). La deuxième étape permet d'additionner le nombre de « oui » et le nombre total d'items afin de déterminer la moyenne correspondant au standard de

performance de l'épreuve. Tous les autres calculs sont effectués de la même façon qu'avec la méthode d'Angoff originale.

Lorsqu'il y a plusieurs rondes de jugement ou plus d'un panéliste, ceux-ci sont informés, après avoir émis leurs jugements sur l'ensemble des items, de la moyenne des résultats avant de se repositionner lors d'une ronde suivante. Le standard de performance est établi à partir de la moyenne des données de la dernière ronde. Cette moyenne est exprimée en pourcentage converti en nombre d'items qui permet de déterminer le nombre exact d'items représentant le seuil de réussite établi pour un candidat jugé minimalement compétent. L'exemple de l'application pratique de la méthode « oui/non » qui suit demande à un des juges-experts de classer 100 items en se demandant si un candidat hypothétique minimalement compétent serait apte à tout juste réussir correctement chacun des items. Le tableau C.2 suivant présente les résultats de cet exemple fictif.

Tableau C.2 – Illustration de l'application de la méthode Oui/Non

Items (# & texte)	Oui	Non
Item #1 ...	Oui	
Item #2 ...		Non
...	...	...
Item # 92 ...	Oui	
Nombre d'items total : 100	Nombre total de Oui : 80	

Le seuil de réussite de notre exemple, établi à 80 %, correspond au nombre de « oui » divisé par le nombre total d'items, soit 80/100. Supposons qu'un deuxième panéliste

ait estimé qu'un candidat hypothétique serait apte à réussir 60 items, le seuil de ce panéliste serait alors de 60 %. Une moyenne devra être calculée pour chacun des items et pour chacun des panélistes afin de leur permettre de se repositionner lors de rondes subséquentes. C'est uniquement la moyenne de la dernière ronde qui permet d'établir le standard de performance du test. Dans notre exemple, le standard de performance serait de 70 %, résultant de la somme des moyennes de la dernière ronde (80 % + 60 %) divisée par le nombre total de panélistes (2); qui représente un seuil de 70 items à réussir par un candidat minimalement compétent.

La méthode du oui/non, quoique plus simple à utiliser que celle d'Angoff, présente un autre désavantage qu'elle partage avec cette dernière. En fait, elle n'utilise pas toute l'information disponible de la part des experts, car ceux-ci, outre leur appréciation du pourcentage de réussite ou de la réussite ou non à chacun des items, ne tiennent pas compte de l'ordonnancement des items selon leur niveau de difficulté. Nous croyons, lorsqu'il y a beaucoup d'items à soupeser, comme c'est justement le cas avec les items I-SMART, qu'il serait avantageux de tenir compte de cet ordonnancement. Ce désavantage, des méthodes d'Angoff et du oui/non, sera tenu en compte avec la méthode alternative proposée dans le cadre de notre recherche.

3 Méthode de Nedelsky

La méthode d'ESP de Nedelsky (1954), comme les deux méthodes d'appréciation présentées précédemment, fait strictement appel aux jugements d'experts. Tout comme les méthodes d'Angoff et du oui/non, plusieurs sessions d'appréciation consécutives peuvent être tenues. Ce qui distingue la méthode Nedelsky des précédentes, c'est qu'avec des items à choix de réponses multiples elle permet de tenir compte des leurres : ce serait même la seule à en tenir compte. Selon Jaeger (1989), la méthode de Nedelsky serait d'ailleurs la plus ancienne procédure qui ait été proposée.

Après avoir établi un consensus sur la notion de personne minimalement compétente, la première étape consiste à demander aux experts d'identifier, parmi les choix de réponses de chacun des items, les leurres qu'une personne minimalement compétente pourrait éliminer dès la première tentative de réponse aux items. Cette opération permet d'identifier les choix de réponses restants parmi lesquels le candidat aura à démontrer sa compétence. Cette approche permet à chaque expert de se prononcer sur chacun des items en répétant le processus jusqu'au dernier item. (Cizek et Bunch, 2007; Raîche, 2005; Scallan, 1981).

Le tableau C.3 présente à la page suivante l'analyse de cinq items à cinq choix de réponses. Par exemple, un expert ayant identifié deux leurres, inscrit qu'il reste trois choix de réponses valables parmi lesquels un candidat minimalement compétent puisse faire un bon choix de réponse. Le niveau de passage minimal à cet item est ici conséquemment établi à $100 \cdot 1/3$, soit un pourcentage égal à 33 %. L'expert qui aurait identifié quatre leurres obtient, pour sa part, un pourcentage égal à 100 %, soit une seule réponse valable possible ($100 \cdot 1/1$). C'est ce que Nedelsky nomme le niveau de passage minimal, *npm*.

Le coefficient *npm* est obtenu à l'aide de l'équation suivante :

$$npm = \frac{1}{N-n_0} \quad (1)$$

où N correspond au nombre de choix de réponse, tandis que n_0 correspond aux leurres plausibles. Pour chacun des items et pour chacun des experts, le coefficient *npm* est calculé.

Tableau C.3 – Niveau de passage minimal calculé en fonction du nombre de choix de réponse restants selon la méthode de Nedelsky (adapté de Hambleton, 1980, p. 106 dans Raïche, 2005, p. 206)

Nombre de choix de réponse possible (N)	Nombre de choix de réponse inappropriés (n _o)	Nombre de choix de réponse restants (appropriés) (N-n _o)	Niveau de passage minimal (npm) $\text{npm} = 100 * \frac{n}{(N-n_o)}$		Probabilité d'obtenir une réponse correcte à un item (%)
5	0	5	1/5	100 * 0,20	20 %
	1	4	1/4	100 * 0,25	25 %
	2	3	1/3	100 * 0,33	33 %
	3	2	1/2	100 * 0,50	50 %
	4	1	1/1	100 * 1,00	100 %

L'étape finale consiste à déterminer le standard de performance du test. À cette fin, il doit calculer la moyenne des pourcentages attribués par chacun des experts. L'exemple d'application de la méthode de Nedelsky, qui est présenté au tableau C.4 de la page suivante, collige les pourcentages déterminés par trois experts à trois items comportant chacun cinq choix de réponses.

Tableau C.4 – Illustration de la méthode d'établissement de standards de Nedelsky

Numéro de l'item	ÉVALUATION DES EXPERTS de la probabilité d'obtenir une réponse correcte à un item			SEUIL DE RÉUSSITE PAR ITEM
	Expert #1	Expert #2	Expert #3	Moyenne des probabilités par item
1	33 %	50 %	33 %	38,67 %
2	25 %	33 %	33 %	30,33 %
3	100 %	100 %	50 %	83,33 %
Moyenne	52,67 %	61 %	38,67 %	
SEUIL DE RÉUSSITE	$\frac{\sum \text{lignes ou colonnes}}{\text{Nombre d'items total}}$		$\left(\frac{152 \%}{3 \text{ experts}} \right) \times 100$	51 %

Dans cet exemple, l'expert 1 a identifié, pour l'item 1, deux leurres parmi les cinq choix de réponse possibles. Conséquemment, il a déterminé qu'un candidat minimalement compétent aurait une chance sur trois d'obtenir une bonne réponse ($100 \times 1/3$), soit 33 %. Chacun des trois experts inscrit, dans les cases centrales, son estimation du pourcentage de réponses correctes à chacun des trois items.

En bas de chaque colonne, on peut observer la moyenne des pourcentages attribués par un expert pour l'ensemble des items. Nous pouvons donc observer que le premier expert a estimé le seuil de réussite de l'ensemble des items à 52,67 %, alors que le deuxième et le troisième ont respectivement estimé le seuil de réussite à 61 % et à 38,67 %. Chacune des lignes présente, à la dernière colonne, la moyenne des pourcentages attribuée à un item par l'ensemble des experts. Nous pouvons donc observer que le seuil de réussite du premier item est fixé à 38,67 % et que celui des items subséquents est fixé à 30,33 % et 83,33 %. La moyenne globale des pourcentages correspond au seuil de réussite de l'ensemble de l'épreuve et est égale à 51 %. Cette valeur représente à la fois la moyenne de l'axe horizontal et celle de l'axe vertical.

Comme pour les autres méthodes d'appréciation, la méthode de Negelsky peut être organisée en plusieurs sessions d'appréciation et suivie d'une discussion entre les experts à chacune de ces sessions. Son principal désavantage est qu'elle ne peut être utilisée qu'avec des items à choix multiples qui présentent suffisamment de leurres. Elle ne peut définitivement pas être appliquée à un test constitué d'items à réponses dichotomiques ou à réponses polytomiques ordonnées. Elle n'est donc pas appropriée à l'établissement de standards à l'I-SMART.

4 Méthode d'Ebel

En 1972, Ebel a proposé une méthode qui a la particularité de tenir compte non seulement du niveau de difficulté des items, mais aussi de leur pertinence. Avec la méthode d'Ebel (1972; Ebel, 1972 : cité dans Cizek et Bunch, 2007; Raîche, 2005; Scallon, 1981; Skakun et Kling, 1980), il s'agit de classer chacun des items d'un test dans une grille qui propose des cases représentant la jonction entre des niveaux de difficulté (par exemple, facile, moyen ou difficile) et des degrés de pertinence (par exemple, discutable, acceptable, important ou essentiel). Dans ce cas-ci, la grille serait constituée de 12 cases. Berk (1986), précise qu'en 1979, on comptait déjà deux modifications à la méthode Ebel (*modified Ebel*).

La procédure proposée par Ebel comporte deux rounds, constitués de trois étapes (Raîche, 2005). À la première étape (*round 1*), les experts doivent classer les items d'une épreuve dans chacune des cases de la grille. Lors de la deuxième étape (*round 2*), les experts déterminent en pourcentage, pour chacune des cases, le pourcentage (p) d'items qu'un candidat minimalement compétent répondrait avec succès. La dernière étape exige d'effectuer un certain nombre d'opérations mathématiques sur le contenu de la grille. On doit d'abord, multiplier, pour chacune des cases, le nombre d'items qui y sont classés (n), par le pourcentage (p) d'items réussi par un candidat minimalement compétent. On doit ensuite, afin de déterminer le standard de performance au test,

additionner tous les résultats obtenus par les experts pour chacun des niveaux de difficulté ainsi que pour chacun des degrés de pertinence. Finalement, on doit faire la somme de tous ces résultats, soit par niveaux de difficulté, soit par niveaux de pertinence. Ce total correspond alors au nombre d'items qu'un candidat minimalement compétent serait en mesure de réussir pour l'ensemble du test. Celui-ci doit ensuite être divisé par le nombre total d'items afin de déterminer le standard de performance déterminé par cet expert (en pourcentage). Lorsque plusieurs experts réalisent l'exercice, une moyenne des estimations réalisées par l'ensemble des experts est calculée et permet de déterminer le seuil de réussite au test.

L'exemple de l'application de la méthode d'Ebel, illustrée à la page suivante (tableau C.5), présente le mode de détermination par un seul expert du seuil de réussite à une épreuve fictive comprenant 72 items.

Tableau C.5 Illustration de la méthode d'établissement de standards d'Ebel

Formule : $(n * p)$ où n = nombre d'items
 où p = pondération (%) de candidats minimalement compétents qui réussiraient ce type d'item

NIVEAU DE PERTI- NENCE	Importance relative (niveau de difficulté)												Total
	Facile				Moyen				Difficile				
	# des items	n	p	(n* p)	# des item s	n	p	(n*p)	# des items	n	p	(n* p)	
Essentiel + + + +	14 à 27	14 item s	80 %	11,2	9,11, 12 et 13	4 item s	70 %	2,8	1 et 2	2 item s	60 %	1,2	15,2
Important + + +	28 à 37	9 item s	70 %	6,3	62 à 73	12 item s	60 %	7,2	3, 4, 5, 6, 7, 8 et 10	7 item s	50 %	3,5	17,0
Acceptable + +	38, 40, 42 et 44	4 item s	60 %	2,4	47 à 59	13 item s	50 %	6,5	39, 41, 43, 45 et 46	5 item s	40 %	2,0	10,9
Discutable +	60	1 item	50 %	0,5	61	1 item	40 %	0,4	Aucu n	0 item	30 %	0,0	0,9
Total des items à réussir				20,4				16,9				6,7	44
Nb d'items total		28				30				14			72
SEUIL DE RÉUSSITE	Σ lignes ou colonnes Nombre d'items total (44/72)												61 %

L'expert amorce l'exercice en déterminant pour chacune des 12 cases de la grille le pourcentage d'items, classés à ce croisement difficulté-pertinence, qu'un candidat minimalement compétent sera apte à répondre convenablement. Dans notre exemple, l'expert a établi que 80 % des items classés au croisement (catégorie) du niveau de difficulté facile et du niveau de pertinence jugés essentiels obtiendraient une bonne réponse, alors que 60 % des items obtiendraient une bonne réponse au croisement (catégorie) du niveau de difficulté moyen et du niveau de pertinence jugé important.

Une fois la première étape terminée, l'expert a classé chacun des 72 items dans l'une

ou l'autre des 12 cases de la grille. Ici, les items 1 et 2 ont été classés dans la catégorie considérée comme essentielle et difficile à réussir. Les items 14 à 27 ont, pour leur part, été classés dans la catégorie considérée comme essentielle, mais facile à réussir.

Pour compléter l'étape finale, on multiplie pour chacune des cases le nombre d'items (n) par le pourcentage d'items (p). Dans la case essentielle et facile, les 14 items qui y sont classés sont multipliés par le pourcentage estimé à 80 %; soit 11,2. Dans la case essentielle et difficile, les deux items qui y sont classés sont multipliés par le pourcentage estimé à 60 %; soit 1,2. Le résultat obtenu (11,2 et 1,2) indique le nombre d'items réussi attendu par un candidat minimalement compétent dans cette catégorie. Enfin, on effectue l'addition, soit pour les degrés de pertinence, soit pour les niveaux de difficulté.

Dans notre exemple, le niveau de pertinence jugé essentiel indique qu'un candidat minimalement compétent devrait être capable de réussir 15,2 items ($11,2 + 2,8 + 1,2$) sur les 20 items jugés essentiels ($14 + 4 + 2$). De la même façon, 17 items sur 28 items devraient être réussis dans la catégorie jugée importante, etc. Au regard du niveau de difficulté, on peut effectuer les calculs exactement de la même façon en additionnant les résultats de chacune des colonnes. Ainsi, pour la catégorie facile, 20,4 items sur 28 items devraient être réussis par un candidat minimalement compétent. L'opération suivante consiste à additionner, soit les totaux des trois colonnes (niveau de difficulté) ou ceux des quatre lignes (niveau de pertinence). Ici, le chiffre 44 représente le nombre d'items qu'un candidat minimalement compétent devrait accomplir avec succès pour réussir le test. Ce résultat (44) doit être divisé par le nombre total d'items (72) afin de déterminer le standard de performance établi ici à 61 % ($100 \times 44/72$).

La dernière opération consistera à calculer la moyenne pondérée des estimations réalisées par chacun des experts impliqués. Afin d'illustrer cette dernière opération,

faisons l'hypothèse selon laquelle trois experts ont respectivement déterminé des seuils de réussite de 61 %, 70 % et 67 %. Il sera alors possible de déterminer que le seuil de performance est égal à 66 %.

La méthode d'Ebel, malgré le fait qu'elle utilise une dimension très importante associée à chacun des items, soit leur pertinence, est très peu utilisée. De plus, elle s'applique difficilement à des items à réponses polytomiques ordonnées tels qu'utilisés dans l'évaluation de performances ou de productions.

5 Méthode de l'appariement au descripteur de l'item

En 2002, Ferrara, Perie et Johnson ont proposé la méthode de l'appariement au descripteur de l'item. Cette méthode sert à établir les points de coupure (« score de césure ») entre les différents niveaux de performance d'une épreuve. La tâche consiste d'abord à appairer chacun des items à un descripteur de performance (*Performance Level Descriptors*). La procédure consiste à fournir aux panélistes une liste ordonnée des items (*Ordered Item Booklet*), du plus facile au plus difficile, auxquels sont associés « les connaissances, les compétences et les processus cognitifs nécessaires pour y répondre correctement » (Ferrara Perie et Johnson, 2008) de même qu'un descripteur de performance. Les auteurs, afin d'appuyer les jugements, présentent dans une colonne spécifique un indice numérique qui fournit une estimation du niveau de difficulté associé à chaque item. Généralement, l'appréciation du niveau de difficulté (variable b) est employée dans la méthode de la théorie de réponse aux items (TRI) où l'habileté d'un candidat est calculée en scores Z (scores centrés et réduits) et illustrée sur l'abscisse d'un graphique où les valeurs de z sont comprises entre -3 et 3 et où la probabilité de réussir un item donné à chacun des scores z est illustrée sur son ordonnée. Il est toutefois possible d'utiliser simplement la proportion de bonnes réponses aux items. Plus cet indice est élevé, plus le niveau de difficulté de l'item est important.

La méthode de TRI ne sera toutefois pas exposée ni illustrée dans le cadre de la recherche puisqu'il existe différents modèles plus ou moins complexes (par ex. le modèle de Rasch). Nous nous contenterons de préciser, comme le soulignent plusieurs auteurs (Raïche *et al.*, 2011, p. 17), qu'en 1980 Lord définit la TRI comme un modèle probabiliste qui « permet de tenir compte des diverses caractéristiques des items (difficulté b_i , discrimination a_i , pseudo-chance c_i) pour l'estimation du niveau d'habileté des sujets » évalués. Cet appariement permet d'identifier les zones d'incertitudes, là où la séquence de jugements des niveaux de performances est irrégulière. Les auteurs, afin de circonscrire les régions délimitant chacun des niveaux de performance et d'assurer une progression cohérente du classement des items par niveau de difficulté, suggèrent d'utiliser la règle départ-arrêt selon laquelle une zone d'incertitude serait située entre deux séries de trois jugements consécutifs univoques. L'étape ultime de cette méthode consiste à réduire chacune des zones d'incertitudes en points de coupure clairs entre chaque niveau de performance. Ceci exige des panélistes qu'ils se repositionnent sur les items situés dans les zones d'incertitudes afin d'obtenir une suite continue et croissante des niveaux de performance. Chaque point de coupure fixe les limites entre chacun des niveaux de performance et introduit une distinction entre chacun des niveaux. Elles permettent d'appuyer, via cette norme, la décision d'attribuer à un candidat un niveau équivalent à la performance réalisée et d'attester du succès ou l'insuccès à l'épreuve.

Selon Cizek et Bunch (2007), les descripteurs de niveaux de « compétence », spécifiques à chaque niveau, représentent les fondements d'une multitude de méthodes de détermination de points de coupure. Selon eux, ces descripteurs détaillés représentent l'un des éléments déterminants pour appuyer les jugements des panélistes, et ce, sans distinctions des préalables exigés par diverses méthodes éprouvées. Ces auteurs font référence à une formulation contextualisée et adaptée aux habiletés attendues à chacun des niveaux de compétence telle qu'« être capable de » faire dans

une situation et un contexte donné.

Tout comme les méthodes précédentes, cette méthode, présentée comme étant relativement récente, s'appuie sur le jugement centré sur les items d'un test. Celle-ci s'applique à des épreuves de nature dichotomiques et polytomiques, généralement à des items à réponses choisies ou à réponses construites. Deux à trois rondes sont généralement requises pour établir les points de coupure, toutefois il arrive qu'une quatrième ronde soit réalisée.

Contrairement aux autres méthodes, la performance s'inscrit cependant dans une approche critériée qui permet de délimiter avec précision chacun des niveaux de performance à l'aide d'une échelle descriptive bien structurée et bien articulée. Côté (2014) précise que cette approche palie relativement bien au caractère subjectif de l'exercice du jugement, en ce sens, qu'il permet de porter un jugement raisonnablement certain, fidèle et limpide.

Pour illustrer l'application de la méthode de l'appariement au descripteur de l'item et pour faciliter la compréhension du processus de détermination des points de coupure, l'exemple de Cizek et Bunch (2007, p. 197-198) a été adapté au contexte de notre recherche. Posons d'abord l'hypothèse que les données de la liste ordonnée des items sont connues et que la grille critériée contient quatre niveaux de performance associés aux libellés suivants : « inapte » (BB) à exercer la profession, « tout juste apte » (B), « apte » (P) et « parfaitement apte » (A).

Le tableau C.6, présenté à la page suivante, propose une épreuve fictive constituée de 33 items, dont 26 à réponses choisies à réponse unique et sept à réponses construites qui offrent chacune trois niveaux de performance possibles. La liste ordonnée des items contient, conséquemment, 47 emplacements pour les 33 items, soit $26 + (7 \times 3)$.

Tableau C.6 – Illustration de la méthode de l'appariement au descripteur de l'item
(Cizek et Bunch, 2007, p. 197-198)

Page 1 OIB

Page 2 OIB

	Rang de l'item dans l'OIB	Emplacement de l'item dans l'épreuve	Difficulté de l'item (b-valeur)	Repère des contenus	Appariement du descripteur de l'item	Jugement des zones de seuil		Rang de l'item dans l'OIB	Emplacement de l'item dans l'épreuve	Difficulté de l'item (b-valeur)	Repère des contenus	Appariement du descripteur de l'item	Jugement des zones de seuil
	1	9	-2,13		BB								
	2	19	-2,11		BB			25	26-1	-1,26		P	
	3	24	-1,96		BB			26	33-2	-1,26	Collecte des données	P	
	4	13	-1,8		BB			27	30	-1,23		P	
	5	33-1	-1,8	Collecte des données	BB			28	16-2	-1,23		P	
	6	27	-1,8		BB			29	7	-1,03		P	
	7	20	-1,74		BB			30	6-2	-1,03		P	
	8	32	-1,7		BB			31	15-2	-1,03		P	
Seuil de la compétence minimale (BB/B)	9	16-1	-1,65		B	B		32	5-3	-0,99		P	
	10	10	-1,62		BB	B		33	3	-0,94		A	P
	11	31	-1,59		B	B		34	18	-0,94		P	P
	12	25-1	-1,59	Interprétation des données	B	B		35	33-3	-0,94	Collecte des données	P	P
	13	28	-1,55		BB	B		36	16-3	-0,88		A	A
	14	2	-1,55		B			37	11	-0,83		A	A
	15	21	-1,47		B			38	26-2	-0,77		P	A
	16	5-1	-1,47		B			39	22	-0,77		A	
	17	4	-1,47		B			40	25-2	-0,71	Interprétation des données	A	
	18	6-1	-1,46		B			41	29	-0,71		A	
Seuil attendu (B/P)	19	15-1	-1,44		B			42	12	-0,66		A	
	20	14	-1,4		P	B		43	6-3	-0,66		A	
	21	5-2	-1,35		B	B		44	26-3	-0,59		A	
	22	8	-1,32		P			45	17	-0,48		A	
	23	1	-1,32		P			46	15-3	-0,33		A	
	24	23	-1,32		P			47	25-3	-0,11	Interprétation des données	A	

Seuil de la compétence optimale

(P/A)

L'illustration de notre exemple nous amène à présenter une rapide vue d'ensemble du tableau, suite à laquelle nous expliquerons plus en détail chacune des colonnes pour finalement préciser le mode de détermination des points de coupure.

Le tableau C.6 présente les six colonnes requises pour synthétiser l'ensemble de l'information nécessaire pour déterminer les trois points de coupure. La première

colonne contient les 47 emplacements correspondant aux 33 items de l'épreuve présentés dans la deuxième colonne. La troisième colonne contient l'appréciation du niveau de difficulté de chacun des items (valeur b), selon la méthode de la TRI. La quatrième colonne, partiellement illustrée, indique succinctement le repère des contenus liés à chacun des items. La cinquième colonne, vide au début de l'exercice, permet aux experts de se prononcer sur le descripteur du niveau de performance correspondant le mieux à l'item. La sixième colonne a été ajoutée pour présenter le deuxième jugement des experts sur les items situés dans chacune des zones d'incertitude.

Le classement des items de la première colonne par niveau de difficulté croissant se fait du plus facile au plus difficile (1 à 47). La deuxième colonne précise l'emplacement exact de chaque item dans l'épreuve. Les items à réponses choisies sont présentés une seule fois et n'ont pas de tiret. Par exemple, tel qu'on peut le constater, l'item 9 se situe au 1^{er} rang dans la liste ordonnée des items (*Ordered Item Booklet* (OIB)). Les items à réponses construites sont, pour leur part, associés à 0, 1, 2 ou 3 points. Ces points ont trois utilités : l'une pour identifier individuellement chacun des pointages possibles associés à cet item, sauf le zéro; l'autre pour déterminer le type d'item et le troisième pour connaître le nombre de fois que cet item apparaîtra dans l'OIB. Par exemple, l'item 5 a une valeur maximale de 3 points (0, 1, 2 et 3). Comme un candidat peut obtenir, 0, 1, 2 ou 3 points, selon sa performance, l'emplacement de cet item dans l'épreuve sera identifié 5-1, 5-2 et 5-3 et sera associé à trois rangs différents, soit le 16^e, le 21^e et le 32^e rang dans l'OIB. Le nombre après le tiret représente la valeur minimale qu'un candidat pourrait obtenir à cet item considérant une probabilité de 0.5 (50 %) de chance d'obtenir une bonne réponse correspondant à au moins 1 point sur 3 à l'item 5-1; à 2 points ou plus à l'item 5-2, ou 3 points à l'item 5-3. Ainsi, l'emplacement de l'item 5-1 dans l'OIB correspond à un rang plus facile que celui de l'item 5-2 et de l'item 5-3. La troisième colonne présente le classement de chaque item.

La quatrième colonne devrait normalement présenter un repère des contenus pour chacun des items. Dans notre exemple, nous remarquons que certains thèmes sont répétés. La collecte des données qui se retrouve par exemple au 5^e, 26^e et 35^e rang illustre les trois pointages possibles de l’item 33 associés à des niveaux de performance distincts. La cinquième, vide au moment de présenter l’OIB, est celle dans laquelle chaque répondant inscrit, pour chaque item, l’abréviation correspondante à l’appariement qu’il considère comme le plus adéquat au regard des connaissances, des compétences et des processus cognitifs sollicités pour réussir l’item. Une sixième colonne a été ajoutée au modèle proposé afin de présenter visuellement les zones d’incertitudes (zones grises) sur lesquelles les répondants devront revisiter leur jugement afin de réduire cette zone à un point de coupure spécifique (lignes rouges).

Chacun des experts amorce individuellement l’exercice en indiquant, pour chacune des 47 cases de l’OIB, l’abréviation correspondante au descripteur du niveau de performance qu’il juge le plus adéquat à chacun des items. L’exemple illustre le jugement d’un expert qui situe les items 1 à 8, 10 et 13 correspondent au descripteur des candidats qualifiés « inaptes » à exercer la profession; les items 9, 11, 12, 14 à 19 et 21 à « tout juste apte », alors que les items 20, 22 à 32, 34, 35, 38, à « apte » et les items 33, 36, 37 et 39 à 47, à « parfaitement apte ». Parmi ces jugements, notre expert estime qu’un score de 3 sur 3 à l’item 5-3 (au 32^e rang de l’OIB) correspond correctement au descripteur « apte » (P) et il estime également qu’obtenir 1 ou 2 (16^e et 21^e rang), à ce même item, correspond à « tout juste apte ». Dans la même logique, il juge que l’item 25-3 (47^e rang), soit le plus difficile de l’épreuve, s’apparie pertinemment au descripteur « parfaitement apte » (A). Après avoir complété les 47 appariements, l’expert est amené à émettre un second jugement sur les zones d’incertitudes où apparaissent des séquences irrégulières. L’exercice est repris jusqu’à ce qu’il obtienne une séquence régulière et des points de coupure précis. Il identifie conséquemment les zones précédées d’une séquence continue de 3 items et succéder

d'autant. En tout, notre expert a identifié trois zones de seuil (*threshold region*) correspondant à cette règle et aux frontières de chaque niveau de performance : du 9^e au 13^e rang à la frontière d'« apte » et « tout juste apte »; du 20^e au 21^e rang, à la frontière de « tout juste apte » et « apte » et finalement du 33^e au 38^e rang, à la frontière de « apte » et « parfaitement compétent ». L'expert, dans cet exemple fictif, a ensuite revu les appariements de chacune des trois zones de seuil comme suit : première et deuxième zone, tous les items sont appariés à « tout juste apte » et la troisième zone, les items 33 à 35, à « apte » et les items 36 à 38, à « parfaitement apte ». Trois points de coupure s'en dégagent : le niveau de performance « inapte » se situe à l'intérieur des frontières du 1^{er} au 8^e item; le niveau « tout juste apte », du 9^e au 21^e item; le niveau « apte », du 22^e au 35^e item et finalement le niveau « parfaitement apte » du 36^e au 47^e item. Chacun des seuils présuppose implicitement que ses prédécesseurs devront être réussis en grand nombre pour atteindre le niveau de performance attendu et nécessite minimalement de réussir un nombre d'items supérieur aux deux tiers du niveau visé.

Selon Ferrara *et al.* (2008), l'efficacité de la méthode de l'appariement au descripteur de l'item repose sur la qualité des descripteurs des niveaux de performance sur lequel s'appuient les jugements. Selon ces auteurs, cette méthode, quoique prometteuse, demeure récente et conséquemment peu appliquée. Selon eux, elle offre une perspective favorable dans divers domaines tels que l'évaluation en Éducation, les certifications d'évaluation de programmes, les permis d'exercice.

En conclusion, les cinq méthodes courantes centrées sur le test que nous avons présentées ne permettent pas de répondre à l'ensemble des besoins liés aux exigences de l'épreuve clinique de l'OAQ, mais offrent certains éléments intéressants desquels s'inspire notre méthode alternative.

