Comparaison des propriétés métriques des scores obtenus avec un test de concordance de script au regard de trois méthodes de détermination des scores

Judith Exantus

Thèse présentée à l'Université d'Ottawa
dans le cadre des exigences du programme Enseignement aux professionnels de la santé
en vue de l'obtention du grade de Maîtrise ès arts en éducation - M.A. (Éd)

Faculté d'Éducation

Université d'Ottawa

« L’homme est capable de faire ce qu’il est incapable d’imaginer »

René Char, 1907-1988

Remerciements

Le périple a été long et pavé de moments de découragement et de sentiment de perte de confiance en soi. Mais grâce au support indéfectible de plus d'uns me voici au jour J. Ainsi je tiens à remercier :

Mon directeur, Eric Dionne, pour tes critiques, tes conseils et les *boli* d'encouragement quand tout sombrait autour de moi.

Les membres de mon jury, Dany Laveault et Michel D. Laurier, merci d'avoir accepté de critiquer mon travail et de me permettre d'être où je suis aujourd'hui.

L'équipe du Département de Médecine Familiale de l'Université de Liège, Belgique. Merci pour votre accord spontané pour l'utilisation de vos données.

Monsieur Bernard Charlin, Monsieur Robert Gagnon et toute l'équipe du CPASS. Merci d'avoir laissé la porte ouverte pour moi dès le premier contact.

Mon époux, Jean Max, pour tes mots d'encouragement, ta patience, ta compréhension et ta présence.

Ma mère, Arthémise, merci mammy pour ton optimisme, tes conseils judicieux et ton amour inconditionnel.

Mes sœurs Rachel Edwige et Fabienne Altagrâce, vraiment sans vous je me serais égarée à maintes reprises. Merci pour votre support sans limite. Un coucou spécial et affectueux à mes trésors Arnaud, Emma, Guessy, Lise et Charles.

Mes amies-sœurs et mes amis qui se reconnaissent ici, j'en suis sûre !

Je tiens aussi à saluer la mémoire de 2 grandes personnes :

- Mon père, feu Maitre Servilus Exantus. Le régime des Duvalier nous a ravis à ton amour mais tes racines sont profondes et fertiles, Papi !

- Ma tante-mère, Erolienne, merci pour Tout, Nennenn. Tu es présente dans mes pensées chaque jour.

RÉSUMÉ

Le test de concordance de script (TCS) est un outil qui vise à mesurer le raisonnement clinique (RC) des apprenants en contexte d'incertitude en comparaison à celui des experts. La nature de cet instrument fait en sorte que le processus de détermination des scores est névralgique et influence ainsi l'interprétation de ces derniers. Le but de cette recherche à caractère exploratoire était d'examiner l'impact de trois méthodes de détermination des scores sur les propriétés métriques des scores obtenus avec un TCS. Nous avons réalisé une analyse secondaire de données recueillies entre 2010 et 2014 auprès de 160 étudiants inscrits dans un programme de médecine familiale à l'Université de Liège en Belgique. Ainsi la théorie classique des tests et le modèle de Rasch ont été mis à contribution. Selon les résultats obtenus, la méthode selon la bonne réponse n'est pas recommandée. En revanche, la méthode des scores combinés très utilisée dans le contexte du TCS n'a pas prouvé, selon nous, sa supériorité sur la méthode avec pénalité de distance. D'autres études s'avèrent nécessaires pour approfondir ces résultats.

MOTS CLÉS

Test de concordance de script, raisonnement clinique, scores, mesure, théorie classique des tests, modèle de Rasch.

ABSTRACT

Script concordance test (SCT) is a tool recently developed for assessing clinical reasoning in context of uncertainty. It measures the concordance between examinees responses (scripts) and those from a panel of expert clinicians. The process of the scoring system is very sensitive and influences the meaning of the scores. The aim of this exploratory research was to examine the impact of three methods of scoring process on the metric properties of the scores. We conducted a secondary analysis over data gathered between 2010 and 2014 from 160 students registered in a program of family medicine at University of Liege, Belgium. Therefore the classical test theory and the Rasch model have been used. After analyzing the results, we conclude that the 3-point single answer is not appropriate in the context of script concordance test. The 5-point aggregate is not superior to the 5-point aggregate with distance penalty. More studies should be done in the future to validate the influence of the scoring methods on the psychometric properties of the script concordance test.

KEYWORDS

Script concordance test, clinical reasoning, scoring process, measurement, classical test theory, Rasch measurement model.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

Le raisonnement clinique[1] (RC) constitue, selon Charlin, Bordage et Van der Vleuten (2003), l'une des trois composantes de la compétence clinique ; les autres étant les connaissances (sciences de base et sciences cliniques) et les habiletés pertinentes (cliniques, techniques et interpersonnelles). Le développement de cette compétence est un long processus qui requiert des stratégies spécifiques d'enseignement, des entraînements supervisés tels que des séances d'apprentissage du raisonnement clinique (ARC) avec rétroaction et avec des évaluations régulières (Carrière, Gagnon, Charlin, Downing et Bordage, 2009 ; Eva, 2005). Dans tout processus d'enseignement et d'apprentissage, l'évaluation est primordiale. Selon Jouquan (2002), l'évaluation partage avec la démarche clinique la double exigence de rigueur et de pertinence. Elle doit répondre à des critères bien définis et surtout on doit utiliser des instruments dont on connait les limites et les forces.

De plus, le processus de correction doit être uniforme pour tous les candidats que ce soit dans un contexte d'évaluation formative ou sommative. Depuis plus d'une vingtaine d'années, la pédagogie médicale s'est enrichie d'un autre dispositif d'évaluation, le test de concordance de script (TCS).

Le TCS est un outil qui vise à mesurer le RC des apprenants en contexte d'incertitude. Ce dernier se définit comme toute situation ambigüe exigeant une réflexion pour interpréter les données et pour progresser vers la résolution du problème (Charlin, Gagnon, Kazi-Tani, Thivierge, 2006). De nombreux auteurs ont avancé que le TCS discriminait adéquatement le RC des cliniciens en fonction de leurs niveaux d'expérience entre autres, Sibert, Charlin, Gagnon, Corcos et Khalaf (2001) et Fournier *et al*. (2006).

Le processus de détermination des scores dans le contexte du test de concordance de script est complexe. Il influence considérablement l'interprétation des scores qui sera faite. En effet, il n'y a pas clairement de bonnes ou de mauvaises réponses avec le TCS;

[1] Nous reviendrons plus loin avec une définition plus formelle du raisonnement clinique.

la structure des scores dépend du jugement des experts qui permet de bâtir l'échelle des scores. Une recension des écrits publiés entre 1998 et 2018 montre que la méthode des scores combinés est la plus commune (92,5 % des 80 études recensées) et elle est très peu remise en question. Pour bien des auteurs (ex. Chang *et al*., 2014 ; Charlin, Brailovsky, Leduc et Blouin, 1998 ; Fournier *et al*., 2006 ; Latreille, 2012 ; Lubarsky, Charlin, Cook, Chalk et Van der Vleuten, 2011), les propriétés métriques des scores obtenus avec le TCS sont généralement bonnes avec cette méthode. Toutefois, selon nous, les analyses dont fait l'objet le TCS sont limitées et incomplètes. Il faut aussi souligner que les analyses en édumétrie relatives au TCS sont sommaires dans le sens que des groupes de participants sont comparés et la fidélité est évaluée avec le coefficient alpha Cronbach. Les travaux de Bland, Kreiter et Gordon, déjà en 2005, révélaient que d'autres méthodes concurrentes pourraient être également utilisées dans le cadre de la correction d'un TCS. Quelques années plus tard en 2014, Wilson, Pike et Humbert ont eux aussi soulevé des interrogations relatives à l'utilisation de la méthode des scores combinés dans ce contexte.

Nous avons réalisé une étude à caractère exploratoire dont l'objectif était de comparer trois méthodes de détermination des scores dont la méthode des scores combinés, largement employée, pour le TCS. Une analyse secondaire des données recueillies auprès d'un échantillon de 160 participants[2] inscrits à un programme de médecine générale a été menée. Le TCS utilisé a une visée certificative et il a un poids de 10% dans le dispositif global d'évaluation en médecine générale. Les participants appartenaient à deux groupes : (1) des étudiants et (2) des experts.

Dans le premier chapitre consacré à la problématique, nous abordons le raisonnement clinique en tant que dimension de la compétence clinique et les enjeux de son évaluation. Nous discutons du test de concordance de script comme outil d'évaluation du RC en priorisant certaines variables. Dans le deuxième chapitre, nous présentons la recension des écrits relative au raisonnement clinique en traitant les différentes façons de l'évaluer dans le domaine médical. Nous enchaînons avec, dans le

[2] Pour faciliter la lecture du document et pour ne pas allonger le texte inutilement, le genre masculin est priorisé.

troisième chapitre, la méthodologie qui précise le contexte d'élaboration du TCS et son administration, les participants et le traitement des données collectées c'est-à-dire les différentes analyses qui ont été menées. Enfin, les derniers chapitres portent sur les résultats, la discussion et la conclusion de l'étude.

CHAPITRE I – PROBLÉMATIQUE

Comme nous l'avons mentionné dans le premier paragraphe de l'introduction, le raisonnement clinique est une des composantes de la compétence clinique des experts et des résidents en formation (Audétat *et al*., 2012 ; Charlin *et al*., 2003; Charlin, 2006; Norman, 2005) et il s'affine avec l'expérience. Psiuk (2012, p. 147) définit le RC, elle-même, comme un « ensemble des processus de pensée et de prise de décisions permettant au clinicien de déterminer les actions les plus appropriées face à un contexte spécifique de résolution de problèmes de santé ».

On reconnaît au raisonnement clinique trois dimensions : (1) l'hypothèse diagnostique, (2) les hypothèses d'investigation complémentaire et (3) les hypothèses thérapeutiques. En général, le clinicien procède d'une manière hypothético-déductive en mettant en relation les signes ou les symptômes cliniques avec les entités cliniques. Il y a tout au cours du processus de multiples micro-jugements qui sont mis en place.

Tout au cours de son cursus, le clinicien construit et organise des réseaux de connaissances appelés scripts par Feltovich et Barrows (1984, cités par Gagnon, Charlin, Coletti, Sauvé et van der Vleuten, 2005). Ces scripts vont s'affiner avec l'expérience (Meterissian, 2006). Selon la théorie des scripts, le médecin réactive les scripts et les utilise en présence d'une situation clinique donnée de façon à émettre un diagnostic précis et à définir un plan de traitement ou d'investigation (Charlin, Boshuizen, Custers et Feltovich, 2007). Les scripts de maladies jouent ainsi un rôle capital en supportant les aptitudes que les apprenants en santé doivent acquérir telles l'aptitude d'émettre des diagnostics différentiels et celle d'interpréter des données cliniques (Lubarsky, Dory, Audetat, Custers et Charlin, 2015).

1.1 Le test de concordance de script

Le test de concordance de script (TCS), basé sur la théorie des scripts[3], évalue le raisonnement clinique c'est-à-dire l'aptitude des cliniciens à résoudre des problèmes mal définis et à raisonner en contexte d'incertitude (Charlin, Gagnon, Sibert et Van der Vleuten (2002), Ramaekers, Kremer, Pilot, Van Beukelen et Van Keulen, 2010). Ces problèmes mal définis sont dits authentiques ressemblant à ceux qu'ils rencontreront dans le cadre de leur pratique. Selon Goulet, Jacques, Gagnon, Charlin et Shabah (2010), le TCS prend en considération un élément spécifique du processus de raisonnement clinique : l'interprétation des données nouvelles par les sujets. Pour d'autres auteurs tels Fournier, Demeester et Charlin (2008), Giet, Massart, Gagnon et Charlin (2013), plus précisément, c'est un micro jugement lié à l'interprétation de ces données qui est évalué avec cet instrument. Charlin et St-Jean (2002, p.4) avancent que « le TCS apprécie la qualité de l'organisation des connaissances et la pertinence de cette organisation pour agir dans une situation donnée ». En captant certains éléments des processus de la pensée, et pas uniquement les connaissances, le TCS peut permettre une évaluation plus valide du raisonnement clinique (Hayward *et al*., 2016). Ceci a été démontré, d'ailleurs, par Lubarsky, Charlin, Cook, Chalk et van der Vleuten en 2011. Il est de plus en plus utilisé en formation médicale initiale ou continue (Hornos *et al*. 2013), en sciences infirmières (Dawson, Comer, Kossick et Neubrander, 2014; Deschênes, Charlin, Gagnon et Goudreau, 2011 ; Latreille, 2012), en physiothérapie (Cohen, Fitzgerald, Lane et Boninger, 2005) et même en éducation médicale.

Le test de concordance de script mesure le degré de concordance entre la performance des étudiants et celle des experts à l'égard des différentes dimensions du raisonnement clinique. Il compare l'organisation de leurs connaissances (les scripts) dans un domaine bien déterminé (Charlin *et al*., 2003). Ainsi le TCS permet de détecter les personnes les plus expérimentées cliniquement. De plus, Kazour, Richa, Zoghbi, El-Hage

[3] La théorie des scripts est une théorie cognitive qui soutient que les connaissances sont organisées, structurées en réseaux qui sont réactivés en situation de résolution de problèmes (Charlin, Tardif et Boshuizen, 2000).

et Haddad (2016) affirment que le TCS est objectif, fidèle et qu'il peut être administré pendant un nombre illimité de fois et à un nombre illimité d'étudiants. L'affirmation de Kazour *et al*. nous parait un peu abusive car le format d'un instrument de mesure ne peut être fidèle. En effet, la démonstration doit être faite à chaque fois dans un contexte de passation donné. Si le contexte change, la démonstration doit être reprise. De nombreux auteurs ont étudié les caractéristiques psychométriques du TCS. Charlin, Brailovsky, Leduc et Blouin (1998), Sibert, Charlin, Gagnon, Corcos et Khalaf (2001) et Fournier *et al*. (2006) ont documenté, par exemple, la validité de construit du TCS.

Brailovsky, Charlin, Beausoleil, Coté, et Van der Vleuten (2001) ont aussi montré que le TCS avait une bonne valeur prédictive c'est-à-dire que le niveau d'organisation optimal des étudiants à un stade précis de leur formation continuait à se maintenir dans le temps. Certains auteurs tels Sibert *et al.* (2002) ont mis en évidence la stabilité des scores d'un TCS en fonction des aspects culturels.

Le TCS permet d'explorer diverses situations cliniques et le temps de passation est en général court. Toujours selon Kazour *et al*. (2016), il améliore considérablement le système d'évaluation dans le contexte clinique. Les tests usuels dont les grilles de correction sont basées sur les consensus entre correcteurs permettent mal cette détection. Charlin *et al*. (2003) avancent que « le TCS part d'une théorie du raisonnement clinique (la théorie des scripts) et vise à mesurer des processus de raisonnement jugés essentiels plutôt que l'issue d'un raisonnement devant une situation qui mime la réalité. Sa structure de correction est tout à fait nouvelle et prend en compte la variabilité de réponses des experts ». En effet c'est une façon peu commune de déterminer des scores avec un instrument de mesure. Toutefois la conception d'un test de concordance de script fait appel à plusieurs variables les unes tout aussi importantes que les autres. Dans la section suivante nous discutons de ces variables.

1.2 Les variables qui peuvent affecter le TCS

Charlin, Brailovsky, Leduc *et al*. (1998) mentionnent que le TCS est facile à construire et à corriger ; mais, en ce qui nous concerne, nous émettons des réserves à cet

égard. En effet, il faut que le concepteur s'assure de respecter, entre autres, la notion de contexte d'incertitude lors de la rédaction des vignettes ce qui n'est pas, selon nous, une mince tâche. Cette incertitude doit être bien calibrée : si elle est trop grande, la variance des scores attribuable aux différences de maitrise du RC des candidats le sera aussi; si l'incertitude est trop restreinte, la variance des scores sera trop petite. La difficulté est d'arriver à un juste équilibre. Ce cas est largement décrié par Lineberry, Kreiter et Bordage (2013). Il est ainsi difficile de justifier qu'un cas clinique présente des solutions valides relativement opposées. Trop peu de variance illustre un cas clinique où tous les sujets s'entendent sur la solution à promulguer ce qui ne permet pas de mettre en valeur adéquatement le jugement.

Plusieurs variables, illustrées dans la Figure 1 ci-dessous, doivent être considérées lors de la construction du TCS pour obtenir les meilleures qualités métriques possibles des scores. On peut s'intéresser au format du test, aux participants qui se verront administrer le test, aux objectifs de l'évaluation ou à son mode d'administration (version papier / crayon ou version électronique selon Sibert, Darmoni, Dahamna, Weber et Charlin en 2005), aux scores ou encore aux propriétés métriques liées à ces scores.

Il nous parait important de discuter de ces différentes variables tout en précisant d'ores et déjà que les méthodes de détermination des scores ont retenu notre attention dans le cadre de cette recherche.

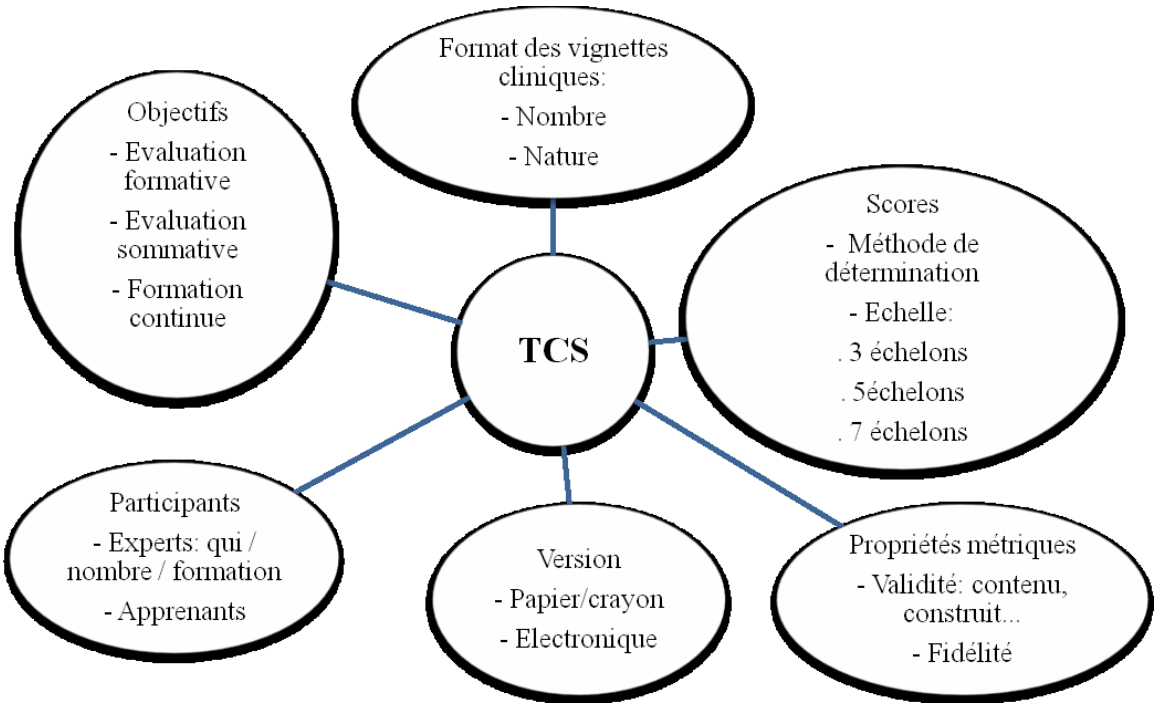

Figure 1. Les différentes variables influençant le TCS.

1.2.1 Le format

Le TCS se présente généralement sous une forme semblable à celle présentée à la Figure 2 ci-dessous. Cette dernière illustre une vignette composée de quatre éléments : (1) un scénario clinique avec des données incomplètes ou encore, par exemple, volontairement ambiguës ; (2) une option qui peut correspondre, par exemple, à un diagnostic initial ; (3) une nouvelle donnée qui permet de juger si la nouvelle information diminue, augmente ou n'a aucun impact sur la décision et (4) une échelle de réponse généralement en cinq échelons allant de -2 à +2. Le candidat choisit l'ancrage « -2 » si l'hypothèse de départ est « totalement rejetée », ou « -1 » si l'hypothèse est « révisée ». Il choisira « 0 » si l'information ajoutée n'a « aucun effet » sur l'hypothèse, « +1 » si l'hypothèse « est renforcée », ou « +2 » si l'hypothèse « est totalement confirmée ». Cette échelle de réponse est généralement uniforme pour l'ensemble du test. Ce format rappelle la façon de procéder des cliniciens quand ils sont confrontés à des situations cliniques complexes et incertaines très proches des cas réels rencontrés (Charlin, Brailovsky, Brazeau-Lamontagne *et al.*, 1998 ; Charlin, Tardif et Boshuizen, 2000 ; Lubarsky,

Gagnon et Charlin, 2013). La gradation de l'hypothèse (ex. probable, fréquente, appropriée etc.) peut changer mais, dans tous les cas, le principe de gradation (moins probable à plus probable) précédemment exposé demeure.

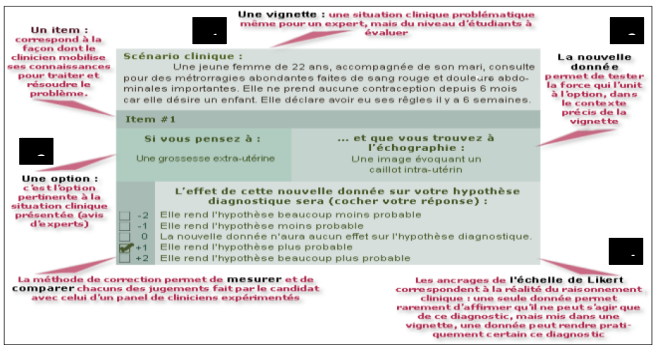

Figure 2. [4]Exemple d'une vignette pour évaluation de la dimension diagnostic.

1.2.2 Les objectifs de l'évaluation

Le TCS répond selon les écrits à différents objectifs d'évaluation. En effet, en 2004, Caire, Sol, Charlin, Isodiri et Moreau ont conçu un TCS pour une évaluation formative en neurochirurgie. Ils ont conclu que « le TCS permet à l'étudiant de se situer par rapport à la référence que constitue le panel d'experts, mais aussi par rapport aux autres étudiants, de même niveau ou de niveaux différents. Il permet surtout à chacun de suivre sa propre évolution au cours de son cursus, pour les différents domaines de la spécialité ». Pour la formation continue, Labelle *et al*. (2003) estiment que le fait de comparer les réponses des évalués à celles d'un groupe d'experts constitue un potentiel considérable dans le contexte d'un atelier de formation à la condition de mettre en perspective non seulement les scores des évalués et ceux des experts mais également les arguments fournis par les évalués pour choisir leur catégorie de réponse. L'identification des forces et faiblesses des participants à l'atelier peut être faite de façon dynamique et immédiate. Par conséquent, cette identification aura un impact positif sur la satisfaction des participants et sur leur évaluation relative au déroulement de l'atelier. Selon les résultats de leur étude, l'atelier de formation basé sur le TCS représente une alternative comme méthode d'enseignement aux ateliers classiques. Quant à l'évaluation sommative, le TCS a été utilisé par des chercheurs tels Nouh *et al*. (2012) en chirurgie à travers le Canada. Plus récemment, Cooke, Lemay, Beran, Sandhu et Amin (2016) ont proposé,

4 Adapté par le centre de pédagogie appliquée aux sciences de la santé (CPASS).

après ajustement et bonification, l'utilisation du test de concordance de script pour l'évaluation sommative en éducation médicale pédiatrique. Toutefois les critiques formulées par Bland et *al*. (2005) ont bien fait comprendre que le consensus n'est toujours pas obtenu. Nous devons noter qu'à date le TCS a du mal à être accepté dans un contexte à enjeux critiques en raison de la paucité des preuves empiriques au regard de sa validité.

1.2.3 Version d'administration

Les écrits ont révélé qu'il y a deux versions d'administration du TCS : 1) une version papier-crayon 2) une version électronique. Dans l'étude de Caire *et al*., précédemment citée, les évalués ont eu accès au test en ligne. De même Sibert, Darmoni Dahmna, Weber et Charlin (2005), Mathieu *et al.* (2013), Faucher, Dufour-Guindon, Lapointe, Gagnon et Charlin (2016) ont également fait l'expérience de l'administration en ligne (respectivement en urologie, en rhumatologie et en optométrie). Dans la majorité des études consultées, la version papier-crayon est la plus utilisée. La comparaison de l'administration du TCS en version papier-crayon et version électronique n'est pas vraiment faite dans les écrits. Nous signalons que les auteurs Hayward *et al*. (2016) ont fait une adaptation du TCS en un outil novateur administré en ligne pour enseigner le raisonnement clinique aux apprenants. En effet ces auteurs ont utilisé la théorie de scripts en décrivant une vignette de cas clinique authentique audio et vidéo avec des images interactives en trois-dimensions (3D). Les réponses des apprenants sont enregistrées tout au cours du processus et sont comparées à celles des experts et des pairs. Les apprenants ont trouvé l'expérience positive, novatrice et l'ont jugée comme une excellente méthode d'apprentissage toujours selon Hayward *et al*. Ces arguments sont toutefois peu convaincants vu l'absence de preuve empirique.

Quant aux propriétés métriques et aux participants, ces variables sont discutées tout au long des différentes études. Nous tenons à rappeler que la méthode de détermination des scores est la variable clé pour nous.

1.2.4. Les scores

Comme nous l'avons mentionné précédemment, l'une des particularités du TCS

est qu'il n'y a pas nécessairement qu'une seule bonne réponse (Charlin, Brailovsky, Leduc *et al*., 1998 ; Charlin *et al*., 2010) comparativement aux questions à choix multiples (QCM) par exemple. En effet, avec les QCM, il s'agit généralement (mais pas toujours) d'un verdict dichotomique. Dans le cas de QCM ayant quatre options, on accorde, par exemple, un crédit (ex. un point) à l'option représentant la réponse adéquate (ex. A). Aucun crédit n'est accordé aux leurres (B, C, D). Ceci est différent avec un TCS où il n'y a pas à strictement parler une seule bonne réponse. Il peut donc y avoir plus d'une option valable donnant chacune un crédit partiel (une partie du score maximal). La correction du TCS peut faire appel à plusieurs méthodes de détermination des scores comme nous le verrons subséquemment.

Vu la diversité des options de réponses des experts confrontés à des situations cliniques d'incertitude, la constitution d'un panel est généralement réalisée pour statuer sur les réponses. Le TCS est donc soumis à un panel d'experts[5] différents de ceux qui l'ont conçu ; leurs réponses vont permettre de déterminer les scores (nous y reviendrons dans la prochaine section). Une seule étude, celle de Gagnon, Charlin, Coletti, Sauvé et van der Vleuten (2005), s'est intéressée au nombre optimal d'experts. Ils ont constaté que ce nombre était compris entre huit et vingt selon le contexte de l'évaluation (formative, sommative, recherche) et des enjeux inhérents à cette dernière. Et voilà pour nous une limite vu le nombre d'études s'y étant intéressées et la variabilité du nombre d'experts.

1.3 Les méthodes de détermination des scores

Pour la détermination des scores, comme nous l'avons précisé plus tôt, la grande majorité des études recensées (92,50 %) utilisent la méthode des scores combinés proposée par Charlin, Brailovsky, Leduc *et al*. (1998). Cette méthode a été décrite en 1985 par Norman et reprise en 1987 par Norcini (cités par Norcini, Shea et Day, 1990) pour les simulations cliniques difficiles dans lesquelles il y avait un grand nombre de réponses probables. Selon cette méthode, l'option choisie par le plus grand nombre d'experts à une question donnée est retenue comme la réponse modale et aura le score de

[5] Gagnon *et al*. (2005) définissent un expert comme un médecin expérimenté dont la présence dans un jury est légitime en considérant le niveau des candidats évalués.

1. On attribue 0 à l'option non choisie par les experts. Un crédit partiel sera accordé aux autres options (égal au nombre d'experts l'ayant choisi divisé par la réponse modale pour la question donnée). Cette méthode est décrite en détail dans le chapitre III traitant de la méthodologie.

De nombreux auteurs dont Cohen, Fitzgerald, Lane et Boninger (2005), Humbert, Besinger et Miech (2011), Lambert, Gagnon, Nguyen et Charlin (2009), Lemay, Donnon et Charlin (2010) ont étudié les propriétés métriques des scores du TCS générés avec la méthode de Charlin[6] et la recommandent pour une utilisation généralisée dans l'évaluation du RC. D'autres chercheurs tels Bland *et al*. (2005), Lineberry *et al*. (2013) ont émis des réserves relatives à la validité de cette méthode. Ils ont aussi décrit d'autres méthodes de détermination des scores.

Quelles sont ces autres méthodes? Nous allons les exposer brièvement ici à savoir la méthode des scores combinés avec pénalité de distance et la méthode selon la bonne réponse. Ces différentes méthodes sont détaillées dans le chapitre suivant à savoir la recension des écrits.

La méthode des scores combinés avec pénalité de distance décrite par Wilson *et al*. (2014) est peu utilisée dans les écrits. Elle fait intervenir la notion de la distance de l'option de réponse du candidat par rapport à la réponse modale et du degré de l'impact sur son score. Il y a comme dans la méthode des scores combinés précédente l'octroi d'un crédit partiel en fonction de cette distance.

Quant à la méthode selon une bonne réponse, largement employée dans les questions à choix multiples par exemple, elle l'est peu dans le domaine du test de concordance de script et elle n'est pas recommandée par Charlin, Desaulniers, Gagnon, Blouin et van der Vleuten (2002). En effet cela s'explique par le fait même que pour les créateurs du TCS (Charlin, Brailovsky, Leduc *et al*., 1998), le clinicien ne peut avoir une bonne réponse devant l'énoncé de signes et symptômes d'un patient, d'où la notion d'incertitude.

[6] On rappelle que cette méthode a été décrite pour la première fois par Norman (1985). Mais elle a été retenue, utilisée, vulgarisée et recommandée par Charlin dans le cadre du TCS. Ceci justifie notre appellation.

Les écrits indiquent que la tendance est d'employer de plus en plus la méthode des scores combinés, et ce, dans pratiquement tous les domaines de la profession médicale. Toutefois, cette méthode a été aussi remise en question dans quelques écrits récents (Bland *et al*., 2005 ; Lineberry *et al*., 2005). Au regard de tout ceci, il nous semble opportun de nous interroger sur les méthodes de détermination des scores et leur impact sur les propriétés métriques des scores.

1.4 Question de recherche

Dans le cadre de cette étude, nous retenons la question de recherche suivante :

En quoi le choix de la méthode de détermination des scores influence-t-il les propriétés métriques des scores obtenus avec un test de concordance de script?

Pour répondre à cette question, nous avons formulé l'hypothèse suivante que nous espérons pouvoir vérifier : il y a des différences à l'égard des propriétés métriques des scores obtenus avec un test de concordance de script entre les trois méthodes de détermination des scores comparées.

1.5 Justification et pertinence

Le test de concordance de script est un outil de plus en plus utilisé pour l'évaluation du raisonnement clinique dans tous les domaines des sciences de la santé. Il est opportun et juste de mieux comprendre les propriétés métriques des scores selon le choix de l'une ou de l'autre des méthodes de détermination des scores. Nous nous intéressons aussi à l'apport de la modélisation Rasch dans l'étude de cet instrument. Nous voulons vérifier si les scores bruts obtenus à l'aide du TCS peuvent être placés sur une échelle à intervalles égaux. En effet ceci constitue l'originalité de cette étude étant donné que cet aspect n'est pas abordé à date.

Le test de concordance de script a fait, certes, l'objet de nombreuses recherches, mais il reste encore beaucoup de points à élucider. La méthodologie est peu explorée et l'intérêt est porté essentiellement sur la comparaison des groupes de participants à savoir les experts, les résidents et les étudiants. De plus, le TCS s'inscrit dans un contexte de mesure particulier. En effet, il se situe entre l'instrument à réponse déterminée et celui

qui évalue des opinions. Cette étude pourrait offrir des pistes pour trouver un consensus en ce qui concerne la détermination des scores pour le TCS et permettre, ainsi, une plus large utilisation. Charlin est l'un des chercheurs qui a le plus publié sur le TCS. D'ailleurs, dans les écrits recensés, son nom figure dans plus de 60% des publications. Sa contribution est importante et indéniable dans l'évaluation du raisonnement clinique. Toutefois, nous devons nous rappeler que la validité des inférences faites à partir des performances des apprenants est directement dépendante de la méthode de détermination des scores ; ces derniers étant eux-mêmes sous l'influence des experts. Affiner le système de notation du TCS est essentiel. Nous pensons qu'il s'agit d'un élément supplémentaire justifiant la pertinence scientifique de cette étude et un regard extérieur jeté sur le test de concordance de script nous parait tant soit peu nécessaire.