APPENDICE D

CODIFICATION DE LA MODÉLISATION MULTIFACETTE DE RASCH

PHASE I

1^{re} ANALYSE DES SCORES BRUTES

```

Title = Etablissement de standards de performance
Facets = 6
Delements = LN
Non-centered = 3
Negatif = 0
Yardstick = 0,15,-2.60,2.60,End
Inter-rater = 3, 1, 4, 5, 6
Pt-biserial = measure
Usort= 4, 3, 2, 1, 6
vertical = 1N, 2N, 3N, 4N, 5A, 6A, SL
arrange = fN
Left = yes
Subsets = Report
Zscore = 1,2
unexpected = 2
*

Models =
?, ?, ?, ?, 4, ?, M
?, ?, ?, ?, ?, 1#, Ponderation
?, ?, ?, ?, ?, 2#, Difficulte
?, ?, ?, ?, ?, 3#, Efficience
*

Rating (or other) scale = Ponderation,R10, General,Ordinal, keep
6 = 91-100 severite,,,91-100
5 = 81-90,,,81-90
4 = 71-80,,,71-80
3 = 61-70,,,61-70
2 = 51-60,,,51-60
1 = 0-50,,,0-50
*

Rating (or other) scale = Difficulte, R4, General, Ordinal
4 = Facile,,,1
3 = Modere,,,2
2 = Difficile,,,3
1 = Tres difficile,,,4
*

Rating (or other) scale = Efficience, R4, Specific, Ordinal
4 = Tres forte,,,1
3 = Forte,,,2
2 = Moyenne,,,3
1 = Faible,,,4
*

```

```

Labels =
1, Taches, G
1-23
*
2, Items, G
1-10 = , 0, 1
11-29 = , 0, 2
30-37 = , 0, 3
38-48 = , 0, 4
49-61 = , 0, 5
62-64 = , 0, 6
65-69 = , 0, 7
70-77 = , 0, 8
78-89 = , 0, 9
90-91 = , 0, 10
92-94 = , 0, 11
95-102 = , 0, 12
103-105 = , 0, 13
106 = , 0, 14
107-108 = , 0, 15
109-113 = , 0, 16
114-118 = , 0, 17
119-126 = , 0, 18
127-131 = , 0, 19
132-137 = , 0, 20
138-141 = , 0, 21
142 = , 0, 22
143-149 = , 0, 23
*
3, Juges
1-6 =<A
7-10 =<B
11-13 =>C
14-20 =>D
*
4, Rondes, D
1 = ronde 1
2 = ronde 2
3 = ronde 3
*
5, Groupes, D
1 = A
2 = B
3 = C
4 = D
*

```

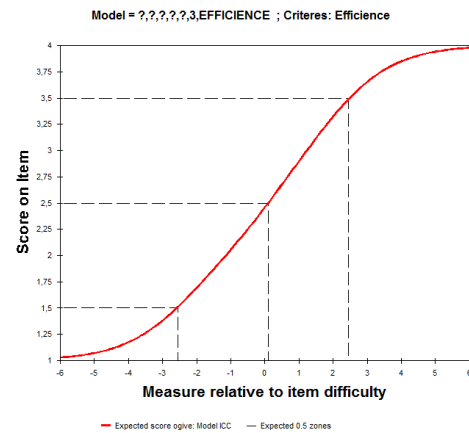
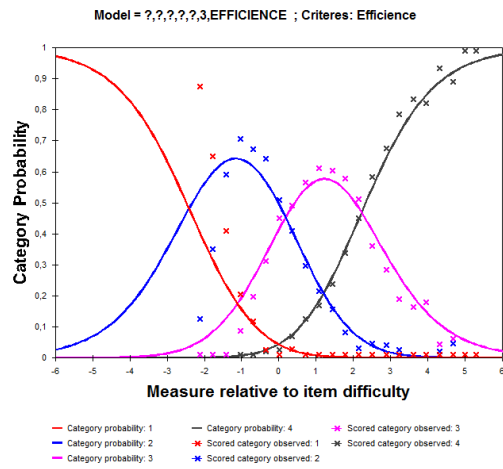
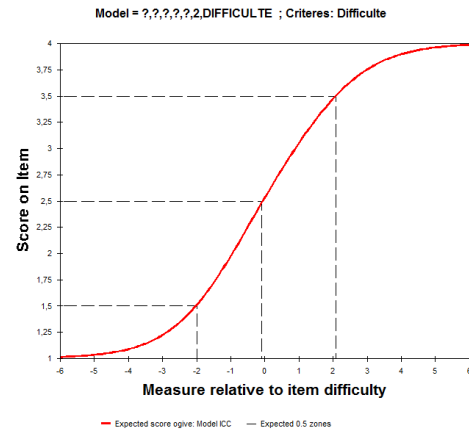
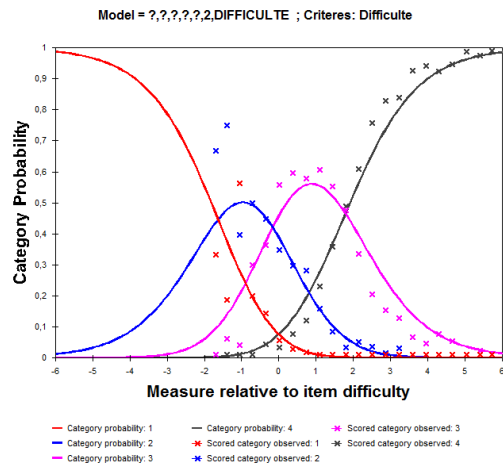
```
6, Criteres
1, Ponderation, R100
2, Difficulte, R4
3, Efficiency, R4
*
Dvalues =
5,3,2,1
*
Data = BD Many-Facet Rasch Measurement.xlsx
```

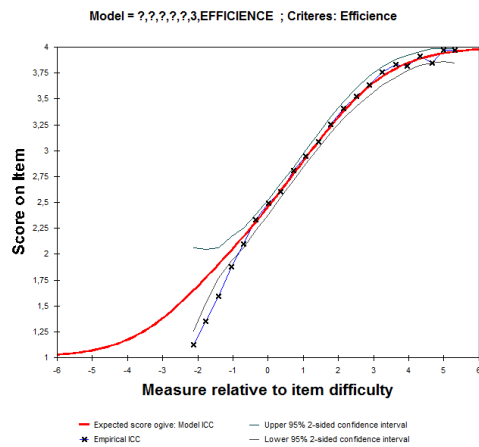
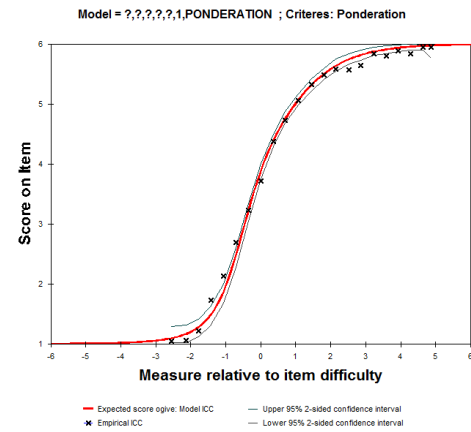
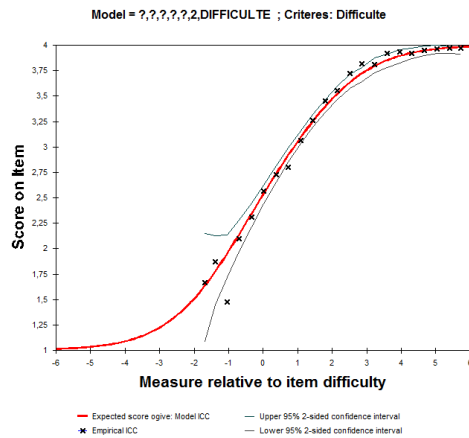
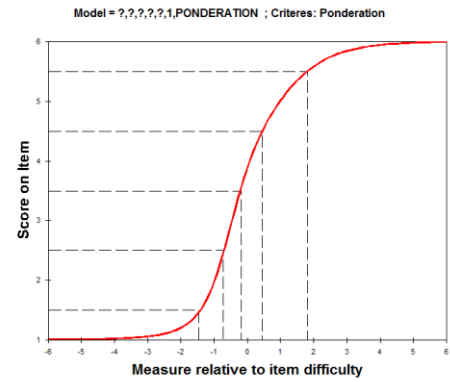
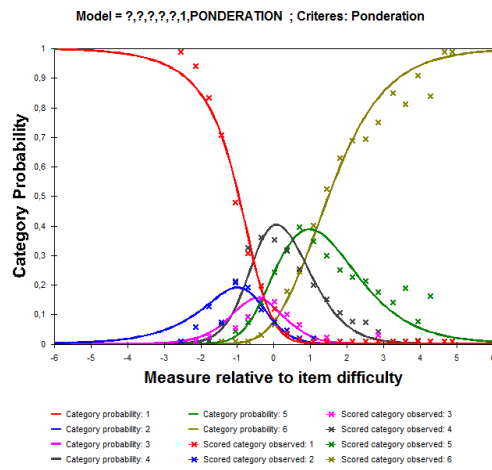
APPENDICE E

RÉSULTATS DE L'ANALYSE SELON LE MODÈLE MULTIFACETTE DE RASCH

PHASE I

1^{re} ANALYSE DES SCORES BRUTES





Etablissement de standards de performance 16/07/20 11:54:50
 Table 5. Measurable Data Summary.

+-----+-----+-----+-----+-----+					
	Cat	Score	Exp.	Resd	StRes
+-----+-----+-----+-----+-----+					
Mean (Count: 11587)	3.46	3.46	3.46	.00	-.01
Mean (Count: 11587)	1.31	1.31	1.06	.80	1.03
Mean (Count: 11587)	1.31	1.31	1.06	.80	1.03
+-----+-----+-----+-----+-----+					
Data log-likelihood chi-square = 22389.8320					
Approximate degrees of freedom = 11389					
Chi-square significance prob. = .0000					
	Count	Mean	S.D.	Params	
Responses used for estimation	= 11587	3,46	1,31	198	
Count of measurable responses	= 11587				
Raw-score variance of observations	= 1,72	100.00%			
Variance explained by Rasch measures	= 1,09	63,14%			
Variance of residuals	= 0,63	36,86%			

Etablissement de standards de performance 16/07/20 11:54:50
Table 6.0 All Facet Vertical "Rulers".

Vertical = (1N,2N,3N,4N,5A,6A,SL) Yardstick (columns lines low high extreme)= 0,15,-2.6,2.6,End

Measr	+Taches	+Items	+Juges	+Rondes	+Groupes	+Criteres	S.1	S.2	S.3
		55	6				(6) 91-100 severit	(4) Facile	(4) Tres fort
		100 105	1						---
		124	13						
2	20		2 3					---	
	21	59	12				---		
	22		11						
		87 109	5						
16		41							
15		1 11 108 114							
14		48 120	4						3 Forte
		24							
1	17	14 47 116	8 10				5 81-90		
		7 58 99						3 Modere	
		27							
		4 25							
		8 72 115 123 127	7						
		30 32 65 85 137 141							
		17 54							
	13	113 136							
		83 94				Difficulte	---		
		69	9						
		34 78 139							
		39 70							
		28 53 63 81 91 92 101 135							
	5	26 49 97 138 143 149							
	18	35 132 148							---
		43 128					4 71-80		
*	0	15 38 52 64 73 106 142 145	*	1 2	A B C	Efficiency	*	*	*
		2 21 88 110 144						---	
		12 20 71 76 130							
	7	9 18 29 31 42 66 67 103 146					---		
	12	62 77 90 121 129 147							
		33 51 74 75 86							
	19	5 36 102 133							
	10	79 84				Ponderation	3 61-70		
		119 131							
		22 44 46							
	11	80 82 89 93 112 122					---		
		68							
		95							
	4 6								
-1	23	6 50					2 51-60	2 Difficile	
		37 57 98							
	3	3 10 117							2 Moyenne
	1	13 40 56 60 126							
	2 9	16 19 45 107							
		23 111 140							
		134					---		
	8	96 125							
		61							
-2		118						---	
		104							
							(1) 0-50	(1) Tres difficile	(1) Faible
Measr	+Taches	+Items	+Juges	+Rondes	+Groupes	+Criteres	S.1	S.2	S.3

S.1: Model = ?,?,?,?,?,1,PONDERATION ; Criteres: Ponderation
S.2: Model = ?,?,?,?,?,2,DIFFICULTE ; Criteres: Difficulte
S.3: Model = ?,?,?,?,?,3,EFFICIENCE ; Criteres: Efficiency

Etablissement de standards de performance 16/07/20 11:54:50
Table 6.1 Taches Facet Summary.

Logit:
 1 21112 1 1111 11 1 1 111 11 1
 +-----Q-----S-----M-----+S-----+Q-----+
 -3 -2 -1 0 1 2 3

Infit MnSq:
 31432221212
 +---Q---S---M---S---Q---+-----+
 0 1 2 3 4 5

Outfit MnSq:
 112732112 11 1
 +---Q---S---M---S---Q---+-----+
 0 1 2 3 4 5

Infit ZStd:
 11 1 1 11 2 1 21 3 2 1 1 2 2
 +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9

Outfit ZStd:
 1 1 1 1 111121 1 1 31 3 1 1 1
 +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9

Etablissement de standards de performance 16/07/20 11:54:50
Table 6.2 Items Facet Summary.

Logit:
 11 1 1 1
 11 1 225743 4433502603074244835225 2 1 12 1
 +-----+---Q-----+S-----+M-----+S-----+Q-----+
 -3 -2 -1 0 1 2 3

Infit MnSq:
 1111111 1
 6744625572642134 1
 +---Q---S---M---S---Q---+-----+
 0 1 2 3 4 5

Outfit MnSq:
 122112
 95205408935223 1 1
 +---Q---S---M---S---Q---+-----+
 0 1 2 3 4 5

Infit ZStd:
 1
 11137555261457765690346437574121 42
 +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9

Outfit ZStd:
 1
 312332634666666574065896732121211 12 11 1
 +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9

Etablissement de standards de performance 16/07/20 11:54:50
Table 6.3 Juges Facet Summary.

Logit:
 1 1 2 1 11 1 3 11
 +-----+---Q-----+S-----+M-----+S-----+Q---+
 -3 -2 -1 0 1 2 3

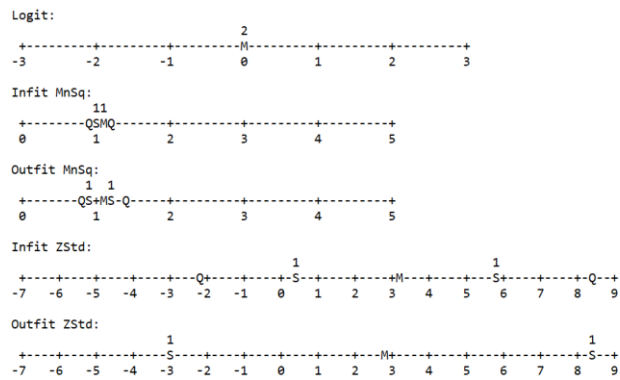
Infit MnSq:
 132131 1 1
 +---Q---S---M---S---Q---+-----+
 0 1 2 3 4 5

Outfit MnSq:
 12211321
 +---Q---S---M---S---Q---+-----+
 0 1 2 3 4 5

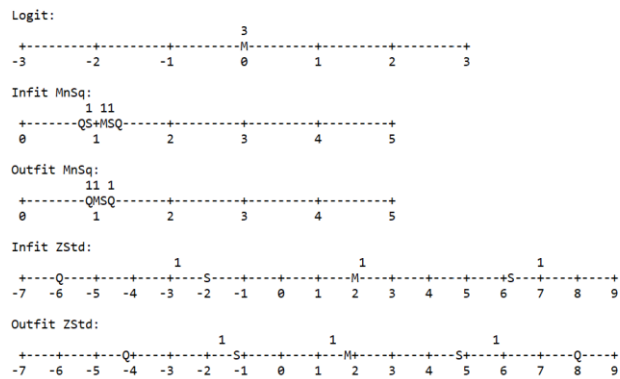
Infit ZStd:
 1 1 11 1 1 1 1 1 1 1
 +---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9

Outfit ZStd:
 1 1 2 1 1 1 1 1 1 21
 +---Q---S---M---S---Q---+-----+
 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9

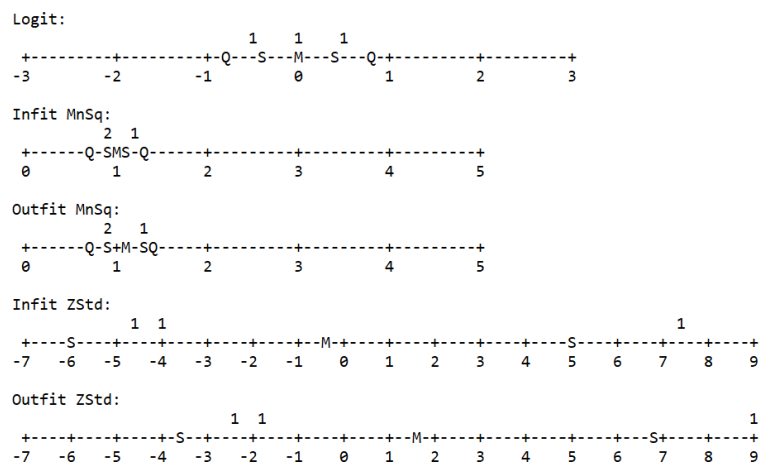
Etablissement de standards de performance 16/07/20 11:54:50
Table 6.4 Rondes Facet Summary.



Etablissement de standards de performance 16/07/20 11:54:50
Table 6.5 Groupes Facet Summary.



Etablissement de standards de performance 16/07/20 11:54:50
Table 6.6 Criteres Facet Summary.



Etablissement de standards de performance 16/07/20 11:54:50
Table 7.1.1 Taches Measurement Report (arranged by fN).

[illegible]

Etablissement de standards de performance 16/07/20 11:54:50
 Table 7.2.1 Items Measurement Report (arranged by fN).

Num Items	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	ZStd	Outfit MnSq	ZStd	Estim. Discrm	Correlation PtMea	PtExp	Group
116 116	351	78	4.50	4.34	G 1.08	.30	1.88	2.6	4.33	5.7	.66	.01	.19	17
115 115	347	78	4.45	4.25	G .77	.26	2.39	4.2	1.28	.9	.88	.40	.22	17
100 100	351	78	4.50	4.56	G 2.33	.30	2.13	3.2	1.58	1.5	.80	.20	.19	12
106 106	338	78	4.33	3.96	G .00	.22	1.81	3.2	2.11	3.8	.41	-.11	.26	14
87 87	311	78	3.99	4.44	G 1.51	.17	2.11	4.3	1.77	3.8	.40	.43	.34	9
112 112	324	78	4.15	3.59	G -.71	.19	2.11	4.3	1.23	1.2	1.03	.63	.31	16
134 134	309	75	4.12	3.13	G -1.40	.19	2.10	4.2	1.29	1.5	.89	.66	.31	20
140 140	320	78	4.10	3.17	G -1.33	.18	2.05	4.1	1.93	4.1	.24	-.03	.32	21
127 127	316	78	4.05	4.26	G .81	.17	2.05	4.1	1.93	4.3	.28	.17	.33	19
136 136	356	78	4.56	4.20	G .61	.37	1.99	2.3	1.29	.7	.90	.23	.16	20
142 142	346	78	4.44	3.96	G .00	.26	1.41	1.6	1.92	2.6	.64	-.14	.22	22
58 58	338	78	4.33	4.32	G 1.02	.22	1.67	2.7	1.82	3.0	.60	-.02	.26	5
128 128	287	78	3.68	3.99	G .09	.15	1.79	3.4	1.67	3.6	.23	.31	.41	19
135 135	353	78	4.53	4.07	G .26	.32	1.75	2.1	1.36	1.0	.85	.17	.18	20
23 23	155	78	1.99	3.17	G -1.34	.14	1.67	2.9	1.39	1.6	.63	.46	.54	2
133 133	345	78	4.42	3.77	G -.38	.25	1.67	2.4	1.60	1.9	.73	.10	.23	20
124 124	354	78	4.54	4.54	G 2.20	.33	1.06	.2	1.66	1.5	.86	-.06	.17	18
117 117	294	78	3.77	3.30	G -1.16	.15	1.64	2.8	1.51	2.8	.49	.36	.39	17
7 7	294	78	3.77	4.33	G 1.03	.15	1.64	2.8	1.30	1.7	.80	.43	.39	1
94 94	296	78	3.79	4.17	G .52	.15	1.63	2.7	1.21	1.2	.94	.60	.38	11
125 125	231	78	2.96	3.06	G -1.47	.13	1.54	2.6	1.60	3.2	.13	-.05	.53	18
137 137	357	78	4.58	4.25	G .75	.39	1.32	.9	1.57	1.1	.88	.07	.15	20
53 53	318	78	4.08	4.07	G .25	.18	1.57	2.5	1.30	1.6	.72	.34	.32	5
16 16	160	78	2.05	3.24	G -1.24	.14	1.56	2.6	1.38	1.6	.73	.49	.54	2
67 67	283	78	3.63	3.86	G -.21	.14	1.55	2.5	1.17	1.1	.98	.60	.42	7
59 59	350	78	4.49	4.48	G 1.77	.29	1.15	.6	1.55	1.5	.76	-.16	.20	5
52 52	309	78	3.96	3.95	G -.01	.17	1.54	2.4	1.06	.3	1.11	.52	.35	5
19 19	158	78	2.03	3.21	G -1.28	.14	1.54	2.5	1.31	1.3	.70	.48	.54	2
47 47	296	75	3.95	4.33	G 1.04	.17	1.52	2.2	1.19	1.1	.84	.34	.36	4
129 129	269	78	3.45	3.83	G -.27	.14	1.51	2.5	1.42	2.4	.50	.40	.46	19
110 110	340	78	4.36	3.94	G -.04	.23	1.51	2.1	.93	-.2	1.04	.48	.25	16
130 130	277	78	3.55	3.90	G -.12	.14	1.50	2.4	1.39	2.3	.52	.37	.43	19
138 138	351	78	4.50	4.04	G .20	.30	1.50	1.6	1.18	.6	.92	.23	.19	21
62 62	239	78	3.06	3.82	G -.28	.13	1.47	2.3	1.49	2.7	.56	.41	.52	6
49 49	316	78	4.05	4.04	G .19	.17	1.48	2.1	1.37	1.9	.68	.23	.33	5
107 107	302	78	3.87	3.20	G -1.29	.16	1.46	2.1	1.04	.3	1.04	.57	.37	15
141 141	356	78	4.56	4.24	G .74	.37	1.28	.8	1.45	1.0	.89	.06	.16	21
91 91	292	78	3.74	4.07	G .26	.15	1.45	2.1	1.23	1.4	.80	.31	.39	10
13 13	162	78	2.08	3.27	G -1.21	.14	1.44	2.2	1.23	1.0	.77	.53	.55	2
41 41	317	78	4.06	4.40	G 1.32	.18	1.35	1.6	1.43	2.2	.44	-.18	.33	4
101 101	302	78	3.87	4.08	G .29	.16	1.43	1.9	1.11	.7	.99	.47	.37	12
111 111	302	78	3.87	3.16	G -1.36	.16	1.42	1.9	1.04	.2	1.01	.51	.37	16
139 139	353	78	4.53	4.12	G .39	.32	1.23	.8	1.41	1.1	.86	.03	.18	21
48 48	312	78	4.00	4.36	G 1.17	.17	1.38	1.7	1.24	1.3	.75	.31	.34	4
37 37	178	78	2.28	3.37	G -1.05	.13	1.36	1.9	1.21	1.1	.75	.46	.56	3
6 6	176	78	2.26	3.41	G -1.00	.13	1.34	1.8	1.36	1.7	.76	.44	.56	1
99 99	313	75	4.17	4.31	G .97	.19	1.35	1.6	1.18	.9	.85	.29	.31	12
34 34	268	78	3.44	4.13	G .42	.13	1.34	1.7	1.15	.9	.84	.52	.46	3
63 63	270	78	3.46	4.07	G .25	.14	1.33	1.7	1.24	1.4	.69	.35	.45	6
105 105	358	78	4.59	4.55	G 2.31	.42	1.32	.8	1.18	.5	.94	.13	.14	13
123 123	332	78	4.26	4.26	G .81	.20	1.23	1.1	1.32	1.4	.85	.27	.28	18
10 10	168	78	2.15	3.31	G -1.14	.13	1.31	1.6	1.17	.8	.75	.39	.55	1
25 25	284	78	3.64	4.28	G .86	.14	1.30	1.5	1.05	.3	1.06	.46	.42	2
51 51	296	78	3.79	3.79	G -.34	.15	1.23	1.1	1.29	1.7	.54	.08	.38	5
14 14	292	78	3.74	4.33	G 1.03	.15	1.28	1.3	1.05	.3	1.06	.52	.39	2
119 119	284	78	3.64	3.68	G -.54	.14	1.28	1.4	1.14	.9	.86	.28	.42	18
132 132	352	78	4.51	4.03	G .16	.31	1.28	.9	1.06	.2	.97	.23	.19	20
50 50	264	78	3.38	3.42	G -.98	.13	1.27	1.4	1.21	1.2	.76	.22	.47	5
4 4	287	78	3.68	4.28	G .88	.15	1.27	1.3	1.05	.3	1.03	.54	.41	1
108 108	356	78	4.56	4.39	G 1.29	.37	1.25	.7	1.18	.5	.95	.14	.16	15

90 90	266	78	3.41	3.83	G	-.26	.13	1.22	1.1	1.24	1.4	.59	.22	.46	10
30 30	274	75	3.65	4.25	G	.75	.15	1.20	1.0	1.20	1.2	.72	.33	.42	3
131 131	255	78	3.27	3.70	G	-.51	.13	1.19	1.0	1.11	.7	.82	.46	.49	19
3 3	170	78	2.18	3.34	G	-1.11	.13	1.19	1.0	1.00	.0	.87	.45	.55	1
65 65	321	78	4.12	4.25	G	.76	.18	1.19	.9	1.16	.8	.84	.32	.32	7
118 118	252	78	3.23	2.62	G	-1.96	.13	1.19	1.0	1.05	.3	.86	.34	.50	17
27 27	287	78	3.68	4.30	G	.93	.15	1.17	.9	1.00	.0	1.12	.48	.41	2
80 80	190	78	2.44	3.58	G	-.72	.13	1.16	.9	1.03	.2	.89	.51	.56	9
121 121	296	78	3.79	3.82	G	-.28	.15	1.16	.8	.89	-.6	1.18	.47	.38	18
11 11	303	78	3.88	4.39	G	1.29	.16	1.07	.3	1.16	.9	.81	.28	.36	2
40 40	182	78	2.33	3.26	G	-1.21	.13	1.16	.9	1.10	.5	.90	.39	.56	4
24 24	296	78	3.79	4.35	G	1.12	.15	1.15	.8	1.09	.5	.91	.44	.38	2
126 126	246	78	3.15	3.25	G	-1.23	.13	1.14	.8	1.05	.3	.90	.28	.51	18
26 26	248	78	3.18	4.05	G	.21	.13	1.14	.8	.94	-.3	1.16	.44	.50	2
97 97	298	78	3.82	4.04	G	.19	.15	1.12	.6	.97	-.1	1.00	.38	.38	12
92 92	283	78	3.63	4.06	G	.24	.14	1.11	.6	1.07	.5	.93	.32	.42	11
1 1	305	78	3.91	4.39	G	1.29	.16	1.10	.5	.99	.0	.98	.41	.36	1
64 64	257	78	3.29	3.97	G	.02	.13	1.10	.5	1.05	.3	.93	.52	.48	6
113 113	350	78	4.49	4.20	G	.60	.29	1.09	.4	1.07	.3	.95	.18	.20	16
69 69	312	78	4.00	4.15	G	.48	.17	1.09	.5	.99	.0	1.05	.40	.34	7
77 77	204	78	2.62	3.85	G	-.24	.13	1.05	.3	1.06	.4	.88	.52	.56	8
143 143	262	78	3.36	4.04	G	.18	.13	1.02	.1	.98	.0	1.16	.45	.47	23
70 70	241	78	3.09	4.11	G	.36	.13	1.01	.0	1.01	.1	.90	.27	.52	8
32 32	283	78	3.63	4.23	G	.71	.14	1.03	.2	.97	-.1	1.03	.43	.42	3
93 93	226	78	2.90	3.56	G	-.75	.13	.96	-.1	1.00	.0	1.03	.39	.54	11
82 82	191	78	2.45	3.59	G	-.70	.13	1.02	.1	.94	-.3	1.06	.28	.56	9
57 57	258	78	3.31	3.35	G	-1.08	.13	1.05	.3	.93	-.3	1.11	.47	.48	5
45 45	179	78	2.29	3.23	G	-1.26	.13	1.01	.0	.93	-.3	.98	.43	.56	4
85 85	277	78	3.55	4.24	G	.73	.14	1.03	.2	.93	-.4	1.04	.40	.43	9
88 88	231	78	2.96	3.93	G	-.06	.13	.92	-.4	1.00	.0	.94	.22	.53	9
38 38	257	78	3.29	3.96	G	.00	.13	.94	-.2	.91	-.5	1.21	.59	.48	4
39 39	275	78	3.53	4.09	G	.32	.14	.91	-.4	.90	-.5	1.20	.45	.44	4
66 66	282	78	3.62	3.85	G	-.23	.14	.93	-.3	.90	-.6	1.06	.37	.42	7
56 56	250	78	3.21	3.26	G	-1.22	.13	.93	-.3	.88	-.7	1.13	.48	.50	5
35 35	250	78	3.21	4.01	G	.11	.13	.96	-.1	.88	-.7	1.28	.75	.50	3
72 72	268	78	3.44	4.27	G	.82	.13	.95	-.2	.87	-.8	1.21	.55	.46	8
122 122	276	78	3.54	3.59	G	-.70	.14	1.00	.0	.87	-.8	1.32	.57	.44	18
103 103	317	78	4.06	3.87	G	-.18	.18	.86	-.6	.89	-.5	1.05	.31	.33	13
17 17	273	78	3.50	4.21	G	.65	.14	1.00	.0	.86	-.9	1.30	.62	.44	2
109 109	344	75	4.59	4.44	G	1.50	.42	.95	.0	.85	-.1	1.01	.20	.14	16
54 54	330	78	4.23	4.22	G	.67	.20	1.01	.1	.85	-.7	1.07	.43	.29	5
114 114	353	78	4.53	4.39	G	1.27	.32	1.10	.4	.85	-.3	1.01	.29	.18	17
146 146	239	78	3.06	3.86	G	-.20	.13	.91	-.4	.84	-1.0	1.28	.37	.52	23
81 81	252	78	3.23	4.08	G	.28	.13	.94	-.2	.83	-1.0	1.19	.42	.50	9
42 42	244	78	3.13	3.85	G	-.22	.13	.94	-.2	.83	-1.1	1.20	.44	.51	4
74 74	197	78	2.53	3.79	G	-.35	.13	.83	-.9	.82	-1.0	1.25	.53	.56	8
46 46	219	78	2.81	3.64	G	-.62	.13	.81	-1.0	.81	-1.2	1.26	.51	.55	4
60 60	252	78	3.23	3.28	G	-1.18	.13	.81	-1.0	.87	-.7	1.01	.21	.50	5
68 68	251	78	3.22	3.54	G	-.79	.13	.95	-.2	.81	-1.2	1.20	.50	.50	7
102 102	268	78	3.44	3.75	G	-.43	.13	.90	-.5	.80	-1.3	1.24	.45	.46	12
95 95	244	78	3.13	3.51	G	-.84	.13	.80	-1.1	.89	-.6	1.08	.24	.51	12
55 55	344	75	4.59	4.58	G	2.63	.42	.94	.0	.80	-.2	1.02	.23	.14	5
33 33	224	78	2.87	3.81	G	-.31	.13	.81	-1.0	.79	-1.3	1.26	.49	.54	3
75 75	199	78	2.55	3.80	G	-.31	.13	.78	-1.3	.80	-1.1	1.29	.53	.56	8
96 96	205	78	2.63	3.07	G	-1.46	.13	.77	-1.3	.80	-1.2	1.11	.52	.56	12
120 120	341	78	4.37	4.38	G	1.23	.23	1.16	.7	.77	-.9	1.12	.47	.25	18
36 36	217	78	2.78	3.75	G	-.42	.13	.77	-1.3	.78	-1.4	1.28	.67	.55	3
73 73	220	78	2.82	3.97	G	.02	.13	.76	-1.3	.81	-1.1	1.21	.51	.55	8
104 104	218	78	2.79	2.44	G	-2.13	.13	.76	-1.3	.79	-1.3	1.25	.50	.55	13
5 5	213	78	2.73	3.76	G	-.40	.13	.82	-1.0	.76	-1.5	1.29	.62	.55	1
83 83	268	78	3.44	4.18	G	.56	.13	.99	.0	.75	-1.7	1.46	.62	.46	9
21 21	233	78	2.99	3.94	G	-.03	.13	.74	-1.5	.79	-1.3	1.34	.61	.53	2
22 22	199	78	2.55	3.66	G	-.58	.13	.73	-1.6	.78	-1.3	1.20	.50	.56	2
2 2	233	78	2.99	3.92	G	-.09	.13	.80	-1.1	.72	-1.8	1.37	.69	.53	1
89 89	187	78	2.40	3.55	G	-.77	.13	.74	-1.5	.72	-1.6	1.19	.44	.56	9
8 8	282	78	3.62	4.25	G	.77	.14	.87	-.6	.72	-1.9	1.37	.57	.42	1

Etablissement de standards de performance 16/07/20 11:54:50
Table 7.2.2 Items Measurement Report (arranged by N).