En ce qui concerne la pertinence sociale, nous jugeons que des médecins mieux formés et compétents avec un raisonnement clinique optimal ne peuvent qu'avoir un impact positif sur l'ensemble de la société.

Après avoir développé la problématique, nous présentons une recension des écrits dans le prochain chapitre en dressant un état de la question s'appuyant sur les écrits décrivant l'utilisation du TCS pour évaluer le raisonnement clinique. Nous discutons d'abord des méthodes de détermination des scores et, par la suite, de la modélisation de Rasch.

CHAPITRE II – RECENSION DES ÉCRITS

Dans ce chapitre, nous faisons une recension des écrits sur l'évaluation des apprenants en sciences de la santé. Nous présentons des outils d'évaluation en premier lieu puis les méthodes de détermination des scores retenues. En fin de chapitre, une synthèse des écrits est faite.

Le raisonnement clinique est multidimensionnel et il dépend du contexte clinique. Nous voulons ainsi dire qu'en fonction du contexte clinique donné, le professionnel de santé va formuler des hypothèses bien spécifiques qui lui permettront d'avancer dans la prise en charge diagnostique ou thérapeutique du patient concerné. De même, ce raisonnement clinique est essentiel dans tous les domaines de santé allant du niveau technicien à celui d'exécution de gestes cliniques et/ou thérapeutiques. L'évaluation du RC a été l'objet de nombreuses recherches, ce depuis les années 60 et 70.

Dans le domaine de la recherche cognitive de la compétence médicale, il y a eu un changement notable se focalisant sur l'organisation et l'utilisation des connaissances, la représentation du problème et surtout l'influence de ce dernier par l'expérience. Elstein, Shulman et Sprafka (1978, cités par Lebeau et Pagonis, 2006) ont suggéré que l'évaluation se concentre sur la qualité des opérations cognitives ou la structure des connaissances en comparant le raisonnement et les choix des apprenants à ceux des cliniciens expérimentés. Le test de concordance de script est un instrument conçu pour répondre à cet objectif.

Le TCS évalue les micros jugements évoqués précédemment en situation d'incertitude. Il consiste en une description simple, volontairement incomplète et problématisée d'une situation clinique susceptible d'être rencontrée par l'étudiant à évaluer (Lebeau et Pagonis, 2006). Plusieurs options de réponses sont proposées et il y a la formulation d'une hypothèse de diagnostic, de traitement ou d'investigation. L'évalué doit apprécier l'effet de la nouvelle donnée sur l'hypothèse initiale (Chang *et al*., 2014 ; Charlin, Brailovsky, Leduc *et al*., 1998 ; Charlin, Tardif et Boshuizen 2000 ; Charlin, Gagnon, Sibert et van der Vleuten 2002 ; Charlin *et al*., 2006 ; Charlin, Gagnon,

Lubarsky *et al*., 2010 ; Lebeau et Pagonis, 2006.). Les réponses sont enregistrées sur une échelle de type Likert ; le degré de similitude entre le script de l'étudiant et celui des experts qui auront répondu au même test va être analysé (d'où le terme de concordance).

Différents moyens d'évaluation du RC sont décrits dans la littérature et certains sont encore d'utilisation fréquente. Le test de concordance de script est un outil plus récent et requiert à date une méthode de détermination des scores tout à fait particulière (méthode des scores combinés de Charlin). Les prochaines sections traitent de l'évaluation en contexte clinique c'est-à-dire les outils d'évaluation mais nous abordons aussi les méthodes de détermination des scores jusque-là employées dans le contexte du TCS dans les écrits.

2.1 Quelques outils d'évaluation

Tout processus de formation inclue inéluctablement une évaluation des connaissances et des compétences. Cette évaluation nécessite des outils appropriés qui fournissent des résultats valides et acceptés de façon universelle. Ainsi il nous a paru essentiel de présenter quelques outils d'évaluation disponibles qui sont concurrents au test de concordance de script dans les paragraphes suivants.

2.1.1 Le *patient management problem* (*PMP*)

Le *patient management problem* évalue le processus suivi par un clinicien en vue de retracer l'histoire de la maladie d'un patient et de collecter les informations par l'examen clinique. Après la description brève d'un cas clinique, l'étudiant a à prendre des décisions liées au diagnostic, à l'investigation ou au traitement. Le jugement de l'étudiant sera comparé à celui d'experts et, ainsi, une détermination des scores sera faite selon des critères bien définis.

Selon Harden (1983), le *PMP* simule la réalité et reproduit les décisions qu'un clinicien aura à prendre face à un patient. Le *PMP* a été très largement employé dans les examens de certification ou d'obtention du droit de pratique médicale. Toutefois, on a

vite soulevé des limites psychométriques ou des appréhensions sur une capacité générale de résolution de problèmes cliniques (Page et Bordage, 1995).

2.1.2 Les grilles d'évaluation globale

Ces grilles sont remplies par les observateurs à la fin de la période de stage clinique des apprenants. Elles comportent une liste de critères dont la multiplicité suggère qu'elles peuvent vraiment apprécier les différentes composantes de la compétence clinique. Cependant, à bien les analyser, les critères en relation avec le raisonnement clinique sont peu nombreux. D'autres points négatifs sont mentionnés également. A titre d'exemple : 1- une fidélité faible due à la collecte d'observations sur une période trop longue, 2- un biais lié à l'analyse d'un critère sur celle d'un autre, 3- l'existence d'une certaine subjectivité (temps de contact entre l'étudiant et l'évaluateur trop important). En dépit de tout, les grilles d'évaluation globales sont encore, de nos jours, largement utilisées.

2.1.3 L'examen clinique objectif et structuré (ECOS)

Élaboré à la fin des années 70 par Harden et Gleeson et très utilisé dans les pays anglo-saxons, l'ECOS permet d'évaluer le raisonnement clinique par observation directe au moyen de situations cliniques standardisées et simulées. Il y a différentes stations d'observation de durée prédéterminée. L'ECOS est standardisé ce qui renforce sa validité selon Alinier (2003). Les candidats ont à effectuer de multiples tâches identiques en présence de différents évaluateurs lesquels ont une grille d'évaluation prédéfinie à remplir. L'ECOS est aussi structuré car on a recours à des patients simulés (patients acteurs). Certains auteurs tels Peden, Cairncross, Harden et Crooks (1985) avancent que l'ECOS est une méthode d'évaluation objective. Ils pensent, en effet, que la variabilité entre les examinateurs est diminuée grâce aux grilles d'évaluation. Toutefois, comme tout instrument, l'ECOS a également des limites. On lui reproche d'évaluer des comportements observables ce qui limite ainsi sa validité comme outil de mesure du RC. De plus, il nécessite beaucoup de ressources humaines et financières toujours selon Alinier.

2.1.4 Les questions à choix multiples (QCM)

Elles sont très prisées dans le monde universitaire médical que ce soit dans les examens d'obtention de licence, dans les concours d'admission ou dans les évaluations sommatives en fin d'année de résidence. La rédaction des QCM obéit à des règles bien expliquées entre autres par Haladyna et Rodriguez (2013). Le système de détermination des scores est objectif (Charlin *et al*., 2003). Par contre, elles ne sont pas toujours appropriées pour apprécier l'organisation des connaissances, des scripts de maladie. De plus, on estime qu'il y a une sur-valorisation des connaissances factuelles au détriment de la capacité de résolution des problèmes. Néanmoins, Jolly et Grant (1997, cités par Charlin et al. en 2003) avancent que le raisonnement clinique peut être évalué en élaborant des questions autour de présentations de patients (QCM à contexte riche).

2.1.5 La question rédactionnelle

Comme les précédents, la question rédactionnelle fait encore partie de la liste des instruments d'évaluation. La fidélité inter-juges est faible et les correcteurs se laissent entrainer facilement par les fautes de grammaire et de stylistique des étudiants. Un point important à signaler est le problème de spécificité de contenu face à l'utilisation de longues réponses relatives à un nombre limité de cas cliniques. Certains auteurs tels Charlin *et al*. (2003) suggèrent l'option de questions à réponses courtes et ouvertes. Les questions rédactionnelles sont faciles à construire et elles minimisent l'effet indice. En demandant aux apprenants de décrire les raisons ayant conduit à la prise de décision on peut évaluer le raisonnement.

2.1.6 L'évaluation orale

L'évaluation orale demeure une composante très employée dans les examens de certification compte tenu de son organisation facile. Selon Charlin *et al*., 2003), elle permet de mesurer l'étendue des connaissances, les capacités de résolution de problème ainsi que des facteurs personnels tels que la tolérance au stress, les valeurs et les attitudes. On lui reconnait une bonne validité apparente. Toujours selon ces auteurs, l'évaluateur

peut se faire une idée sur le jugement clinique du candidat. La fidélité entre les cas (spécificité de contenu) est basse par opposition à la fidélité inter-juges qui est élevée. Il faut noter que préalablement à l'évaluation, le candidat va interroger et examiner un patient. Par la suite, son évaluateur aura à le questionner au sujet de son patient (l'histoire de la maladie, la physiopathologie, la démarche diagnostique ou la prise en charge thérapeutique).

2.1.7 Synthèse des outils d'évaluation

Les différents outils présentés dans la section précédente sont destinés à évaluer les connaissances et/ou les compétences des apprenants en sciences de la santé. Certains diront qu'ils conviennent à toutes les compétences attendues d'un professionnel de la santé d'autres en diront le contraire. Ainsi s'explique le développement du test de concordance de script comme outil d'évaluation spécifique du raisonnement clinique. Nous abordons dans la section suivante les différentes méthodes de détermination des scores décrites dans le contexte du TCS et au prime abord celles que nous mettons en application dans le cadre de cette étude exploratoire du raisonnement clinique.

2.2 Les méthodes de détermination des scores

Toute évaluation en éducation sous-entend à la fois pour l'évalué et pour l'évaluateur la formulation de l'appréciation d'un résultat sous forme de scores. Cette dernière constitue ce qu'on appelle la détermination des scores. Ce principe est tout à fait particulier dans le contexte du TCS. En effet, selon la recension des écrits, les scores sont déterminés selon la méthode des scores combinés. Le score de l'étudiant est rapporté à celui de tous les experts ; ceci permet de prendre en compte la variabilité des réponses de ces derniers (Charlin, Brailovsky, Leduc *et al*., 1998 ; Charlin, Gagnon, Sibert *et al*., 2002). Plus le score de l'étudiant est élevé, plus il a des réponses identiques à celles d'un grand nombre d'experts (Charlin, Gagnon, Sibert e*t al*.; Lebeau et Pagonis, 2006 ; Wilson, Pike et Humbert, 2014). Mais ceci soulève aussi des interrogations qui pourraient remettre en question la validité des scores obtenus avec le TCS (Bland *et al*., 2005 ;

Lemay *et al*., 2010 ; Lineberry *et al.,* 2013). Nous nous sommes fixés comme objectif d'apporter quelques pistes avec cette étude.

Une fois que le test est administré aux candidats et que les scores sont calculés en fonction des méthodes retenues, il nous faut comparer les résultats pour répondre à notre question de recherche.

Il faut certes évaluer les compétences des apprenants mais il est aussi impératif de choisir la meilleure méthode de détermination des scores ; ceci est au cœur de notre recherche. Nous présentons successivement la méthode des scores combinés (que nous appelons la méthode de Charlin), en deuxième lieu la méthode des scores combinés avec pénalité de distance et en dernier lieu la méthode selon une bonne réponse. Les autres méthodes décrites dans le contexte du test de concordance de script sont exposées à la fin de cette recension.

2.2.1 La méthode des scores combinés

La méthode des scores combinés est basée sur le principe que chaque réponse reflète une opinion valable ; le nombre de réponses à chaque catégorie doit être enregistré. Elle prend en compte la notion importante de variabilité du raisonnement chez les experts confrontés à des problèmes mal définis selon Charlin et van der Vleuten (2004). Pour chaque item, la réponse de l'évalué reçoit un crédit correspondant à la proportion des membres du panel d'experts ayant fait le même choix. Le score maximal pour chaque question est 1 pour la réponse modale. Les autres choix des membres du panel reçoivent un crédit partiel proportionnel à leur nombre. On accorde zéro aux réponses non choisies par le panel.

Dans le cadre du TCS, chaque réponse donnée par un clinicien membre du panel a une valeur intrinsèque. Donc le score d'un évalué est le reflet de la concordance avec le score des experts (Charlin *et al*., 2010). Prenons l'exemple d'un item répondu par neuf experts sur une échelle en cinq échelons comme l'illustre le tableau 1. Un étudiant ayant choisi l'option -2 recevra le score modal soit 1.

Tableau 1 *Exemple de la détermination des scores avec la méthode des scores combinés*

	Catégorie					Commentaires
	-2	-1	0	+1	+2	
N	4	3	2	0	0	Identifier la catégorie la plus choisie : -2
Pondération	4/4	¾	2/4	0/4	0/4	Diviser par la FM établie à 4
Score A	1,00	0,75	0,50	0,00	0,00	Différents scores possibles

Note. N = nombre de réponses des experts par catégorie. Score A = score déterminé. FM = fréquence modale.

Les résultats du test sont représentés par la somme des crédits obtenus à chaque question. Pour faciliter l'interprétation, Charlin et ses collaborateurs ont suggéré en 2000 de transformer le score total de façon à avoir une note totale maximale de 100.

Bland *et al*. (2005) remettent en question la validité discriminante du TCS avec la méthode des scores combinés. En 2013, Lineberry *et al*. avancent que cette méthode introduit des incohérences logiques dans la clé des scores. En effet, ils mentionnent, par exemple, la possibilité que certains évalués soient plus performants que les experts pour certaines questions[7] et qu'ainsi ils fassent preuve d'une meilleure interprétation des données que ces derniers (Lubarsky, Gagnon et Charlin, 2013). De plus, selon eux tout désaccord au niveau du panel relatif à l'effet d'une information sur une hypothèse qui ne peut être résolue par une discussion rend cet item non approprié pour une évaluation de réussite éducationnelle. Tout ceci constitue une faiblesse quant à la validité de contenu du TCS.

Quant à la fidélité des résultats au TCS en se basant sur la méthode des scores combinés, Lineberry *et al.* (2013) estiment que les erreurs inter panel et les erreurs de mesure test-retest sont généralement ignorées. En outre, en 2007, Vanbelle, Massart, Giet et Albert avancent que l'effet du hasard sur les scores obtenus avec cette méthode doit être considéré. Selon Gagnon et Charlin (2007), ce système de notation du TCS tient

[7] On peut penser, par exemple, à des questions qui concernent des notions ou des concepts récemment développés ou des questions pour lesquelles les étudiants sont plus à l'aise que les experts.

compte de l'éventualité d'une connaissance partielle. Lineberry *et al*. pensent que les évalués choisissant le milieu de l'échelle (ex. 0) pour chaque item auront toujours une meilleure performance que ceux qui utilisent tous les échelons de l'échelle. Pour Vanbelle *et al*., il y a la tendance de certains sujets à choisir soit les catégories extrêmes (biais des extrêmes) soit les catégories intermédiaires (biais de la tendance centrale). Ils proposent alors de déterminer la fréquence avec laquelle les sujets, étudiants ou experts, choisissent les différentes catégories sur l'échelle de réponses indépendamment des items. Ainsi pourra-t-on calculer le degré d'accord dû au hasard et le degré d'accord corrigé en soustrayant le degré d'accord dû au hasard au degré d'accord observé.

Dans la plupart des études recensées ayant utilisé la méthode de Charlin, les auteurs ont essayé de suivre les recommandations de Charlin, Gagnon, Sibert *et al*. (2002), de Charlin et van der Vleuten (2004), de Gagnon, Charlin, Coletti, Sauvé et van der Vleuten, (2005) pour la construction du test de concordance de script de façon à avoir un coefficient de cohérence interne d'au moins de 0,75. En ce qui a trait au niveau de formation des sujets, ce n'est point homogène dans le sens que certains auteurs ont eu comme population soit des étudiants en médecine soit des résidents en spécialité soit les deux, par exemple pour des études de médecine, en comparaison avec des experts d'un domaine donné. Le nombre d'items varie aussi entre 40 et 140. Par exemple, Carrière, Gagnon, Charlin, Downing et Bordage (2009) et Lemay Donnon et Charlin (2010) ont respectivement dans leurs études 50 et 40 items avec un coefficient alpha de Cronbach de 0,77. Par contre dans les études avec un nombre d'items supérieur ou égal à 70, ce coefficient est supérieur ou égal à 0,80. Il est à noter cependant dans l'étude de Deschênes, Charlin, Gagnon et Goudreau en 2011 le score moyen des experts est le plus faible comparé aux autres études avec un coefficient alpha de Cronbach de 0,86. Si nous considérons le nombre d'experts, il est recommandé d'avoir un panel constitué d'au moins 10 experts. Et il faut retenir aussi qu'un nombre de 20 experts n'augmente pas de façon significative la fidélité du TCS (Gagnon *et al*., 2005). Toutefois Lambert, Gagnon, Nguyen et Charlin (2009) ont dans leur étude obtenu un coefficient alpha de Cronbach de 0,90 avec le plus grand nombre d'experts à date c'est-à-dire 47 experts en radio-oncologie. Il est important de rappeler quelque soit le coefficient de cohérence interne obtenu, quel que soit le nombre d'experts, la majorité des études a montré qu'il y avait

une différence relative au degré d'expertise des participants au test de concordance de script. Nous tenons à rappeler que le coefficient alpha de Cronbach est influencé aussi par le nombre d'items Le tableau 2 ci-dessous est un récapitulatif des données de quelques études recensées.

Tableau 2 *Récapitulatif de quelques études relatives au TCS selon la méthode de Charlin*

Etude	N d'items	N d'apprenants[8] Etudiants/résidents	N d'experts	Coefficient α Cronbach	SM des experts/100	SM des apprenants/100 Etudiants/résidents
Boulouffe *et al*. (2013)		21/19 (R)	12	0,80	77,88	59,01/69,53
Bursztein *et al*. (2011)	132	34 (R)	16	0,80	75,60	65,00 (R)
Carrière *et al*. (2009)	50	53 (R)	12	0,77	75,90	69,90
Cooke *et al.* (2016)	137	91	21	0,85	80,00	68,00
Deschênes *et al*. (2011)	73	31	15	0,86	61,60	53,30
Ducos *et al*. (2015)	60	60	10	0,63	82,00	69,90
Kazour *et al.* (2016)	100	47	10	0,79	79,42	58,47
Lambert *et al*. (2009)	90	70 /38 (R)	47	0,90	76,67	51,62/71,20
Latreille (2012)	45	30	15	0,61	77,40	62,30
Lemay *et al*. (2010)	40	53/42	11	0,77	80,00	45,75/54,00
Mathieu *et al*. (2013)	60	15	19	0,82	76,60	61,50

N = nombre ; SM = score moyen

[8] Quand un seul nombre est reporté, il représente le nombre d'étudiants.

2.2.2 La méthode des scores combinés avec pénalité de distance

La méthode de détermination des scores avec pénalité de distance a été décrite par Wilson, Pike et Humbert (2014) qui ont voulu démontrer que le TCS peut effectivement différencier le niveau des apprenants au moyen d'autres méthodes de détermination des scores. Cette méthode est peu utilisée dans la littérature. Elle nous parait particulièrement intéressante car elle tient compte à la fois de la distance de l'option de réponse du candidat par rapport à la réponse modale et du degré de l'impact sur son score. Les scores sont donc déterminés selon la distance par rapport à la réponse modale avec un crédit partiel. Nous appelons cette méthode M2.

Pour cette étude, deux types de TCS ont été mis à profit : l'un avec résolution de problèmes et l'autre en médecine d'urgence. L'échelle de réponses est constituée de cinq catégories (-2, -1, 0, +1 et +2). La réponse modale des experts équivaut à 1 point. Une pénalité selon la distance par rapport à cette réponse modale est faite selon la formule $C = 1-(\delta/\Delta)$: C = méthode de détermination[9] ; δ= distance entre la réponse du candidat et la réponse modale ; Δ = distance maximale entre la réponse modale et les extrêmes (ex. 2, 3, 4). Cette méthode 2 est comparée aux cinq autres dont la méthode de Charlin et la méthode selon la bonne réponse. Nous avons illustré avec un exemple dans le paragraphe suivant.

Nous avons considéré un groupe de neuf experts participant à un TCS. Quatre experts ont choisi l'échelon -2 comme réponse à un item donné; trois experts ont choisi -1 et deux autres ont choisi 0 pour le même item. Ainsi la réponse modale à cet item est -2 ; un candidat choisissant cette réponse obtient 1 point. Un autre candidat ayant choisi +1 va obtenir 0,25 comme score car la catégorie +1 est à une distance de 3 catégories par rapport à la réponse modale en utilisant la formule décrite plus haut. L'exemple est rapporté dans le tableau 3 ci-dessous. Il est à noter qu'avec cette méthode, Wilson *et al*. ont avancé qu'on pouvait bien distinguer les experts et les apprenants en fonction de leur niveau d'apprentissage comme décrit dans la majorité des études. Cette méthode décrite seulement par Wilson, Pike et Humbert (2014) est peu étudiée. La méthode des scores

[9] Le terme est inadéquat. Toutefois nous avons voulu rester fidèles aux propos des auteurs pour la définition de la formule.

combinés de Charlin ne prend pas en compte les choix des candidats ni trop proches ni trop éloignés de la réponse modale en effet. Est-ce pourquoi il nous a paru judicieux de retenir une autre méthode considérant à la fois la distance et l'impact du choix des évalués.

Tableau 3 *Exemple de détermination des scores avec la méthode M2*

Echelle	-2	-1	0	+1	+2
Réponse des experts	4	3	2	0	0
Description	Scores attribués				
- Échelle en 5 échelons : -2, -1, 0, +1, +2	1,00	0,75	0,50	0,25	0,00
- Pénalité selon la distance par rapport à la réponse modale : C = 1-(δ/Δ)*.					

* C : méthode de détermination ; δ = distance entre la réponse du candidat et la réponse modale ; Δ : distance maximale entre la réponse modale et les extrêmes (ex. 2, 3, 4)

2.2.3 La méthode selon une bonne réponse

Comme autre méthode de détermination de scores, il y a la méthode selon une bonne réponse. Cette méthode est classiquement celle utilisée tous les jours en évaluation en sciences de la santé à titre d'exemple pour les questions à choix multiples. En 2002, Charlin, Desaulniers, Gagnon, Blouin et Van der Vleuten se sont interrogés sur les propriétés métriques des scores d'un test de concordance de script corrigé selon cette méthode. Ils ont soumis un TCS de 45 items à des étudiants en médecine et à des experts en obstétrique-gynécologie. Ce TCS était associé à une échelle type Likert en sept échelons (-3, -2, -1, 0, +1, +2, +3). Ils ont comparé les scores obtenus avec la méthode des scores combinés et la méthode des scores selon une bonne réponse (par consensus).

Ils ont observé en fait que les réponses des experts variaient en fonction du contexte – réponse individuelle versus réponse par consensus soit 59 % de désaccord. Quant à la comparaison entre les scores des apprenants et ceux des experts, la différence était statistiquement significative avec la méthode des scores combinés contrairement à

celle obtenue avec la méthode par consensus. La méthode des scores combinés a permis d'obtenir des scores plus élevés et une plus grande moyenne ; ceci est attribuable au crédit partiel accordé avec cette méthode. Avec la méthode par consensus, les meilleurs scores ont été également observés dans le groupe des experts. Des scores très faibles et une grande variabilité de réponses ont été notés. Charlin, Desaulniers *et al*. (2002) ont conclu que, pour un TCS respectant la notion de problèmes mal définis et le contexte d'incertitude, la méthode des scores combinés était supérieure à celle par consensus (Charlin et Van der Vleuten, 2004 ; Norcini, Shea et Day, 1990).

Toutefois, Bland *et al*. (2005) et Lineberry *et al*. (2013) ne partagent pas ces conclusions relatives à la supériorité de la méthode des scores combinés quant au pouvoir discriminant du TCS entre les cliniciens expérimentés et les apprenants. Selon eux, le fait d'avoir plusieurs bonnes réponses constitue un écueil à la validité des inférences faites. Bland *et al*. mentionnent aussi qu'on évalue peut-être la même aptitude, mais il paraît évident qu'on obtient beaucoup plus d'erreurs dues au hasard avec la méthode des scores combinés. Ces auteurs rejettent les recommandations de Charlin, Brailovsky, Leduc *et al*., 1998 ; Charlin, Roy, Brailovsky et van der Vleuten, 2000 ; Charlin, Gagnon, Sibert et van der Vleuten 2002 ; Charlin et van der Vleuten, 2004) et concluent que la méthode des scores combinés n'est pas toujours adéquate.

Nous avons retenu la méthode de détermination des scores selon une bonne réponse que nous avons appelée M3 décrite aussi par Wilson, Pike et Humbert (2014). A noter que l'échelle de réponses préalablement avec 5 échelons a été recodée par ces auteurs pour devenir une échelle en 3 échelons. Nous avons ainsi les catégories négatives (-2 et -1) sont devenues -1. Les catégories positives (+1 et +2) sont devenues +1. Il n'y a pas de crédit partiel. En voici un exemple, considérons toujours 9 comme le nombre d'experts participant au TCS. Quatre experts choisissent l'échelon -2 comme réponse à un item donné; trois experts choisissent -1. Après recodage, la catégorie -2 devient -1 ; donc 7 experts ont choisi l'échelon -1 laquelle devient la réponse modale. Les deux autres experts choisissent l'échelon 0 comme réponse pour le même item. Tout candidat qui choisit la catégorie -1 (la réponse modale) obtient 1 comme score et tout autre choix aura le score de 0. Le tableau 4 ci-dessous est une illustration de la méthode M3.

Il n'y a pas de consensus entre Bland *et al*. (2005) et Wilson, Pike *et al*. (2014) dans l'efficacité ou du moins la supériorité de la M3 comparée à la méthode des scores combinés de Charlin. La fidélité était bien moins élevée selon Wilson, Pike *et al*. avec les méthodes avec une échelle à trois échelons. De plus pour ces derniers, la valeur d'opinions différentes des experts est minimisée. Vu le désaccord entre les suggestions de ces 2 auteurs relatives à la méthode selon la bonne réponse (avec échelle en trois échelons), nous avons voulu retenir aussi cette méthode avec le TCS à l'étude pour en tirer nos propres conclusions.

Tableau 4 *Exemple de détermination des scores avec la méthode selon une bonne réponse*

Description	Nombre d'experts par échelons			Réponse modale	Scores déterminés		
	-1	0	+1				
- Échelle en 5 échelons : (-2, -1), 0, (+1, +2) - Recodage des réponses d'une échelle en 5 échelons à une autre en 3 échelons - Pas de crédit partiel	7	2	0	-1	1,00	0,00	0,00

2.2.4 Les autres méthodes

Toujours selon Bland *et al.* (2005), le fait d'avoir plusieurs bonnes réponses n'est pas idéal et l'étude de Charlin, Desaulniers *et al*. (2002) n'a pas su prouver l'inefficacité de la méthode de détermination des scores selon un consensus avec une seule bonne réponse. De plus, toujours selon ces auteurs, la taille des groupes de sujets de Charlin, Desaulniers *et al*. était inégale et petite. Ils s'interrogent, donc, sur l'application d'autres méthodes de correction et leurs effets sur les qualités métriques des résultats au TCS. Voilà pourquoi ils développent un TCS avec 50 questions en néphrologie selon les recommandations formulées en 2000 par Charlin, Roy *et al.* Cependant pour la notation,

Bland *et al.* proposent quatre autres méthodes en plus de la méthode des scores combinés avec une échelle en cinq échelons, appelée M1 décrite précédemment. La deuxième méthode (2) est également une méthode des scores combinés avec une échelle en trois échelons, avec un crédit partiel et absence de bonne réponse. La troisième méthode (3) décrite par Bland *et al*. prend en compte la réponse modale avec une échelle en trois échelons. La quatrième méthode (4) se rapproche de la M2 retenue car elle génère des scores selon la distance par rapport à la réponse modale mais elle a une échelle en trois échelons. La cinquième et dernière méthode a aussi trois échelons et elle considère la distance par rapport à la moyenne. Il faut bien noter que les trois dernières méthodes de Bland ont un paramètre commun : une bonne réponse. Le tableau 5 illustre les méthodes de détermination des scores de Bland *et al*. Les propriétés métriques des scores du TCS ont été comparées avec ces cinq méthodes.

Tableau 5 *Méthode de détermination des scores selon Bland et al.* (2005)

Méthodes	Description
1. Scores combinés (5 échelons)	- Échelle en 5 échelons : -2, -1, 0, +1, +2 - Aucune bonne réponse - Crédit partiel selon la distribution des scores des experts
2. Scores combinés (3 échelons)	- Échelle en 3 échelons : (-2, -1), 0, (+1, +2) - Aucune bonne réponse - Crédit partiel selon la distribution des scores des experts
3. Scores modaux (3 échelons)	- Échelle en 3 échelons : (-2, -1), 0, (+1, +2) - Bonne réponse - Une réponse qui correspond à la valeur modale vaut 1 point et les autres 0.
4. Scores selon la distance par rapport à la réponse modale (3 échelons)	- Échelle en 3 échelons : (-2, -1), 0, (+1, +2) - Bonne réponse - Crédit partiel selon l'écart entre la bonne réponse et l'option du répondant
5. Scores selon la distance par rapport à la moyenne (3 échelons)	- Échelle en 3 échelons : (-2, -1), 0, (+1, +2) - Bonne réponse - Crédit partiel selon l'écart entre la bonne réponse et l'option du répondant

La troisième méthode basée sur la réponse unique a montré une corrélation du score avec le niveau d'expérience similaire à celles des méthodes 1 et 2 présentées précédemment. Cependant, le coefficient alpha de Cronbach est moins important avec la troisième méthode (Bland *et al*., 2005).

En 2010, Lemay, Donnon et Charlin ont évalué la validité et la fidélité des résultats au TCS en utilisant trois autres méthodes de détermination des scores en pédiatrie. Ils ont construit et soumis un TCS de pédiatrie de 40 items à un panel d'experts (11) et à deux groupes d'apprenants (53 étudiants et 42 résidents en pédiatrie). La

première méthode est basée sur la meilleure réponse ; la deuxième prend en compte la meilleure réponse avec des crédits partiels et la troisième considère seulement les crédits partiels. Contrairement à la méthode des scores combinés, pour la deuxième méthode, ils font le rapport entre le nombre d'experts ayant choisi la réponse et le nombre d'experts du panel. A titre d'exemple, trois experts ont choisi la catégorie -2 sur un total de 11 experts ; ainsi on obtient un score de 0,27 (3/11). Le tableau 6 ci-dessous synthétise la méthode décrite par Lemay *et al*.

Tableau 6 *Méthode de détermination des scores selon Lemay et al.* (2010)

	Catégories					STM	Alpha*
	- 2	- 1	0	+ 1	+ 2		
Score des experts (n=11)	3/11	6/11	2/11	0/11	0/11		
Une bonne réponse	0	1	0	0	0	40	0,74
Une bonne réponse et crédit partiel	0,27	1,00	0,18	0	0	40	0,77
Crédit partiel	0,27	0,55	0,18	0	0	29,3	0,78

Note. n = nombre d'experts. * = coefficient alpha Cronbach. STM = score total maximal.

On retrouve une distribution des scores identique à celle que nous avons décrite précédemment à savoir les étudiants avaient les scores les plus faibles, suivis par les résidents et les experts ont eu les meilleurs scores. La fidélité estimée avec le coefficient alpha de Cronbach et la validité sont relativement élevées. Il n'y a pas de différence significative du coefficient alpha obtenu avec les trois méthodes (Lemay *et al*., 2010). Toutefois, la méthode de la meilleure réponse associée à des crédits partiels montrait une plus grande validité apparente. Par contre, ils n'ont pas pu mettre en évidence une grande différence dans les scores en fonction des trois années de résidence.