Num Items	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq ZStd	Outfit MnSq ZStd	Estim. Discrm	Correlation PtMea PtExp	Group
Mean (Count: 149)	268.9	77.8	3.46	3.88	.00	.17	1.16 .6	1.07 .1		.40	
S.D. (Population)	53.6	1.2	.69	.41	.88	.07	.38 1.7	.41 1.7		.19	
S.D. (Sample)	53.8	1.2	.69	.41	.89	.07	.38 1.7	.42 1.7		.19	
Mean (Count: 10)	234.9	77.8	3.02	3.88	.00	.14	1.09 .4	.96 -.4		.52	1
S.D. (Population)	51.2	.6	.65	.40	.89	.01	.30 1.6	.25 1.6		.10	1
S.D. (Sample)	53.9	.6	.69	.42	.93	.01	.32 1.7	.27 1.7		.11	1
Mean (Count: 19)	233.4	78.0	2.99	3.89	.00	.14	1.04 .0	.95 -.5		.53	2
S.D. (Population)	47.8	.0	.61	.40	.83	.01	.35 1.9	.26 1.5		.10	2
S.D. (Sample)	49.2	.0	.63	.41	.85	.01	.36 1.9	.26 1.6		.10	2
Mean (Count: 8)	240.5	77.6	3.10	3.93	.00	.13	1.02 .1	.96 -.3		.51	3
S.D. (Population)	32.8	1.0	.44	.27	.58	.01	.24 1.3	.19 1.2		.13	3
S.D. (Sample)	35.1	1.1	.47	.29	.62	.01	.26 1.4	.21 1.3		.13	3
Mean (Count: 11)	247.5	76.6	3.23	3.89	.00	.14	1.03 .1	.98 -.2		.38	4
S.D. (Population)	49.2	3.5	.60	.39	.86	.02	.27 1.4	.22 1.3		.19	4
S.D. (Sample)	51.6	3.6	.63	.41	.90	.02	.29 1.5	.23 1.4		.20	4
Mean (Count: 13)	295.7	77.8	3.80	3.82	.00	.19	1.18 .8	1.13 .5		.27	5
S.D. (Population)	40.8	.8	.54	.52	1.23	.08	.30 1.4	.31 1.4		.20	5
S.D. (Sample)	42.5	.8	.56	.54	1.28	.08	.31 1.4	.33 1.4		.21	5
Mean (Count: 3)	255.3	78.0	3.27	3.95	.00	.13	1.30 1.6	1.26 1.5		.43	6
S.D. (Population)	12.7	.0	.16	.10	.22	.00	.16 .7	.18 1.0		.09	6
S.D. (Sample)	15.6	.0	.20	.12	.27	.00	.19 .9	.22 1.2		.09	6
Mean (Count: 5)	289.8	78.0	3.72	3.93	.00	.15	1.14 .7	1.00 .0		.44	7
S.D. (Population)	24.8	.0	.32	.25	.55	.02	.23 1.1	.14 .9		.10	7
S.D. (Sample)	27.7	.0	.36	.28	.62	.02	.25 1.2	.16 1.0		.11	7
Mean (Count: 8)	218.5	78.0	2.80	3.94	.00	.13	.83 -1.0	.84 -1.0		.50	8
S.D. (Population)	22.8	.0	.29	.15	.37	.00	.15 1.0	.13 .9		.09	8
S.D. (Sample)	24.4	.0	.31	.16	.40	.00	.16 1.0	.14 1.0		.10	8
Mean (Count: 12)	232.8	78.0	2.99	3.92	.00	.13	.99 -.2	.90 -.8		.44	9
S.D. (Population)	38.5	.0	.49	.28	.68	.01	.37 1.7	.29 1.7		.11	9
S.D. (Sample)	40.2	.0	.52	.29	.71	.01	.39 1.7	.31 1.7		.11	9
Mean (Count: 2)	279.0	78.0	3.58	3.95	.00	.14	1.33 1.7	1.24 1.5		.26	10
S.D. (Population)	13.0	.0	.17	.12	.26	.01	.12 .5	.00 .0		.05	10
S.D. (Sample)	18.4	.0	.24	.16	.36	.01	.17 .7	.00 .0		.07	10
Mean (Count: 3)	268.3	78.0	3.44	3.93	.00	.14	1.23 1.1	1.09 .6		.44	11
S.D. (Population)	30.4	.0	.39	.26	.55	.01	.29 1.3	.09 .5		.12	11
S.D. (Sample)	37.2	.0	.48	.32	.67	.01	.35 1.5	.11 .6		.14	11
Mean (Count: 8)	276.5	77.6	3.56	3.84	.00	.16	1.14 .3	1.00 -.3		.38	12
S.D. (Population)	45.1	1.0	.60	.47	1.15	.05	.46 1.8	.27 1.3		.12	12
S.D. (Sample)	48.3	1.1	.64	.50	1.23	.06	.49 1.9	.29 1.4		.13	12
Mean (Count: 3)	297.7	78.0	3.82	3.62	.00	.24	.98 -.4	.96 -.5		.31	13
S.D. (Population)	58.8	.0	.75	.88	1.82	.13	.25 .9	.16 .7		.15	13
S.D. (Sample)	72.0	.0	.92	1.08	2.22	.16	.30 1.2	.20 .9		.18	13
Mean (Count: 1)	338.0	78.0	4.33	3.96	.00	.22	1.81 3.2	2.11 3.8		-.11	14
S.D. (Population)	.0	.0	.00	.00	.00	.00	.00 .0	.00 .0		.00	14
Mean (Count: 2)	329.0	78.0	4.22	3.80	.00	.26	1.36 1.5	1.11 .4		.35	15
S.D. (Population)	27.0	.0	.35	.59	1.29	.11	.10 .7	.07 .1		.22	15
S.D. (Sample)	38.2	.0	.49	.84	1.83	.15	.15 .9	.10 .2		.31	15
Mean (Count: 5)	332.0	77.4	4.29	3.86	.00	.26	1.42 1.8	1.02 .3		.40	16
S.D. (Population)	17.3	1.2	.25	.45	1.00	.09	.40 1.5	.13 .5		.18	16
S.D. (Sample)	19.3	1.3	.29	.50	1.12	.10	.45 1.7	.14 .6		.20	16
Mean (Count: 5)	319.4	78.0	4.09	3.78	.00	.23	1.64 2.2	1.81 1.9		.28	17
S.D. (Population)	40.2	.0	.52	.70	1.31	.08	.47 1.4	1.28 2.2		.14	17
S.D. (Sample)	44.9	.0	.58	.79	1.46	.09	.53 1.5	1.43 2.4		.16	17
Mean (Count: 8)	295.0	78.0	3.78	3.82	.00	.18	1.19 1.0	1.16 .6		.28	18
S.D. (Population)	41.8	.0	.54	.50	1.20	.07	.15 .7	.31 1.4		.22	18
S.D. (Sample)	44.6	.0	.57	.53	1.29	.07	.16 .8	.34 1.5		.23	18
Mean (Count: 5)	280.8	78.0	3.60	3.94	.00	.14	1.61 2.7	1.51 2.7		.34	19
S.D. (Population)	20.5	.0	.26	.19	.45	.02	.29 1.0	.28 1.2		.10	19
S.D. (Sample)	22.9	.0	.29	.21	.50	.02	.32 1.2	.31 1.4		.11	19
Mean (Count: 6)	345.3	77.5	4.45	3.91	.00	.30	1.68 2.2	1.36 1.1		.24	20
S.D. (Population)	16.7	1.1	.16	.38	.72	.07	.31 1.1	.18 .5		.19	20
S.D. (Sample)	18.3	1.2	.17	.42	.79	.08	.34 1.2	.20 .6		.21	20
Mean (Count: 4)	345.0	78.0	4.42	3.90	.00	.29	1.51 1.9	1.49 1.7		.07	21
S.D. (Population)	14.5	.0	.19	.42	.79	.07	.33 1.4	.27 1.4		.10	21
S.D. (Sample)	16.8	.0	.22	.49	.92	.08	.38 1.6	.31 1.6		.11	21
Mean (Count: 1)	346.0	78.0	4.44	3.96	.00	.26	1.41 1.6	1.92 2.7		-.14	22
S.D. (Population)	.0	.0	.00	.00	.00	.00	.00 .0	.00 .0		.00	22
Mean (Count: 7)	250.0	77.6	3.22	3.95	.00	.13	.82 -1.1	.71 -2.1		.51	23
S.D. (Population)	12.6	1.0	.13	.08	.18	.00	.11 .7	.14 1.1		.07	23
S.D. (Sample)	13.6	1.1	.14	.09	.19	.00	.12 .8	.15 1.2		.08	23

Etablissement de standards de performance 16/07/20 11:54:50
Table 7.3.1 Juges Measurement Report (arranged by fN).

Nu Juges	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq ZStd	Outfit MnSq ZStd	Estim. Discrm	Correlation PtMea PtExp	Exact Agree. Obs % Exp %
8 <B	2781	894	3.11	3.67	1.03	.04	1.69 9.0	1.24 4.1	.98	.69 .69	39.7 33.3
4 <A	2838	885	3.21	3.75	1.18	.04	1.52 8.1	1.32 5.3	.84	.71 .67	32.6 32.2
2 <A	3349	888	3.77	4.16	2.08	.05	1.06 .9	1.40 5.2	.75	.46 .57	34.7 36.8
11 >C	3190	894	3.57	4.02	1.73	.04	1.21 3.3	1.34 5.1	.70	.48 .61	42.0 36.1
3 <A	3357	894	3.76	4.15	2.06	.05	.84 -2.8	1.21 2.9	.89	.65 .57	37.1 36.7
9 <B	2391	892	2.68	3.27	.39	.04	1.13 2.1	.94 -1.0	1.12	.79 .73	38.4 33.4
7 <B	2623	891	2.94	3.53	.79	.04	.92 -1.5	1.12 2.1	.96	.71 .70	34.5 33.8
5 <A	3082	885	3.48	3.96	1.60	.04	1.00 .0	1.05 .8	.93	.64 .63	34.0 34.9
10 <B	2772	894	3.10	3.66	1.02	.04	1.08 1.4	.87 -2.5	1.15	.78 .69	37.5 33.4
6 <A	3531	891	3.96	4.28	2.48	.05	.83 -2.9	1.20 2.5	1.07	.53 .52	38.2 37.2
1 <A	3504	891	3.93	4.27	2.42	.05	.92 -1.2	.82 -2.5	1.15	.58 .53	39.5 37.1
12 >C	3263	894	3.65	4.07	1.87	.04	.81 -3.3	.79 -3.6	1.19	.64 .60	43.2 36.5
13 >C	3384	894	3.79	4.17	2.11	.05	.66 -6.4	.69 -5.1	1.24	.62 .56	43.0 36.7
Mean (Count: 13)	3081.9	891.3	3.46	3.92	1.60	.04	1.05 .5	1.07 1.0		.64	
S.D. (Population)	350.5	3.2	.39	.30	.63	.00	.28 4.3	.22 3.5		.10	
S.D. (Sample)	364.8	3.4	.41	.31	.66	.00	.29 4.4	.23 3.7		.10	

Model, Populn: RMSE .04 Adj (True) S.D. .63 Separation 14.34 Strata 19.45 Reliability (not inter-rater) 1.00
 Model, Sample: RMSE .04 Adj (True) S.D. .66 Separation 14.92 Strata 20.23 Reliability (not inter-rater) 1.00
 Model, Fixed (all same) chi-square: 2673.4 d.f.: 12 significance (probability): .00
 Model, Random (normal) chi-square: 11.9 d.f.: 11 significance (probability): .37
 Inter-Rater agreement opportunities: 198516 Exact agreements: 73947 = 37.2% Expected: 70111.9 = 35.3%

Etablissement de standards de performance 16/07/20 11:54:50
Table 7.4.1 Rondes Measurement Report (arranged by fN).

N Rondes	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq ZStd	Outfit MnSq ZStd	Estim. Discrm	Displ.	Correlation PtMea PtExp
1 ronde 1	20399	5788	3.52	3.96 A	.00	.02	1.14 5.7	1.22 8.4	.82	.11	.63 .67
2 ronde 2	19666	5799	3.39	3.96 A	.00	.02	1.01 .4	.93 -2.9	1.18	-.11	.74 .67
Mean (Count: 2)	20032.5	5793.5	3.46	3.96	.00	.02	1.07 3.1	1.07 2.7			.68
S.D. (Population)	366.5	5.5	.07	.00	.00	.00	.06 2.7	.15 5.7			.05
S.D. (Sample)	518.3	7.8	.09	.00	.00	.00	.09 3.8	.21 8.0			.08

Model, Populn: RMSE .02 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
 Model, Sample: RMSE .02 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
 Model, Fixed (all same) chi-square: .0 d.f.: 1 significance (probability): 1.00

Etablissement de standards de performance 16/07/20 11:54:50
Table 7.5.1 Groupes Measurement Report (arranged by fN).

N Groupes	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq ZStd	Outfit MnSq ZStd	Estim. Discrm	Correlation PtMea PtExp
2 B	10567	3571	2.96	3.96 A	.00	.02	1.20 6.9	1.04 1.4	1.06	.75 .71
1 A	19661	5334	3.69	3.96 A	.00	.02	1.05 2.1	1.16 5.7	1.04	.68 .61
3 C	9837	2682	3.67	3.96 A	.00	.03	.90 -2.8	.94 -1.6	1.04	.52 .59
Mean (Count: 3)	13355.0	3862.3	3.44	3.96	.00	.02	1.05 2.1	1.05 1.9		.65
S.D. (Population)	4469.0	1102.1	.34	.00	.00	.00	.12 4.0	.09 3.0		.07
S.D. (Sample)	5473.3	1349.8	.41	.00	.00	.00	.15 4.9	.11 3.7		.09

Model, Populn: RMSE .02 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
 Model, Sample: RMSE .02 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
 Model, Fixed (all same) chi-square: .0 d.f.: 2 significance (probability): 1.00

Etablissement de standards de performance 16/07/20 11:54:50
Table 7.6.1 Criteres Measurement Report (arranged by fN).

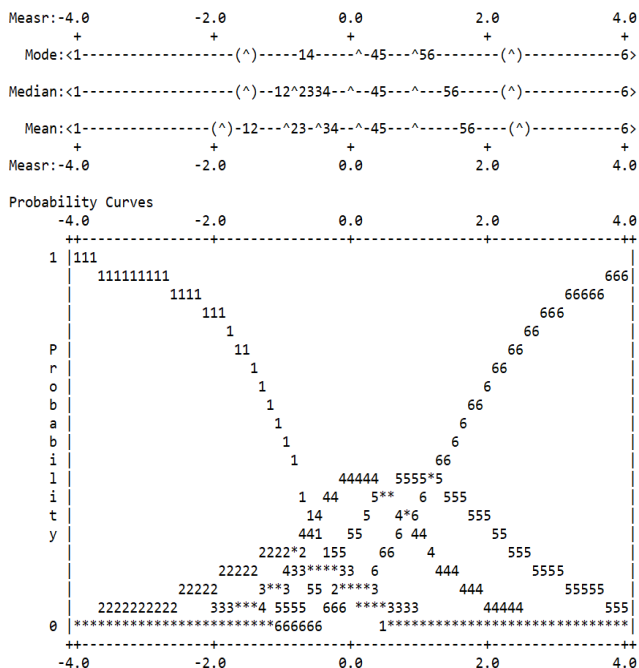
[illegible]

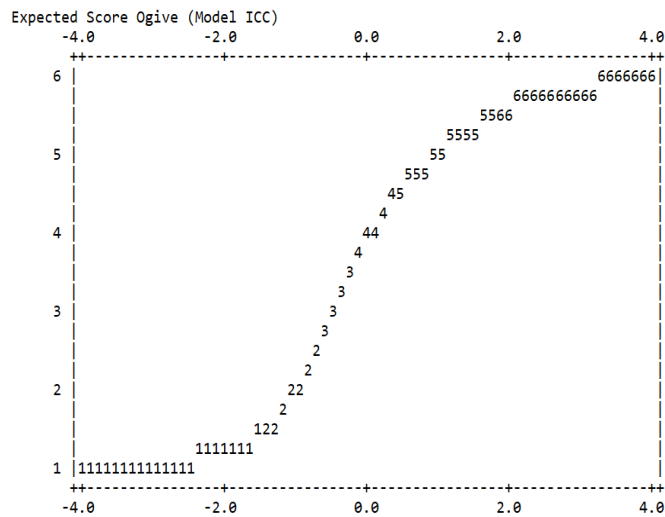
Etablissement de standards de performance 16/07/20 11:54:50
Table 8.1 Category Statistics.

Model = ?,?,?,?,1, PONDERATION ; Criteres: Ponderation
Rating (or partial credit) scale = PONDERATION,R10,G,0

Response	DATA					QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat	
Category	Category Counts		Cum.	Avg	Exp.	OUTFIT	Thresholds	Measure at	PROBABLE	THURSTONE	PEAK					
Name	Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob	
0-50 (0-50)	1	539	539	14%	14%	-1.00	-1.04	1.3			(-1.85)		low	low	100%	
51-60 (51-60)	2	235	235	6%	20%	-.48	-.63	1.6	-.01	.06	-.98	-1.42		-.96	19%	
61-70 (61-70)	3	251	251	6%	27%	-.03	-.21	1.6	-.49	.06	-.46	-.70		-.64	15%	
71-80 (71-80)	4	819	819	21%	48%	.21	.25	1.3	-1.17	.05	.07	-.21	-.56	-.40	41%	
81-90 (81-90)	5	848	848	22%	70%	.93	.91	1.5	.52	.04	.99	.45	.52	.41	39%	
91-100 severite (91-100)	6	1170	1170	30%	100%	2.11	2.18	1.1	1.15	.05	(2.52)	1.81	1.15	1.49	100%	
(Mean)											(Modal)					(Median)

Scale structure for "PONDERATION"





Etablissement de standards de performance 16/07/20 11:54:50

Table 8.2 Category Statistics.

Model = ?,?,?,?,?,2,DIFFICULTE ; Criteres: Difficulte
Rating (or partial credit) scale = DIFFICULTE,R4,G,0

Response Category Name	DATA					QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
	Category Counts		Cum.		Avg	Exp.	OUTFIT	Thresholds	Measure at	PROBABLE	THURSTONE	PEAK			
	Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
Tres difficile (4)	1	109	109	3%	3%	-0.35	-0.25	1.4	-1.63	.10	(-2.85)		low	low	100%
Difficile (3)	2	592	592	15%	18%	.41	.39	1.1	-0.17	.05	-0.95 -2.02	-1.63	-1.81	50%	
Moderne (2)	3	1529	1529	40%	58%	1.11	1.22	.7	-0.17	.05	.90 -0.06	-0.17	-0.11	56%	
Facile (1)	4	1633	1633	42%	100%	2.78	2.67	.9	1.80	.04	(2.97) 2.08	1.80	1.91	100%	
(Mean)											(Modal) -- (Median)				

Scale structure for "DIFFICULTE"

Measr:-4.0	-2.0	0.0	2.0	4.0
+	+	+	+	+
Mode:<1-----(^)-----12-----^-----23-----^-----34-----(^)-----4>				
Median:<1-----(^)-----12-----^-----23-----^-----34-----(^)-----4>				
Mean:<1-----(^)-----12-----^-----23-----^-----34-----(^)-----4>				
+	+	+	+	+
Measr:-4.0	-2.0	0.0	2.0	4.0

Etablissement de standards de performance 16/07/20 11:54:50
Table 8.3 Category Statistics.

Model = ?,?,?,?,?,3,EFFICIENCE ; Criteres: Efficiency
Rating (or partial credit) scale = EFFICIENCE,R4,S,0

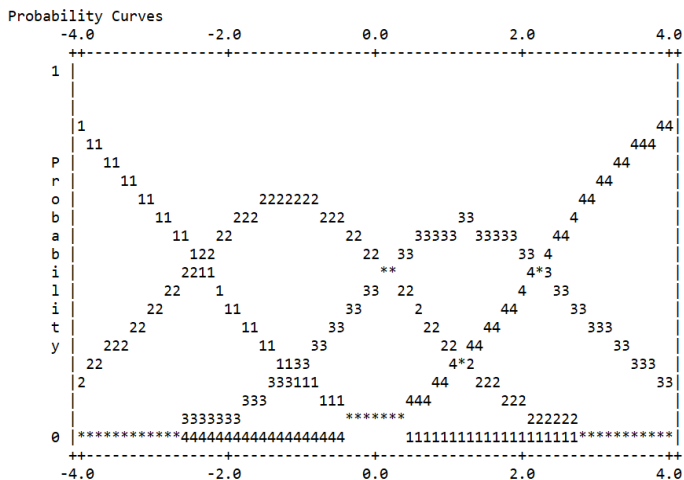
Response Category Name	DATA Category Score	Counts Total	Used	Cum. %	QUALITY CONTROL Avg	EXP. Meas	OUTFIT MnSq	RASCH-ANDRICH Thresholds Measure S.E.	EXPECTATION Measure at Category -0.5	MOST PROBABLE from	RASCH-THURSTONE Thresholds	Cat PEAK Prob
Faible (4)	1	116	116	3%	3%	-.90	-.55	.8	(-3.53)	low	low	100%
Moyenne (3)	2	1064	1064	28%	31%	.15	.17	1.1	-2.42 .10	-1.13 -2.58	-2.42	64%
Forte (2)	3	1617	1617	42%	72%	1.10	1.10	.8	.19 .04	1.23 .10	.19	58%
Tres forte (1)	4	1065	1065	28%	100%	2.71	2.66	.9	2.23 .05	(3.40) 2.49	2.23	100%
										(Mean)	(Modal)	(Median)

Scale structure for "EFFICIENCE"

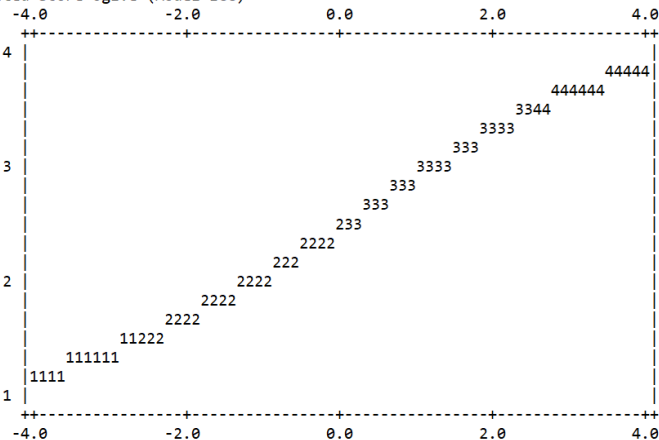
```

Measr:-4.0      -2.0      0.0      2.0      4.0
+-----+-----+-----+-----+
Mode:<1--(^)-----12-----^-----23-----^-----34-----(^)---4>
Median:<1--(^)-----12-----^-----23-----^-----34-----(^)---4>
Mean:<1--(^)-----12-----^-----23-----^-----34-----(^)---4>
+-----+-----+-----+-----+
Measr:-4.0      -2.0      0.0      2.0      4.0

```



Expected Score Ogive (Model ICC)



Etablissement de standards de performance 16/07/20 11:54:50

Table 4.1 Unexpected Responses (100 residuals sorted by 4,3,2,1,6).

Nu	Ta	Num	Ite	Nu	Ju	N	Rondes	N	G	N	Criteres	Cat	Score	Exp.	Resd	StRes
20	20	133	133	1	<A	1	ronde	1	1	A	2 Difficulte	3	3	3.9	-0.9	-3.8
5	5	49	49	2	<A	1	ronde	1	1	A	2 Difficulte	2	2	3.7	-1.7	-3.7
5	5	59	59	2	<A	1	ronde	1	1	A	2 Difficulte	3	3	3.9	-0.9	-3.8
7	7	65	65	2	<A	1	ronde	1	1	A	3 Efficience	1	1	3.6	-2.6	-4.5
9	9	87	87	2	<A	1	ronde	1	1	A	2 Difficulte	2	2	3.7	-1.7	-3.5
12	12	99	99	2	<A	1	ronde	1	1	A	2 Difficulte	2	2	3.8	-1.8	-4.3
18	18	123	123	2	<A	1	ronde	1	1	A	2 Difficulte	2	2	3.8	-1.8	-4.8
20	20	137	137	2	<A	1	ronde	1	1	A	1 Ponderation	5	5	6.0	-1.0	-4.7
20	20	137	137	2	<A	1	ronde	1	1	A	3 Efficience	2	2	3.9	-1.9	-7.5
21	21	138	138	2	<A	1	ronde	1	1	A	3 Efficience	2	2	3.9	-1.9	-5.4
21	21	139	139	2	<A	1	ronde	1	1	A	1 Ponderation	5	5	5.9	-0.9	-3.6
21	21	139	139	2	<A	1	ronde	1	1	A	3 Efficience	2	2	3.9	-1.9	-5.9
21	21	141	141	2	<A	1	ronde	1	1	A	1 Ponderation	5	5	6.0	-1.0	-4.3
21	21	141	141	2	<A	1	ronde	1	1	A	3 Efficience	2	2	3.9	-1.9	-7.0
5	5	58	58	3	<A	1	ronde	1	1	A	1 Ponderation	3	3	5.8	-2.8	-6.4
14	14	106	106	3	<A	1	ronde	1	1	A	1 Ponderation	3	3	5.8	-2.8	-6.4
13	13	105	105	3	<A	1	ronde	1	1	A	3 Efficience	2	2	3.9	-1.9	-6.4
20	20	133	133	5	<A	1	ronde	1	1	A	3 Efficience	2	2	3.7	-1.7	-3.5
20	20	135	135	5	<A	1	ronde	1	1	A	2 Difficulte	3	3	3.9	-0.9	-3.4
9	9	87	87	6	<A	1	ronde	1	1	A	2 Difficulte	2	2	3.8	-1.8	-4.2
17	17	117	117	6	<A	1	ronde	1	1	A	2 Difficulte	2	2	3.7	-1.7	-3.5
18	18	124	124	6	<A	1	ronde	1	1	A	1 Ponderation	5	5	6.0	-1.0	-4.7
18	18	124	124	6	<A	1	ronde	1	1	A	2 Difficulte	3	3	4.0	-1.0	-5.7
18	18	124	124	6	<A	1	ronde	1	1	A	3 Efficience	3	3	3.9	-0.9	-3.6
19	19	127	127	6	<A	1	ronde	1	1	A	2 Difficulte	2	2	3.8	-1.8	-4.5
2	2	11	11	11	>C	1	ronde	1	1	C	2 Difficulte	2	2	5.7	-3.7	-6.9
15	15	108	108	7	<B	1	ronde	1	2	B	1 Ponderation	4	4	5.8	-1.8	-4.4
17	17	114	114	7	<B	1	ronde	1	2	B	1 Ponderation	4	4	5.8	-1.8	-3.6
17	17	115	115	7	<B	1	ronde	1	2	B	1 Ponderation	2	2	5.6	-3.6	-5.9
20	20	132	132	7	<B	1	ronde	1	2	B	1 Ponderation	4	4	5.7	-1.7	-3.4
20	20	133	133	7	<B	1	ronde	1	2	B	1 Ponderation	3	3	5.5	-2.5	-3.9
20	20	135	135	7	<B	1	ronde	1	2	B	1 Ponderation	3	3	5.8	-2.8	-5.6
20	20	136	136	7	<B	1	ronde	1	2	B	1 Ponderation	3	3	5.8	-2.8	-6.8
20	20	136	136	7	<B	1	ronde	1	2	B	3 Efficience	2	2	3.7	-1.7	-3.8
1	1	6	6	8	<B	1	ronde	1	2	B	1 Ponderation	4	4	1.3	-1.7	-3.7
4	4	40	40	8	<B	1	ronde	1	2	B	1 Ponderation	4	4	1.4	2.6	3.3
4	4	45	45	8	<B	1	ronde	1	2	B	1 Ponderation	4	4	1.4	2.6	3.5
8	8	74	74	8	<B	1	ronde	1	2	B	1 Ponderation	5	5	1.6	3.4	3.5
8	8	75	75	8	<B	1	ronde	1	2	B	1 Ponderation	5	5	1.6	3.4	3.4
6	6	62	62	9	<B	1	ronde	1	2	B	1 Ponderation	6	6	1.6	4.4	4.5
16	16	110	110	9	<B	1	ronde	1	2	B	1 Ponderation	2	2	5.2	-3.2	-3.8
21	21	138	138	9	<B	1	ronde	1	2	B	1 Ponderation	3	3	5.6	-2.6	-4.0
2	2	11	11	11	>C	1	ronde	1	3	C	2 Difficulte	1	1	3.5	-2.5	-4.5
5	5	50	50	11	>C	1	ronde	1	3	C	2 Difficulte	1	1	3.2	-2.2	-3.4
14	14	106	106	11	>C	1	ronde	1	3	C	1 Ponderation	4	4	5.7	-1.7	-3.4
14	14	106	106	11	>C	1	ronde	1	3	C	2 Difficulte	2	2	3.8	-1.8	-4.7
20	20	132	132	11	>C	1	ronde	1	3	C	2 Difficulte	3	3	3.9	-0.9	-3.5
20	20	135	135	11	>C	1	ronde	1	3	C	2 Difficulte	3	3	3.9	-0.9	-3.7
21	21	139	139	11	>C	1	ronde	1	3	C	2 Difficulte	3	3	3.9	-0.9	-3.7
22	22	142	142	11	>C	1	ronde	1	3	C	1 Ponderation	4	4	5.8	-1.8	-4.4
22	22	142	142	11	>C	1	ronde	1	3	C	2 Difficulte	2	2	3.9	-1.9	-5.8
23	23	143	143	11	>C	1	ronde	1	3	C	2 Difficulte	1	1	3.2	-2.2	-3.4
16	16	113	113	13	>C	1	ronde	1	3	C	2 Difficulte	3	3	3.9	-0.9	-3.9
18	18	124	124	13	>C	1	ronde	1	3	C	1 Ponderation	5	5	5.9	-0.9	-3.8
18	18	123	123	1	<A	2	ronde	2	1	A	1 Ponderation	4	4	5.8	-1.8	-4.3
21	21	140	140	1	<A	2	ronde	2	1	A	1 Ponderation	4	4	5.7	-1.7	-3.4
5	5	59	59	2	<A	2	ronde	2	1	A	2 Difficulte	3	3	3.9	-0.9	-3.8
15	15	108	108	2	<A	2	ronde	2	1	A	3 Efficience	3	3	3.9	-0.9	-3.3
5	5	55	55	3	<A	2	ronde	2	1	A	1 Ponderation	5	5	6.0	-1.0	-4.9
5	5	58	58	3	<A	2	ronde	2	1	A	1 Ponderation	4	4	5.8	-1.8	-4.1
13	13	105	105	3	<A	2	ronde	2	1	A	1 Ponderation	5	5	6.0	-1.0	-5.1
15	15	108	108	3	<A	2	ronde	2	1	A	1 Ponderation	5	5	6.0	-1.0	-4.3
16	16	109	109	3	<A	2	ronde	2	1	A	1 Ponderation	5	5	6.0	-1.0	-5.0
17	17	114	114	3	<A	2	ronde	2	1	A	1 Ponderation	5	5	5.9	-0.9	-3.5
18	18	124	124	3	<A	2	ronde	2	1	A	1 Ponderation	5	5	5.9	-0.9	-3.7
20	20	133	133	3	<A	2	ronde	2	1	A	1 Ponderation	4	4	5.9	-1.9	-5.1
20	20	135	135	3	<A	2	ronde	2	1	A	1 Ponderation	5	5	5.9	-0.9	-3.5
20	20	136	136	3	<A	2	ronde	2	1	A	1 Ponderation	5	5	6.0	-1.0	-4.3
20	20	137	137	3	<A	2	ronde	2	1	A	1 Ponderation	5	5	6.0	-1.0	-4.6
21	21	139	139	3	<A	2	ronde	2	1	A	1 Ponderation	5	5	5.9	-0.9	-3.5
21	21	141	141	3	<A	2	ronde	2	1	A	1 Ponderation	5	5	6.0	-1.0	-4.3
10	10	91	91	4	<A	2	ronde	2	1	A	1 Ponderation	1	1	4.5	-3.5	-3.3
19	19	127	127	4	<A	2	ronde	2	1	A	1 Ponderation	1	1	5.1	-4.1	-4.7
21	21	140	140	4	<A	2	ronde	2	1	A	1 Ponderation	1	1	5.2	-4.2	-5.0
2	2	13	13	5	<A	2	ronde	2	1	A	1 Ponderation	5	5	1.5	3.5	3.7
2	2	16	16	5	<A	2	ronde	2	1	A	1 Ponderation	5	5	1.5	3.5	3.8
2	2	19	19	5	<A	2	ronde	2	1	A	1 Ponderation	5	5	1.5	3.5	4.0
2	2	23	23	5	<A	2	ronde	2	1	A	1 Ponderation	5	5	1.4	3.6	4.3
21	21	140	140	5	<A	2	ronde	2	1	A	1 Ponderation	3	3	5.4	-2.4	-3.3
17	17	116	116	6	<A	2	ronde	2	1	A	1 Ponderation	4	4	5.9	-1.9	-8.2
17	17	116	116	6	<A	2	ronde	2	1	A	2 Difficulte	1	1	4.0	-3.0	-9.0
20	20	132	132	7	<B	2	ronde	2	2	B	1 Ponderation	4	4	5.7	-1.7	-3.4
4	4	47	47	8	<B	2	ronde	2	2	B	1 Ponderation	1	1	4.7	-3.7	-3.9
4	4	48	48	8	<B	2	ronde	2	2	B	1 Ponderation	1	1	4.9	-3.9	-4.1
5	5	49	49	8	<B	2	ronde	2	2	B	1 Ponderation	1	1	5.0	-4.0	-4.4
5	5	52	52	8	<B	2	ronde	2	2	B	1 Ponderation	1	1	4.8	-3.8	-4.0
12	12	99	99	8	<B	2	ronde	2	2	B	1 Ponderation	2	2	5.2	-3.2	-3.9
12	12	101	101	8	<B	2	ronde	2	2	B	1 Ponderation	1	1	4.6	-3.6	-3.5
15	15	107	107	8	<B	2	ronde	2	2	B	1 Ponderation	1	1	4.6	-3.6	-3.5
16	16	111	111	8	<B	2	ronde	2	2	B	1 Ponderation	1	1	4.6	-3.6	-3.5
16	16	112	112	8	<B	2	ronde	2	2	B	1 Ponderation	1	1	5.2	-4.2	-5.0
20	20	134	134	8	<B	2	ronde	2	2	B	1 Ponderation	2	2	5.1	-3.1	-3.7
16	16	112	112	9	<B	2	ronde	2	2	B	1 Ponderation	1	1	4.6	-3.6	-3.5
17	17	115	115	9	<B	2	ronde	2	2	B	1 Ponderation	2	2	5.4	-3.4	-4.8
20	20	134	134	9	<B	2	ronde	2	2	B	1 Ponderation	1	1	4.5	-3.5	-3.4
16	16	112	112	10	<B	2	ronde	2	2	B	1 Ponderation	2	2	5.1	-3.1	-3.8
20	20	134	134	10	<B	2	ronde	2	2	B	1 Ponderation	2	2	5.1	-3.1	-3.6
2	2	19	19	11	>C	2	ronde	2	3	C	1 Ponderation	5	5	1.6	3.4	3.4
15	15	108	108	11	>C	2										

PHASE II

1^{re} ANALYSE DES SCORES BRUTS DES ZONES D'INCERTITUDE

Etablissement de standards de performance 28/06/20 19:43:50
Table 5. Measurable Data Summary.