Il reste évident que malgré l'utilisation du TCS depuis quelques années, il y a d'autres aspects à étudier quant à la méthode de détermination des scores. En 2014, Wilson, Pike *et al*. estiment que le TCS peut différencier les niveaux d'expérience et ceci avec diverses méthodes de détermination. Deux TCS ont été construits par ces auteurs : un en médecine d'urgence et l'autre en résolution de problèmes. Le TCS en médecine

d'urgence concernait 988 étudiants, 40 résidents et 12 experts. Par contre, le TCS en résolution de problèmes était administré à 522 étudiants et 13 experts. Après analyse et optimisation des items, 49 items ont été retenus. Le tableau 7 ci-dessous résume les méthodes utilisées par Wilson, Pike *et al*. (2014). Les quatre premières méthodes sont basées sur une échelle en cinq échelons. La première méthode (A) est la méthode Charlin. La deuxième (B) fait intervenir une bonne réponse (modale). La troisième (C) attribue à l'évalué un score selon la distance par rapport à la réponse modale avec une pénalité, déjà décrite aussi. La quatrième (D), une méthode des scores combinés, regroupe les méthodes A et C ; D = (A+C)/2. La cinquième méthode (E) selon une échelle en trois échelons est une méthode des scores combinés ; on accorde un crédit partiel. Pour la sixième méthode de Wilson, Pike *et al*. (F), il y a une bonne réponse et une absence de crédit partiel.

Tableau 7 *Méthodes de détermination des scores selon Wilson, Pike et al.* (2014)

Méthode	Description
A. Scores combinés (5 échelons)	- La méthode de Charlin.
B. Une bonne réponse (5 échelons)	- Échelle en 5 échelons : -2, -1, 0, +1, +2. - Bonne réponse. - Une réponse qui correspond à la valeur modale vaut 1 point et les autres 0.
C. Scores selon la distance par rapport à la réponse modale (5 échelons)	- Échelle en 5 échelons : -2, -1, 0, +1, +2. - Pénalité selon la distance par rapport à la réponse modale : $C = 1 - (\delta/\Delta)$*.
D. Scores combinés (5 échelons)	- Échelle en 5 échelons : -2, -1, 0, +1, +2. - Crédits partiel et total sont accordés. De plus, pénalité selon la distance par rapport à la réponse modale. - Association des méthodes A et C ; D= (A+C)/2.
E. Scores combinés (3 échelons)	- Échelle en 5 échelons : (-2, -1), 0, (+1, +2) recodée en 3. - Les réponses +1 et +2 sont recodées en +1. Crédit partiel accordé.
F. Une bonne réponse (3 échelons)	- Méthode E, mais sans crédit partiel.

* C : méthode de détermination ; δ= distance entre la réponse du candidat et la réponse modale ; Δ : distance maximale entre la réponse modale et les extrêmes (ex. 2, 3, 4).

Wilson, Pike *et al. (2014)* ont conclu que les six méthodes discriminent bien le niveau d'expérience des différents groupes. La fidélité obtenue avec les méthodes basées sur une échelle en trois échelons était, de façon significative, plus faible contrairement aux conclusions de Bland *et al*. (2005). De plus, parmi les méthodes avec une échelle en cinq échelons, on avait de meilleurs coefficients de corrélation item-total avec les méthodes A et D, donc une plus grande discrimination.

Nous rappelons que la méthode des scores combinés, largement employée, a deux caractéristiques particulières. Elle ne donne pas de point selon le degré d'exactitude de la réponse des évalués et aucun point sur l'échelle n'est considéré comme étant la meilleure ou la réponse la plus correcte.

Bland *et al*. (2005) pensent que les résultats des scores obtenus avec la méthode des scores combinés et ceux avec la méthode basée sur une seule bonne réponse sont très similaires. En 2012, Dory, Gagnon, Vanpee et Charlin avançaient que les études portant sur des méthodes alternatives de détermination des scores n'étaient pas concluantes et que la méthode traditionnelle de notation (les scores combinés) est satisfaisante. La recension des écrits nous a montré que la question est soulevée et qu'il n'y a toujours pas eu de consensus. Voilà pourquoi nous pensons que cette étude descriptive exploratoire méritait d'être menée afin d'essayer d'éclaircir cet aspect.

Dans le tableau 8 ci-dessous, nous faisons un rappel des méthodes retenues dans le cadre de cette étude. Nous avons pris le cas d'un item avec cinq échelons (-2, -1, 0, +1, +2) et pour lequel 4 experts ont choisi la réponse -2, trois ont choisi -1, deux ont choisi 0. Les autres réponses +1 et +2 n'ont pas été choisies. A partir de ce cas de figure, nous démontrons comment on détermine les scores pour les étudiants.

Tableau 8 *Synthèse des trois méthodes de détermination des scores retenues*

	Echelle	-2	-1	0	+1	+2
	Réponse des experts	4	3	2	0	0
Méthode	Description	Scores déterminés				
1. Scores selon la méthode des scores combinés	- Identifier la catégorie la plus choisie : -2 - Diviser par la FM° établie à 4 - Différents scores possibles	1,00	0,75	0,50	0,00	0,00
2. Scores selon la distance par rapport à la réponse modale (5 échelons, Wilson *et al.*)	- Échelle en 5 échelons : -2, -1, 0, +1, +2 - Pénalité selon la distance par rapport à la réponse modale : C = 1-(δ/Δ)*.	1,00	0,75	0,50	0,25	0,00
3. Bonne réponse (3 échelons, Wilson *et al.*)	- Recodage des réponses d'une échelle en 5 échelons à une autre en 3 échelons : -1, 0, +1 - Nombre d'experts par échelons respectivement : 7, 2, 0 - Réponse modale: -1 - Pas de crédit partiel	1		0		0

° FM: fréquence modale; * C : méthode de détermination; δ = distance entre la réponse du candidat et la réponse modale; Δ : distance maximale entre la réponse modale et les extrêmes (ex. 2, 3, 4).

La partie précédente de la recension ne concerne que la théorie classique des tests. Cette étude exploratoire s'intéresse aussi à l'apport de la modélisation de Rasch pour évaluer les propriétés métriques des scores d'un test de concordance de script. A date très peu de données sont disponibles dans la littérature à ce sujet.

2.3 La famille des modèles de Rasch

En dehors de la théorie classique des tests[10], il existe des modèles permettant d'apprécier les propriétés métriques des scores obtenus avec les instruments de mesure. Nous pouvons citer quelques-uns tels la théorie de la généralisabilité, l'analyse factorielle, la théorie des réponses aux items et la modélisation de Rasch. Cette dernière a retenu notre attention et est au cœur même de cette étude

Les modèles de mesure de Rasch sont très utilisés en édumétrie de façon générale. Pour certains auteurs, il y a beaucoup d'avenir avec ces modèles dans l'évaluation en éducation ou en sciences de la santé (Bond et Fox, 2015; Downing, 2003). En effet, la modélisation de Rasch a été employée dans le contexte d'examens écrits (Yang, Tsou, Chen, Chan et Chang, 2011), d'examens cliniques (McManus, Thompson et Mollon, 2006) et d'examens cliniques objectifs structurés, ECOS (Iramaneerat, Yudkowsky, Myford et Downing, 2007). En revanche, peu d'auteurs se sont intéressés à date à l'apport des modèles de mesure de Rasch dans le contexte du test de concordance de script. Il a fallu attendre qu'en 2017 Dionne, Grondin et Latreille s'y intéressent de plus près. En effet ces auteurs ont modélisé un test de concordance de script conçu pour évaluer le raisonnement clinique des étudiantes infirmières par Latreille (2012). De même, Grondin, Dionne, Savard et Casimiro (2017) ont utilisé la modélisation de Rasch pour évaluer les propriétés métriques d'un instrument de mesure de l'offre active de services sociaux et de santé en français dans les communautés francophones en situation minoritaire.

Il faut noter qu'il y a des points communs entre la théorie classique des tests et la modélisation de Rasch ainsi une comparaison entre ces deux théories est présentée dans le tableau 9 ci-dessous.

[10] La théorie classique des tests n'est pas expliquée en détail dans ce document. Toutefois le document de Laveault et Grégoire (2002) peut être consulté pour en savoir plus.

Tableau 9 *Comparaison entre la TCT et la modélisation de Rasch**

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

**tiré de Grondin, Dionne, Savard et Casimiro 2017, traduit et adapté de Cano et al., 2013*

En faisant une modélisation de Rasch, on vise à créer une échelle linéaire (en logit) à intervalles égaux. Ainsi la position des sujets et la difficulté de l'item sont placées sur une échelle commune (Wilson, Allen et Corser, 2006). Les erreurs standard de mesure sont aussi prises en compte. Selon Cavanagh et Waugh en 2011 (p. 123), avec les modèles de Rasch, on calcule la probabilité de réussite d'un sujet à un item donné à partir de la difficulté de cet item (D) et de la compétence du sujet (B). Dans la modélisation selon Rasch, il y a deux paramètres statistiques, *infit* et *outfit*. Ces derniers sont reportés sous deux formes à savoir une version non standardisée (carré moyen) et une version standardisée, (statistique t).

Les indices *outfit* sont très sensibles, selon Smith et Smith (2004), à la présence de données aberrantes. Ainsi les indices *infit* sont plus appropriés pour l'étude de l'ajustement des données au modèle. De plus, avec les *outfit*, les réponses inattendues éloignées du score du sujet sont identifiées tandis que les *infit* permettent de déceler les réponses inattendues qui sont près du score du sujet (Wright et Masters, 1982). Rappelons que, selon Wright et Linacre (1994), les problèmes identifiés avec les indices *infit* sont en général difficiles à diagnostiquer et à corriger contrairement aux indices *outfit*.

Les résidus représentent la différence entre les scores observés et les réponses attendues (Bond et Fox, 2007). Pour les résidus standardisés *outfit* t, Bond et Fox recommandent les valeurs entre -2 et +2. Quant au résidu standardisé *infit* t, la valeur doit se situer entre 0,75 et 1,30. Si les statistiques pour un item ne sont pas comprises dans ces intervalles, cet item devrait être mis de côté temporairement pour manque d'ajustement au modèle.

Il y a différents modèles unidimensionnels dans la famille des modèles de Rasch, entre autres, le modèle dichotomique, le modèle *Rating Scale* et le modèle *Partial Credit.* Ce dernier modèle est celui qui sera utilisé dans le cadre de cette étude.

2.3.1 Le modèle à crédit partiel

Dans ce modèle, les intervalles entre chaque catégorie sont sujets à des variations d'un item à l'autre. Comme nous l'avions dit précédemment, les seuils[11] ne sont pas identiques avec le TCS.

Le modèle à crédit partiel permet de faire des analyses et les chercheurs espèrent avoir des résultats invariants. Selon la propriété d'invariance avec Bertrand et Blais (2004, p. 110-114), on peut affirmer que : « 1) l'estimation de l'habileté d'un individu est indépendante des items auxquels il doit répondre ; 2) l'estimation des caractéristiques des items est indépendante des caractéristiques des individus qui répondent aux items. Pour

[11] On appelle seuil (*threshold*) le point où la probabilité qu'une personne choisisse la catégorie suivante plutôt que la précédente.

que la propriété d'invariance soit établie, le modèle doit d'abord s'ajuster aux données. Pour cela, les hypothèses de base des modèles de Rasch et, en ce qui nous concerne particulièrement, les postulats du modèle à crédit partiel doivent être vérifiés. Nous voulons parler de l'unidimensionnalité et de l'indépendance locale.

Avec l'unidimensionnalité, selon Blais (1987), le chercheur doit pouvoir prouver que les réponses des évalués sont sous la dépendance d'une dimension unique dominante (par exemple, le raisonnement clinique en ce qui nous concerne) laquelle permet d'expliquer la performance des évalués à chacun des items. Nous tenons à préciser que dans la famille de Rasch il existe des modèles multidimensionnels aussi. Mais pour les besoins de cette étude, nous avons opté pour l'utilisation d'un modèle à structure unidimensionnelle. Qu'en est-il de l'indépendance locale?

Selon le postulat de l'indépendance locale, toutes les réponses d'un évalué sont statistiquement indépendantes l'une de l'autre. C'est ainsi que pour un TCS de 135 items, la réponse à l'item 12 ne permet pas de prédire l'option de réponse que l'évalué prendra pour les items 54, 70 ou 84. En général, la vérification de l'unidimensionnalité confirme celle de l'indépendance locale mais ce n'est pas toujours le cas. Il faut faire donc une démonstration raisonnable du caractère unidimensionnel et de l'indépendance des items entre eux. Pour Dionne, Grondin et Latreille (2017), l'essence même du TCS dans le sens que plusieurs items sont associés l'un à l'autre pourrait expliquer que le postulat d'indépendance locale soit difficile à vérifier. Après cette étape, l'analyse se poursuit avec l'ajustement des données du test au modèle à crédit partiel.

Dans la section 2.4 suivante, une synthèse de la recension des écrits est présentée. Puis la méthodologie appliquée dans le cadre de cette étude exploratoire est décrite dans le prochain chapitre.

2.4 Synthèse de la recension des écrits

Les écrits recensés ont permis de peindre le portrait de l'utilisation du TCS dans la recherche en éducation médicale pour évaluer le raisonnement clinique. Le test de concordance de script a été, depuis sa conception, l'objet de plusieurs études. En effet il a

été mis en application dans différentes spécialités médicales. En faisant référence à la pratique clinique réelle, la position des auteurs tels Lubarsky, Charlin, Cook, Chalk et van der Vleuten (2011) semble tout à fait adaptée. Malheureusement en 2002, Charlin, Desaulniers *et al*. ont réalisé une étude comparative de la méthode de scores combinés en utilisant une échelle en sept échelons avec la méthode de consensus. Nous pensons que l'interprétation des résultats était biaisée par l'échantillonnage. Nous rappelons que nous avons discuté des recommandations faites par les auteurs du TCS pour sa construction (cf section 1.2).

L'étude de Bland *et al*. (2005) a été la première à remettre en question la suprématie de la méthode Charlin pour mesurer le RC ; mais l'échantillon était trop faible selon nous pour en tirer des conclusions. Lemay *et al*. (2010) ont aussi contesté cette méthode en prenant l'option de meilleure réponse tout en accordant des crédits partiels. L'étude la plus récente relative au TCS et aux méthodes de détermination des scores est celle menée par Wilson, Pike *et al*. (2014) qui semble consacrer la méthode des scores combinés en cinq échelons. Ils ont introduit une nouvelle méthode des scores combinés avec pénalité de distance (aussi en cinq échelons), bien prometteuse.

Nous nous posons la question de trouver une alternative à tout cela en réalisant une analyse secondaire des données issues d'un TCS. Un aspect nouveau dans le TCS sera abordé dans cette étude à savoir l'apport de la modélisation Rasch à l'étude des propriétés métriques des scores. En effet, très peu d'études s'y sont intéressées, à ce jour.

La recension des écrits a permis de faire le tour du test de concordance de script et des différentes méthodes de détermination des scores utilisées par les nombreux auteurs ayant recours au TCS. La modélisation de Rasch a été décrite en mettant l'emphase sur le modèle à crédit partiel qui nous parait le plus approprié dans le cadre de cette étude. Dans le prochain chapitre est présentée la méthodologie qui décrit le contexte de l'étude, la population étudiée et la description du calcul des scores avec chacune des méthodes retenues. Puis nous présentons le recueil et le plan d'analyse en exposant aussi les indices statistiques et la modélisation. En dernier lieu une synthèse de la méthodologie est faite.

CHAPITRE III – MÉTHODOLOGIE

3.1 Contexte de l'étude

En vue de réaliser cette étude exploratoire, un contact a été établi avec le CPASS de l'Université de Montréal d'abord par un échange de courriels puis par vidéoconférence. Par la suite, l'Université de Liège en Belgique a été proposée comme lieu d'étude car cette équipe souhaitait une analyse plus approfondie des données de leur test de concordance de script en utilisation depuis peu.

En Belgique, l'accès au Master Complémentaire en Médecine Générale (MC en MG) est conditionné à la réussite d'une évaluation sommative à la fin d'une année préparatoire en Médecine Générale. Pour la plupart des étudiants, cette année coïncide avec la dernière année des études médicales de base (7^{e} année du Master en Médecine). Les étudiants qui s'orientent vers la Médecine Générale font le choix de suivre des enseignements optionnels visant l'acquisition de compétences et de ressources nécessaires à l'exercice de cette discipline.

Le dispositif d'enseignement comporte six stages d'un mois dans diverses pratiques de Médecine générale ainsi que 200 heures d'enseignement dispensées sous forme de cours *ex cathedra*, d'ateliers en petits groupes et d'enseignements interactifs. Au terme de ces enseignements, l'examen de Médecine Générale est composé de stations d'ECOS (Examen Clinique Objectif Structuré), de postes d'EMS (Entrevues Médicales Simulées), de questions Vrai ou Faux et de QCM (Questions à Choix Multiples).

Chaque dispositif d'évaluation présente toutefois des avantages et des limites, en termes notamment de catégories de performances évaluées (connaissances, raisonnement clinique, habiletés etc.). C'est la multiplicité (raisonnable) et la complémentarité des évaluations qui tendent à apprécier le plus justement possible les capacités et les compétences des étudiants. Cependant en prenant en considération les situations auxquelles les professionnels de santé sont confrontés, nous pouvons dire qu'elles sont de deux types. Il s'agit certaines fois de problèmes dits simples ou encore bien structurés

avec toutes les données nécessaires pour les résoudre avec un haut degré de certitude. Toutefois nous savons que dans la majorité des cas ces professionnels font face à des problèmes complexes ou mal structurés. Alors intervient le raisonnement clinique c'est-à-dire le processus de résolution de problèmes qui doit être évalué chez les apprenants.

Le dispositif d'évaluation actuellement mis en place à l'Université de Liège présente selon les auteurs des faiblesses. Les questions Vrai/Faux permettent d'apprécier des connaissances factuelles liées à des problèmes simples tandis que les ECOS et les EMS évaluent des éléments d'une compétence professionnelle dans des situations relativement complexes mais pour lesquelles la conduite à tenir est prédéfinie.

D'autres modalités d'évaluation pourraient adéquatement compléter ce dispositif, notamment dans le but d'explorer le raisonnement clinique et la gestion de situations plus complexes qui ne peuvent être résolues de manière univoque avec un haut degré de certitude. Ainsi, le Département de Médecine Générale (DMG) a décidé d'introduire une modalité d'évaluation complémentaire au dispositif actuel. Il s'agit du test de concordance de script (TCS). Le TCS constitue donc une évaluation sommative, certificative, ayant un poids de 10 % dans le dispositif global d'évaluation en Médecine Générale.

3.2 Rédaction du TCS

Sept professeurs du DMG, par ailleurs médecins généralistes en exercice, ont rédigé chacun 10 questions, basées sur leur expérience clinique quotidienne, de type TCS à trois hypothèses : diagnostique, d'investigation et thérapeutique. Secondairement, 13 experts, maîtres de stage médecins généralistes, ont répondu individuellement au questionnaire et ils ont ainsi constitué le panel de référence. Un expert a été supprimé parce que ses réponses étaient systématiquement différentes de celles des autres experts. Les questions bimodales des experts c'est-à-dire celles pour lesquelles les experts répondaient aux deux extrêmes de l'échelle ont été annulées ; cela faisait évoquer une opposition radicale des opinions des experts. Les auteurs du TCS ont vérifié également que les choix des experts ne se positionnaient pas systématiquement sur le même échelon

de l'échelle de Likert. Finalement, le panel était constitué de 12 experts et leurs réponses ont permis de bâtir la grille de correction. Cinquante vignettes ont donc été retenues pour le test de concordance de script composé ainsi de 24 vignettes de type diagnostic, 13 vignettes de type thérapeutique et 13 autres de type investigation.

Dans un premier temps, ce TCS a été administré, en 2009, à un groupe de 31 étudiants de 7ième année d'études médicales de base se dirigeant vers la spécialisation de médecine générale. Les scores ont été attribués aux étudiants selon la méthode de correction validée par l'Université de Montréal, à savoir la méthode des scores combinés. Des analyses statistiques selon la théorie classique des tests ont été réalisées telles des analyses descriptives, la mesure de la fidélité (coefficient alpha de Cronbach). Aussi une procédure d'optimisation du test a été menée par les auteurs du TCS ; cela signifie l'élimination des items ayant une corrélation item-total inférieure à 0,10. Ceci a permis d'identifier cinq vignettes problématiques et elles ont été écartées. Ainsi la version finale du TCS est constituée de 45 vignettes soit 135 items (21 vignettes à visée diagnostique, 12 vignettes à visée d'investigation et 12 à visée thérapeutique).

3.3 L'échantillon utilisé dans le cadre de cette recherche

Préalablement l'accord de l'Université de Liège pour l'analyse secondaire des données a été obtenu et également celui du Centre de Pédagogie Appliquée aux Sciences de la Santé (CPASS) de l'Université de Montréal. Les données brutes ont été transférées sous la forme d'un dossier Excel, de façon anonyme, après obtention du certificat éthique de l'Université d'Ottawa.

Le TCS ainsi décrit plus haut a été administré à cinq groupes de candidats entre 2010 et 2014. Ces candidats sont au même niveau dans leur cursus ainsi aucune stratification basée sur le niveau d'étude n'a été faite contrairement à d'autres études. La taille de la population varie d'une année à l'autre. En 2010, 27 étudiants ont passé le test; en 2011, ils étaient au nombre de 40 et en 2012, au nombre de 22 étudiants. En 2013, le TCS a été administré à 25 étudiants et en 2014, à 46 autres étudiants. Au total, l'échantillon est composé de 160 étudiants.

3.4 Détermination des scores

La présente étude visait à comparer les propriétés métriques des scores du test de concordance de script de Liège au regard de trois méthodes de détermination des scores. Pour rappel voici les méthodes retenues : (a) M1, la méthode des scores combinés avec une échelle en cinq échelons (Charlin, Brailovsky, Leduc *et al*., 1998) ; (b) M2, la méthode attribuant les scores selon la distance par rapport à la réponse modale (Wilson, Pike *et al*., 2014) et (c) M3, la méthode des scores selon la meilleure réponse en trois échelons, sans crédit partiel (Wilson *et al*., 2014). Ce choix s'explique par le fait que certains auteurs opposés à la M1 tels Bland *et al*. (2005) ont avancé qu'il n'y avait aucune supériorité de cette méthode par rapport à la méthode des scores combinés à trois échelons. De plus, pour eux, la méthode selon le consensus prenant en considération la réponse modale donnait des résultats fiables permettant de discriminer les participants selon leur expérience (Bland *et al*., Wilson *et al*., 2014).

Nous expliquons maintenant la démarche suivie pour avoir les scores. Le fichier Excel contenant les choix de réponses des 160 étudiants pour les 135 items a été envoyé par l'équipe du Département de Médecine Générale de Liège. De plus, il y a eu un fichier Excel pour chacune des 5 cohortes tel expliqué plus haut (section 3.3) ce qui nous a poussé à nous intéresser aux résultats des étudiants par cohorte. Nous avons comparé les cohortes entre elles pour déceler ainsi un éventuel effet de cohorte ; à noter cependant qu'il y avait peu d'informations concernant ces étudiants dans ce fichier. Nous avons aussi déterminé les scores des étudiants en fonction des trois dimensions du test de concordance de script à savoir la dimension diagnostique, la dimension investigation et la dimension thérapeutique. Ceci a été fait pour chacune des méthodes de détermination des scores à l'étude.

Les choix de réponse ont été codés de façon brute en : 1, 2, 3, 4, 5. Pour M1, les scores bruts ont été calculés pour chaque étudiant en fonction du vecteur de réponse des experts à partir du logiciel automatique du CPASS (largement employé dans les études relatives au TCS et disponible sur le site du CPASS). La somme de ces scores a été faite et a été exprimée en pourcentage. La moyenne, les valeurs minimales et maximales, l'écart-type et la variance ont été calculés.

Pour les deux autres méthodes à savoir la M2 et la M3, nous tenons à souligner la non-utilisation du logiciel du CPASS. A partir du vecteur de réponse des experts, les scores bruts de chacun des 160 étudiants ont été déterminés en fonction de la description faite de la M2 ou de la M3 (pour rappel, le tableau 8). Le score total des étudiants est calculé et les analyses statistiques sont réalisées.

3.5 Plan d'analyse

Plusieurs logiciels permettent de traiter les données tant pour la théorie classique des tests (Excel, SPSS, Jmetrik par exemple) que pour la modélisation Rasch (Facets, Winsteps, RUMM30). Le traitement des données de cette étude a été fait avec Excel principalement et aussi avec SPSS en ce qui a trait à la théorie classique des tests. Vu l'importance que revêt le crédit partiel dans la détermination des scores, la modélisation des données a été réalisée selon le modèle de crédit partiel de Rasch avec le logiciel Winsteps 3.90.0.

À la lumière de la théorie classique des tests (TCT), nous avons fait une analyse descriptive et inférentielle des scores des deux groupes (experts et étudiants). En ce qui a trait aux étudiants, cette analyse a été réalisée en trois temps. Nous avons considéré d'abord l'échantillon total des 160 étudiants puis l'analyse a été faite pour chaque cohorte d'étudiants selon l'année de passation du TCS bien que nous ne disposions pas de toutes les caractéristiques de ces étudiants. Notre but était de détecter un éventuel effet de cohorte. En dernier lieu, nous avons poussé plus loin notre analyse en tenant compte des trois dimensions du test de concordance de script. Il faut noter aussi que certaines propriétés métriques de la TCT n'ont pas été prises en compte ; en effet, la TCT a été largement utilisée dans le contexte du test de concordance de script par les auteurs. Nous avions placé de préférence la modélisation de Rasch au centre de cette étude. Il est admis que la théorie classique des tests et la modélisation de Rasch sont complémentaires. En effet, si on considère les statistiques d'ajustement pour les sujets, elles n'existent pas dans la TCT tandis que dans la modélisation de Rasch elles nous permettent d'identifier les sujets à enlever de la modélisation pour améliorer la mesure. De même quand un sujet obtient un score élevé à la modélisation cela sous-entend que ce sujet a un construit élevé.

Un autre point important de la complémentarité de ces deux modèles est le fait de pouvoir vérifier le postulat d'indépendance locale des items. La Figure 3 ci-dessous résume les différentes analyses effectuées. La comparaison des indices statistiques obtenus a été effectuée et est présentée dans le chapitre des résultats.

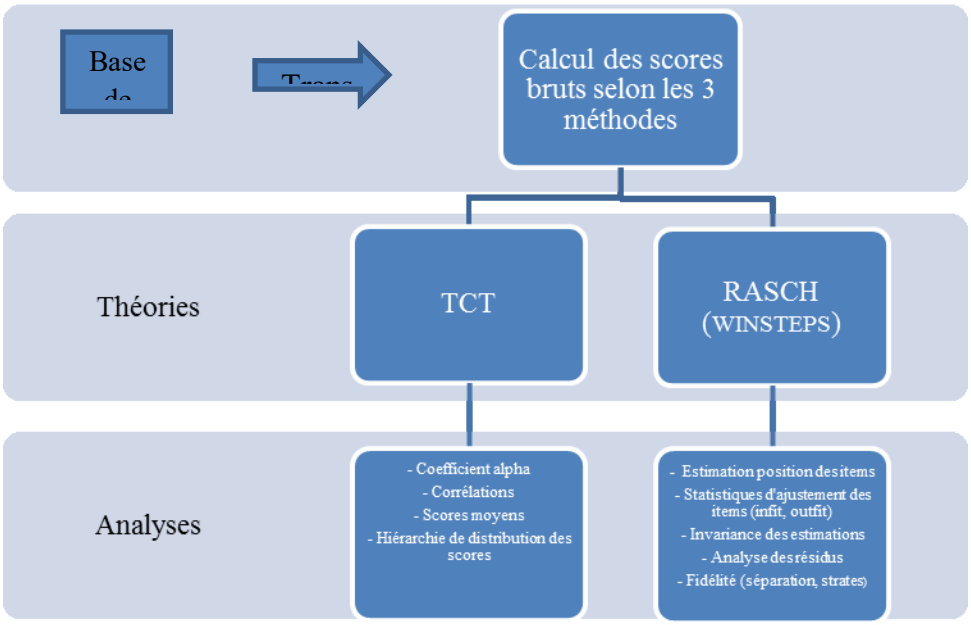

Figure 3. Récapitulatif des analyses.

3.5.1 Les indices statistiques

Un score a été attribué à chaque item et le score total est, comme expliqué par Caire *et al*. (2004), la somme des scores de tous les items rapportés sur 100. Pour ce faire, le calculateur TCS disponible sur le site suivant http://www.cpass.umontreal.ca a été utilisé. C'est un logiciel automatisé de correction sous format Excel. Ainsi pour la théorie classique des tests, les indices de statistique descriptive suivants ont été calculés: la moyenne, l'écart-type, la médiane, le mode, la corrélation item-total et la cohérence interne. Dans le but d'évaluer la distribution des scores des experts, la variance a été calculée. Ainsi avons-nous considéré toute valeur de variance inférieure à 0,50 comme une variance faible. Toute valeur comprise entre 0,50 et 1,00 est considérée comme variance modérée et pour la variance élevée, nous avons retenu la valeur supérieure à 1,00.

3.5.2 Modélisation de Rasch

La modélisation Rasch a été aussi mise à profit pour évaluer les propriétés métriques du test de concordance de script. Les scores bruts obtenus selon la TCT sont sur une échelle ordinale donc une échelle non linéaire.

Avec la modélisation Rasch, les réponses aux items sont utilisées pour créer une échelle linéaire qui représente une variable latente telle une compétence particulière. La position de chaque item sur une échelle représente la difficulté de l'item en unités standard, appelée *logit* (pour *log odds of answering successfully*) (Cavanagh et Waugh, 2011, p. 58). Ainsi la position du sujet et celle de l'item pourront être comparées sur une même échelle pour cette variable (Wilson *et al*., 2006), illustrée en général dans la carte sujet-item de Wright. L'échelle est généralement centrée à zéro logit représentant l'item de difficulté moyenne pour cette échelle donnée. En 2007, Tennant et Conaghan ont élaboré une démarche à suivre en sept étapes pour la modélisation synthétisée comme suit :

1. Fournir une description du modèle choisi et expliciter les raisons qui ont motivé le choix du modèle.
2. Effectuer une analyse de la qualité de l'ajustement des données au modèle choisi, autant pour les sujets que pour les items, et une justification des choix (méthode d'analyse, retrait ou non de sujets ou d'items, intervalle d'ajustement).
3. S'assurer d'avoir des échelles de réponses bien ordonnées (c'est-à-dire monotone croissante), effectuer un recodage de celles-ci si nécessaire (c'est-à-dire peut-être regrouper deux ou trois catégories afin d'améliorer l'échelle et de corriger les problèmes trouvés) et expliciter les choix.
4. Fournir une démonstration de la vérification de la condition d'indépendance locale, ainsi qu'une vérification de la dépendance des réponses et de la condition d'unidimensionnalité nécessaires à l'utilisation de ce type de modèle.
5. Effectuer une vérification de la présence ou non d'un fonctionnement différentiel d'items (FDI) et expliciter les actions prises pour les corriger s'il y a lieu.

6. Fournir une description de la qualité de l'échelle de mesure par une bonne mise en correspondance entre les items et les sujets.
7. Fournir une analyse des indices de fidélité pour les sujets et les items.

3.5.2.1 Explication du modèle choisi

Tous les items du test de concordance de script à l'étude sont conçus avec le même nombre de catégories de réponse. Comme nous l'avions dit précédemment, les seuils[12] ne sont pas identiques avec le TCS. Chaque item possède ses paramètres de seuil F_x. Dans le modèle à crédit partiel retenu pour l'analyse, les intervalles entre chaque catégorie sont sujets à des variations d'un item à l'autre. L'équation du modèle à crédit partiel est la suivante :

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

Dans cette équation, P_{nijk} est la probabilité que la personne n avec une opinion B_n réussisse la catégorie x (où x = 0 à m-1, pour m catégories de réponse offertes) d'un item i dont le degré de difficulté est D_i. Le paramètre F_{ix} correspond au point, pour cet item, où la probabilité de choisir l'une ou l'autre des catégories est égale.

3.5.2.2 Description des analyses faites avec le modèle

Selon Bertrand et Blais (2004, p.110-114), l'appréciation d'un modèle de Rasch doit être dirigée vers une évaluation de l'ajustement global des données au modèle : celle de l'ajustement pour chaque item (*item fit*) ou de l'ajustement pour chaque sujet (*person fit*). En 1980, Rasch avançait que le non-ajustement entre le modèle et les données pouvait être lié soit à un problème au niveau des données soit à un problème au niveau du modèle. Cette étape de vérification est essentielle pour les conclusions portées avec le modèle. Wright et Mok (2004) expliquent que l'ajustement global dépend de la qualité des données et de la validité du construit.