	Cat	Score	Exp.	Resd	StRes
Mean (Count: 3576)	3.43	3.43	3.43	.00	.00
Mean (Count: 3576)	1.06	1.06	.88	.56	.99
Mean (Count: 3576)	1.06	1.06	.88	.56	.99

Data log-likelihood chi-square = 5782.2568

Approximate degrees of freedom = 3476

Chi-square significance prob. = .0000

	Count	Mean	S.D.	Params
Responses used for estimation	= 3576	3,43	1,06	100
Count of measurable responses	= 3576			
Raw-score variance of observations	= 1,12	100.00%		
Variance explained by Rasch measures	= 0,80	71,64%		
Variance of residuals	= 0,32	28,36%		

Etablissement de standards de performance 28/06/20 19:43:50
Table 6.0 All Facet Vertical "Rulers".

Vertical = (1N,2N,3N,4N,5A,6A,7A,SL) Yardstick (columns lines low high extreme)= 0,15,-2.6,2.6,End

Measr	+Taches	+Items	+Juges	+Rondes	+Groupes	+Zones	+Criteres	S.1	S.2	S.3
	20 16	48 49 58 87	14					(6) 91-100 severit 5 81-90	(4) Facile	(4) Tres fort
2	+	+	+	+	+	+	+	+	+	+
		11 101 25 112							---	3 Forte
5			18 15 17 20							
		7 24 27								
1	21	121 122 30 32 143	16	+	+	+	+	---	+	+
	7		19							
	15	94 97								
4										
1		4 127 134 76 17								
17							Difficulte Efficiency	4 71-80	3 Modere	---
*	0	11	* 67 74 91 104 107 117 140 *	* 1 2 3 *	0	* 1 2 3 4 *				
	18 10 19	8 77 34 119 133 20 128					Ponderation			
		29 31 39						---		
3										
		92								
2		147								
12		12 98								
23										
-1	+	18	+	+	+	+	+			
		33 86 21 43 80 9						3 61-70		
		96						---		
		111 125								
		51							---	2 Moyenne
-2	+	+	+	+	+	+	+			
		46 22								
	13 9							2 51-60		
	8	61						(1) 0-50	(2) Difficile	(1) Faible
Measr	+Taches	+Items	+Juges	+Rondes	+Groupes	+Zones	+Criteres	S.1	S.2	S.3

S.1: Model = ?,?,?,?,?,1,PONDERATION ; Criteres: Ponderation
S.2: Model = ?,?,?,?,?,2,DIFFICULTE ; Criteres: Difficulte
S.3: Model = ?,?,?,?,?,3,EFFICIENCE ; Criteres: Efficiency

Etablissement de standards de performance 28/06/20 19:43:50
Table 7.2.1 Items Measurement Report (arranged by fN).

Num Items	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq ZStd	Outfit MnSq ZStd	Estim. Discrm	Correlation PtMea PtExp	Group
87 Z2	196	54	3.63	4.27 G	2.40	.24	2.42 4.9	2.84 6.3	-.63	-.01 .16	9
96 Z4	159	62	2.56	3.03 G	-1.62	.18	1.92 4.0	2.61 6.8	-1.16	.17 .26	12
127 Z2	225	62	3.63	3.73 G	.33	.23	2.24 4.7	2.60 6.0	-.48	.00 .18	19
76 Z4	158	62	2.55	3.69 G	.21	.18	1.71 3.2	2.36 6.0	-.82	.16 .27	8
117 Z2	228	62	3.68	3.63 G	.00	.23	1.71 3.0	1.90 3.8	.19	.02 .18	17
8 Z2	225	62	3.63	3.57 G	-.19	.23	1.72 3.1	1.81 3.5	.33	.13 .18	1
77 Z4	145	62	2.34	3.56 G	-.22	.19	1.30 1.5	1.58 3.0	.27	.20 .28	8
39 Z2	210	59	3.56	3.45 G	-.53	.23	1.49 2.1	1.51 2.3	.51	.14 .18	4
22 Z4	145	62	2.34	2.78 G	-2.17	.19	1.26 1.3	1.45 2.4	.37	.21 .28	2
74 Z4	152	62	2.45	3.63 G	.01	.18	1.08 .4	1.40 2.2	.51	.10 .27	8
51 Z2	213	62	3.44	2.88 G	-1.94	.21	1.32 1.5	1.39 1.9	.65	.13 .18	5
111 Z2	239	62	3.85	2.98 G	-1.72	.24	1.34 1.6	1.18 .9	.93	.22 .18	16
30 Z2	231	62	3.73	3.90 G	.96	.23	1.28 1.3	1.34 1.6	.68	.15 .18	3
24 Z2	232	62	3.74	3.97 G	1.22	.23	1.22 1.0	1.31 1.5	.79	.10 .18	2
143 Z2	222	62	3.58	3.87 G	.84	.22	1.31 1.4	1.30 1.5	.69	-.02 .18	23
25 Z2	241	62	3.89	4.10 G	1.74	.24	1.29 1.4	1.30 1.5	.69	.05 .19	2
94 Z2	237	62	3.82	3.84 G	.72	.24	1.25 1.2	1.27 1.3	.69	.13 .18	11
27 Z2	230	62	3.71	3.94 G	1.12	.23	1.26 1.3	1.17 .9	.91	.09 .18	2
58 Z1	278	62	4.48	4.25 G	2.40	.36	1.22 .9	.84 -.1	.90	.12 .16	5
133 Z1	276	62	4.45	3.54 G	-.26	.34	1.21 1.0	.93 .0	.86	.00 .17	20
164 Z4	166	62	2.68	3.63 G	.00	.18	1.02 .1	1.20 1.1	.71	.18 .25	13
121 Z2	240	62	3.87	3.91 G	.99	.24	1.16 .8	1.20 1.0	.80	.04 .18	18
80 Z4	127	62	2.05	3.17 G	-1.30	.20	1.18 .9	1.03 .2	.96	.23 .28	9
101 Z2	241	62	3.89	4.12 G	1.83	.24	1.15 .7	1.06 .3	1.06	.08 .19	12
140 Z2	242	62	3.90	3.63 G	.00	.25	1.10 .5	1.09 .5	.88	.05 .19	21
122 Z2	240	62	3.87	3.91 G	.99	.24	1.07 .3	.95 -.2	1.16	.08 .18	18
107 Z2	238	62	3.84	3.63 G	.00	.24	1.00 .0	1.05 .3	.96	.03 .18	15
32 Z2	231	62	3.73	3.90 G	.96	.23	.96 -.1	.97 -.1	.98	.16 .18	3
4 Z2	235	62	3.79	3.73 G	.35	.24	.95 -.1	.96 -.1	1.02	.04 .18	1
67 Z2	240	62	3.87	3.63 G	.00	.24	.87 -.6	.89 -.5	1.11	.08 .18	7
86 Z4	132	62	2.13	3.24 G	-1.10	.20	.96 -.1	.84 -.9	1.23	.23 .28	9
11 Z2	243	62	3.92	4.13 G	1.86	.25	.76 -1.2	.76 -1.3	1.24	.12 .19	2
49 Z1	281	62	4.53	4.34 G	2.84	.41	.99 .0	.74 -.2	1.02	.05 .14	5
112 Z1	281	62	4.53	4.09 G	1.72	.41	.99 .0	.74 -.2	1.02	.06 .14	16
48 Z1	281	62	4.53	4.51 G	3.88	.41	.99 .0	.74 -.2	1.02	.06 .14	4
134 Z1	280	62	4.52	3.71 G	.26	.39	.97 .0	.69 -.4	1.05	.10 .14	20
119 Z2	216	62	3.48	3.54 G	-.27	.22	.59 -2.3	.63 -2.2	1.32	.17 .18	18
7 Z2	249	62	4.02	3.96 G	1.19	.25	.62 -2.2	.58 -2.5	1.42	.08 .19	1
97 Z2	221	62	3.56	3.84 G	.74	.22	.58 -2.4	.57 -2.6	1.36	.10 .18	12
91 Z2	220	62	3.55	3.63 G	.00	.22	.56 -2.6	.55 -2.8	1.38	.09 .18	10
29 Z3	194	62	3.13	3.46 G	-.50	.19	.52 -2.8	.49 -3.4	1.55	.19 .20	2
128 Z2	211	62	3.40	3.52 G	-.33	.21	.49 -3.1	.53 -3.0	1.41	.10 .18	19
20 Z3	198	62	3.19	3.52 G	-.35	.20	.51 -2.9	.47 -3.6	1.58	.17 .20	2
33 Z3	173	59	2.93	3.22 G	-1.16	.19	.49 -3.1	.46 -3.7	1.59	.29 .23	3
92 Z2	208	62	3.35	3.39 G	-.72	.21	.48 -3.1	.46 -3.6	1.54	.20 .19	11
9 Z3	199	62	3.21	3.14 G	-1.35	.20	.51 -2.9	.45 -3.7	1.59	.19 .20	1
61 Z3	178	62	2.87	2.22 G	-3.29	.18	.46 -3.4	.45 -4.0	1.60	.26 .23	5
17 Z2	210	62	3.39	3.67 G	.15	.21	.46 -3.3	.43 -3.9	1.56	.11 .18	2
31 Z3	199	62	3.21	3.46 G	-.52	.20	.49 -3.0	.42 -4.0	1.61	.27 .20	3
46 Z3	181	62	2.92	2.78 G	-2.15	.19	.45 -3.5	.42 -4.2	1.63	.40 .23	4
147 Z3	181	62	2.92	3.34 G	-.84	.19	.41 -3.8	.44 -4.1	1.59	.31 .23	23
18 Z3	181	62	2.92	3.30 G	-.97	.19	.41 -3.8	.44 -4.1	1.59	.31 .23	2
125 Z3	171	59	2.90	2.99 G	-1.71	.19	.44 -3.6	.41 -4.2	1.64	.41 .23	18
43 Z2	206	62	3.32	3.20 G	-1.20	.21	.46 -3.2	.41 -4.2	1.60	.20 .19	4
12 Z3	174	59	2.95	3.32 G	-.90	.19	.40 -3.9	.41 -4.2	1.62	.43 .23	2
34 Z2	206	62	3.32	3.55 G	-.24	.21	.43 -3.5	.40 -4.3	1.61	.31 .19	3
98 Z3	179	62	2.89	3.30 G	-.95	.19	.39 -4.1	.42 -4.3	1.61	.33 .23	12
21 Z3	174	62	2.81	3.20 G	-1.21	.18	.37 -4.4	.35 -5.1	1.71	.42 .24	2
Mean (Count: 58)	211.5	61.7	3.43	3.58	.00	.23	.97 -.5	.99 -.5		.16	
S.D. (Population)	38.5	1.3	.61	.43	1.34	.06	.48 2.5	.59 3.1		.11	
S.D. (Sample)	38.8	1.3	.62	.43	1.35	.06	.48 2.5	.60 3.1		.11	

Model, Populn: RMSE .24 Adj (True) S.D. 1.32 Separation 5.52 Strata 7.69 Reliability .97
Model, Sample: RMSE .24 Adj (True) S.D. 1.33 Separation 5.57 Strata 7.76 Reliability .97
Model, Fixed (all same) chi-square: 1776.9 d.f.: 57 significance (probability): .00
Model, Random (normal) chi-square: 54.7 d.f.: 56 significance (probability): .52

Etablissement de standards de performance 28/06/20 19:43:50
Table 7.3.1 Juges Measurement Report (arranged by fn).

Nu Juges	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq ZStd	Outfit MnSq ZStd	Estim. Discrm	Correlation PtMea PtExp	Exact Agree. Obs % Exp %
20 >D	1783	522	3.42	3.61	1.28	.07	.92 -1.0	1.16 1.8	.83	.68 .68	59.0 49.5
19 >D	1707	522	3.27	3.48	.87	.07	1.14 1.9	1.16 2.0	.86	.71 .70	61.5 48.0
14 >D	1937	519	3.73	3.90	2.25	.08	1.12 1.6	1.01 .0	.99	.56 .63	58.2 48.1
18 >D	1798	516	3.48	3.68	1.48	.08	1.03 .4	1.05 .5	.83	.74 .67	56.6 49.8
16 >D	1732	522	3.32	3.53	1.00	.07	.92 -1.1	.97 -.3	1.03	.73 .69	61.7 48.6
17 >D	1795	522	3.44	3.64	1.34	.07	.76 -3.7	.87 -1.5	1.15	.67 .68	59.8 49.7
15 >D	1517	453	3.35	3.51	1.33	.08	.54 -7.3	.62 -4.8	1.36	.75 .67	66.1 50.8
Mean (Count: 7)	1752.7	510.9	3.43	3.62	1.37	.08	.92 -1.3	.98 -.3		.69	
S.D. (Population)	117.6	23.7	.14	.13	.41	.00	.20 3.0	.17 2.2		.06	
S.D. (Sample)	127.0	25.6	.15	.14	.44	.00	.21 3.3	.19 2.4		.06	

Model, Populn: RMSE .08 Adj (True) S.D. .40 Separation 5.32 Strata 7.43 Reliability (not inter-rater) .97
Model, Sample: RMSE .08 Adj (True) S.D. .44 Separation 5.76 Strata 8.01 Reliability (not inter-rater) .97
Model, Fixed (all same) chi-square: 197.1 d.f.: 6 significance (probability): .00
Model, Random (normal) chi-square: 5.8 d.f.: 5 significance (probability): .32
Inter-Rater agreement opportunities: 25035 Exact agreements: 15106 = 60.3% Expected: 12313.8 = 49.2%

Etablissement de standards de performance 28/06/20 19:43:50
Table 7.4.1 Rondes Measurement Report (arranged by fn).

N Rondes	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq ZStd	Outfit MnSq ZStd	Estim. Discrm	Correlation Displ. PtMea PtExp
1 ronde 1	3879	1152	3.37	3.59 A	.00	.05	1.28 5.3	1.26 4.0	.74	-.11 .66 .68
3 ronde 3	4202	1209	3.48	3.64 A	.00	.05	.77 -5.4	.85 -2.7	1.12	.09 .71 .68
2 ronde 2	4188	1215	3.45	3.64 A	.00	.05	.76 -5.5	.85 -2.7	1.13	.02 .71 .68
Mean (Count: 3)	4089.7	1192.0	3.43	3.63	.00	.05	.94 -1.9	.99 -.5		.69
S.D. (Population)	149.1	28.4	.05	.02	.00	.00	.24 5.1	.19 3.2		.03
S.D. (Sample)	182.6	34.8	.06	.03	.00	.00	.29 6.3	.23 3.9		.03

Model, Populn: RMSE .05 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Sample: RMSE .05 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Fixed (all same) chi-square: .0 d.f.: 2 significance (probability): 1.00

Etablissement de standards de performance 28/06/20 19:43:50
Table 7.6.1 Zones Measurement Report (arranged by fn).

N Zones	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq ZStd	Outfit MnSq ZStd	Estim. Discrm	Correlation PtMea PtExp
4 4	1184	496	2.39	3.63 A	.00	.07	1.31 4.3	1.56 8.2	.24	.28 .36
2 2	7026	1911	3.68	3.63 A	.00	.04	1.06 1.5	1.10 2.9	.92	.31 .35
1 1	1677	372	4.51	3.63 A	.00	.16	1.08 .7	.78 -.9	.98	.10 .17
3 3	2382	797	2.99	3.63 A	.00	.05	.45 -9.0	.43 -9.0	1.61	.40 .28
Mean (Count: 4)	3067.3	894.0	3.39	3.63	.00	.08	.98 -.6	.97 .3		.27
S.D. (Population)	2324.9	607.2	.79	.00	.00	.05	.32 5.1	.42 6.3		.11
S.D. (Sample)	2684.6	701.1	.91	.00	.00	.05	.37 5.8	.48 7.3		.12

Model, Populn: RMSE .09 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Sample: RMSE .09 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Fixed (all same) chi-square: .0 d.f.: 3 significance (probability): 1.00

tablissement de standards de performance 28/06/20 19:43:50
 Table 7.7.1 Criteres Measurement Report (arranged by FN).

N Criteres	Total Score	Total Count	Obsvd Average	Fair(M) Average	Measure	Model S.E.	Infit MnSq	Outfit ZStd	Estim. MnSq	Correlation ZStd	Discrm	PtMea	PtExp
3 Efficiency	3444	1211	2.84	2.93	.07	.06	1.47	9.0	1.46	8.7	.56	.55	.70
1 Ponderation	4838	1154	4.19	4.55	-.20	.04	.78	-5.3	.88	-2.8	1.15	.80	.77
2 Difficulte	3987	1211	3.29	3.41	.13	.06	.66	-9.0	.60	-6.2	1.45	.75	.63
Mean (Count: 3)	4089.7	1192.0	3.44	3.63	.00	.05	.97	-1.8	.98	-.2		.70	
S.D. (Population)	573.7	26.9	.56	.68	.14	.01	.36	7.8	.36	6.4		.11	
S.D. (Sample)	702.6	32.9	.69	.83	.17	.01	.44	9.5	.44	7.9		.13	

Model, Populn: RMSE .05 Adj (True) S.D. .13 Separation 2.51 Strata 3.69 Reliability .86
 Model, Sample: RMSE .05 Adj (True) S.D. .17 Separation 3.16 Strata 4.55 Reliability .91
 Model, Fixed (all same) chi-square: 27.3 d.f.: 2 significance (probability): .00
 Model, Random (normal) chi-square: 1.9 d.f.: 1 significance (probability): .17

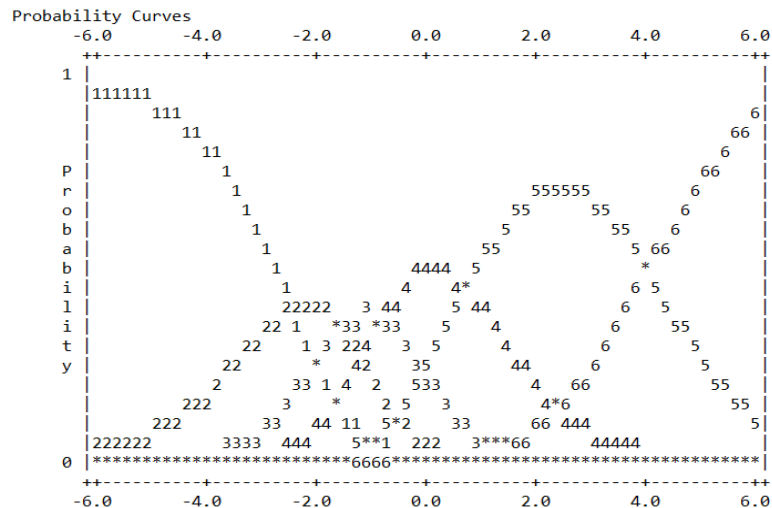
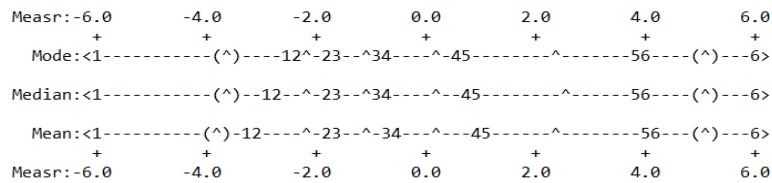
tablissement de standards de performance 28/06/20 19:43:50
 Table 8.1 Category Statistics.

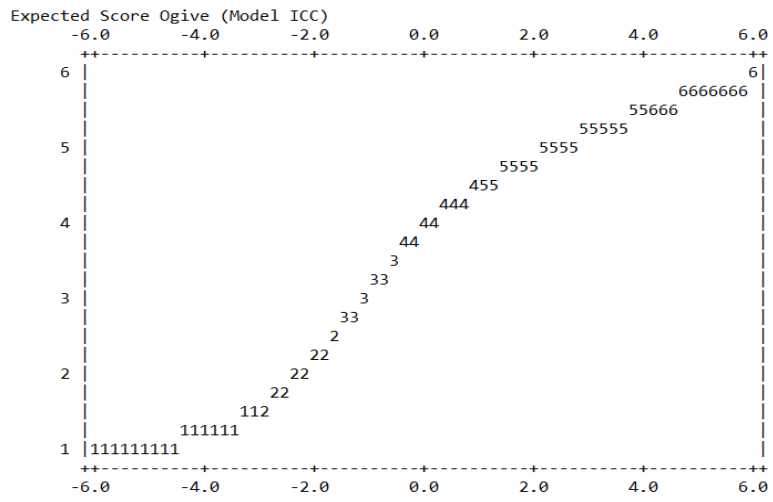
Model = ?,?,?,?,?,1,PONDERATION ; Criteres: Ponderation
 Rating (or partial credit) scale = PONDERATION,R6,G,0

Response Category Name	DATA Category Score	Counts Total	Used	%	Cum. %	QUALITY CONTROL Avg	EXP. OUTFIT	RASCH-ANDRICH Thresholds	EXPECTATION Measure at	MOST PROBABLE from	RASCH-THURSTONE Thresholds	Cat PEAK Prob
0-50 (0-50)	1	40	40	3%	3%	-2.18	-1.91	.6	(-3.76)	low	low	100%
51-60 (51-60)	2	89	89	8%	11%	-1.32	-1.35	1.4	-2.21	-3.03	-2.45	41%
61-70 (61-70)	3	156	156	14%	25%	-.80	-.58	.5	-1.10	-1.63	-1.55	38%
71-80 (71-80)	4	322	322	28%	53%	.45	.46	.7	-.14	-.54	-.80	50%
81-90 (81-90)	5	418	418	36%	89%	1.81	1.66	1.0	2.43	1.05	.76	72%
91-100 severeite (91-100)	6	129	129	11%	100%	4.13	4.29	1.1	4.03	4.13	4.03	100%

(Mean) (Modal) (Median)

Scale structure for "PONDERATION"





Etablissement de standards de performance 28/06/20 19:43:50

Table 8.3 Category Statistics.

Model = ?,?,?,?,?,3,EFFICIENCE ; Criteres: Efficience

Rating (or partial credit) scale = EFFICIENCE,R4,S,0

Response	DATA					QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat	
Category	Category Counts			Cum.		Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK	
Name	Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob	
Faible (4)	1	27	27	2%	2%	.76	-1.37	2.7			(-4.50)		low	low	100%	
Moyenne (3)	2	338	338	28%	30%	-.13*	-.28	1.2	-3.41	.21	-1.77	-3.48	-3.41	-3.44	72%	
Forte (2)	3	643	643	53%	83%	1.26	1.27	1.1	-.16	.08	1.69	-.11	-.16	-.14	76%	
Tres forte (1)	4	203	203	17%	100%	3.42	3.92	1.7	3.57	.11	(4.66)	3.62	3.57	3.58	100%	
										(Mean)	(Modal)		(Median)			

Scale structure for "EFFICIENCE"

```

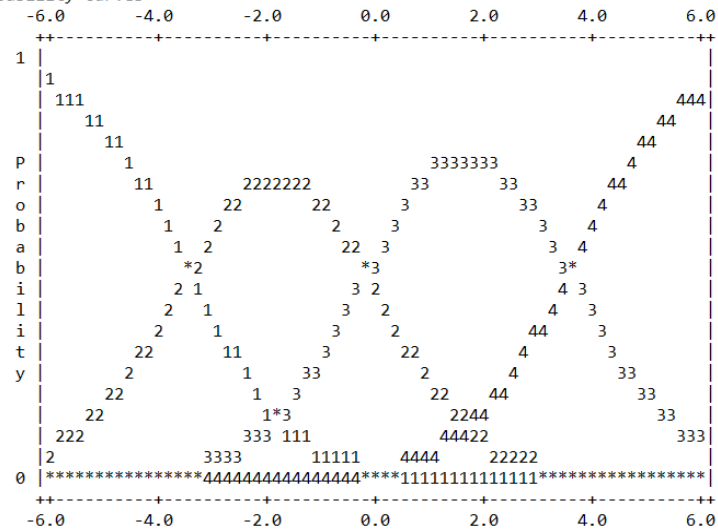
Measr:-6.0      -4.0      -2.0      0.0      2.0      4.0      6.0
      +          +          +          +          +          +
      Mode:<1-----(^)---12-----^-----23-----^-----34-----(^)---4>

Median:<1-----(^)---12-----^-----23-----^-----34-----(^)---4>

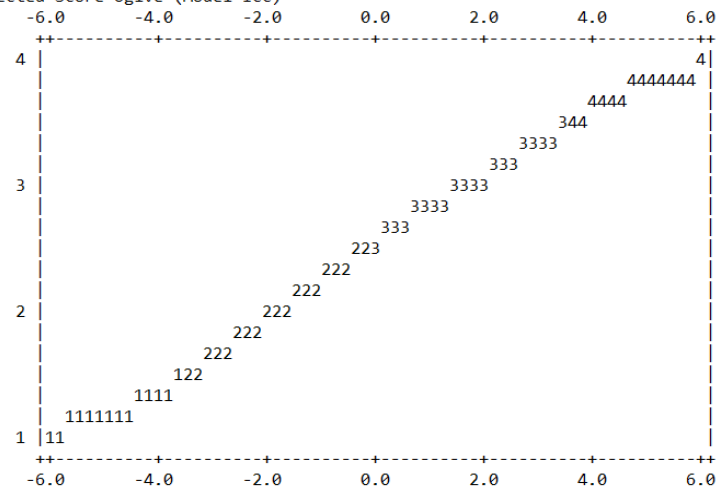
Mean:<1-----(^)---12-----^-----23-----^-----34-----(^)---4>
      +          +          +          +          +          +
Measr:-6.0      -4.0      -2.0      0.0      2.0      4.0      6.0

```

Probability Curves



Expected Score Ogive (Model ICC)



Etablissement de standards de performance 28/06/20 19:43:50

Table 4.1 Unexpected Responses (100 residuals sorted by 3,2,1,4,5,6,7).