Linacre (2015) recommande aux chercheurs pour l'analyse de l'ajustement

[12] On appelle seuil (*threshold*) le point où la probabilité qu'une personne choisisse la catégorie suivante plutôt que la précédente.

d'étudier d'abord :

1- les données ayant des corrélations négatives,
2- les problèmes d'ajustement *outfit* avant ceux d'*infit*,
3- Les indices d'ajustement basés sur le carré moyen (CM) avant ceux basés sur la version standardisée (STD)
4- Les valeurs élevées des indices d'ajustement avant les valeurs les plus faibles.

3.5.2.2.1 Analyse de l'ajustement

Nous avons analysé en suivant les étapes sus-décrites l'ajustement des sujets puis celui des items et en dernier lieu l'ajustement global des données au modèle car ainsi le nombre d'items conservé peut être maximisé. Nous nous sommes basés sur les valeurs recommandées par Linacre (2002) en optant pour -1,9 et +1,9 pour une prévisibilité raisonnable des données.

Dans le tableau 10 ci-dessous nous avons reporté les valeurs des indices d'ajustement basées sur la version standardisée qui ont servi de référence selon Linacre (2002). La comparaison de la position du score moyen obtenu par un sujet à celle de la valeur zéro pour un item va renseigner sur la façon dont les items sont en adéquation avec les sujets. Une valeur moyenne positive pour un sujet indique que l'échantillon représentant un ensemble était situé à un niveau plus élevé que la moyenne de l'échelle, par exemple (Tennant et Conaghan, 2007).

Tableau 10 *Les indices d'ajustement basés sur la version standardisée – Interprétation*

Intervalle	Qualité de la mesure
≥ 3,00	Données très inattendues. Certains facteurs ont pu contrevenir à la mesure (inattention, réponses au hasard, etc).
2,00 – 2,90	Données très difficiles à prédire et ce, de façon notable.
-1,90 – 1,90	Données raisonnablement prédites.
≤ -2,00	Données trop faciles à prédire. D'autres dimensions (facteurs) ou des problèmes de dépendance locale peuvent contraindre les données dans des patrons de réponse qui les rendent excessivement prévisibles.

Source : Tiré de Grondin *et al.* (2017) traduit et adapté de Linacre (2002)

3.5.2.2.2 Vérification du fonctionnement différentiel d'items

La vérification de la présence ou non d'un fonctionnement différentiel d'items n'a pas pu être faite car les données contextuelles n'étaient pas disponibles.

3.5.2.2.3 Analyse de la fidélité

La fidélité de séparation des sujets (R_p) varie entre 0 et 1 et elle est exprimée en unités d'erreur standard (Bond et Fox, 2001, p. 207). Selon Bond et Fox (2007) et Linacre (2002), l'indice de séparation des sujets (G_p) équivaut à la racine carrée du rapport entre la variance vraie du sujet et l'erreur de variance des données et il s'étend de 0 à l'infini. Nous pouvons évaluer les différentes strates de l'échantillon de sujets avec l'indice de séparation des sujets dont une valeur élevée est souhaitable. Pour cet indice, Boone, Staver et Yale (2014) proposent les valeurs suivantes : 1,50 comme indice acceptable; 2,00 comme indice bon et 3,00 comme indice excellent. Une valeur faible indique que le test ne permet pas de différencier les sujets performants de ceux qui ne le sont pas.

Comme nous l'avons signalé précédemment, la fidélité de séparation des items (R_i) et l'indice de séparation des items (G_i) ont aussi été estimés de la même manière. Quant à l'indice de séparation des items, nous avons retenu la valeur minimale proposée par Boone *et al*. (2014, p. 227) à savoir 2,5 pour une utilisation de groupe. Toutefois, l'essentiel est de se rappeler qu'une valeur élevée renseigne mieux qu'une valeur basse

(Wilson *et al*., 2006). Il y a aussi l'indice de fidélité (*reliability*) dont la valeur doit être supérieure à 0,90.

3.5.2.2.4 Vérification des deux postulats fondamentaux

Les hypothèses d'indépendance locale et d'unidimensionalité sont aussi appréciées. La dépendance locale se définit par le fait qu'une réponse à un item détermine celle à un autre item (Tennant et Conaghan, 2007). Nous voulions détecter ainsi l'existence d'items dépendants l'un de l'autre avec la matrice de corrélation résiduelle. En ce qui a trait à l'unidimensionnalité, nous avons voulu nous assurer de l'absence de sous-dimensions qui pourraient influencer les réponses des sujets et ainsi avoir un impact sur l'échelle de mesure comme expliqué par Grondin *et al*. (2017). Pour cela, nous avons effectué une analyse en composantes principales sur les résidus standardisés, en d'autres termes après l'extraction du facteur Rasch. Il y a une forte probabilité de présence de sous dimensions quand la valeur propre des différents contrastes est supérieure à 2 ; ceci signifie qu'il y a regroupement d'au moins deux items. Nous avons analysé le degré de corrélation des items avec les différents contrastes de l'analyse en composantes principales. Une attention particulière est donnée aux items avec un poids d'au moins 0,40 associé au facteur. Par la suite, nous avons relevé les trois groupes d'items identifiés par le logiciel Winsteps en fonction de leur corrélation avec le facteur : fortement, moyennement et faiblement. En nous rapportant à Grondin *et al*., la valeur de corrélation entre les mesures des répondants modélisés par les deux groupes les plus opposés c'est-à-dire la plus forte et la plus faible corrélation est étudiée. Nous avons synthétisé dans le tableau 11 ci-dessous la stratégie d'analyse de l'indépendance locale et de l'unidimensionnalité retenue.

Tableau 11 *Stratégie d'analyse de l'indépendance locale et de l'unidimensionnalité**

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

Une valeur de 0,71 correspond à une variance commune de 50,4% et donc à des items un peu plus dépendants qu'indépendants (Linacre 2005).
*Source : tiré de Grondin *et al*. Après traduction et adaptation de Linacre (2015).

3.6 Synthèse de la méthodologie

Pour résumer, nous avons traité les réponses des experts et des étudiants selon un vecteur de réponse. Il a fallu coder les options de réponse choisies et déterminer les scores pour chaque méthode M1, M2 ou M3. Puis les analyses ont été faites selon la théorie classique des tests et selon le modèle de crédit partiel. Pour la TCT, nous avons analysé les scores de trois façons en considérant soit l'échantillon total des étudiants, soit les cohortes d'étudiants selon l'année de passation du test soit selon les trois dimensions du TCS (diagnostique, investigation et thérapeutique). Les indices suivants ont été calculés: les statistiques descriptives, les statistiques d'ajustement des items et des sujets (*infit* et *outfit* standardisé et carré moyen, indice de séparation des items et des sujets, indice de fidélité des items et sujets. Nous avons étudié aussi la position des sujets et des items sur la carte Wright, la présence et le nombre de strates.

Dans le chapitre IV qui suit nous présentons les résultats obtenus pour les trois méthodes en premier lieu selon la théorie classique des tests et en second lieu selon la modélisation de Rasch.

CHAPITRE IV – LES RÉSULTATS

Tout au long des chapitres précédents, nous avons discuté des analyses qui seraient faites avec nos données. Nous avons précisé les valeurs de référence des différents indicateurs statistiques retenus. Les données ont été recueillies entre 2010 et 2014 auprès de 160 étudiants. Elles ont été anonymisées préalablement avant le transfert. En premier lieu, les scores bruts ont été analysés au regard de la théorie classique des tests (TCT) et secondairement la modélisation a été faite. La première section de ce chapitre traite les scores des experts.

4.1 Les scores des experts

Nous rappelons que la grille de correction se base sur les réponses des 12 experts de la spécialité c'est-à-dire des experts en médecine générale. Nous avons dénombré neuf valeurs manquantes pour les réponses aux items (6,70%) au total pour un nombre de trois experts. Aucune décision n'a été prise pour le remplacement ou non de ces valeurs. Quand il y avait égalité du nombre d'options de réponses pour un item, en d'autres termes deux réponses modales, nous avons gardé les deux dans la grille de correction pour déterminer les scores des étudiants. Ces valeurs ont été incluses dans le logiciel du CPASS et les calculs ont été faits pour chacun des 12 experts pour les 135 items. Par la suite les indices statistiques ont été calculés à savoir la moyenne, l'écart-type, la variance. La moyenne des scores des experts est de 80,15% ± 7,54 avec une variance de 70,20.

L'étendue des scores se situe entre 70 et 88 pour les experts. Nous avons dénombré 64 items (47,40%) de variance faible dont 4 items avec une variance inférieure à 0,10. 52 items (38,52%) ont une variance modérée et 19 items (14,08%) ont une variance élevée. Nous nous sommes aussi intéressés au patron de réponse des experts. L'option 3 (c'est-à-dire 0) a été choisie pour 46 items et l'option 4 (+1) pour 48 items. Les deux extrêmes à savoir l'option 1 (-1) et l'option 5 (+2) ont été choisies très rarement. En ce qui a trait à la réponse à 2 catégories d'écart par rapport à la réponse

modale, nous avons trouvé 10 items (7,41%). Toutes ces valeurs sont présentées dans l'annexe 1.

Nous continuons avec les scores des étudiants lesquels, nous le rappelons, ont été analysés en trois temps : 1) avec l'échantillon total, 2) par cohorte d'étudiants selon l'année de passation et 3) selon les dimensions du TCS. Pour faciliter la compréhension de la section des résultats, nous avons exposé en premier lieu les résultats des étudiants avec la théorie classique des tests pour les trois méthodes retenues. En second lieu, sont donnés les résultats obtenus après modélisation en suivant le même plan c'est-à-dire d'abord avec M1 puis avec M2 et finalement avec M3.

4.2 Théorie classique des tests

4.2.1 La méthode des scores combinés (M1)

La méthode 1 est la méthode des scores combinés, décrite à la section 2.2.1 sous le nom de méthode de Charlin[13]. Les options de réponse des candidats selon une échelle de Likert (-2, -1, 0, +1, +2) ont été recodées de la façon suivante en vue de faciliter le traitement par les logiciels d'analyse : - 2 = 1; -1 = 2; 0 = 3; +1 = 4 et +2 = 5. On a comptabilisé 18 valeurs manquantes sur les 21600 réponses attendues (soit 0,06%) pour huit étudiants sur 160 (5,00%). Nous signalons que, comme la plupart des auteurs l'ont fait selon les écrits scientifiques, aucune action spécifique n'a été posée pour les valeurs manquantes. Le faible nombre de données manquantes ne justifiait pas de procéder au remplacement de ces dernières. Dans un second temps, les options de réponse recodées ont été transférées dans le calculateur automatique du CPASS[14] de l'Université de Montréal. Ce calculateur largement utilisé donne les paramètres suivants : la réponse modale, les scores de tous les candidats pour chaque item, la variance totale du test, la

[13] Nous désignons ainsi la méthode des scores combinés recommandée par Charlin et utilisée dans la majorité des écrits sur le TCS.
[14] http://www.cpass.umontreal.ca/sct.html

variance par item, le coefficient alpha de Cronbach, la corrélation item-total, les statistiques descriptives pour les experts et les étudiants (moyenne, minimum, maximum, médiane et écart-type). Nous présentons les paramètres décrits dans l'ensemble des études relatives au TCS à titre d'exemple Latreille, 2012 ; Marie *et al*., 2005 ; Mathieu *et al*. 2013.

4.2.1.1 Résultats pour l'échantillon total des 160 étudiants avec M1

Le score moyen des étudiants est de 72,40 ± 6,10. Les valeurs des différents indicateurs statistiques calculés sont reportées dans le tableau 12 ci-dessous.

Tableau 12 *Statistiques descriptives des scores bruts des étudiants avec M1*

Statistiques	Etudiants (n=160)
Moyenne (%)	72,40
Écart-type	6,10
IC 95 %	71,40 – 73,30
Mode	44,50
Médiane	73,30
Variance	35,60
Asymétrie	-1,40
Aplatissement	3,30

Le score moyen des étudiants (72,40%) est inférieur au score moyen des experts (80,00%) avec une différence de 7,60%. En considérant l'étendue de ces scores, elle varie entre 45 et 84 pour les étudiants. Il y a une asymétrie négative de -1,40 et un aplatissement de 3,30. Ceci porte à croire que la distribution des données n'est pas normale.

Le coefficient de cohérence interne alpha Cronbach a été calculé pour l'ensemble du test ; ce coefficient est de 0,79. Pour la corrélation item-total, la valeur moyenne est de 0,18 (± 0,13). Les 135 items du TCS ont été regroupés selon trois groupes en fonction de

la valeur de la corrélation item-total[15] (désignée par r) : $r \leq 0,10$; $0,10 < r \leq 0,20$ et $r > 0,20$. Ainsi nous dénombrons 31 items (22,96%) dans le groupe avec $r \leq 0,10$ et 38 items (28,15%) dans le groupe $0,10 < r \leq 0,2$. Soixante-six items (48,90%) ont un coefficient de corrélation item-total supérieur à 0,20. Les valeurs de corrélation item-total sont présentées en annexe 2.

Nous avons procédé, comme mentionné précédemment, à l'analyse des résultats des scores par cohorte d'étudiants selon l'année de passation. Ces résultats vont être présentés dans la section suivante.

4.2.1.2 Résultats par cohortes d'étudiants selon l'année de passation avec M1

Entre 2010 et 2014, la moyenne des scores varie entre 51,23 % et 74,72%. En 2013, nous avons la plus faible moyenne avec un coefficient alpha de Cronbach à 0,35. La plus forte moyenne est observée en 2014 et elle est de 74,72%, par contre, avec un coefficient alpha de Cronbach à 0,14. Dans le tableau 13 ci-dessous les statistiques descriptives des scores de ces différentes cohortes sont reportées.

[15] Ces balises de corrélation item-total sont largement employées dans les études du TCS (Lambert, 2005 ; Latreille, 2012)

Tableau 13 *Statistiques descriptives des scores selon l'année de passation avec M1*

Année	2010	2011	2012	2013	2014
Nombre d'étudiants	27	40	22	25	46
Moyenne (%)	72,72	73,06	69,45	51,23	74,72
Étendue	48,85-80,87	58,67-83,92	44,48-78,27	38,63-57,03	65,30-81,53
Médiane	74,51	73,18	70,81	53,17	75,54
Variance	47,51	28,49	47,01	23,51	15,58
Écart type	6,89	5,34	6,86	4,85	3,95
Asymétrie	-1,82	-0,46	-2,50	-1,03	-0,72
Aplatissement	4,38	0,08	8,32	0,54	0,01
IC 95%	2,60	1,65	2,87	1,90	1,14
Cronbach	0,69	0,50	0,67	0,33	0,14

Une analyse comparative des résultats obtenus avec la méthode 1 selon les cohortes des étudiants de juin 2010 à juin 2014 a été faite. Nous avons voulu chercher de possibles effets de cohorte mais nous n'avions pas les caractéristiques propres à chaque cohorte pour procéder à cette recherche. Un test F d'indépendance des échantillons a fourni les résultats suivants :

- Les cohortes 2010 et 2011 proviennent de deux échantillons indépendants à variances égales (F < *Critical* F) et ils ont donc la même dispersion.
- Les distributions des moyennes pour les cohortes 2010 et 2014 n'ont pas la même dispersion (F > *Critical* F).
- Les distributions des moyennes pour les cohortes 2011 et 2014 n'ont pas la même dispersion (F > *Critical* F).
- Les distributions 2012-2013 n'ont pas la même dispersion (F > *Critical* F).

Par conséquent, de peur que les différences de variances ne soient confondues avec des différences de moyennes (problème d'hétérocedasticité), le test de comparaison

de moyennes (*test T*) n'a été réalisé que sur les distributions de 2010 et 2011. Et voici les résultats obtenus : un degré de liberté 65 avec un t stat de 0,25 ; *p* (T ≤ t) unilatéral 0,40 et *p* (T ≤ t) bilatéral 0,80.