Nu	Ta	Num	It	Nu	Ju	N	Rondes	N	G	N	Z	N	Criteres	Cat	Score	Exp.	Resd	StRes	
2	2	22	Z4	14	>D	2	ronde	2	4	D	4	4	3	Efficiency	4	4	2.3	1.7	3.0
2	2	25	Z2	14	>D	2	ronde	2	4	D	2	2	3	Efficiency	2	2	3.4	-1.4	-2.7
3	3	30	Z2	14	>D	3	ronde	3	4	D	2	2	3	Efficiency	2	2	3.2	-1.2	-2.4
4	4	48	Z1	14	>D	1	ronde	1	4	D	1	1	1	Ponderation	5	5	5.9	-.9	-3.1
5	5	49	Z1	14	>D	1	ronde	1	4	D	1	1	1	Ponderation	5	5	5.9	-.9	-3.1
5	5	58	Z1	14	>D	1	ronde	1	4	D	1	1	1	Ponderation	5	5	5.9	-.9	-2.5
8	8	76	Z4	14	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.5	1.5	2.7
9	9	80	Z4	14	>D	1	ronde	1	4	D	4	4	1	Ponderation	5	5	2.5	2.5	2.5
9	9	86	Z4	14	>D	1	ronde	1	4	D	4	4	1	Ponderation	5	5	2.7	2.3	2.3
9	9	87	Z2	14	>D	2	ronde	2	4	D	2	2	3	Efficiency	2	2	3.2	-1.2	-2.3
9	9	87	Z2	14	>D	3	ronde	3	4	D	2	2	3	Efficiency	2	2	3.2	-1.2	-2.3
12	12	96	Z4	14	>D	1	ronde	1	4	D	4	4	3	Efficiency	4	4	2.5	1.5	2.7
12	12	96	Z4	14	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.5	1.5	2.7
16	16	112	Z1	14	>D	1	ronde	1	4	D	1	1	1	Ponderation	5	5	5.9	-.9	-3.1
17	17	117	Z2	14	>D	2	ronde	2	4	D	2	2	3	Efficiency	2	2	3.2	-1.2	-2.4
19	19	127	Z2	14	>D	3	ronde	3	4	D	2	2	3	Efficiency	1	1	3.2	-2.2	-4.3
20	20	133	Z1	14	>D	2	ronde	2	4	D	1	1	3	Efficiency	3	3	3.9	-.9	-3.2
20	20	134	Z1	14	>D	1	ronde	1	4	D	1	1	1	Ponderation	5	5	5.9	-.9	-2.8
21	21	140	Z2	14	>D	3	ronde	3	4	D	2	2	3	Efficiency	2	2	3.4	-1.4	-2.7
5	5	49	Z1	15	>D	1	ronde	1	4	D	1	1	3	Efficiency	3	3	3.9	-.9	-2.8
8	8	76	Z4	15	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.2	1.8	3.3
12	12	96	Z4	15	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.2	1.8	3.3
21	21	140	Z2	15	>D	3	ronde	3	4	D	2	2	3	Efficiency	2	2	3.2	-1.2	-2.3
1	1	8	Z2	16	>D	1	ronde	1	4	D	2	2	3	Efficiency	1	1	2.9	-1.9	-3.7
3	3	30	Z2	16	>D	1	ronde	1	4	D	2	2	1	Ponderation	3	3	4.6	-1.6	-2.4
4	4	39	Z2	16	>D	3	ronde	3	4	D	2	2	3	Efficiency	4	4	2.8	1.2	2.3
8	8	76	Z4	16	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.1	1.9	3.5
12	12	96	Z4	16	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.1	1.9	3.5
13	13	104	Z4	16	>D	1	ronde	1	4	D	4	4	3	Efficiency	4	4	2.2	1.8	3.4
13	13	104	Z4	16	>D	2	ronde	2	4	D	4	4	3	Efficiency	4	4	2.2	1.8	3.4
16	16	111	Z2	16	>D	1	ronde	1	4	D	2	2	1	Ponderation	3	3	4.7	-1.7	-2.9
17	17	117	Z2	16	>D	1	ronde	1	4	D	2	2	3	Efficiency	1	1	2.9	-1.9	-3.8
17	17	117	Z2	16	>D	3	ronde	3	4	D	2	2	3	Efficiency	1	1	2.9	-1.9	-3.8
18	18	119	Z2	16	>D	1	ronde	1	4	D	2	2	3	Efficiency	1	1	2.7	-1.7	-3.3
19	19	127	Z2	16	>D	3	ronde	3	4	D	2	2	3	Efficiency	1	1	2.9	-1.9	-3.7
23	23	143	Z2	16	>D	1	ronde	1	4	D	2	2	3	Efficiency	1	1	2.8	-1.8	-3.6
1	1	7	Z2	17	>D	1	ronde	1	4	D	2	2	3	Efficiency	2	2	3.3	-1.3	-2.5
4	4	48	Z1	17	>D	1	ronde	1	4	D	1	1	3	Efficiency	3	3	3.9	-.9	-2.8
5	5	49	Z1	17	>D	1	ronde	1	4	D	1	1	3	Efficiency	3	3	3.9	-.9	-2.8
8	8	76	Z4	17	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.2	1.8	3.3
9	9	87	Z2	17	>D	3	ronde	3	4	D	2	2	3	Efficiency	1	1	2.9	-1.9	-3.9
12	12	96	Z4	17	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.2	1.8	3.3
16	16	112	Z1	17	>D	1	ronde	1	4	D	1	1	3	Efficiency	3	3	3.9	-.9	-2.8
19	19	127	Z2	17	>D	3	ronde	3	4	D	2	2	3	Efficiency	1	1	2.9	-1.9	-3.9
2	2	11	Z2	18	>D	2	ronde	2	4	D	2	2	3	Efficiency	2	2	3.2	-1.2	-2.4
2	2	17	Z2	18	>D	1	ronde	1	4	D	2	2	3	Efficiency	4	4	2.8	1.2	2.3
2	2	24	Z2	18	>D	1	ronde	1	4	D	2	2	1	Ponderation	3	3	4.8	-1.8	-3.0
5	5	51	Z2	18	>D	2	ronde	2	4	D	2	2	3	Efficiency	1	1	2.8	-1.8	-3.6
5	5	58	Z1	18	>D	1	ronde	1	4	D	1	1	3	Efficiency	3	3	3.9	-.9	-2.4
7	7	67	Z2	18	>D	3	ronde	3	4	D	2	2	3	Efficiency	2	2	3.2	-1.2	-2.3
8	8	76	Z4	18	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.2	1.8	3.2
11	11	92	Z2	18	>D	1	ronde	1	4	D	2	2	3	Efficiency	4	4	2.8	1.2	2.4
12	12	96	Z4	18	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.2	1.8	3.2
13	13	104	Z4	18	>D	2	ronde	2	4	D	4	4	3	Efficiency	1	1	2.3	-1.3	-2.4
18	18	121	Z2	18	>D	2	ronde	2	4	D	2	2	3	Efficiency	2	2	3.2	-1.2	-2.3
19	19	128	Z2	18	>D	1	ronde	1	4	D	2	2	3	Efficiency	4	4	2.8	1.2	2.3
21	21	140	Z2	18	>D	3	ronde	3	4	D	2	2	3	Efficiency	2	2	3.2	-1.2	-2.4
1	1	8	Z2	19	>D	1	ronde	1	4	D	2	2	1	Ponderation	2	2	4.3	-2.3	-3.3
2	2	17	Z2	19	>D	1	ronde	1	4	D	2	2	1	Ponderation	2	2	3.9	-1.9	-2.3
2	2	22	Z4	19	>D	2	ronde	2	4	D	4	4	3	Efficiency	4	4	1.9	2.1	3.9
2	2	27	Z2	19	>D	1	ronde	1	4	D	2	2	1	Ponderation	2	2	4.5	-2.5	-3.6
4	4	39	Z2	19	>D	1	ronde	1	4	D	2	2	1	Ponderation	2	2	4.2	-2.2	-3.0
5	5	51	Z2	19	>D	1	ronde	1	4	D	2	2	1	Ponderation	2	2	4.0	-2.0	-2.5
5	5	51	Z2	19	>D	2	ronde	2	4	D	2	2	3	Efficiency	1	1	2.7	-1.7	-3.1
5	5	58	Z1	19	>D	1	ronde	1	4	D	1	1	1	Ponderation	4	4	5.6	-1.6	-3.1
8	8	76	Z4	19	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.1	1.9	3.6

9	9	80	Z4	19	>D	1	ronde	1	4	D	4	4	3	Efficiency	3	3	1.7	1.3	2.3
9	9	87	Z2	19	>D	3	ronde	3	4	D	2	2	3	Efficiency	1	1	2.8	-1.8	-3.6
12	12	96	Z4	19	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.1	1.9	3.6
12	12	101	Z2	19	>D	1	ronde	1	4	D	2	2	1	Ponderation	2	2	4.7	-2.7	-4.5
16	16	111	Z2	19	>D	1	ronde	1	4	D	2	2	1	Ponderation	2	2	4.7	-2.7	-4.3
18	18	122	Z2	19	>D	1	ronde	1	4	D	2	2	1	Ponderation	2	2	4.7	-2.7	-4.4
19	19	127	Z2	19	>D	1	ronde	1	4	D	2	2	3	Efficiency	4	4	2.8	1.2	2.3
19	19	127	Z2	19	>D	3	ronde	3	4	D	2	2	3	Efficiency	1	1	2.8	-1.8	-3.6
19	19	128	Z2	19	>D	1	ronde	1	4	D	2	2	3	Efficiency	4	4	2.6	1.4	2.5
20	20	133	Z1	19	>D	1	ronde	1	4	D	1	1	1	Ponderation	4	4	5.5	-1.5	-2.9
23	23	143	Z2	19	>D	1	ronde	1	4	D	2	2	1	Ponderation	2	2	4.3	-2.3	-3.1
23	23	143	Z2	19	>D	2	ronde	2	4	D	2	2	3	Efficiency	4	4	2.8	1.2	2.3
1	1	8	Z2	20	>D	1	ronde	1	4	D	2	2	3	Efficiency	1	1	2.9	-1.9	-3.9
2	2	24	Z2	20	>D	1	ronde	1	4	D	2	2	3	Efficiency	1	1	3.0	-2.0	-4.1
2	2	25	Z2	20	>D	3	ronde	3	4	D	2	2	3	Efficiency	1	1	3.1	-2.1	-4.3
4	4	39	Z2	20	>D	1	ronde	1	4	D	2	2	3	Efficiency	1	1	2.9	-1.9	-3.7
4	4	48	Z1	20	>D	1	ronde	1	4	D	1	1	3	Efficiency	3	3	3.9	-.9	-2.7
5	5	51	Z2	20	>D	1	ronde	1	4	D	2	2	3	Efficiency	4	4	2.8	1.2	2.4
5	5	51	Z2	20	>D	2	ronde	2	4	D	2	2	3	Efficiency	4	4	2.8	1.2	2.4
8	8	74	Z4	20	>D	1	ronde	1	4	D	4	4	3	Efficiency	4	4	2.1	1.9	3.5
8	8	76	Z4	20	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.2	1.8	3.4
8	8	77	Z4	20	>D	1	ronde	1	4	D	4	4	3	Efficiency	4	4	2.0	2.0	3.6
9	9	80	Z4	20	>D	1	ronde	1	4	D	4	4	3	Efficiency	4	4	1.9	2.1	4.0
9	9	87	Z2	20	>D	1	ronde	1	4	D	2	2	3	Efficiency	1	1	2.9	-1.9	-3.9
9	9	87	Z2	20	>D	2	ronde	2	4	D	2	2	3	Efficiency	1	1	2.9	-1.9	-3.9
9	9	87	Z2	20	>D	3	ronde	3	4	D	2	2	3	Efficiency	1	1	2.9	-1.9	-3.9
12	12	96	Z4	20	>D	2	ronde	2	4	D	4	4	3	Efficiency	4	4	2.2	1.8	3.3
12	12	96	Z4	20	>D	3	ronde	3	4	D	4	4	3	Efficiency	4	4	2.2	1.8	3.3
16	16	112	Z1	20	>D	1	ronde	1	4	D	1	1	3	Efficiency	3	3	3.9	-.9	-2.7
19	19	127	Z2	20	>D	1	ronde	1	4	D	2	2	3	Efficiency	1	1	2.9	-1.9	-3.9
19	19	128	Z2	20	>D	1	ronde	1	4	D	2	2	3	Efficiency	4	4	2.8	1.2	2.4
20	20	134	Z1	20	>D	1	ronde	1	4	D	1	1	3	Efficiency	3	3	3.9	-.9	-2.5
20	20	134	Z1	20	>D	2	ronde	2	4	D	1	1	3	Efficiency	3	3	3.9	-.9	-2.5
21	21	140	Z2	20	>D	2	ronde	2	4	D	2	2	3	Efficiency	2	2	3.2	-1.2	-2.3
Nu Ta Num It Nu Ju N Rondes N G N Z N Criteres															Cat	Score	Exp.	Resd	StRes

2°ANALYSE DES SCORES BRUTS DE TOUS LES ITEMS

Etablissement de standards de performance 28/06/20 17:26:44
Table 5. Measurable Data Summary.

	Cat	Score	Exp.	Resd	StRes
Mean (Count: 9218)	3.46	3.46	3.46	.00	.00
Mean (Count: 9218)	1.10	1.10	1.01	.45	.98
Mean (Count: 9218)	1.10	1.10	1.01	.45	.98

Data log-likelihood chi-square = 10449.0391

Approximate degrees of freedom = 9027

Chi-square significance prob. = .0000

	Count	Mean	S.D.	Params
Responses used for estimation	= 9218	3,46	1,10	191
Count of measurable responses	= 9218			
Raw-score variance of observations	= 1,22	100.00%		
Variance explained by Rasch measures	= 1,02	83,65%		
Variance of residuals	= 0,20	16,35%		

Etablissement de standards de performance 28/06/20 17:26:44
Table 6.0 All Facet Vertical "Rulers".

Vertical = (1H,2H,3H,4H,5A,6A,5L) Yardstick (columns lines low high extreme) = 0,15,-2.6,2.6,End

Measr +Taches	+Items	+Juges	+Rondes +Groupes +Criteres	S.1	S.2	S.3
14 15 16 20 21 22	1 4 7 11 14 24 25 27 41 47 48 49 53 54 55 58 59 87 100 105 114 115 116 120 123 124	14 15 16 17 18 20		(6) 91-100 severit	(4) Facile	(6) Tres fort
17	30 32 108 99 94 8 70 72	19				
2 +						
	65 69 78 81 83 85 88 101 143			---		
5	127 138 139 141					
18	17					
1 +	100 110 113 39					
	67 15 26 28 103 34 91 112 35 2			Efficiency		
13						
	71 73 132 135 136 137 20 31 52 9 128			4 71-80	3 Moderate	---
0 +	62 63 64 106 142 144 145 146 148 149 29 92 97 43 38 42 79 84 134 76 129 130 131	1 2 3 0				
7				Difficulte		
	98 121 122 12 74 18 33 133			Ponderation		
12						
-1 +	77 21			---		
	36 5 95 102 147					
10	46			3 61-70		
19						
-2 +	66 68					
4 11	44 86 93					
6 23						
	22 107 75 80 119			---	(2) Difficile	(1) Faible
1 2 3 8 9	3 6 10 13 16 19 23 37 40 45 50 51 56 57 60 61 82 89 96 98 104 111 117 118 125 126 140			(1) 0-50		(1)
Measr +Taches	+Items	+Juges	+Rondes +Groupes +Criteres	S.1	S.2	S.3

S.1: Model = 3,3,3,3,3,1, PONDARATION ; Criteres: Ponderation
S.2: Model = 3,3,3,3,3,2, DIFFICULTE ; Criteres: Difficulte
S.3: Model = 3,3,3,3,3,3, EFFICIENCE ; Criteres: Efficiency

Etablissement de standards de performance 28/06/20 17:26:44
Table 7.1.1 Taches Measurement Report (arranged by fN).

[illegible]

Etablissement de standards de performance 28/06/20 17:26:44
Table 7.2.1 Items Measurement Report (arranged by #N).

Num Items	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq ZStd	Outfit MnSq ZStd	Estim. Discrm	Correlation PtMea PtExp	Group
87 87	196	54	3.63	4.36 G	3.16	.30	4.10 9.0	6.95 9.0	-2.83	-.38 .16	9
127 127	225	62	3.63	4.00 G	1.15	.28	3.64 8.7	5.88 9.0	-2.30	-.16 .17	19
117 117	228	62	3.68	3.17 G	-3.01	.28	2.78 6.8	4.31 7.6	-1.24	-.22 .16	17
8 8	225	62	3.63	4.14 G	2.04	.28	2.92 7.0	4.22 7.4	-1.02	-.18 .17	1
96 96	159	62	2.56	2.94 G	-3.86	.21	2.40 5.6	3.61 9.0	-2.95	.50 .38	12
76 76	158	62	2.55	3.74 G	-.43	.21	2.15 4.8	3.38 9.0	-2.54	.42 .38	8
39 39	210	59	3.56	3.96 G	.93	.29	2.48 5.2	3.15 5.5	-.39	.06 .21	4
51 51	213	62	3.44	3.19 G	-2.92	.28	2.29 4.3	2.71 4.8	-.12	.09 .23	5
24 24	232	62	3.74	4.28 G	2.81	.29	1.86 3.9	2.70 4.9	-.16	-.20 .13	2
30 30	231	62	3.73	4.24 G	2.55	.28	1.98 4.4	2.68 4.8	-.29	-.11 .14	3
143 143	222	62	3.58	4.03 G	1.33	.28	2.12 4.4	2.50 4.3	-.06	-.01 .19	23
77 77	145	62	2.34	3.62 G	-1.02	.22	1.53 2.4	2.26 5.0	-.27	.56 .44	8
94 94	237	62	3.82	4.18 G	2.22	.29	1.83 3.7	2.19 3.9	.07	.13 .11	11
22 22	145	62	2.34	3.31 G	-2.41	.22	1.58 2.6	2.18 4.8	-.42	.43 .44	2
25 25	241	62	3.89	4.41 G	3.56	.29	1.88 3.7	2.12 3.8	-.18	-.27 .09	2
121 121	240	62	3.87	3.70 G	-.61	.29	1.66 3.0	2.09 3.7	.03	-.26 .09	18
27 27	230	62	3.71	4.25 G	2.64	.28	1.97 4.3	2.05 3.3	.18	-.07 .14	2
74 74	152	62	2.45	3.68 G	-.70	.21	1.40 1.9	2.03 4.6	-.16	.36 .41	8
111 111	239	62	3.85	2.99 G	-3.68	.29	1.95 4.0	1.78 2.8	.33	.02 .10	16
107 107	166	62	3.84	3.31 G	-2.39	.29	1.48 2.3	1.82 2.9	.38	-.16 .10	15
104 104	268	62	2.68	2.53 G	-5.12	.21	1.40 2.0	1.76 3.8	-.29	.20 .25	13
32 32	231	62	3.73	4.24 G	2.55	.28	1.47 2.4	1.73 2.5	.42	.04 .14	3
140 140	242	62	3.90	3.09 G	-3.34	.29	1.57 2.5	1.68 2.5	.18	-.27 .09	21
101 101	241	62	3.89	4.05 G	1.47	.29	1.68 3.0	1.60 2.3	.59	-.08 .09	12
4 4	235	62	3.79	4.29 G	2.85	.29	1.43 2.1	1.61 2.2	.47	-.14 .12	1
122 122	240	62	3.87	3.70 G	-.61	.29	1.56 2.6	1.37 1.5	.71	-.17 .09	18
80 80	127	62	2.05	3.30 G	-2.45	.25	1.42 1.8	1.44 1.7	.62	.52 .50	9
67 67	240	62	3.87	3.94 G	.76	.29	1.25 1.3	1.42 1.7	.56	-.20 .09	7
133 133	276	62	4.45	3.65 G	-.87	.36	1.38 1.9	1.24 .5	.53	-.10 .14	20
58 58	278	62	4.48	4.31 G	2.98	.38	1.33 1.5	.91 .1	.69	.03 .13	5
86 86	132	62	2.13	3.37 G	-2.16	.23	1.21 1.0	1.08 .4	.88	.45 .48	9
11 11	243	62	3.92	4.44 G	3.73	.30	1.08 .4	1.18 .8	.79	-.22 .08	2
119 119	216	62	3.48	3.28 G	-2.55	.28	1.02 .1	1.11 .5	.89	.12 .22	18
49 49	281	62	4.53	4.40 G	3.46	.43	1.06 .3	.86 .1	.94	.01 .12	5
97 97	221	62	3.56	3.79 G	-.16	.28	.91 -.4	.86 -.5	1.04	.12 .20	12
112 112	281	62	4.53	3.92 G	-.61	.43	1.06 .3	.85 .1	.94	.02 .12	16
48 48	281	62	4.53	4.63 G	6.70	.43	1.06 .3	.85 .1	.94	.02 .12	4
91 91	220	62	3.55	3.91 G	.58	.28	.87 -.6	.83 -.6	1.07	.14 .20	10
7 7	249	62	4.02	4.48 G	4.06	.30	.86 -.6	.81 -.7	1.09	-.35 .07	1
134 134	280	62	4.52	3.76 G	-.30	.41	1.05 .2	.81 .0	.96	.03 .13	20
29 29	194	62	3.13	3.81 G	-.04	.25	.84 -.7	.81 -.8	1.22	.21 .27	2
33 33	173	59	2.93	3.65 G	-.88	.23	.78 -1.1	.75 -1.3	1.25	.13 .30	3
61 61	178	62	2.87	2.53 G	-5.11	.22	.75 -1.3	.77 -1.3	1.23	.04 .31	5
20 20	198	62	3.19	3.85 G	.22	.25	.82 -.7	.72 -1.2	1.27	.25 .26	2
106 106	283	62	4.56	3.81 G	.00	.49	.97 .0	.71 .1	1.03	.14 .11	14
142 142	283	62	4.56	3.81 G	.00	.49	.97 .0	.71 .1	1.03	.14 .11	22
132 132	283	62	4.56	3.86 G	.29	.49	.97 .0	.71 .1	1.03	.14 .11	20
135 135	283	62	4.56	3.86 G	.29	.49	.97 .0	.71 .1	1.03	.14 .11	20
136 136	283	62	4.56	3.86 G	.29	.49	.97 .0	.71 .1	1.03	.14 .11	20
137 137	283	62	4.56	3.86 G	.29	.49	.97 .0	.71 .1	1.03	.14 .11	20
100 100	283	62	4.56	4.62 G	6.00	.49	.97 .0	.71 .1	1.03	.14 .11	12
138 138	283	62	4.56	3.99 G	1.11	.49	.97 .0	.71 .1	1.03	.14 .11	21
139 139	283	62	4.56	3.99 G	1.11	.49	.97 .0	.71 .1	1.03	.14 .11	21
141 141	283	62	4.56	3.99 G	1.11	.49	.97 .0	.71 .1	1.03	.14 .11	21
108 108	283	62	4.56	4.21 G	2.39	.49	.97 .0	.71 .1	1.03	.14 .11	15
109 109	283	62	4.56	3.98 G	1.02	.49	.97 .0	.71 .1	1.03	.14 .11	16
110 110	283	62	4.56	3.98 G	1.02	.49	.97 .0	.71 .1	1.03	.14 .11	16
113 113	283	62	4.56	3.98 G	1.02	.49	.97 .0	.71 .1	1.03	.14 .11	16
120 120	283	62	4.56	4.48 G	4.01	.49	.97 .0	.71 .1	1.03	.14 .11	18
123 123	283	62	4.56	4.48 G	4.01	.49	.97 .0	.71 .1	1.03	.14 .11	18
124 124	283	62	4.56	4.48 G	4.01	.49	.97 .0	.71 .1	1.03	.14 .11	18
53 53	283	62	4.56	4.46 G	3.88	.49	.97 .0	.71 .1	1.03	.14 .11	5
54 54	283	62	4.56	4.46 G	3.88	.49	.97 .0	.71 .1	1.03	.14 .11	5
55 55	283	62	4.56	4.46 G	3.88	.49	.97 .0	.71 .1	1.03	.14 .11	5
58 58	283	62	4.56	4.46 G	3.88	.49	.97 .0	.71 .1	1.03	.14 .11	5
114 114	283	62	4.56	4.24 G	2.60	.49	.97 .0	.71 .1	1.03	.14 .11	17
115 115	283	62	4.56	4.24 G	2.60	.49	.97 .0	.71 .1	1.03	.14 .11	17
116 116	283	62	4.56	4.24 G	2.60	.49	.97 .0	.71 .1	1.03	.14 .11	17
105 105	283	62	4.56	4.52 G	4.43	.49	.97 .0	.71 .1	1.03	.14 .11	13
9 9	199	62	3.21	3.83 G	.07	.26	.80 -.8	.70 -1.3	1.29	.29 .26	1
99 99	250	62	4.03	4.18 G	2.27	.30	.73 -1.4	.69 -1.4	1.26	-.35 .08	12
65 65	250	62	4.03	4.08 G	1.64	.30	.73 -1.4	.69 -1.4	1.26	-.35 .08	7
69 69	250	62	4.03	4.08 G	1.64	.30	.73 -1.4	.69 -1.4	1.26	-.35 .08	7
41 41	250	62	4.03	4.38 G	3.38	.30	.73 -1.4	.69 -1.4	1.26	-.35 .08	4
47 47	250	62	4.03	4.38 G	3.38	.30	.73 -1.4	.69 -1.4	1.26	-.35 .08	4
103 103	250	62	4.03	3.93 G	.69	.30	.73 -1.4	.69 -1.4	1.26	-.35 .08	13
14 14	250	62	4.03	4.52 G	4.36	.30	.73 -1.4	.69 -1.4	1.26	-.35 .08	2
1 1	250	62	4.03	4.49 G	4.15	.30	.73 -1.4	.69 -1.4	1.26	-.35 .08	1
52 52	250	62	4.03	3.84 G	-.14	.30	.73 -1.4	.69 -1.4	1.26	-.35 .08	5
92 92	208	62	3.35	3.80 G	-.08	.27	.74 -1.1	.67 -1.4	1.23	.36 .25	11
46 46	181	62	2.92	3.47 G	-1.73	.22	.72 -1.5	.67 -1.9	1.31	.23 .30	4
128 128	211	62	3.40	3.82 G	.05	.28	.66 -1.5	.70 -1.2	1.12	.40 .24	19
125 125	171	59	2.90	2.61 G	-4.89	.23	.66 -1.8	.64 -2.1	1.35	.26 .31	18
147 147	181	62	2.92	3.53 G	-1.42	.22	.64 -2.0	.69 -1.8	1.27	.19 .30	23
18 18	181	62	2.92	3.67 G	-.75	.22	.64 -2.0	.69 -1.8	1.27	.19 .30	2
12 12	174	59	2.95	3.69 G	-.65	.23	.63 -2.0	.63 -2.1	1.31	.31 .30	2
98 98	179	62	2.89	3.19 G	-2.94	.22	.61 -2.3	.67 -1.9	1.30	.18 .30	12
71 71	173	62	2.79	3.86 G	.25	.22	.61 -2.4	.64 -2.3	1.38	.04 .33	8
73 73	173	62	2.79	3.86 G	.25	.22	.61 -2.4	.64 -2.3	1.38	.04 .33	8
93 93	173	62	2.79	3.37 G	-2.13	.22	.61 -2.4	.64 -2.3	1.38	.04 .33	11
17 17	210	62	3.39	3.99 G	1.07	.28	.66 -1.5	.60 -1.8	1.29	.34 .24	2
36 36	172	62	2.77	3.55 G	-1.36	.22	.59 -2.5	.62 -2.5	1.42	.02 .33	3
79 79	172	62	2.77	3.76 G	-.28	.22	.59 -2.5	.62 -2.5	1.42	.02 .33	9
84 84	172	62	2.77	3.76 G	-.28	.22	.59 -2.5	.62 -2.5	1.42	.02 .33	9
5 5	172	62	2.77	3.54 G	-1.39	.22	.59 -2.5	.62 -2.5	1.42	.02 .33	1
44 44	172	62	2.77	3.37 G	-2.16	.22	.59 -2.5	.62 -2.5	1.42	.02 .33	4
21 21	174	62	2.81	3.60 G	-1.09	.22	.59 -2.5	.60 -2.6	1.42	.15 .32	2

31 31	199	62	3.21	3.83	G	.10	.26	.78	-.9	.58	-1.9	1.33	.37	.26	3
43 43	206	62	3.32	3.78	G	-.21	.27	.71	-1.2	.57	-1.9	1.31	.36	.25	4
34 34	206	62	3.32	3.91	G	.59	.27	.61	-1.8	.54	-2.1	1.36	.59	.25	3
40 40	119	62	1.92	2.62	G	-4.86	.27	.98	.0	.48	-2.4	1.34	.61	.51	4
45 45	119	62	1.92	2.62	G	-4.86	.27	.98	.0	.48	-2.4	1.34	.61	.51	4
3 3	119	62	1.92	2.87	G	-4.09	.27	.98	.0	.48	-2.4	1.34	.61	.51	1
6 6	119	62	1.92	2.87	G	-4.09	.27	.98	.0	.48	-2.4	1.34	.61	.51	1
10 10	119	62	1.92	2.87	G	-4.09	.27	.98	.0	.48	-2.4	1.34	.61	.51	1
13 13	119	62	1.92	2.94	G	-3.88	.27	.98	.0	.48	-2.4	1.34	.61	.51	2
16 16	119	62	1.92	2.94	G	-3.88	.27	.98	.0	.48	-2.4	1.34	.61	.51	2
19 19	119	62	1.92	2.94	G	-3.88	.27	.98	.0	.48	-2.4	1.34	.61	.51	2
23 23	119	62	1.92	2.94	G	-3.88	.27	.98	.0	.48	-2.4	1.34	.61	.51	2
37 37	119	62	1.92	2.88	G	-4.06	.27	.98	.0	.48	-2.4	1.34	.61	.51	3
75 75	119	62	1.92	3.29	G	-2.50	.27	.98	.0	.48	-2.4	1.34	.61	.51	8
82 82	119	62	1.92	3.17	G	-2.98	.27	.98	.0	.48	-2.4	1.34	.61	.51	9
89 89	119	62	1.92	3.17	G	-2.98	.27	.98	.0	.48	-2.4	1.34	.61	.51	9
70 70	205	62	3.31	4.15	G	2.08	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	8
72 72	205	62	3.31	4.15	G	2.08	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	8
78 78	205	62	3.31	4.07	G	1.60	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	9
81 81	205	62	3.31	4.07	G	1.60	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	9
83 83	205	62	3.31	4.07	G	1.60	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	9
85 85	205	62	3.31	4.07	G	1.60	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	9
88 88	205	62	3.31	4.07	G	1.60	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	9
62 62	205	62	3.31	3.81	G	.00	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	6
63 63	205	62	3.31	3.81	G	.00	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	6
64 64	205	62	3.31	3.81	G	.00	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	6
129 129	205	62	3.31	3.74	G	-1.40	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	19
130 130	205	62	3.31	3.74	G	-1.40	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	19
131 131	205	62	3.31	3.74	G	-1.40	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	19
144 144	205	62	3.31	3.82	G	.02	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	23
145 145	205	62	3.31	3.82	G	.02	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	23
146 146	205	62	3.31	3.82	G	.02	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	23
148 148	205	62	3.31	3.82	G	.02	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	23
149 149	205	62	3.31	3.82	G	.02	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	23
90 90	205	62	3.31	3.71	G	-1.58	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	10
35 35	205	62	3.31	3.90	G	.52	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	3
15 15	205	62	3.31	3.93	G	.69	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	2
26 26	205	62	3.31	3.93	G	.69	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	2
28 28	205	62	3.31	3.93	G	.69	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	2
66 66	205	62	3.31	3.40	G	-2.02	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	7
68 68	205	62	3.31	3.40	G	-2.02	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	7
2 2	205	62	3.31	3.90	G	.49	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	1
38 38	205	62	3.31	3.76	G	-.28	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	4
42 42	205	62	3.31	3.76	G	-.28	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	4
95 95	205	62	3.31	3.54	G	-1.39	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	12
102 102	205	62	3.31	3.54	G	-1.39	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	12
126 126	205	62	3.31	3.07	G	-3.38	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	18
118 118	205	62	3.31	2.64	G	-4.80	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	17
50 50	205	62	3.31	3.04	G	-3.52	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	5
56 56	205	62	3.31	3.04	G	-3.52	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	5
57 57	205	62	3.31	3.04	G	-3.52	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	5
60 60	205	62	3.31	3.04	G	-3.52	.27	.52	-2.4	.35	-3.4	1.46	.52	.25	5
Mean (Count: 149)	214.1	61.9	3.46	3.76		.00	.31	1.00	-.2	.97	-.6		.22		
S.D. (Population)	48.5	.8	.78	.50		2.53	.09	.59	2.4	.99	3.0		.30		
S.D. (Sample)	48.7	.8	.78	.50		2.54	.09	.60	2.4	1.00	3.0		.30		

Model, Populn: RMSE .32 Adj (True) S.D. 2.51 Separation 7.83 Strata 10.77 Reliability .98
Model, Sample: RMSE .32 Adj (True) S.D. 2.52 Separation 7.85 Strata 10.80 Reliability .98
Model, Fixed (all same) chi-square: 10588.4 d.f.: 148 significance (probability): .00
Model, Random (normal) chi-square: 145.7 d.f.: 147 significance (probability): .52

Etablissement de standards de performance 28/06/20 17:26:44
Table 7.3.1 Juges Measurement Report (arranged by fN).

Nu Juges	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	Outfit MnSq	Estim. Zstd	Correlation PtMea	Exact Agree. Obs %	Exp %
14 >D	5026	1338	3.76	4.07	4.55	.06	1.63	9.0	1.43	.73	42.2	41.4
19 >D	4434	1341	3.31	3.68	2.29	.06	1.28	5.5	1.15	.87	43.3	40.6
18 >D	4643	1335	3.48	3.83	3.10	.06	1.00	.0	.93	.89	42.3	42.0
17 >D	4614	1341	3.44	3.80	2.91	.06	.85	-3.2	.88	1.09	42.8	41.8
16 >D	4529	1341	3.38	3.74	2.61	.06	.83	-3.7	.87	1.10	43.2	41.3
20 >D	4646	1341	3.46	3.82	3.03	.06	.73	-5.9	.96	1.04	42.7	42.0
15 >D	4008	1181	3.39	3.71	3.11	.07	.46	-9.0	.47	1.36	46.1	43.5
Mean (Count: 7)	4557.1	1316.9	3.46	3.81	3.09	.06	.97	-1.1	.96	.84		
S.D. (Population)	281.8	55.5	.13	.12	.66	.00	.36	5.9	.27	.05		
S.D. (Sample)	304.4	60.0	.14	.13	.71	.00	.39	6.4	.29	.05		

Model, Populn: RMSE .06 Adj (True) S.D. .65 Separation 10.66 Strata 14.54 Reliability (not inter-rater) .99
Model, Sample: RMSE .06 Adj (True) S.D. .71 Separation 11.52 Strata 15.69 Reliability (not inter-rater) .99
Model, Fixed (all same) chi-square: 787.4 d.f.: 6 significance (probability): .00
Model, Random (normal) chi-square: 6.0 d.f.: 5 significance (probability): .31
Inter-Rater agreement opportunities: 252814 Exact agreements: 109164 = 43.2% Expected: 105617.7 = 41.8%

Etablissement de standards de performance 28/06/20 17:26:44

Table 7.4.1 Rondes Measurement Report (arranged by fN).

N Rondes	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	Outfit ZStd	Estim. MnSq	Correlation Discrm	PtMea	PtExp
1 ronde 1	10034	2972	3.38	3.78 A	.00	.04	1.48 9.0	1.40 4.1	.68	-.29	.83	.83
2 ronde 2	10926	3126	3.50	3.83 A	.00	.04	.74 -8.7	.77 -3.0	1.15	.11	.85	.83
3 ronde 3	10940	3120	3.51	3.83 A	.00	.04	.74 -8.7	.75 -3.3	1.15	.16	.85	.83
Mean (Count: 3)	10633.3	3072.7	3.46	3.81	.00	.04	.99 -2.8	.97 -.7			.84	
S.D. (Population)	423.8	71.2	.06	.02	.00	.00	.35 8.4	.30 3.5			.01	
S.D. (Sample)	519.1	87.2	.07	.03	.00	.00	.43 10.3	.37 4.2			.01	

Model, Populn: RMSE .04 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Sample: RMSE .04 Adj (True) S.D. .00 Separation .00 Strata .33 Reliability .00
Model, Fixed (all same) chi-square: .0 d.f.: 2 significance (probability): 1.00

Etablissement de standards de performance 28/06/20 17:26:44

Table 7.6.1 Criteres Measurement Report (arranged by fN).

N Criteres	Total Score	Total Count	Obsvd Average	Fair(M) Average	+ Measure	Model S.E.	Infit MnSq	Outfit ZStd	Estim. MnSq	Correlation Discrm	PtMea	PtExp
3 Efficience	9264	3122	2.97	3.08	.59	.05	1.19 5.9	1.28 5.4	.75	.80	.84	
1 Ponderation	12475	2974	4.19	4.78	-.40	.03	1.06 1.8	1.25 7.2	.91	.86	.87	
2 Difficulte	10161	3122	3.25	3.56	-.19	.05	.56 -9.0	.38 -6.2	1.40	.87	.81	
Mean (Count: 3)	10633.3	3072.7	3.47	3.81	.00	.04	.94 -.4	.97 2.1			.85	
S.D. (Population)	1352.8	69.8	.52	.72	.43	.01	.27 6.3	.42 6.0			.03	
S.D. (Sample)	1656.8	85.4	.64	.88	.52	.01	.33 7.7	.51 7.3			.04	

Model, Populn: RMSE .04 Adj (True) S.D. .42 Separation 9.86 Strata 13.48 Reliability .99
Model, Sample: RMSE .04 Adj (True) S.D. .52 Separation 12.09 Strata 16.46 Reliability .99
Model, Fixed (all same) chi-square: 307.7 d.f.: 2 significance (probability): .00
Model, Random (normal) chi-square: 2.0 d.f.: 1 significance (probability): .16

Etablissement de standards de performance 28/06/20 17:26:44

Table 8.1 Category Statistics.

Model = ?,?,?,?,1,PONDERATION ; Criteres: Ponderation

Rating (or partial credit) scale = PONDERATION,R6,G,0

Response Category Name	DATA Category Counts	Cum. %	QUALITY CONTROL Avg	RASCH-ANDRICH Outfit	EXPECTATION Measure at	MOST PROBABLE from	RASCH-THURSTONE Thresholds	Cat
	Score Total Used	% %	Meas Meas MnSq	Measure S.E.	Category -0.5			
0-50 (0-50)	1 183 183	6% 6%	-4.32 -4.16 .7		(-5.04)	low	low	100%
51-60 (51-60)	2 219 219	7% 14%	-2.60 -2.92 1.8	-3.82 .11	-3.24 -4.22	-3.82	-4.01	49%
61-70 (61-70)	3 324 324	11% 24%	-1.14 -1.11 .7	-2.38 .10	-1.80 -2.50	-2.38	-2.45	43%
71-80 (71-80)	4 898 898	30% 55%	.15 .34 .9	-1.40 .07	.14 -1.00	-1.40	-1.18	68%
81-90 (81-90)	5 810 810	27% 82%	3.52 3.08 1.4	1.53 .06	3.77 1.62	1.53	1.56	83%
91-100 severite (91-100)	6 540 540	18% 100%	7.00 7.42 1.8	6.06 .09	(7.14) 6.08	6.06	6.06	100%

(Mean) (Modal) (Median)

Scale structure for "PONDERATION"

Measr:-9.0 -6.0 -3.0 0.0 3.0 6.0 9.0
 + + + + + +

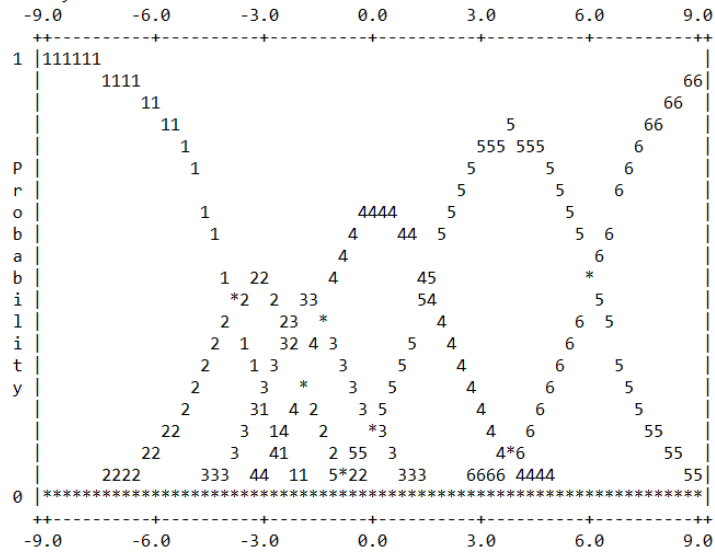
Mode:<1-----(^)-12--^23--^34-----^--45-----^-----56--(^)----6>

Median:<1-----(^)12--^23--^34-----^--45-----^-----56--(^)----6>

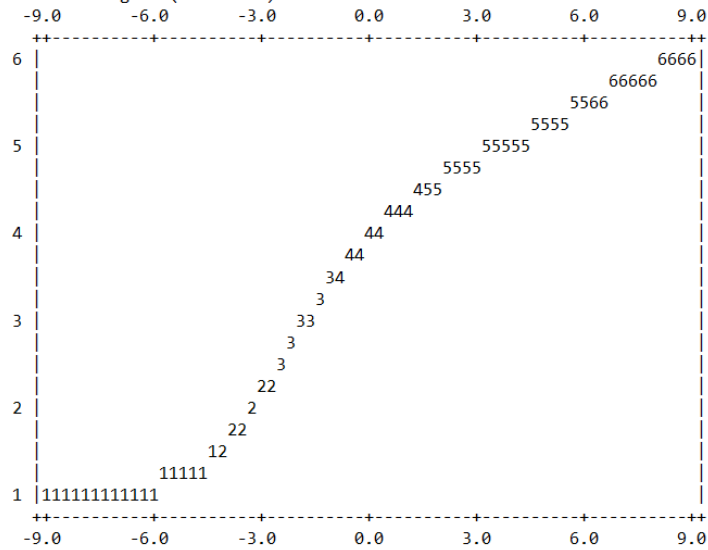
Mean:<1-----(^)12--^23--^34-----^--45-----^-----56--(^)----6>

+ + + + + +
 Measr:-9.0 -6.0 -3.0 0.0 3.0 6.0 9.0

Probability Curves



Expected Score Ogive (Model ICC)

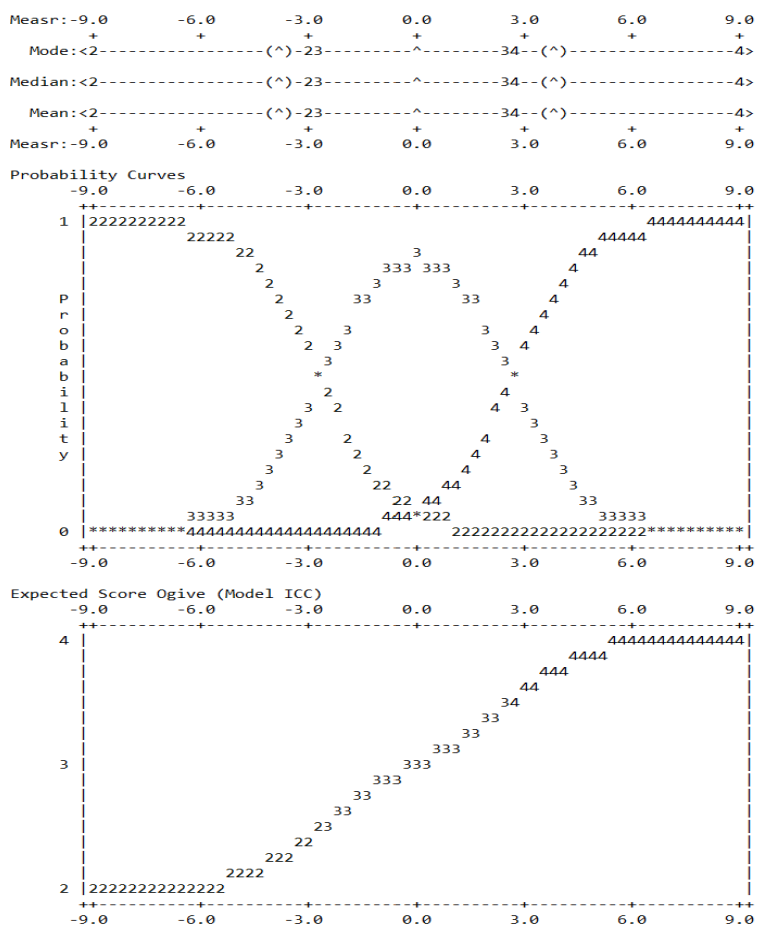


Etablissement de standards de performance 28/06/20 17:26:44
Table 8.2 Category Statistics.

Model = ?,?,?,?,2,DIFFICULTE ; Criteres: Difficulte
Rating (or partial credit) scale = DIFFICULTE,R4,G,0

Response Category Name	DATA					QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST		RASCH- THURSTONE		Cat
	Score	Total	Counts Used	%	Cum. %	Avg Meas	Exp. Meas	OUTFIT MnSq	Thresholds	S.E.	Measure at Category	-0.5	PROBABLE from	Thresholds	THURSTONE Thresholds	PEAK Prob	
Tres difficile (4)	1	0															
Difficile (3)	2	402	402	13%	13%	-3.53	-2.93	.4			(-3.71)		low		low	100%	
Moderate (2)	3	1523	1523	49%	62%	.29	.30	.3	-2.65	.08	.00	-2.64	-2.65	-2.65	-2.65	88%	
Facile (1)	4	1197	1197	38%	100%	5.77	5.57	.5	2.65	.06	(3.72)	2.65	2.65	2.64	2.64	100%	
(Mean)------(Modal)---(Median)-----																	

Scale structure for "DIFFICULTE"

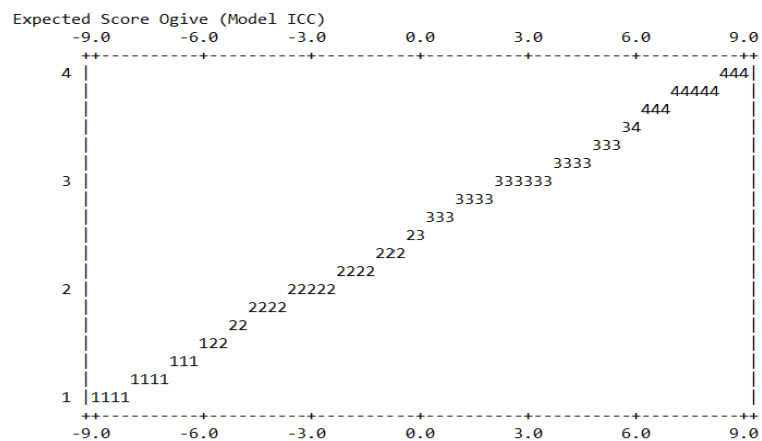
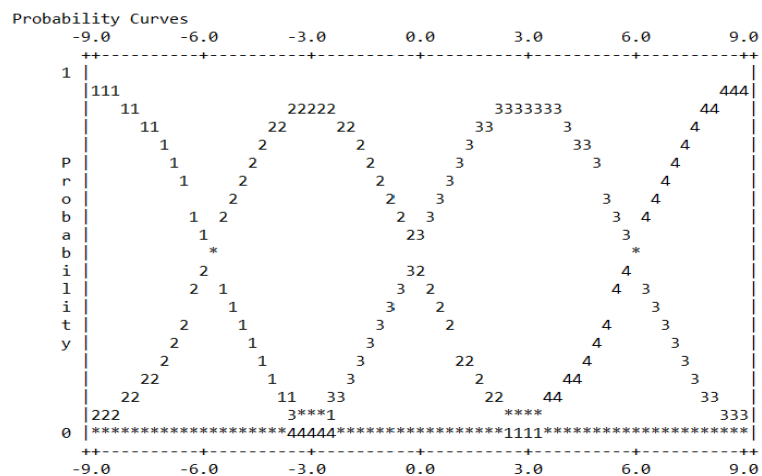
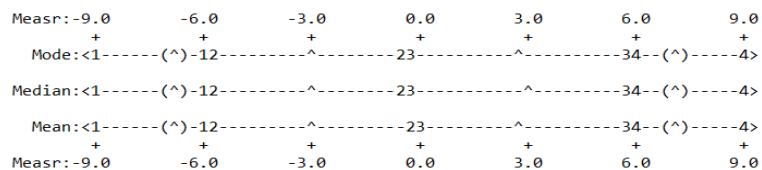


Etablissement de standards de performance 28/06/20 17:26:44
Table 8.3 Category Statistics.

Model = ?,?,?,?,3,EFFICIENCE ; Criteres: Efficiency
Rating (or partial credit) scale = EFFICIENCE,R4,S,0

Response	DATA					QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat	
Category	Category		Counts	Cum.		Avg	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK	
Name	Score	Total	Used	%	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob	
Faible (4)	1	27	27	1%	1%	1.75	-3.23	4.2			(-6.83)		low	low	100%	
Moyenne (3)	2	776	776	25%	26%	-1.12*	-1.09	1.2	-5.77	.20	-2.88	-5.75	-5.77	-5.77	89%	
Forte (2)	3	1591	1591	51%	77%	2.22	2.22	.6	-.14	.06	2.80	-.12	-.14	-.14	91%	
Tres forte (1)	4	728	728	23%	100%	7.79	7.93	1.8	5.91	.09	(6.99)	5.90	5.91	5.90	100%	
											(Mean)	(Modal)		(Median)		

Scale structure for "EFFICIENCE"



Etablissement de standards de performance 28/06/20 17:26:44

Table 4.1 Unexpected Responses (100 residuals sorted by 3,2,1,6,4,5).

Nu	Ta	Num	Ite	Nu	Ju	N	Rondes	N	G	N	Criteres	Cat	Score	Exp.	Resd	StRes	
4	4	48	48	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	5.9	-.9	-3.9
5	5	49	49	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	5.9	-.9	-3.9
5	5	53	53	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
5	5	54	54	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
5	5	55	55	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
5	5	59	59	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
12	12	100	100	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
13	13	105	105	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
14	14	106	106	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
15	15	108	108	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
16	16	109	109	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
16	16	110	110	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
16	16	112	112	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	5.9	-.9	-3.9
16	16	113	113	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
17	17	114	114	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
17	17	115	115	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
17	17	116	116	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
18	18	120	120	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
18	18	123	123	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
18	18	124	124	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
19	19	127	127	14	>D	3	ronde	3	4	D	3	Efficiency	1	1	3.2	-2.2	-5.5
20	20	132	132	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
20	20	133	133	14	>D	2	ronde	2	4	D	3	Efficiency	3	3	4.0	-1.0	-4.7
20	20	134	134	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	5.9	-.9	-3.6
20	20	135	135	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
20	20	136	136	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
20	20	137	137	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
21	21	138	138	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
21	21	139	139	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
21	21	141	141	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
22	22	142	142	14	>D	1	ronde	1	4	D	1	Ponderation	5	5	6.0	-1.0	-4.8
8	8	76	76	15	>D	3	ronde	3	4	D	3	Efficiency	4	4	2.3	1.7	3.8
12	12	96	96	15	>D	3	ronde	3	4	D	3	Efficiency	4	4	2.3	1.7	3.8
17	17	117	117	15	>D	3	ronde	3	4	D	3	Efficiency	2	2	3.0	-1.0	-3.4
1	1	4	4	16	>D	2	ronde	2	4	D	3	Efficiency	2	2	3.0	-1.0	-3.4
1	1	8	8	16	>D	1	ronde	1	4	D	3	Efficiency	1	1	3.0	-2.0	-6.3
2	2	24	24	16	>D	2	ronde	2	4	D	3	Efficiency	2	2	3.0	-1.0	-3.4
8	8	76	76	16	>D	3	ronde	3	4	D	3	Efficiency	4	4	2.2	1.8	4.5
12	12	96	96	16	>D	3	ronde	3	4	D	3	Efficiency	4	4	2.2	1.8	4.4
13	13	104	104	16	>D	1	ronde	1	4	D	3	Efficiency	4	4	2.2	1.8	4.0
13	13	104	104	16	>D	2	ronde	2	4	D	3	Efficiency	4	4	2.2	1.8	4.0
16	16	111	111	16	>D	1	ronde	1	4	D	1	Ponderation	3	3	4.8	-1.8	-3.6
17	17	117	117	16	>D	1	ronde	1	4	D	3	Efficiency	1	1	3.0	-2.0	-6.6
17	17	117	117	16	>D	3	ronde	3	4	D	3	Efficiency	1	1	3.0	-2.0	-6.6
18	18	119	119	16	>D	1	ronde	1	4	D	3	Efficiency	1	1	2.9	-1.9	-5.2
19	19	127	127	16	>D	3	ronde	3	4	D	3	Efficiency	1	1	3.0	-2.0	-6.3
23	23	143	143	16	>D	1	ronde	1	4	D	3	Efficiency	1	1	2.9	-1.9	-6.0
3	3	30	30	17	>D	3	ronde	3	4	D	3	Efficiency	2	2	3.0	-1.0	-3.4
8	8	76	76	17	>D	3	ronde	3	4	D	3	Efficiency	4	4	2.2	1.8	4.1
9	9	87	87	17	>D	1	ronde	1	4	D	3	Efficiency	4	4	3.0	1.0	3.4
9	9	87	87	17	>D	3	ronde	3	4	D	3	Efficiency	1	1	3.0	-2.0	-6.7
12	12	96	96	17	>D	3	ronde	3	4	D	3	Efficiency	4	4	2.2	1.8	4.0
19	19	127	127	17	>D	2	ronde	2	4	D	3	Efficiency	4	4	3.0	1.0	3.4
19	19	127	127	17	>D	3	ronde	3	4	D	3	Efficiency	1	1	3.0	-2.0	-6.6
2	2	24	24	18	>D	1	ronde	1	4	D	1	Ponderation	3	3	4.7	-1.7	-3.5
2	2	27	27	18	>D	3	ronde	3	4	D	3	Efficiency	2	2	3.0	-1.0	-3.4
4	4	39	39	18	>D	1	ronde	1	4	D	3	Efficiency	4	4	3.0	1.0	3.4
5	5	51	51	18	>D	2	ronde	2	4	D	3	Efficiency	1	1	2.9	-1.9	-5.6
8	8	76	76	18	>D	3	ronde	3	4	D	3	Efficiency	4	4	2.3	1.7	3.8
12	12	96	96	18	>D	3	ronde	3	4	D	3	Efficiency	4	4	2.3	1.7	3.8
17	17	117	117	18	>D	2	ronde	2	4	D	3	Efficiency	2	2	3.0	-1.0	-3.4
23	23	143	143	18	>D	1	ronde	1	4	D	3	Efficiency	4	4	3.0	1.0	3.4
1	1	8	8	19	>D	1	ronde	1	4	D	1	Ponderation	2	2	4.3	-2.3	-4.1
2	2	22	22	19	>D	2	ronde	2	4	D	3	Efficiency	4	4	2.0	2.0	5.8
2	2	27	27	19	>D	1	ronde	1	4	D	1	Ponderation	2	2	4.5	-2.5	-4.4
3	3	30	30	19	>D	2	ronde	2	4	D	3	Efficiency	4	4	3.0	1.0	3.4
4	4	39	39	19	>D	1	ronde	1	4	D	1	Ponderation	2	2	4.2	-2.2	-3.8
5	5	51	51	19	>D	2	ronde	2	4	D	3	Efficiency	1	1	2.8	-1.8	-4.2
8	8	76	76	19	>D	3	ronde	3	4	D	3	Efficiency	4	4	2.1	1.9	5.0

9	9	87	87	19	>D	3	ronde	3	4	D	3	Efficiency	1	1	2.9	-1.9	-5.9
12	12	96	96	19	>D	3	ronde	3	4	D	3	Efficiency	4	4	2.1	1.9	4.9
12	12	101	101	19	>D	1	ronde	1	4	D	1	Ponderation	2	2	4.7	-2.7	-5.4
15	15	107	107	19	>D	1	ronde	1	4	D	3	Efficiency	2	2	3.0	-1.0	-3.4
16	16	111	111	19	>D	1	ronde	1	4	D	1	Ponderation	2	2	4.7	-2.7	-5.2
18	18	121	121	19	>D	2	ronde	2	4	D	3	Efficiency	2	2	3.0	-1.0	-3.4
18	18	122	122	19	>D	1	ronde	1	4	D	1	Ponderation	2	2	4.7	-2.7	-5.3
19	19	127	127	19	>D	3	ronde	3	4	D	3	Efficiency	1	1	2.9	-1.9	-5.8
23	23	143	143	19	>D	1	ronde	1	4	D	1	Ponderation	2	2	4.2	-2.2	-3.9
1	1	8	8	20	>D	1	ronde	1	4	D	3	Efficiency	1	1	3.0	-2.0	-6.7
2	2	24	24	20	>D	1	ronde	1	4	D	3	Efficiency	1	1	3.0	-2.0	-6.6
2	2	25	25	20	>D	3	ronde	3	4	D	3	Efficiency	1	1	3.1	-2.1	-5.7
2	2	27	27	20	>D	2	ronde	2	4	D	3	Efficiency	2	2	3.0	-1.0	-3.4
4	4	39	39	20	>D	1	ronde	1	4	D	3	Efficiency	1	1	3.0	-2.0	-6.4
4	4	39	39	20	>D	2	ronde	2	4	D	3	Efficiency	4	4	3.0	1.0	3.4
4	4	39	39	20	>D	3	ronde	3	4	D	3	Efficiency	4	4	3.0	1.0	3.4
8	8	74	74	20	>D	1	ronde	1	4	D	3	Efficiency	4	4	2.2	1.8	4.3
8	8	76	76	20	>D	3	ronde	3	4	D	3	Efficiency	4	4	2.2	1.8	3.9
8	8	77	77	20	>D	1	ronde	1	4	D	3	Efficiency	4	4	2.1	1.9	4.8
9	9	80	80	20	>D	1	ronde	1	4	D	3	Efficiency	4	4	2.0	2.0	6.0
9	9	87	87	20	>D	1	ronde	1	4	D	3	Efficiency	1	1	3.0	-2.0	-6.7
9	9	87	87	20	>D	2	ronde	2	4	D	3	Efficiency	1	1	3.0	-2.0	-6.7
9	9	87	87	20	>D	3	ronde	3	4	D	3	Efficiency	1	1	3.0	-2.0	-6.7
12	12	96	96	20	>D	2	ronde	2	4	D	3	Efficiency	4	4	2.3	1.7	3.9
12	12	96	96	20	>D	3	ronde	3	4	D	3	Efficiency	4	4	2.3	1.7	3.9
12	12	97	97	20	>D	1	ronde	1	4	D	3	Efficiency	4	4	3.0	1.0	3.4
17	17	117	117	20	>D	2	ronde	2	4	D	3	Efficiency	2	2	3.0	-1.0	-3.4
17	17	117	117	20	>D	3	ronde	3	4	D	3	Efficiency	2	2	3.0	-1.0	-3.4
19	19	127	127	20	>D	1	ronde	1	4	D	3	Efficiency	1	1	3.0	-2.0	-6.7
23	23	143	143	20	>D	2	ronde	2	4	D	3	Efficiency	4	4	3.0	1.0	3.4
23	23	143	143	20	>D	3	ronde	3	4	D	3	Efficiency	4	4	3.0	1.0	3.4
-----+-----													Cat	Score	Exp.	Resd	StRes
-----+-----																	

APPENDICE F

STRUCTURE DE LA GRILLE D'ANALYSE I-SMART

Dimension professionnelle

(3 composantes; 5 compétences; 23 tâches = 23 éléments de compétence; 92 critères; X items = X indicateurs)

1. Composante de l'évaluation de la situation clinique

↳ Compétence : _____

1. Tâche : _____

○ Critère : _____

○ Critère : _____

○ Critère : _____

Indicateurs (items) :

Indicateur :

Code	Fortement en désaccord	En désaccord	Partiellement en accord	En accord	Fortement en accord
T1					

Indiquez le code de l'indicateur dans la case correspondant à votre appréciation.

À quel point trouvez-vous que cet indicateur permet de ...

		Pas du tout	Un peu	Modérément	Beaucoup
Intégrateur	mobiliser des ressources appropriées aux situations complexes				
Spécifique	cibler de façon explicite ce qui doit être accompli				
Mesurable	d'observer et d'apprécier la performance liée au critère associé				
Acceptable	susceptible d'être accepté par des experts du domaine				
Réalisable	tenir compte des ressources disponibles, des rôles et responsabilités, de la mission et de la vision de l'acupuncture au Québec				
Temporel	canaliser les efforts dans le temps imparti				

Inscrire ici les commentaires ou suggestions :

--

APPENDICE G

SONDAGE D'OPINION



FORMULAIRE D'APPRÉCIATION RELATIVE À LA RECHERCHE

L'évaluation des compétences professionnelles au moyen de standards de performance : un outil de mesure et d'évaluation en situation de stage clinique pour la profession d'acupuncteur au Québec

IDENTIFICATION

Panéliste :

ÉVALUATION

Consigne : S.V.P., indiquez, pour chacun des énoncés, votre niveau d'accord en cochant la case qui correspond à votre niveau d'appréciation. Ensuite, indiquez vos commentaires relatifs au processus dans l'espace prévu à cet effet à la fin du formulaire. Merci !

		Fortement en désaccord	En désaccord	En accord	Fortement en accord
1	Les explications m'ont permis de clairement comprendre le but de la rencontre.	☐	☐	☐	☐
2	La tâche à accomplir a été clairement expliquée.	☐	☐	☐	☐
3	La formation et l'exercice pratique m'ont aidé à comprendre comment la façon de réaliser la tâche.	☐	☐	☐	☐
4	Les descriptions des niveaux de difficulté et des niveaux d'efficience étaient claires et utiles.	☐	☐	☐	☐
5	Les discussions, en petit groupe, m'ont aidé(e) à comprendre le processus d'établissement de standard.	☐	☐	☐	☐
6	Le temps alloué aux discussions était suffisant.	☐	☐	☐	☐
7	Chacun des membres du groupe a eu une opportunité réelle et équitable d'exprimer ses idées et ses opinions.	☐	☐	☐	☐
8	J'ai été capable de suivre les instructions et de compléter, avec précision, l'ensemble des tableaux.	☐	☐	☐	☐
9	Les discussions, après le premier et le deuxième round, ont été utiles.	☐	☐	☐	☐
10	Les informations sur les résultats des autres membres du groupe m'ont été utiles.	☐	☐	☐	☐
11	Je suis confiant quant à la pertinence, la cohérence et la validité des données recueillies pour établir les standards finaux recommandés.	☐	☐	☐	☐
12	Les installations et les commodités ont permis créer un environnement de travail propice à la productivité et à l'efficacité.	☐	☐	☐	☐
13	Commentaires :				

Formulaire d'évaluation du processus d'établissement de standard destiné aux participants et aux acteurs. Adapté de Cizek et Bunch (2007, p.171).

Merci !

Prière de remettre votre formulaire complété à la chercheuse.

APPENDICE H

**Lettre d'autorisation d'utilisation
et de reproduction
du contenu des ouvrages
remise par l'OAQ
à la chercheuse**



Ordre des acupuncteurs du Québec

Montréal le 14 juin 2018

Madame Isabelle Dion
Consultante et chargé de projet
En mesure et évaluation des
Compétences professionnelles

Objet: Mémoire de maîtrise - UQAM – Madame Isabelle Dion

A qui de droit

Par la présente nous avons le plaisir de vous informer que le conseil d'administration de l'Ordre des acupuncteurs du Québec, lors de sa réunion du 18 janvier 2018, a autorisé madame Isabelle Dion à utiliser sans aucune limitation ni restriction, dans le cadre de son mémoire de maîtrise, les figures et les tableaux ainsi que les ouvrages auxquels elle a contribué.

Espérant le tout conforme, veuillez agréer, l'expression de nos sentiments les meilleurs.

Me François Houle
Le directeur général et conseiller juridique

505, boul. René-Lévesque Ouest, bureau 1106, Montréal (Québec) H2Z 1Y7

☎ 514.523.2882 / 1.800.474.5914 ☎ 514.523.9669

info@o-a-q.org www.o-a-q.org

RÉFÉRENCES

- Academy of Medical Royal Colleges (2015). *Guidance for standard setting: A framework for high stakes postgraduate competency-based examinations*. London : l'auteur. Récupéré de http://www.aomrc.org.uk/wp-content/uploads/2016/05/Standard_setting_framework_postgrad_exams_1015.pdf
- Andrich, D. et Styles, I. (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)*. Perth, WA: Murdoch University.
- Angoff, W. H. (1984). *Scales, Norms, and Equivalent Scores*. Princeton, NJ: Educational testing service. Récupéré de <http://www.ets.org/Media/Research/pdf/Angoff.Scales.Norms.Equiv.Scores.pdf>
- Angoff, W. H. (1971). Scales, Norms, and Equivalent Scores. Dans R.L. Thorndike (dir.), *Educational Measurement* (2^e éd., p. 508-600). Washington, DC: American Council on Education.
- Béland, Sébastien (2015). *Étude comparative de nouveaux indices de détection de réponses inappropriées*. (thèse de doctorat non publiée). Université du Québec à Montréal.
- Berk, R. A. (1984a). *A guide to criterion-referenced test construction*. Baltimore : Johns Hopkins University Press.
- Berk, R. A. (1984b). A Method for Reaching a Compromise between Absolute and Relative Standards in Examinations. *Journal of Educational Measurement*, 21(2), 147-152. Récupéré de <http://www.jstor.org/stable/1434539>
- Berk, R. A. (1986). A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests. *Review of Educational Research*, 56(1), 137-172. Récupéré de <http://www.jstor.org/stable/1170289>
- Blais, J.-G. (2008). Les standards de performance en éducation ». *Mesure et évaluation en éducation*. 31(2), 93-105. Récupéré de <http://id.erudit.org/iderudit/1025009ar>
- Block, A. (1979). La taxonomie des objectifs éducatifs. *Revue française de pédagogie*. 48(48), 5-11. <http://dx.doi.org/10.3406/rfp.1979.1702>

- Bloom, B. S. et Heyns, D. V. (1956, février). Development and Applications of Tests of Educational Achievement. *Review of Educational Research*. 26(1). 72-88. Récupéré de <http://www.jstor.org/stable/1169309>
- Böhning, W.R. et Zegers de Beijl, R. (1996). *L'intégration des travailleurs migrants sur le marché du travail : Les politiques et leur impact*. (Cahiers de migrations internationales no 8). Genève : Bureau international du travail (BIT). Récupéré le 5 octobre 2016 de ILO http://ilo.org/wcmsp5/groups/public/---ed_protect/---protrav/---migrant/documents/publication/wcms_201037.pdf et dans Google Scholar http://natlex.ilo.ch/wcmsp5/groups/public/---ed_protect/---protrav/---migrant/documents/publication/wcms_201037.pdf
- Bond, T. G. et Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3^e éd.). New York : Routledge.
- Bond, T. G. et Fox, C. M. (2012). *Applying the Rasch model: fundamental measurement in the human sciences* (2^e éd.). New York : Routledge.
- Boudarbat, B. et Grenier, G. (2014). *L'impact de l'immigration sur la dynamique économique du Québec*. (Rapport de recherche). Montréal : Ministère de l'Immigration, de la Diversité et de l'Inclusion du Québec (MIDI). Récupéré le 5 octobre 2016 de http://www.midi.gouv.qc.ca/publications/fr/recherches-statistiques/ETU_ImmigrProsperite_BoudarbatGrenier.pdf
- Bourguignon, C. (2011, septembre). *Conférence HDC en didactique - L'évaluation des compétences des élèves : vers un changement de paradigme?*. Communication présentée à l'Université Rouen, Académie de Royen.
- Buckendahl, C., Smith R., Impara, J. C. et Plake, B. S (2002). A Comparison of Angoff and Bookmark Standard Setting Methods. *Journal of Educational Measurement*, 39(3), 253-263. Récupéré de <http://onlinelibrary.wiley.com.proxy.bibliotheques.uqam.ca :2048/doi/10.1111/j.1745-3984.2002.tb01177.x/epdf>
- Bureau du commissaire à l'équité. (2013, avril). *Permis de succès : Vers un programme d'accès équitable : Rapport annuel 2012–2013*. Toronto : l'auteur. Récupéré de http://www.fairnesscommissioner.ca/files_docs/content/pdf/fr/OFC_%20Ann%20Report_%202012-13_%20French.pdf
- Carbonneau, M. (1993). Modèles de formation et professionnalisation de l'enseignement : analyse critique de tendances nord-américaines. *Revue des sciences de l'éducation*, 19 (1), 33–57. <http://dx.doi.org.proxy.bibliotheques.uqam.ca :2048/10.7202/031599ar>

- Caseau, Y. (2011). *Processus et entreprise 2.0 : innover par la collaboration et le lean management*. [Document électronique]. Paris : Dunod.
- Cerqueira, M. D., Arrighi, J. A. et Geiser, E. A. (2008). Physician Certification in Cardiovascular Imaging: Rationale, Process, and Benefits. *JACC: Cardiovascular Imaging*. 1(6), 801-808. <http://dx.doi.org/10.1016/j.jcmg.2008.09.004>
- Chênevert, P. (2015). *La variabilité du concept de personne raisonnable dans les décisions de la Cour suprême du Canada* (Mémoire de maîtrise). Université Laval.
- Cizek, G. J. (2012). *Setting Performance Standards: Foundations, Methods, and Innovations* (2^e éd.). New York : Routledge.
- Cizek, G. J. et Bunch, M. B. (2007). *Standard Setting: A guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA : Sage Publications.
- Cizek, J. G., Bunch, M. B. et Koons, H. (2004). Setting Performance Standards: Contemporary Methods. *Educational Measurement : Issues and Practice*. 23(4), 31-50. <http://dx.doi.org/10.1111/j.1745-3992.2004.tb00166.x>
- Cizek, G.J. (1993). Reconsidering Standards and Criteria. *Journal of Educational Measurement*, 30(2), 93-106. www.jstor.org/stable/1435458
- Comité d'adaptation de la main-d'œuvre – Personnes immigrantes (2013). *Guide d'intervention à l'intention des intervenants et intervenantes auprès des personnes immigrantes – Prix canadien de la reconnaissance des acquis*. Montréal : l'auteur. Récupéré de http://www.camo-pi.qc.ca/pdf/guide_interactif/CAMO_guide_intervention_Interactif.pdf
- Conseil interprofessionnel du Québec (2015, janvier). *Mémoire relatif au document Vers une politique québécoise en matière d'immigration, de diversité et d'inclusion*. Montréal : l'auteur. Récupéré de https://professions-quebec.org/wp-content/uploads/2015/01/CIQ_memoire_immigration_2015-01.pdf
- Conseil interprofessionnel du Québec. Fontaine, France et Bernard, Huguette. (2010, février). *Reconnaissance des compétences développées par l'expérience professionnelle : Document de soutien à la réflexion des ordres professionnels* [Ouvrage de référence imprimé]. Ouvrage de référence déposé en février 2010 au CIQ. Québec : Gouvernement du Québec.

- Conseil interprofessionnel du Québec (2009). *Mémoire sur le Projet de loi no.53 : Loi instituant le poste de Commissaire aux plaintes concernant les mécanismes de reconnaissance des compétences professionnelles*. Montréal : l'auteur.
Récupéré de <http://www.professions-quebec.org/index.php/fr/element/visualiser/id/12#804>
- Conseil interprofessionnel du Québec (2006). *Principes en matière de reconnaissance d'une équivalence de diplôme et de formation acquis hors Québec*. Montréal : l'auteur.
Récupéré de <http://www.professions-quebec.org/index.php/fr/element/visualiser/id/12>
- Côté, F. (2014). *Construire des grilles d'évaluation descriptives au collégial : Guide d'élaboration et exemples de grille* (Collection Mesure et évaluation;4). Québec : Presses de l'Université du Québec.
- De Ketele, J.M. (2006). La notion émergente de compétence dans la construction des apprentissages. [Document électronique]. Dans G. Figari et Lucie Mottier Lopez (dir.), *Recherche sur l'évaluation en éducation* (p. 17-24). Paris : L'Harmattan.
- De Ketele, J.M. et al. (2001). Chapitre 1. Renouvellement des formes d'évaluation des apprentissages : Place de la notion de compétence dans l'évaluation des apprentissages. Dans G. Figari et al. (dir.), *L'activité évaluative réinterrogée* (p. 39-43). Bruxelles : De Boeck Supérieur « Pédagogies en développement ». DOI 10.3917/dbu.figar.2001.01.0037
- Doran, G. T. (1981, novembre). There's a S.M.A.R.T. way to write management's goals and objectives. *Management Review : AMA FORUM*, 70(11), 35-36.
- Downing, S. M., Lieska, N. G., et Raible, M. D. (2003). Who gets to pass? Establishing Passing Standards for Classroom Achievement Tests in Medical Education: A Comparative Study of Four Methods. *Academic Medicine : Journal of the Association of American Medical Colleges*, 78(10), S85-S87. Récupéré de http://journals.lww.com/academicmedicine/Fulltext/2003/10001/Establishing_Passing_Standards_for_Classroom.27.aspx
- Ebel, R. L. (1972, été). Essentials of Educational Measurement: Realistic and Practical Guidelines for Better Classroom Testing. *Journal of Educational Measurement* : National Council on Measurement in Education, 9(2), [s. p.]
Récupéré de <http://www.jstor.org.proxy.bibliotheques.uqam.ca:2048/stable/pdfplus/1433814.pdf?&acceptTC=true&jpdConfirm=true>

- Egan, K. L., Ferrara, S. Schneider, M. C. et Barton, K. E. (2009). Writing Performance Level Descriptors and Setting Performance Standards for Assessments of Modified Achievement Standards: The Role of Innovation and Importance of Following Conventional Practice. *Peabody Journal of Education*, 84(4). 552-577. <http://dx.doi.org/10.1080/01619560903241028>
- Eisner, E. W. (2000). Benjamin Bloom, 1913-1999 : Elliot W. Eisner*. *Perspectives: revue trimestrielle d'éducation comparée*, 30(3), 437-446. Récupéré de <http://www.ibe.unesco.org/publications/ThinkersPdf/bloomf.pdf>
- Emploi et du Développement social Canada (EDSC). (2013). *Ressources humaines et Développement des compétences Canada : Rapport ministériel sur le rendement 2012-2013*. [Document PDF]. Rapport déposé en 2013 aux Travaux publics et Services gouvernementaux Canada (TPSGC). Ottawa : Emploi et Développement social Canada : Sa Majesté la Reine du chef du Canada. Récupéré de http://www.edsc.gc.ca/fra/publications/rmr/2012_2013/hrsdc_rmr_2012_2013.pdf
- Environnement Canada, Bureau du développement durable (2013). *Planifier un avenir durable : stratégie fédérale de développement durable pour le Canada 2013-2016*. Gatineau : l'auteur. Récupéré de http://ec.gc.ca/dd-sd/A22718BA-0107-4B32-BE17-A438616C4F7A/1339_FSDS_2013-2016_f_v10.pdf
- Faucher, C. (2012). Texte lauréat du prix Jean-Marie-Van-der-Maren – Concours 2011 : Développement d'un modèle de raisonnement clinique par son explication auprès d'optométristes de deux niveaux d'expertise contrastants. Dans Association pour la recherche qualitative. *Recherches qualitatives* (Trois-Rivières) : Association pour la recherche qualitative, 31(2), 79-107. Récupéré de http://www.recherche-qualitative.qc.ca/documents/files/revue/edition_reguliere/numero31_%282_%29/faucher.pdf
- Ferrara, S., Perie, M. et Johnson, E. (2008). Matching the Judgmental Task with Standard Setting Panelist Expertise: The Item-Descriptor (ID) Matching Method. *Journal of Applied Testing Technology*, 9(1/2). 1-22. Récupéré de <http://web.b.ebscohost.com.proxy.bibliotheques.uqam.ca:2048/ehost/pdfviewer/pdfviewer?sid=342a58ff-1877-4d76-aa50-18c642f524ed%40sessionmgr113&vid=4&hid=125>

- Forum des ministres du marché du Travail (Canada). (2011). *Cadre pancanadien d'évaluation et de reconnaissance des qualifications professionnelles acquises à l'étranger*. (Rapport d'étape novembre 2009 – décembre 2010). Ottawa : l'auteur. Récupéré de [http://www.flmm-fmmt.ca/CMFiles/FLMM %20Reports %20and %20Releases/Pan-Canadian %20Framework %20First %20Progress %20Report %20- %20FR.pdf](http://www.flmm-fmmt.ca/CMFiles/FLMM%20Reports%20and%20Releases/Pan-Canadian%20Framework%20First%20Progress%20Report%20-%20FR.pdf)
- Gagné, C. et Godin, G. (1999, février). *Les théories sociales cognitives : Guide pour la mesure des variables et le développement de questionnaire*. Groupe de recherche sur les aspects psychosociaux de la santé. Québec : Université Laval : École des sciences infirmières. Récupéré de http://www.godin.fsi.ulaval.ca/Fichiers/Rapp/Guide_mesure_variables.pdf
- Gaudreau, L. (2011). *Guide pratique pour créer et évaluer : une recherche scientifique en éducation*. [Manuel de l'élève]. Montréal : Guérin.
- Gérard, F.-M. (2010). Chapitre 15. L'évaluation des compétences par des situations complexes. Dans G. Baillat, D. Niclot et D. Ulma (dir.). *La formation des enseignants en Europe : Approche comparative* (p. 231-241). Bruxelles : De Boeck.
- Gérard, F.-M. (2009). *Évaluer des compétences - Guide pratique*. (2^e éd.). Bruxelles : De Boeck.
- Gouvernement du Québec (2009). *Glossaire des indicateurs : Références, définitions, typologie, introduction, schéma*. Rédigé par Bouchard, É. N. Québec : Secrétariat du Conseil du Trésor, Sous-secrétariat à la modernisation de l'État. Récupéré de <http://collections.banq.qc.ca/ark:/52327/bs2218575>
- Haertel, E. H. et Lorie, W. A. (2004). Validating Standards-Based Test Score Interpretations. *Measurement: Interdisciplinary Research and Perspectives*, 2(2), 61-103. http://dx.doi.org/10.1207/s15366359mea0202_1
- Hambleton, R. K., Jaeger, R. M., Plake, B. S. et Mills, C. (2000a). Setting Performance Standards on Complex Educational Assessments. *Applied Psychological Measurement*, 24(4), 355-366. <http://dx.doi.org/10.1177/01466210022031804>

- Hambleton, R. K., Brennam, R. L., Brown, W. Dodd, B., Forsyth, R. A., Mehrens, W. A., Nelhaus, J., Reckase, M., Rindone, D., van der Linden, W. J. et Zwick, R. (2000b). A Response to Setting Reasonable and Useful Performance Standards dans le National Academy of Sciences: Grading the Nations Report Card. *Educational Measurement : Issues and Practice*, 19(2), 5-14. <http://dx.doi.org/10.1111/j.1745-3992.2000.tb00024.x>
- Hambleton, R. K. et Plake, B. S. (2000). A standard-Setting Method Designed for Complex Performance Assessments: Categorical Assignments of Student Work. *Educational Assessment*, 6(3), 197-215. http://dx.doi.org/10.1207/S15326977EA0603_2
- Hambleton, R. K. et Plake, B. S. (1995). Using an Extended Angoff Procedure to Set Standards on Complex Performance Assessments. *Measurement in Education*, 8(1), 41-55. Récupéré de <http://search.ebscohost.com/login.aspx?direct=true&db=afh&AN=7364029&site=ehost-live>
- Hambleton, R. K., Swaminathan, H. et Rogers, H. J. (1991). *Fundamentals of Item Response Theory* (Ser. Measurement methods for the social sciences series, 2). Thousand Oaks, CA : Sage Publications.
- Hambleton, R. K. (1980). Test Score Validity and Standard-Setting Methods. Dans R. A. Berk (dir.), *Criterion-referenced Measurement: The state of the art* (p. 80-123). Baltimore, MD : Johns Hopkins University Press.
- Houle, F. et Roux, D. (2012). Le droit des professionnels et des gens de métier qui ont le statut de résident permanent de gagner leur vie au Québec en vertu de l'article 6 (2) (b) de la Charte canadienne des droits et libertés. *Les Cahiers de droit*, 53(1), 79-108. <http://id.erudit.org/iderudit/1007826ar>
- Impara, J. C. et Plake, B. S. (1997). Standard Setting: An Alternative Approach. *Journal of Educational Measurement*, 34(4), 353-366. <http://dx.doi.org/10.1111/j.1745-3984.1997.tb00523.x>
- INDUSTRIE CANADA. (2008). *La mobilité internationale des travailleurs hautement qualifiés : synthèse des principaux résultats et de leurs incidences sur les politiques*. Ottawa : Direction générale de l'analyse de politique micro-économique, l'auteur. Récupéré de <http://www.ic.gc.ca/eic/site/eas-aes.nsf/fra/ra02070.html>
- Jaeger, R. (1989). Certification of Student Competence. Dans R. Linn (dir.), *Educational Measurement* (3^e éd., p. 485-514). New York : American Council on Education and Macmillan.

- Kane, M. T., Crooks, T. J. et Cohen, A. S. (1999). Designing and Evaluating Standard-Setting Procedures for Licensure and Certification Tests. *Advances in Health Sciences Education*, 4(3), 195–207.
- Kane, M. (1998). Choosing Between Examinee-Centered and Test-Centered Standard-Setting Methods. *Educational Assessment*, 5(3). 129-145. http://dx.doi.org/10.1207/s15326977ea0503_1
- Kaftandjieva, F. (2010). *Methods for Setting cut Scores in Criterion-Referenced Achievement Tests: A comparative analysis of six recent methods with an application to tests of reading in EFL*. (European Association for Language Testing and Assessment, trad.). Arnhem : Cito. Récupéré de http://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf
- Kelley, T., Ebel, R. et Linacre, J. M. (2002). Item Discrimination Indices. *Rasch Measurement Transactions*, 16(3), 883–884.
- Korotkina, M. et Neuwahl, N. (2012, juillet). *La reconnaissance des qualifications professionnelles comme condition à l'immigration au Québec? Cadre juridique et enjeux politiques d'une réforme de procédure en amont*. (Mémoire de maîtrise). Université de Montréal. Récupéré de https://papyrus.bib.umontreal.ca/xmlui/bitstream/handle/1866/9669/Korotkina_Maia_2012_memoire.pdf;jsessionid=D4EE4522B989992A165B05F0294512DD?sequence=2
- Kustec, S. (2012). *Le rôle de l'offre de travailleurs migrants sur le marché du travail canadien*. Ottawa : Citoyenneté et Immigration Canada. Récupéré le 5 octobre 2016 de <http://www.cic.gc.ca/francais/ressources/recherche/2012-migrant/documents/pdf/migrant2012-fra.pdf>
- Lainey, S. et Dion, I. (2014). *Validation des tâches et des indicateurs de performance liés à l'évaluation de la compétence des candidats demandant le droit d'exercer la profession d'acupuncteur au Québec*. [Document non publié]. Montréal : Ordre des acupuncteurs du Québec.
- Lamari, M. (2011). Mesurer l'incommensurable : quels défis et quels rendements pour les investissements en capital humain dans les organisations gouvernementales?, *Télescope*, vol. 17(3), 1-30.
- Le Boterf, G. (2007). Qu'est-ce qu'un professionnel compétent?. Dans G. Le Boterf (dir.) (p. 23-58). *Professionnaliser : le modèle de la navigation professionnelle*. Paris : Éditions Eyrolles.

- Le Boterf, G. (2002). *Ingénierie et Évaluation des compétences (4^e éd. entièrement renouvelée)*. Paris : Éditions d'Organisation.
- Le Boterf, G. (1998). Évaluer les compétences : Quels jugements? Quels critères? Quelles instances?. *Éducation permanente*. 135(2). 143-152. Récupéré de http://s3.amazonaws.com/academia.edu.documents/4282546/135_LE_BOTERF.pdf?AWSAccessKeyId=AKIAJ56TQJRTWSMTNPEA&Expires=1480638100&Signature=pouQ9yGMEt%20FXVdaw9Z5E2N0tht0%3D&response-content-disposition=inline%3B%20filename%3DEvaluer_les_compences_quels_jugements.pdf
- Le Boterf, G. (1997). *De la compétence à la navigation professionnelle*. Paris : Les Éditions d'Organisation.
- Legendre, R. (2005). *Dictionnaire actuel de l'Éducation (3^e éd.)*. Montréal : Guérin.
- Leprohon, J., Lessard, L.-M., Lévesque-Barbès et Bellavance, M. (2009). *Mosaïque des compétences cliniques de l'infirmière : Compétences initiales*. Montréal : Ordre des infirmières et infirmiers du Québec.
- Leprohon, J. Lessard, L.-M., Lévesque-Barbès, H. (2001). *Mosaïque des compétences cliniques de l'infirmière : compétences initiales*. Montréal : Ordre des infirmières et infirmiers du Québec.
- Leroux, J. L. (2010). *L'évaluation des compétences au collégial : un regard sur des pratiques évaluatives - PAREA 2010*. (Rapport de recherche PAREA). Saint-Hyacinthe : Cégep de Saint-Hyacinthe. Récupéré de <http://www.cdc.qc.ca/parea/787400-leroux-evaluation-compences-st-hyacinthe-PAREA-2010.pdf>
- Leroux, J. L. et N. Bigras (2003). *L'évaluation des compétences : une réalité accessible dans nos collèges*. Saint-Hyacinthe : Cégep de Saint-Hyacinthe. Récupéré de https://www.usherbrooke.ca/performa/fileadmin/sites/performa/documents/Recherches_subventionnees/Leroux_et_Bigras_2003.pdf
- Linacre, J. M. (2019, 4 mars). Standard Error of Measurement [Message publié sur un forum de discussion]. Récupéré de <https://raschforum.boards.net/thread/1059/standard-error-measurement>
- Linacre, J. M. (2016, 14 février). OSCE, Disjoint Subsets, Examiner Pairs - The Approach I Took [Message publié sur un forum de discussion]. Récupéré de <https://raschforum.boards.net/thread/414/disjoint-subsets-examiner-pairs-approach>

- Linacre, J. M. (2012). Many-Facet Rasch Measurement: Facets tutorial [Tutoriel, 4]. Récupéré de <https://winsteps.com/facetman/tutorialpdfs.htm>
- Linacre, J. M. (2010). Transitional Categories and Usefully Disordered Thresholds. *Online Educational Research Journal*, 1(3). 1-10. Récupéré de <http://community.dur.ac.uk/p.b.tymms/oerj/publications/2.pdf>
- Linacre, J. M. (1994). *Many-Facet Rasch Measurement*. Chicago, Illinois, MESA press
- Livingston, S. A. et Zieky, M. J. (1983). *A Comparative Study of Standard-Setting Methods - Research Report Series* (RR-83-38). Princeton, N. J. : Educational Testing Service. <http://dx.doi.org/10.1002/j.2330-8516.1983.tb00038.x>
- Loomis, S. C. Hanick, P. L., Bay, L. et Crouse, J. D. (2000, décembre). *Developing Achievement Levels for the 1998 NAEP in Civics Interim Report: Field Trials*. Washington : National Assessment Governing Board . Récupéré de <http://www.nagb.org/content/nagb/assets/documents/publications/achievement/civics-achievement-levels-ft-report-2000.pdf>
- Louis, R., Justras, F. et Hensler, H. (1996). Des objectifs aux compétences : implications pour l'évaluation de la formation initiale des maîtres. *Revue canadienne de l'éducation*, 21(4), 414-432.
- McGahee, T. W. et Ball, J. (2009). How to Read and Really Use an Item Analysis. *Nurse educator*, 34(4), 166–171.
- Ministère de la Jeunesse, de l'Éducation nationale et de la Recherche (MJÉNR). (2003). Dix-huit questions sur le système éducatif : Synthèses de travaux de la DEP. *Éducation & Formations*. 66 (juillet-décembre 2003). 201-213. Récupéré de <http://cache.media.education.gouv.fr/file/27/1/5271.pdf>
- Ministère de l'Immigration et des Communautés Culturelles. (2013). *Résumé des projets*. Montréal : l'auteur. Récupéré de <http://www.micc.gouv.qc.ca/publications/fr/divers/ResumeProjetsOrdres-Juin2013.pdf>
- Ministère de l'Immigration et des Communautés culturelles, Direction des affaires publiques et des communications. (2005). *Les personnes immigrantes formées à l'étranger et l'accès aux professions et métiers réglementés : rapport*. Montréal : Gouvernement du Québec. Récupéré de <http://collections.banq.qc.ca/ark:/52327/bs47266>

- Nedelsky, L. (1954). Absolute Grading Standards for Objective Tests. *Educational and Psychological Measurement*, 14(1), 3–19.
<https://doi.org/10.1177/001316445401400101>
- Norcini, J. (2003). Setting Standards on Educational Tests. *Medical Education*, (37), 464-469.
- Noël, C. (2010). *Retour aux études universitaires et acquis expérientiels : Perceptions d'acteurs du communautaire*. (Mémoire de maîtrise). Université de Sherbrooke.
 Récupéré de
<http://www.collectionscanada.gc.ca/obj/thesescanada/vol2/002/MR65636.PDF>
- Noijons, J., Béréšová, J., Breton, G. et Szabó G. (2011). *Relier les examens de langues au Cadre européen commun de référence pour les langues : Apprendre, enseigner, évaluer (CECR) – Les points essentiels du Manuel*. Strasbourg : Éditions du Conseil de l'Europe. Récupéré de
<http://www.ecml.at/tabid/277/PublicationID/67/Default.aspx>
- Office of Professional Examination Services (OPES): California Department of Consumer Affairs. (2016, janvier). *Review of the National Certification Commission for Acupuncture and Oriental Medicine Examinations*. Californie : l'auteur. Récupéré de
http://www.acupuncture.ca.gov/pubs_forms/nccaom_audit.pdf
- Office des professions du Québec. (2017, 27 mars). *Liste des ordres professionnels*. Québec : l'auteur. Récupéré de <https://www.opq.gouv.qc.ca/Rapports/>
- Office québécois de la langue française. (2012a). Fiche terminologique : Évaluation (domaine de l'éducation). Dans Le grand dictionnaire terminologique (GDT). Récupéré le 3 mars 2017 de
http://www.granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=1199565
- Office québécois de la langue française. (2012b). Fiche terminologique : Standard (domaine de l'éducation). Dans Le grand dictionnaire terminologique (GDT). Récupéré le 3 mars 2017 de
http://www.granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=8358598
- Office québécois de la langue française. (2012c). Fiche terminologique : Standard (domaine de la normalisation). Dans Le grand dictionnaire terminologique (GDT). Récupéré le 3 mars 2017 de
http://www.granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=8355205

- Ordre des acupuncteurs du Québec. (2010, 2015). *Polygramme des compétences liées à l'exercice de la profession d'acupuncteur au Québec*. Montréal : l'auteur. Récupéré de <https://www.o-a-q.org/getmedia/7550f51d-e415-4ab6-a1bb-560c1d1b7da4/Polygramme-des-competences.aspx>
- Ordre des acupuncteurs du Québec. (2014a). *Cahier d'observation, des comportements cliniques, destiné aux maîtres cliniques dument accrédités par l'Ordre des acupuncteurs du Québec*. Montréal : l'auteur. Récupéré de <http://o-a-q.org/getmedia/2fa38994-ee0-4da3-b674-f875b8e99f99/Cahier-d-observaiton-des-comportements-cliniques.aspx>
- Ordre des acupuncteurs du Québec (2014b, mai). *Grilles d'observation des savoir-faire liés aux comportements cliniques*. [Document non publié]. Montréal : l'auteur.
- Ordre des infirmières et infirmiers de l'Ontario. (2014, avril). *Élaborer des objectifs d'apprentissage SMART*. (Rapport no. 54047). Toronto : l'auteur. Récupéré de <http://www.cno.org/Global/docs/qa/SMART%20Goals%20Intro%20French.pdf>
- Ordre des physiothérapeutes de l'Ontario. (2012). *Guide sur le portefeuille professionnel : Programme d'assurance de la qualité*. Toronto : l'auteur. Récupéré de http://www.collegept.org/Assets/qualitymanagement/QM%20Resources%200FRENCH/QM_Portfolio_Guide_FR_130222.pdf
- Ostermeier, M, Linde, S., Lay, J. et Prediger, S. (2015). Des indicateurs « SMART » pour mesurer le travail décent : proposition dans la perspective du programme de développement pour l'après-2015. *Revue internationale du Travail*, 154(3), 315-335.
- Ouellette, M. (2006). *La reconnaissance des acquis dans les universités québécoises : état de la situation*. [Rapport de recherche]. [s. l.]. Montréal : Association canadienne d'éducation des adultes des universités de langue française. Récupéré de <http://www.ageefep.qc.ca/pdf/rda/acdeaulf.pdf>
- Perrenoud, P. (2004, mars). Évaluer des compétences. *Éducateur*, N° spécial « La note en pleine évaluation », mars, 8-11. Récupéré de <http://darous.blogvie.com/files/2012/12/evaluer-des-competnces.pdf>
- Perrenoud, P. (2002). D'une métaphore à l'autre : transférer ou mobiliser ses connaissances?. Dans J. Dolz et Ollagnier, E. (dir.). *L'énigme de la compétence en éducation* (p. 45-60). Bruxelles : De Boeck Supérieur. DOI 10.3917/dbu.dolz.2002.01.0045

- Plake, B. S. et Giraud, G. (1998, avril). *Effect of a Modified Angoff Strategy for Obtaining Item Performance Estimates in a Standard Setting Study*. San Diego, CA : l'auteur. Récupéré de <http://files.eric.ed.gov/fulltext/ED422354.pdf>
- Plake, B. S. (1998). Setting Performance Standards for Professional Licensure and Certification. *Applied Measurement in Education*, 11(1), 65-80. Récupéré de <http://search.ebscohost.com.proxy.bibliotheques.uqam.ca:2048/login.aspx?direct=true&db=a9h&AN=7366417&lang=fr&site=ehost-live>
- Poggio, J.P., Glasnap, D.R. et Eros, D.S. (1981, avril). *An Empirical Investigation of the Angoff, Ebel, and Nedelsky Standard Setting Methods*. Communication présentée à the annual meeting of the American Educational Research Association, Los Angeles. Récupéré de <http://files.eric.ed.gov/fulltext/ED205552.pdf>
- Raïche, G., Paquette-Côté, K. et D., Magis. (dir.). (2011). *Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation : Volume 1 – La mesure*. Sainte-Foy : Presses de l'Université du Québec. Récupéré de Google books https://books.google.ca/books?id=c10nDwAAQBAJ&pg=PT218&dq=Des+m%C3%A9canismes+pour+assurer+la+validit%C3%A9+de+l%27interpr%C3%A9tation+de+la+mesure+en+%C3%A9ducation&hl=fr&sa=X&ved=2ahUK_EwiNyuTF_PsAhXFJt8KHS60AXoQ6AEwAHoECAMQAg#v=onepage&q=2011&f=false
- Raïche, G. (2005). *Évaluation des compétences à l'enseignement supérieur : cheminement du développement de la compétence (ch. 14)*. [Document non publié]. Université du Québec à Montréal. Récupéré de http://www.er.uqam.ca/nobel/r17165/ENSEIGNEMENT/.../CH14_Seuil.pdf
- Roegiers, X. (2005). L'évaluation selon la pédagogie de l'intégration : est-il possible d'évaluer les compétences des élèves?. Dans K. Toualbi-Thaâlibi et S. Tawil (dir.), *La refonte de la pédagogie en Algérie – Défis et enjeux d'une société en mutation* (p. 107-124). Alger : UNESCO-ONPS. <http://unesdoc.unesco.org/images/0014/001498/149828f.pdf>
- Roegiers, X. (2000). *Une pédagogie de l'intégration : compétences et intégration des acquis dans l'enseignement*. Bruxelles : De Boeck.

- Santé Canada et l'Agence de la santé publique du Canada. (2013, septembre). *Évaluation du Programme de contributions pour les politiques en matière de soins de santé*. Ottawa : Gouvernement du Canada. Récupéré de http://www.hc-sc.gc.ca/ahc-asc/alt_formats/pdf/performance/eval/hcpc-evaluation-pcpms-fra.pdf
- Scallon, G. (2015). *Des savoirs aux compétences : Explorations en évaluation des apprentissages*. Montréal, Québec : Éditions du Renouveau Pédagogique Inc.
- Scallon, G. (2004a). L'évaluation des compétences et l'importance du jugement. *Pédagogie collégiale*. 18(1). 14-20. Récupéré de http://aqpc.qc.ca/sites/default/files/revue/Scallon_18_1.pdf
- Scallon, G. (2004b). *L'évaluation des apprentissages dans une approche par compétences*. Saint-Laurent, Québec : Éditions du Renouveau Pédagogique Inc.
- Scallon, G. (1981, Hiver). La généralisabilité des seuils de performance établis par des appréciateurs : étude de trois procédés. *Revue des sciences de l'éducation*. 7(1). 3-21. Récupéré de <http://dx.doi.org/10.7202/900313ar>
- Scholz, C. et V. Stein (2011). « Les universités allemandes en mutation et les leçons à tirer par les administrations publiques pour leur gestion axée sur les connaissances », *Télescope*, vol. 17(3), 31-53.
- Service public fédéral – Personnel et Organisation (Belgique). (2010, septembre). *Gestion des compétences au sein de l'Administration fédérale : Articuler la gestion des compétences et les cercles de développement – Guide méthodologique*. Bruxelles : Jacky Leroy. Récupéré de http://www.fedweb.belgium.be/fr/binaries/1009_CM_CD_guide_final_FR_tcm119-235769.pdf

- Shields, A. (2007, 17 novembre). Reconnaissance des acquis – Le Québec connaît un énorme retard. *Le Devoir*. B6. Récupéré de http://www.biblio.eureka.cc.proxy.bibliotheques.uqam.ca:2048/WebPages/Pdf/SearchResult.aspx?Date=2007-11-17 00:00:00&Source=LE_P
- Skakun, E. N. et Kling, S. (1980, Automne). Comparability of Methods for Setting Standards. *Journal of Educational Measurement*. 17(3). 229-235. Récupéré de <http://dx.doi.org/10.1111/j.1745-3984.1980.tb00830.x>
- Tardif, J. (2006). L'évaluation des compétences. Documenter le parcours de développement. Montréal : Chenelière éducation.
- Taube, K. T. (1997, décembre). The Incorporation of Empirical Item Difficulty Data into the Angoff Standard-Setting Procedure. *Evaluation & the Health Professions*, 20(4), 479-498.
- Tremblay, G. (1994, mars). À propos de l'approche par compétences appliquée à la formation générale. *Pédagogie collégiale*. 7(3). 12-16. Récupéré de http://aqpc.qc.ca/sites/default/files/revue/tremblay_gilles_07_3.pdf
- Wiggins, G. (1989, mai). A True Test: Toward More Authentic and Equitable Assessment. *The Phi Delta Kappan*. 70(9). 703-713. Récupéré de <http://www.jstor.org.proxy.bibliotheques.uqam.ca:2048/stable/pdf/20404004.pdf>

GLOSSAIRE

L'abondance et la diversité des définitions que l'on recense dans la littérature nous amènent à présenter quelques fragments des définitions retenues telles que proposées par Legendre dans la 3e édition du *Dictionnaire actuel de l'éducation* publiée en 2005.

Acquis expérientiels : « Dans le contexte de la reconnaissance des acquis, expérientiel qualifie le plus souvent des apprentissages qui se sont déroulés au cours d'expériences de travail et de vie, et généralement sans que ces apprentissages aient été planifiés et recherchés en eux-mêmes, les expériences vécues ayant alors d'autres finalités » (Legendre, 2005, p. 657).

Accréditation : « Reconnaissance et approbation des normes curriculaires d'un établissement d'enseignement par un organisme externe, une association, une corporation professionnelle ou un établissement prestigieux » (CEPES : UNESCO, 1992, cité dans Legendre, 2005, p. 36).

Approche par compétences : « [...] consiste à définir les compétences inhérentes à l'exercice d'une profession et à les formuler, dans les programmes, en objectifs et en standards » (D.G.F.P.T., MEQ, 1995, cité dans Legendre, 2005, p. 115).

Certification : « un mécanisme volontaire de reconnaissance de la compétence d'une personne pratiquant un métier ou une profession » (Landry, 1987, cité dans Legendre, 2005, p. 197).

Compétence professionnelle : « Aptitude à exercer efficacement un métier, une fonction ou certaines tâches spécifiques, avec toutes les qualifications requises à cet effet » (BIT, 1987, cité dans Legendre, 2005, p. 257).

Comportement observable : « Comportement externe d'une personne, pouvant être suffisamment manifeste pour que deux ou plusieurs observateurs s'entendent sur la présence ou l'absence de ce comportement » (Legendre, 2005, p. 260).

Critère d'évaluation : « Repère observable tant pour soutenir le développement de la compétence que pour en juger » (MEQ, 2001, cité dans Legendre, 2005, p. 314).

Critère de performance : « Seuil minimal à partir duquel on considère qu'un objectif est atteint. [...] Les critères de performance indiquent la qualité du travail attendu, c'est-à-dire le seuil d'acceptabilité ou de réussite d'une performance : limites de temps ou d'espace; proportion d'exercices à réussir; nature d'une production, d'un comportement, d'un acte moteur. [S'exprime] de diverses façons : par des unités de temps (4 minutes), des pourcentages (10 %), des proportions (1 sur 5), des normes de précisions (à 4 décimales près), des quantités (5 ou moins), des adverbes (correctement), sans erreur » (NADEAU, 1980, cité dans Legendre, 2005, p. 314).

Critère de rendement : « Critère qui sert à préciser et à expliciter le succès d'une intervention, d'une opération, d'une organisation, d'un projet, d'un programme au regard de ses productions, de ses résultats et de ses retombés en vue d'une prise de décision et d'un jugement de valeur » (Legendre, 2005, p. 314).

Domaine : « Sphère où s'exercent les activités d'une personne. [...] Univers d'éléments auquel un programme, un objectif, une épreuve ou un item fait référence » (MEQ, D.G.D.P., mars 1984, cité dans Legendre, 2005, p. 435)

Échelle : « [...] Instrument [de mesure] utilisé pour assigner un nombre qui indique la qualité de l'information ou la quantité de la caractéristique mesurée. [...] Une échelle existe en toute situation de mesure; toute mesure fait appel à un ensemble de règles pour assigner des nombres à des objets ou à des événements; ce sont ces règles qui donnent un sens aux mesures (nombres assignés). [...] Selon le type de règles, on distingue souvent quatre types principaux d'échelles de mesure en éducation [soit une qualitative nommée] échelle nominale [et trois échelles quantitatives nommées :], échelle ordinale (ou ordonnée), échelle d'intervalle (ou intervalles égaux) et échelle de rapport (ou proportionnelle ou de proportions » (Legendre, 2005, p. 470).

Échelle d'appréciation (échelle descriptive, échelle d'évaluation, échelle de notation) : « Instrument [de mesure] qui permet de consigner son opinion à propos des qualités d'un apprentissage ou d'une performance. [...] Il y a plusieurs formes d'échelles d'appréciation[. Par ex. une] échelle descriptive » (MEQ, Direction générale du développement pédagogique : Département d'éducation et pédagogique (DGDP : DEP) (03.84), cité dans Legendre, 2005, p. 471).

Échelle d'intervalle (échelle quantitative) : « Échelle de mesure permettant de préciser quantitativement les différences qui séparent les entités mesurées [où] le point zéro (origine) est fixé arbitrairement [...] » (Legendre, 2005, p. 471).

Échelle de Likert : « Échelle de mesure des attitudes, des croyances et des jugements, comprenant de trois à onze catégories égales, s'étendant d'une extrémité inférieure à une extrémité supérieure de réponses possibles. Par ex. : 1. très important; 2. important; 3. neutre; 4. peu important; 5. aucune importance » (Legendre, 2005, p. 471).

Échelle de mesure : « Il existe quatre échelles de mesure : nominale, ordinale, à intervalle et de rapport » (Legendre, 2005, p. 868).

Échelle de rapport (échelle quantitative, échelle de proportion/proportionnelle) : « Échelle de mesure jouissant des mêmes propriétés que l'échelle d'intervalle, mais avec une valeur absolue (plutôt que conventionnelle) de zéro exprimant l'absence totale de la caractéristique mesurée et permettant l'utilisation de rapports entre les valeurs de l'échelle. [...] Les échelons sont égaux [et] permettent d'effectuer des relations significatives entre les mesures (par ex. une personne de 2 mètres est deux fois plus grande qu'une autre de 1 mètre). [...] De toutes les échelles, l'échelle de rapport est la plus précise; elle est cependant la moins utilisée en éducation » (Legendre, 2005, p. 472).

★Échelle dichotomique (échelle nominale ou artificielle) : « Échelle de mesure ne comprenant que deux catégories. [Elle] peut être nominale [et] également être artificielle, c'est-à-dire lorsque la variable est continue, mais que seules deux catégories sont utilisées pour regrouper deux étendues de cette variable, séparées par un point de césure (par ex. : réussite-échec) » (Legendre, 2005, p. 472).

★Échelle discrète (échelle nominale ou dichotomique) : « Échelle de mesure comprenant deux catégories ou plus [qui] peut-être nominale [ou] ordonnée » (Legendre, 2005, p. 472).

Échelle nominale (échelle qualitative, échelle discrète) : « Échelle de mesure permettant de classer, de nommer ou d'énumérer des sujets ou des objets selon des propriétés, qui n'ont pas de relation quantitative entre elles » (Legendre, 2005, p. 472).

★Échelle ordinale (échelle quantitative) : « Échelle de mesure permettant d'indiquer l'ordre ou le rang des sujets ou des objets au regard d'une caractéristique définie sur un continuum. Il n'existe [...] pas de différence constante de la caractéristique entre tous les rangs consécutifs. [...] Les seules relations permises entre les nombres sont celles de plus grand ($>$), plus petit ($<$) ou égal ($=$), ce dernier exprimant l'équivalence des sujets recevant le même rang » [(par ex. une échelle d'appréciation de la difficulté d'items à réponses polytomiques rangées par niveau de difficulté croissant, du plus facile au plus difficile] (Legendre, 2005, p. 473).

Échelle quantitative (échelle d'intervalle, de rapport ou ordinale) : « Ensemble des échelles où les valeurs numériques possèdent une signification mathématique » (Legendre, 2005, p. 473).

Échelle unidimensionnelle : « Échelle qui ne mesure qu'une seule variable » (Legendre, 2005, p. 473).

Épreuve : « Exercice écrit, oral ou pratique, que subit une personne au moment d'un examen ou d'un concours afin d'être évaluée selon ses capacités ou ses connaissances » (O.L.F., 1998, cité dans Legendre, 2005, p. 608).

« Instrument de mesure composé d'un ensemble d'items choisis en fonction des objectifs dont on veut mesurer l'atteinte » (MEQ, D.G.D.P. : DEP, mars 1984, cité dans Legendre, 2005, p. 608).

Évaluation : « Démarche permettant de porter un jugement, à partir de normes ou de critères établis, sur la valeur d'une situation, d'un processus ou d'un élément donné, en vue de décisions pédagogiques ou administratives » (MEQ, 2004, cité dans Legendre, 2005, p. 631).

« Opération de confrontation, de mise en rapport, entre des attentes et une réalité » (HADJI, 1997, cité dans Legendre, 2005, p. 631) ».

« Ensemble d'activités d'appréciation et de mesure de résultats obtenus par la poursuite d'un objectif » (BIBEAU, 1987, cité dans Legendre, 2005, p. 631).

« Activité consistant à accorder une valeur aux résultats provenant d'une mesure en les situant par rapport à un critère ou à une norme » (C.T.E.Q., 1988, cité dans Legendre, 2005, p. 631).

« Dans le sixième niveau de la taxonomie des objectifs d'apprentissage du domaine cognitif de Benjamin S.BLOOM et coll. [...] » (1956, cité dans Legendre, 2005, p. 630).

« Les principales étapes du processus d'évaluation sont les suivantes : [1.] énoncer clairement les objectifs de l'évaluation et en vérifier la pertinence par rapport à la décision à prendre [2.] recherche et choisir des critères opérationnels valides, ainsi que des indicateurs pour ces critères [3.] déterminer les informations pertinentes à recueillir en regard des critères opérationnels choisis [4.] déterminer une stratégie du recueil de l'information [5.] recueillir l'information de façon fiable (mesure, observations, entretiens...) [6.] confronter l'information recueillie aux critères élaborés [7.] formuler les conclusions de façon précise et claire pour l'auteur de la prise de décision » (DE GETELE; ROEGIERS, 1996, cité dans Legendre, 2005, p. 631).

Évaluation certificative : « Processus d'évaluation débouchant sur une décision dichotomique de réussite ou d'échec relatif à une période d'apprentissage, d'acceptation ou de rejet d'une promotion, de poursuite d'une action ou de l'arrêt de celle-ci » (De Ketele et Roegiers, 1996, cité dans Legendre, 2005, p. 633).

Évaluation critériée (criterion-referenced, évaluation normative) : « Mode d'évaluation où la performance du sujet dans l'accomplissement d'une tâche spécifique est jugée par rapport à un seuil ou à un critère de réussite, déterminé dans la formulation du ou des objectifs explicitement visés, indépendamment de la performance de tout autre sujet » (Legendre, 2005, p. 634).

Évaluation de classement (évaluation certificative) : « Processus de mise en rang des personnes selon leurs performances » (De Ketele et Roegiers, 1996, cité dans Legendre, 2015, p. 633).

« Évaluation diagnostique dont le but est d'identifier le niveau de compétence ainsi que les obstacles pédagogiques d'un sujet avant qu'il entreprenne l'étude d'une nouvelle matière ou le développement de nouvelles habiletés » (Legendre, 2005, p. 635).

★Évaluation de la capacité professionnelle : « Méthode qui consiste à évaluer systématiquement les tâches accomplies par un candidat, dans des conditions réelles ou simulées, pour s'assurer de sa compétence professionnelle ainsi que pour l'aider à planifier sa carrière » (Legendre, 2005, p. 636).

« Un agent spécialisé observe, note et évalue des échantillons de la performance du sujet au travail; ces observations sont comparées avec des critères comme déterminants d'un comportement professionnel adéquat. Toutes les données médicales, psychologiques, sociales, éducatives, professionnelles et culturelles possibles sont rassemblées aux fins d'une évaluation d'ensemble » (Legendre, 2005, p. 636).

Évaluation de la performance (rendement professionnel) : « Jugement porté sur l'attitude et sur les capacités techniques d'une personne dans l'exercice de son activité professionnelle ainsi que sur les résultats obtenus et les progrès réalisés » (BIT, 1987, cité dans Legendre, 2005, p. 636).

Évaluation de performance (mesure des performances) : « Ensemble de techniques de contrôle servant à s'assurer que les réalisations des centres de responsabilités de l'entreprise coïncident avec les normes ou les objectifs quantifiés établis pour chacun d'eux » (Villers, 1981, cité dans Legendre, 2005, p. 636).

★Évaluation des acquis : « Action de jauger, à partir d'outils de mesure, le degré de maîtrise de connaissances, d'habiletés ou d'aptitudes dans des activités et dans un cadre déterminé » (MEQ, 1982, cité dans Legendre, 2005, p. 639).

Évaluation des ressources humaines (évaluation du personnel) : « Action conduisant à porter un jugement de valeur sur les compétences et les performances des personnels » (Legendre, 2005, p. 640).

« Parmi les principaux moyens d'évaluation, nous retrouvons l'entrevue, la notation écrite, l'évaluation collective et l'autoévaluation [et] de bandes vidéo » (Legendre, 2005, p. 640).

Évaluation professionnelle (de la capacité professionnelle) : « Méthode qui consiste à évaluer systématiquement les tâches accomplies par un candidat, dans des conditions réelles ou simulées, pour s'assurer de sa compétence professionnelle ainsi que pour l'aider à planifier sa carrière » (Legendre, 2005, p. 636).

Évaluation normative (norm-referenced) : « Mode d'évaluation où la performance d'un sujet est comparée, au moyen de rang centile ou de stanine, à celle des autres personnes d'un groupe de référence d'après un même instrument. [Cela signifie] qu'aucun jugement n'est porté sur la mesure observée » (Legendre, 2005, p. 646).

Expert : « Personne qui a acquis des compétences exceptionnelles, a connu des expériences et des succès remarquables dans un domaine, qui est donc en mesure de contribuer significativement et efficacement à la résolution de problèmes ou à la réalisation de projets dans sa sphère d'activités » (Legendre, 2005, p. 657).

★Grille d'observation (instrument d'observation) : « Instrument de mesure qui permet de constater les particularités d'une action, d'un produit ou d'un processus en fournissant une liste d'éléments observables ainsi qu'une façon d'enregistrer les observations [(par ex.] comportement observable; technique d'observation[] » (MEQ, DGDP : DEP (03.84), cité dans Legendre, 2005, p. 724).

★Indicateurs : « Symptôme, manifestation qui signale volontairement et précisément la présence, l'absence, l'état ou l'évolution d'une chose, d'un objet, d'une situation, etc. [Cette] manifestation [est] observable et mesurable. [C'est un] facteur quantitatif ou qualitatif qui, à la suite de l'observation rigoureuse et systématique, permet d'identifier les changements apportés à la suite d'une intervention et, par le fait même, d'évaluer le travail d'un agent de développement » (Legendre, 2005, p. 762).

Instrument d'évaluation : « servent à consigner des données qui sont obtenues à la suite d'une appréciation faite par [...] l'application d'instruments de mesure » (Legendre, 2005, p. 782).

★Instrument de mesure : « Ensemble d'items regroupés pour recueillir des données pertinentes quant au jugement à poser et à la décision à prendre » (Legendre, 2005, p. 782).

« [Il peut] s'agir [...] d'épreuve, de grille d'observation, d'échelle d'appréciation ou autres » (MEQ;D.G.D.P., 1984, cité dans Legendre, 2005, p. 782).

« Pour qu'un instrument soit bon, il doit répondre aux caractéristiques suivantes : être valide, commode, efficace, avoir une portée, une bonne présentation et être réaliste » (NADEAU, 1981, cité dans Legendre, 2005, p. 782).

Item : « Élément d'un instrument d'évaluation ou de mesure. Banque d'items [d'un test ou d'une épreuve qui] peut être un problème à résoudre, une question ou une tâche plus ou moins complexe » (Legendre, 2005, p. 810).

Item à réponse choisie (question fermée) : « Item dont les réponses sont formulées à l'avance et entre lesquelles [un candidat] doit choisir celle qu'il croit être la réponse correcte ou admise » (MEQ, DGDP : DEP (03.84), cité dans Legendre, 2005, p. 810). « [Divers types possibles :] item fermé, item à choix multiple, exercice d'appariement, item vrai ou faux ».

Item à réponse construite (question ouverte) : « Item dont la réponse doit être formulée par [le candidat] MEQ, DGDP : DEP (03.84), cité dans Legendre, 2005, p. 810). [Divers types possibles :] item ouvert; item à réponse courte; item à réponse élaborée; énumération; identification; tâche; production » (Legendre, 2005, p. 810).

Licence : « Un mécanisme obligatoire de reconnaissance de la compétence, qui est défini dans un texte de loi » (Landry, 1987, cité dans Legendre, 2005, p. 197).

Mesure : « Processus d'attribution de nombres aux résultats d'un test ou d'une autre forme d'évaluation selon une règle précise » (Linn et Grönlund, 1995, cité dans Legendre, 2005, p. 867).

« Lors de l'opération de mesure, on décrit quantitativement les observations faites de phénomènes choisis. C'est lors de l'évaluation que l'on se prononcera sur la compétence » (Legendre, 2005, p. 868).

« La mesure en éducation est plus large que le « testing » [...], quoique la mesure emploie fréquemment les tests, elle peut aussi utiliser des instruments d'observation, des listes de vérification et d'autres méthodes qui ne peuvent pas être interprétées comme du testing » (Husén *et al.*, 1994, cité dans Legendre, 2005, p. 868).

Mesure normative (norm-referenced) : « Score qui permet de comparer la maîtrise ou la performance d'un individu avec celle des autres individus de son groupe. [...] Ce score (norm-referenced) situe l'individu dans son groupe particulier et se traduit par des normes comme le rang-centile [...], le score T, etc. [...] Cette approche se traduit souvent en classe par la comparaison des résultats de l'élève avec ceux des autres élèves en classe, avec la moyenne, entre autres » (Legendre, 2005, p. 868).

« Mesure critériée (criterion-referenced, mesure de maîtrise) : « La mesure critériée (criterion-referenced) permet d'obtenir un score représentant la maîtrise de chacun des comportements individuels exigés pour la réalisation des objectifs spécifiques. [...] Elle permet d'établir pour chacun des [candidats] une fiche individuelle identifiant où il est rendu dans son apprentissage, le temps qu'il a mis pour s'y rendre, et les objectifs qu'il maîtrise et ceux qu'il ne maîtrise pas » (Lavallée *et al.*, 1974, cité dans Legendre, 2005, p. 868). [...] La performance cible [...] est décrite par un ensemble de critères qui se rapportent à l'apprentissage. L'instrument de mesure est élaboré en fonction du domaine, des critères et un seuil de réussite est fixé quand cela est pertinent. On situera le résultat [du candidat] par rapport aux critères et, s'il y a lieu, au seuil de réussite » (Legendre, 2005, p. 868).

★ Niveau taxonomique : « Chacun des objectifs d'habileté placé dans un ordre hiérarchique croissant sur une échelle où chaque échelon intègre le précédent » (Legendre, 2005, p. 936).

Norme (standard) : « Ensemble des règles, des principes et des critères qui expriment les attentes d'un groupe par rapport à ses membres et qui orientent l'action, les décisions et les idées de ceux-ci » (Legendre, 2005, p. 939).

Notation : « Action d'attribuer un score (chiffre ou pourcentage) ou une cote (lettre ou symbole), ou encore de donner une appréciation ou un commentaire sur les apprentissages ou le développement d'un élève » (Legendre, 2005, p. 939).

Note de passage (seuil de réussite) : « Seuil de réussite appliqué à la totalité d'un instrument de mesure qui porte sur un large domaine [...] et imposé à tous les répondants » (Legendre, 2005, p. 940).

Novice : « Personne qui débute dans un domaine, qui y possède peu de savoir et d'expertise ainsi qu'une vision morcelée » (Legendre, 2005, p. 942).

Objectif : « Résultats précis, circonscrit et vérifiable dont l'atteinte exige une focalisation d'actions cohérentes et d'efforts concertés pendant une certaine période de temps » (Legendre, 2005, p. 943).

« Expression tangible et réaliste d'un résultat (quoi?), délimité dans le temps (quand?) et dans l'espace (où?), qualifié et comportant des indices d'appréciation, quantité mesurable et comparable qui, à un moment déterminé, devrait avoir été réalisée par des personnes qualifiées (qui?) pour satisfaire un besoin déterminé (pourquoi?) » (BIBEAU, 1987, cité dans Legendre, 2005, p. 943).

Objectif comportemental (Behavioriste) : « Objectif décrivant les actions observables, attendues du Sujet, [...] devant être les manifestations concrètes des compétences acquises » (Legendre, 2005, p. 946).

Objectif de performance : « Objectif opératoire qui précise, avant la situation [...], ce que [le candidat] devra être capable de faire après, les conditions inhérentes à la réalisation du comportement désiré ainsi que les critères de réussite » (Legendre, 2005, p. 949).

Performance : « Manifestation concrète d'une compétence » (SCALLON, 2004, cité dans Legendre, 2005, p. 1029).

« C'est le comportement. C'est une manifestation externe, observable, mesurable (sur laquelle nous pouvons porter un jugement) » (CAUCHY, 1995, cité dans Legendre, 2005, p. 1028).

« [C'est la] mise en œuvre d'une activité orientée vers la réalisation d'une tâche comme indicateur d'une compétence » (Legendre, 2005, p. 1029).

Profession : « Activité professionnelle de caractère intellectuel, dont l'exercice exige des connaissances et des compétences de haut niveau, généralement acquises dans le cadre d'études supérieures, universitaires ou autres, dans des domaines [...] » (BIT, 1987, cité dans Legendre, 2005, p. 1087).

Professionnel : « Celui qui, face à une situation complexe, mobilise divers savoirs pour en arriver à prendre la bonne décision au bon moment » (SCHÖN, 1994 dans Gauthier et Tardif, 1996, cité dans Legendre, 2005, p. 1088).

Reconnaissance de la compétence (des compétences professionnelles) : Reconnaissance de la capacité à pratiquer un métier ou à exercer une profession » (LANDRY, 1987, cité dans Legendre, 2005, p. 1159).

Reconnaissance des acquis : « Principe selon lequel les apprentissages expérimentiels et les autres cours suivis possèdent une valeur réelle de formation qui doit être reconnue dans un programme d'études » (Legendre, 2005, p. 1159).

Rendement professionnel : « Exécution d'une ou de plusieurs tâches professionnelles selon des critères préétablis de qualification de compétence, de rapidité et d'efficacité » (BIT, 1987, cité dans Legendre, 2005, p. 1177).

Score : « Nombre exprimant le résultat obtenu à un test [ou] à un examen [...] » (Legendre, 2005, p. 1220).

Score critérié : « Mesure extérieure servant de critère indépendant d'appréciation de la variable que l'on désire mesurer par un test » (Legendre, 2005, p. 1221).

Score de césure (voir seuil de réussite).

«**Seuil** : « Entrée, commencement, début, départ. [...] Limite inférieure ou supérieure, niveau critique » (Legendre, 2005, p. 1233).

Seuil de maîtrise (standard/seuil de performance/de réussite, score/point de césure) : « Sur une échelle de mesure, point qui distingue l'atteinte ou non d'un objectif ou la réussite ou l'échec de la réalisation d'une tâche. Obtenir 60 % à un examen; réussir 75 % ou plus des tâches assignées. [C'est le] pourcentage de réussite exigé au test à partir duquel l'enseignant juge que l'élève maîtrise les apprentissages » (Legendre, 2005, p. 1233).

Seuil de réussite (seuil/standard de performance) : « Niveau de qualité à partir duquel on considère une performance comme réussie » (GOUVERNEMENT DU QUÉBEC, 1985, cité dans Legendre, 2005, p. 1233).

« Critère quantitatif exprimé en nombre (par ex. : 4 items sur 5) ou en pourcentage (par ex. : 80 %) des items auxquels on a répondu correctement, parmi un ensemble de tâches, de problèmes, ou de questions, relatif à un objectif ou à un domaine déterminé permettant de diviser la distribution des scores en deux catégories mutuellement exclusives : une catégorie contient les scores égaux ou supérieurs au critère à partir desquels la maîtrise ou la compétence des répondants est inférée, une autre catégorie comprend les scores inférieurs au critère à partir desquels la non-maîtrise ou l'incompétence des répondants est inférée » (Legendre, 2005, p. 1233).

Standard de performance : (voir seuil de maîtrise/de réussite)

Standard : « Qui est conforme à une norme commune, à un archétype, à un modèle uniforme, à des exigences précises prédéterminées » (Legendre, 2005, p. 1256).

Tâche : « Activité observable et mesurable qui, à l'intérieur d'une séquence temporelle, constitue une démarche logique et nécessaire pour la réalisation d'un travail ou l'atteinte d'un but [(objectif)] » (Legendre, 2005, p. 1317).

Tâche évaluative : « Tâche spécifique assignée à un [candidat] dans le but d'apprécier sa performance et de porter un jugement » (Legendre, 2005, p. 1318).

★**Test critérié** : « Test servant à comparer la performance d'un sujet à un critère de réussite prédéterminé, où le score obtenu par le sujet indique s'il a satisfait au critère exigé ou exprime l'écart entre sa performance et un seuil de réussite absolu, indépendamment de ses performances antérieures, de celles de tout autre individu ou d'un groupe » (Legendre, 2005, p. 1377).

Test de capacité : « Évaluation objective de la compétence d'un stagiaire dans l'accomplissement d'une tâche professionnelle, de manière à contrôler et à diriger sa progression tout au long du programme de formation » (BIT, 1987, cité dans Legendre, 2005, p. 1378). « Ce procédé peut également être appliqué au poste de travail afin de déterminer le niveau de qualification qui doit être atteint pour accomplir efficacement une tâche, ce qui peut, dans certains cas, conduire à une promotion ou appeler une formation complémentaire » (*ibid.*).

★Test de performance : « Épreuve non verbale dont l'objectif est de connaître certaines aptitudes du sujet [qui peut, dans certains cas, permettre de] déterminer le degré d'aptitude acquis et, l'exécution de tâches spécifiques caractéristiques d'un métier ou d'une profession qu'il désire exercer. [...] Le test de performance demande une réponse manuelle ou motrice [...], par opposition au test qui exige une réponse verbale ou écrite » (Legendre, 2005, p. 1379).

Test normatif : « Test servant à comparer la performance d'un sujet à celle de ses pairs, dans lequel le score obtenu par le sujet est exprimé par sa position relative sur une échelle de normes construites à partir de performances étalonnées d'un groupe de référence » (Legendre, 2005, p. 1380).

Testing : « Opération de mesure ou d'évaluation prenant place avant ou après une situation pédagogique; action d'administrer un test » (Legendre, 2005, p. 1380).

INDEX

A

Acquis expérientiels · xvii, 1, 2, 8, 10, 11, 13, 14, 16, 30, 39, 84, 85, 268, 273
Approche par compétences · 2, 8, 17, 18, 19, 20, 24, 75, 170, 271, 272, 273

C

Cadre procédural d'intégration · 16, 37, 38, 40
Cahier d'observation des comportements cliniques · 17, 29, 37, 74, 269
Classification des méthodes · 2, 63, 64
Comportements observables · 46, 273
Critères · 4, 20, 21, 24, 36, 37, 41, 42, 46, 47, 48, 49, 51, 55, 56, 57, 58, 68, 86, 87, 88, 89, 96, 97, 98, 99, 152, 154, 156, 164, 165, 169, 173, 266, 273, 274, 276, 277, 280, 281, 282

D

Degré d'efficience · 28, 29, 49, 79, 80, 82, 89, 90, 102, 106, 107, 108, 111, 128, 129, 132, 137, 149, 152, 155, 168

E

Efficience · xvii, 7, 20, 28, 29, 49, 52, 79, 80, 81, 82, 89, 90, 93, 95, 100, 101, 102, 105, 106, 107, 108, 111, 120, 121, 128, 129, 132, 135, 137, 148, 149, 150, 152, 155, 156, 161, 162, 163, 164, 165, 168, 169, 170, 171, 172, 174, 281
Épreuve clinique · xvii, 1, 2, 3, 15, 16, 29, 43, 44, 47, 50, 65, 66, 74, 75, 76, 83, 84, 85, 86, 87, 91, 93, 94, 100, 104, 111, 112, 121, 125, 126, 139, 141,

144, 147, 148, 152, 153, 164, 168, 169, 170, 172, 205

Établissement de standards de performance · vi, xv, xvii, xviii, 2, 3, 14, 16, 17, 26, 29, 43, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 61, 62, 63, 64, 65, 66, 67, 71, 74, 75, 80, 81, 83, 85, 89, 90, 92, 95, 96, 99, 100, 101, 103, 104, 105, 106, 107, 108, 112, 115, 116, 117, 122, 124, 132, 134, 141, 144, 147, 148, 151, 152, 154, 155, 161, 166, 172, 173, 174, 178, 182, 183, 184, 188, 194, 195, 197, 265

Évaluation des compétences · xviii, xx, 2, 5, 10, 13, 14, 15, 17, 18, 20, 21, 22, 23, 28, 30, 46, 87, 93, 104, 161, 178, 259, 263, 266, 270, 271, 272

Évaluation des compétences professionnelles · 5, 14, 15, 30, 104, 178

G

Grille d'analyse I-SMART · 50
Grilles d'observation des savoir-faire · 16, 37, 38, 41, 42, 43, 47, 269

H

Historique des méthodes · 2, 50, 67, 96

I

Indicateurs de performance · 47, 49, 52, 74, 75, 178, 265

I-SMART · xvii, 46, 47, 48, 50, 84, 86, 89, 116, 144, 147, 149, 168, 170, 191, 195, 251

M

Méthode Angoff · xviii, xx, 56, 184
 Méthode de l'appariement au descripteur de l'item
 · xv, xviii, xx, 15, 66, 71, 73, 102, 183, 199, 201,
 202, 205
 Méthode Ebel · xviii, xx, 15, 50, 63, 66, 67, 69, 70,
 72, 76, 81, 83, 102, 112, 183, 195, 196, 197,
 199, 261, 270
 Méthode oui/non · 71, 72, 83, 188, 189, 190
 Méthodes centrées sur les items · 63
 Modèle à multifacettes de Rasch · xviii, 4, 148, 154,
 158, 159, 161, 165, 172, 173, 175, 258, 259
 Modèle I-SMART · xvii, 46, 47, 48, 50, 84, 86, 89,
 116, 144, 147, 149, 168, 170, 191, 195, 251

N

Neuf Principes · 2, 50, 53, 54, 55, 58, 59, 60, 61, 62,
 63, 67, 69, 76, 81, 100, 104
 Niveau de difficulté · 19, 27, 49, 51, 52, 56, 69, 70,
 72, 73, 75, 76, 77, 78, 79, 80, 82, 89, 90, 100,
 101, 102, 106, 107, 108, 120, 128, 129, 130,
 132, 133, 134, 148, 150, 155, 160, 163, 165,
 168, 169, 172, 175, 178, 189, 191, 195, 196,
 197, 198, 199, 200, 203, 275

O

Ordre des acupuncteurs du Québec · v, xv, xvii,
 xviii, 1, 2, 3, 10, 15, 16, 17, 18, 25, 27, 29, 30,
 31, 32, 33, 35, 36, 37, 38, 40, 41, 42, 43, 44, 46,
 47, 48, 49, 65, 66, 73, 74, 75, 76, 77, 82, 83, 84,
 86, 87, 89, 92, 93, 96, 97, 98, 102, 105, 113,
 115, 146, 151, 152, 154, 162, 168, 171, 177,
 178, 179, 180, 205, 255, 265, 269
 Ordres professionnels · xvii, xviii, xx, 1, 5, 7, 8, 10,
 12, 13, 14, 15, 29, 30, 45, 63, 76, 83, 149, 151,
 152, 154, 162, 165, 171, 172, 174, 178, 260, 268

P

Permis d'exercice · xvii, 2, 5, 10, 11, 12, 16, 50, 67,
 82, 83, 84, 85, 93, 95, 97, 98, 162, 178, 186, 205

Polygramme des compétences · 17, 29, 30, 31, 34,
 36, 37, 38, 48, 179, 269

R

Reconnaissance des acquis · xvii, 6, 9, 10, 11, 13,
 14, 260, 269, 272, 273, 281
 Reconnaissance des compétences · xvi, xvii, 1, 3, 7,
 8, 9, 11, 12, 13, 14, 27, 29, 43, 50, 116, 260,
 261, 269, 272, 273, 281
 Reconnaissance des compétences professionnelles
 · xvi, xviii, xx, 1, 3, 7, 8, 10, 11, 12, 13, 14, 27, 29,
 43, 50, 116, 260, 261
 Référentiel des compétences · xviii, xx, 1, 2, 3, 7,
 13, 14, 16, 17, 29, 30, 34, 37, 47, 65, 66, 74, 86,
 91

S

Standards de performance · vi, xv, xvii, xviii, xx, 2,
 3, 14, 15, 16, 17, 18, 23, 24, 25, 26, 27, 29, 34,
 42, 43, 49, 50, 51, 52, 53, 54, 55, 57, 58, 60, 61,
 62, 63, 65, 67, 70, 72, 74, 75, 76, 81, 82, 83, 84,
 85, 86, 93, 94, 96, 99, 102, 103, 104, 106, 107,
 111, 112, 115, 116, 117, 118, 121, 122, 125,
 126, 127, 132, 134, 135, 139, 140, 141, 142,
 143, 144, 146, 147, 148, 149, 151, 152, 154,
 155, 156, 168, 170, 171, 172, 174, 178, 179,
 183, 184, 186, 187, 188, 190, 191, 193, 194,
 195, 196, 198, 258, 260, 261, 262, 263, 264,
 265, 268, 270, 272, 273, 280, 282

T

Taxonomie de Bloom · xvii, 27, 77, 150, 171

Z

Zones d'incertitudes · xvii, 71, 73, 74, 81, 101, 102,
 105, 107, 108, 109, 110, 111, 117, 129, 130,
 132, 135, 137, 138, 149, 160, 168, 169, 173,
 175, 200, 204