Dans la section suivante, nous présentons les résultats obtenus selon M1 en considérant les trois dimensions du test de concordance de script à savoir la dimension diagnostique, la dimension investigation et la dimension thérapeutique (cf. section 3.2). Ces analyses ont été faites pour l'ensemble des étudiants.

4.2.1.3 Résultats selon les dimensions du TCS et l'échantillon total avec M1

Pour rappel le TCS à l'étude a 21 vignettes pour la dimension diagnostique, 12 pour la dimension investigation et 12 vignettes pour la dimension thérapeutique. Chacune de ces vignettes regroupe 3 items. Pour la dimension diagnostique, la moyenne est de 74,55% avec une variance au test de 44,67. Le coefficient alpha de Cronbach est de 0,87. Pour la dimension investigation, la moyenne est de 70,70% avec une variance au test de 62,13 et un coefficient alpha de Cronbach de 0,96. La dernière dimension est la dimension thérapeutique et les résultats obtenus sont les suivants : une moyenne de 70,17% avec une variance au test de 65,21 et un coefficient alpha de Cronbach de 0,96. Tous ces résultats sont reportés dans le tableau 14 ci-dessous.

Tableau 14 *Statistiques descriptives du TCS avec M1 par dimension*

	Dimension		
	Diagnostique (63 items)	Investigation (36 items)	Thérapeutique (36 items)
Moyenne	74,55	70,70	70,17
Étendue	41,40 - 85,22	42,46 - 88,68	40,41 - 89,77
Médiane	75,76	71,09	71,96
Asymétrie	-1,86	-0,87	-0,98
Aplatissement	5,84	0,49	1,41
Variance	44,67	62,13	65,21
Variance items	6,24	4,27	4,18
Écart type	6,68	7,88	8,08
Cronbach	0,87	0,96	0,96

Toute la section précédente traitait des résultats avec la méthode de Charlin appelée dans cette étude M1. Qu'en est-il de la méthode des scores combinés avec pénalité de distance? La section suivante lui est consacrée.

4.2.2 La méthode des scores combinés avec pénalité de distance (M2)

Le score est déterminé avec une pénalité selon la distance par rapport à la réponse modale avec cette méthode. Les valeurs manquantes n'ont pas été l'objet de codage spécifique comme préalablement mentionné à la section 4.2.1. L'échelle est constituée de cinq échelons comme pour la méthode 1. Le score moyen des experts est de 80,40% ± 5,40 avec une étendue de 71,35 et 91,20. Nous présentons dans les paragraphes suivants les résultats obtenus par les étudiants.

4.2.2.1 Statistiques descriptives pour l'échantillon total des 160 étudiants avec M2

Les étudiants ont un score moyen de 77,10 ± 4,40. La distribution des scores est entre 57,90 et 84,80. L'asymétrie est positive de 1,50 et il y a un aplatissement de 4,90. Dans le tableau 15 ci-dessous sont reportées les statistiques descriptives des scores des 160 étudiants.

Tableau 15 *Statistiques descriptives des scores bruts des étudiants avec M2*

Statistiques	Étudiants (n=160)
Moyenne	77,10
Écart-type	4,40
IC 95 %	77,40 – 77,80
Mode	77,71
Médiane	77,70
Variance	18,70
Asymétrie	1,50
Aplatissement	4,90

Nous avons également évalué la cohérence interne du TCS avec la méthode M2 en calculant le coefficient alpha de Cronbach pour l'ensemble du test. Ce coefficient est de 0,73. Pour la corrélation item-total, la valeur moyenne est de 0,17 ± 0,13. Les 135 items ont été aussi regroupés comme pour la M1 selon trois groupes en fonction de la valeur de la corrélation item-total (désignée par r) : $r \leq 0,10$; $0,10 < r \leq 0,20$ et $r > 0,20$. Ainsi trente-quatre items (25,19%) ont un coefficient de corrélation inférieur ou égal à 0,10. Quarante items (29,62%) ont un coefficient de corrélation compris entre 0,10 et 0,20. Par contre, le coefficient de corrélation est supérieur à 0,20 pour 61 items (45,19%). Les valeurs des indices de corrélation item-total sont présentées en annexe 3.

Nous avons procédé également à l'analyse des scores par cohorte d'étudiants au regard de la méthode 2. Ceci est présenté dans la section suivante.

4.2.2.2 Résultats par cohortes d'étudiants selon l'année de passation avec M2

Les cinq cohortes d'étudiants sont réparties sur les années 2010, 2011, 2012, 2013 et 2014. Ces résultats sont reportés dans le tableau 16 ci-dessous. La moyenne de ces scores varie entre 74,60% et 79,19% et la plus grande valeur est notée en 2014. Quant au coefficient alpha de Cronbach, sa valeur varie entre 0,15 et 0,70. La plus faible valeur du

coefficient de cohérence interne est retrouvée pour la cohorte de 2014 et la plus élevée (0,70) est pour la cohorte de 2012.

Tableau 16 *Statistiques descriptives des scores bruts selon l'année de passation avec M2*

Année	2010	2011	2012	2013	2014
Nombre d'étudiants	27	40	22	25	46
Moyenne (%)	76,97	75,78	75,02	74,60	79,19
Étendue	61,85-83,58	66,23-83,27	57,90-81,67	60,56-81,36	72,47-84,81
Médiane	78,02	76,98	76,05	75,68	79,66
Variance	24,31	17,93	21,52	25,12	8,00
Écart type	4,93	4,23	4,64	5,01	2,83
Asymétrie	-1,17	-0,33	-2,55	-1,35	-0,50
Aplatissement	1,99	-0,44	8,92	1,28	-0,24
IC 95%	-1,86	1,14	2,21	1,70	0,82
Cronbach	0,67	0,42	0,70	0,56	0,15

Les tests F d'égalité de la variance pour les différentes cohortes 2010, 2011, 2012, 2013 et 2014 ont montré qu'il y avait égalité de variance pour les distributions 2010/2013 et 2011/2012. En effet, pour ces deux couples d'années, le F calculé est inférieur à la valeur F critique. Par conséquent, un t-test a été réalisé pour détecter les différences au niveau des moyennes pour ces distributions. Pour le couple 2010-2013, la différence n'est pas statistiquement significative (Test t, t(24) =1,32, ddl = 48, p=0,09, n=26. Pour le couple 2011-2012, la différence est significative statistiquement (Test t, t(21) =2,19, ddl = 58, p= 0,02, n=39).

Toutefois, il faut préciser que leurs caractéristiques étant inconnues, l'on ne peut tirer de conclusions seulement à partir de ces tests.

Comme pour la méthode 1, voici dans la prochaine section les résultats obtenus selon les dimensions du test de concordance de script au regard de la méthode 2.

4.2.2.3 Résultats selon les trois dimensions du TCS avec M2

Le score moyen des étudiants est de 77,40% pour la dimension diagnostique ; il est de 77,74% et de 75,96% respectivement pour la dimension investigation et la dimension thérapeutique. Le coefficient alpha de Cronbach est de 0,87 pour la dimension diagnostique, de 0,97 pour la dimension investigation et de 0,94 pour la dimension thérapeutique. Les différentes statistiques descriptives obtenues sont reportées dans le tableau 17 ci-dessous.

Tableau 17 *Statistiques descriptives des scores selon les dimensions du TCS avec M2*

	Dimension		
	Diagnostique (63 items)	Investigation (36 items)	Thérapeutique (36 items)
Moyenne	77,40	77,74	75,96
Étendue	50,13 - 86,11	56,02 - 91,90	58,33 - 84,95
Médiane	77,84	77,66	76,62
Asymétrie	-1,83	-0,51	-0,76
Aplatissement	6,94	0,82	0,62
Variance	23,44	36,68	26,33
Variance items	3,48	2,26	2,15
Écart type	4,84	6,06	5,13
Cronbach	0,87	0,97	0,94

Des analyses comparatives ont été faites également en fonction des dimensions et des cohortes d'étudiants. La valeur critique est supérieure à 0,05 ; ainsi il n'y a pas de différence statistiquement significative entre les scores des étudiants selon les trois dimensions du TCS.

Nous allons à présent rapporter les résultats obtenus en utilisant la méthode 3, la méthode selon une bonne réponse dans le cadre de cette étude à la section suivante. Par la suite un tableau comparatif des résultats de ces trois méthodes au regard de la théorie classique des tests est présenté.

4.2.3 La méthode selon une bonne réponse (M3)

Nous rappelons que la méthode M3 prend en considération la réponse modale comme la bonne réponse. Il n'y a pas de crédit partiel, elle est donc dichotomique (soit 0 pour les autres réponses, soit 1 pour la réponse modale). Le score moyen des experts est de 79,30% ± 6,90 avec une étendue de 67,50 et 82,70 selon M3.

4.2.3.1 Résultats pour l'échantillon total des 160 étudiants avec M3

Avec la méthode M3, comme rapporté dans le tableau 18 ci-dessous, le score moyen des étudiants est de 64,60% ± 4,50 avec une distribution entre 38,50 et 77,03. Nous avons un aplatissement de 7,20 et une asymétrie négative de -1,40.

Tableau 18 *Statistiques descriptives des scores bruts des étudiants avec M3*

Statistiques	Étudiants (n=160)
Moyenne (%)	64,60
Écart-type	4,50
IC 95 %	63,90 – 65,30
Mode	66,67
Médiane	65,20
Variance	19,90
Asymétrie	-1,40
Aplatissement	7,20

De la même façon que pour la M1 et la M2, le coefficient alpha de Cronbach a été calculé avec la méthode 3 pour l'ensemble du test en vue d'évaluer la cohérence interne. Nous avons obtenu une valeur de 0,38. Pour la corrélation item-total, la valeur moyenne est de 0,11 ± 0,09. Les 135 items de notre test ont été regroupés selon trois groupes en fonction de la valeur de la corrélation item-total (désignée par r) : $r \leq 0,10$; $0,10 < r \leq 0,20$ et $r > 0,20$. Ainsi cinquante-sept items (42,20%) ont un coefficient de corrélation inférieur ou égal à 0,10. Cinquante-cinq items (40,70%) ont un coefficient de corrélation compris entre 0,10 et 0,20. Par contre, le coefficient de corrélation est supérieur à 0,20 pour 23 items (17,03%). Les valeurs des indices de corrélation item-total sont également présentées en annexe comme pour les 2 méthodes précédentes.

Nous avons procédé aussi au calcul des statistiques descriptives des scores selon la méthode 3 en fonction de l'année de passation du test de concordance de script conformément à notre plan pour les méthodes 1 et 2.

4.2.3.2 Résultats par cohortes selon l'année de passation avec M3

Voici les résultats obtenus dans le tableau 19 ci-dessous. La moyenne des scores varie entre 62,66% et 66,09%. Mais il faut noter qu'avec la méthode 3, nous avons obtenu des valeurs négatives du coefficient alpha de Cronbach (2010, 2011 et 2014). La variance la plus élevée est de 43,36 et correspond à l'année 2012 et le coefficient alpha de Cronbach est de 0,47.

Tableau 19 *Statistiques descriptives des scores selon l'année de passation avec M3*

Année	2010	2011	2012	2013	2014
Nombre d'étudiants	27	40	22	25	46
Moyenne	64,99	66,09	62,66	63,53	64,51
Etendue	60,00-74,81	60,00-77,04	38,52-69,63	48,89-70,37	56,30-74,07
Médiane	64,44	65,93	64,44	64,44	64,44
Variance	13,19	12,39	43,36	23,15	14,97
Ecart type	3,63	3,52	6,58	4,81	3,87
Asymétrie	0,68	0,78	-2,46	-1,24	0,00
Aplatissement	0,41	1,08	8,30	2,32	-0,05
IC 95%	1,37	1,09	2,75	1,89	1,12
Cronbach	-0,75	-0,78	0,47	0,04	-0,54

Pour la méthode 3, le test F d'égalité de variance a montré que les variances sont statistiquement égales pour les cohortes suivantes prises deux à deux 2010/2011, 2010/2012, 2011/2012, 2011/2013, 2012/2013, 2013/2014. L'analyse des données a retrouvé une différence statistiquement significative seulement pour les couples 2011-2012 et 2011-2013 avec le Test t de student.

Les dimensions du test de concordance de script ont été aussi analysées pour la M3 ; Les résultats sont présentés dans la section suivante.

4.2.3.3 Statistiques descriptives des scores selon les dimensions avec M3

Avec la méthode selon une bonne réponse, le score moyen est de 63,77 pour la dimension diagnostique tandis qu'il est de 67,17 pour la dimension investigation et de 63,40 pour la dimension thérapeutique. Le coefficient alpha de Cronbach pour apprécier la fidélité du test de concordance de script est de 0,72 pour la dimension diagnostique. Respectivement pour la dimension investigation et la dimension thérapeutique, le coefficient est de 0,89 et de 0,90. Avec la dimension diagnostique, les valeurs les plus faibles de la variance au test et du coefficient alpha de Cronbach ont été observées. Les valeurs de ces statistiques descriptives sont présentées dans le tableau 20 ci-dessous. De plus l'analyse comparative des scores moyens selon les dimensions et selon les années de passation n'a pas révélé de différence statistiquement significative.

Tableau 20 *Statistiques descriptives des scores selon les dimensions avec M3*

	Dimension		
	Diagnostique (63 items)	Investigation (36 items)	Thérapeutique (36 items)
Moyenne	63,77	67,17	63,40
Étendue	33,33-79,37	50,00-88,89	33,33-80,56
Médiane	63,49	66,67	63,89
Asymétrie	-0,70	0,03	-0,60
Aplatissement	3,59	0,27	1,56
Variance au test	35,45	44,44	51,40
Variance items	10,40	5,92	6,35
Écart type	5,95	6,67	5,13

Cronbach	0,72	0,89	0,90

Analyse comparative des résultats des différentes méthodes

En considérant les cinq cohortes de l'échantillon total et les trois dimensions du TCS, les résultats des analyses descriptives ne sont pas différents. L'hypothèse nulle d'égalité des variances en fonction d'une dimension donnée n'est pas vérifiée ; ainsi les scores ne sont pas identiques. Au vu des résultats de T-test, nous avons poussé plus loin nos analyses en réalisant des comparaisons multiples avec le Tukey test des différentes cohortes en prenant en considération chacune des méthodes de détermination des scores. La valeur critique est hautement significative donc inférieure à 0,05 et nous avons ce profil : M2>M1>M3 pour la cohorte de 2010 et notamment pour toutes les autres cohortes.

Nous avons eu une meilleure variance des scores des étudiants en considérant la M1. Le test *T* a été effectué pour vérifier si les variances dans les scores en fonction des deux méthodes étaient égales. Ainsi, avons-nous pu voir que les variances ne sont pas significativement égales (valeur critique inférieure à 0,05, rejet de l'hypothèse d'égalité des variances des scores). Nous avons aussi comparé les moyennes obtenues avec chacune de ces deux méthodes à savoir M1 et M2. Et la valeur critique pour le test *T* était égal à 0,000 donc inférieur à 0,05 ; ceci permet d'avancer que les scores sont significativement différents avec un intervalle de confiance (IC = -5,90666; -3,59121) à 95%. De plus, pour la corrélation des échantillons appariés, entre M1 et M2 il y a une forte corrélation de 0,952 avec une valeur critique de 0,000.

Pour la comparaison M1- M3, les scores des étudiants étaient moins dispersés avec la méthode 3. Certes on peut voir que les scores attribués aux étudiants ne sont pas égaux ; mais pour confirmer ou non cette inégalité on a eu recours au test *T*. Avec le test *T* la valeur critique est inférieure à 0,05 ce qui prouve que les scores moyens sont significativement différents avec un intervalle de confiance (IC=6,60 ; 8,94) à 95% en fonction de la méthode de détermination des scores choisie (M1 ou M3). Le test de Levene montre que les variances des scores avec les deux méthodes ne sont pas

significativement égales (valeur critique inférieure à 0,05). Pour cette paire M1-M3, la corrélation est de 0,580 avec une valeur critique inférieure à 0,05.

Pour la comparaison M2-M3, les scores des étudiants étaient moins dispersés avec la méthode 2. Les scores attribués aux étudiants ne sont pas égaux et le test *T* a été utilisé pour confirmer ou non cette inégalité. Avec le test *T*, la valeur critique est inférieure à 5% (0,000) ce qui prouve que les scores moyens sont significativement différents avec un intervalle de confiance (IC =11,56 ; 13,49) à 95% en fonction de la méthode de détermination des scores choisie (M2 ou M3). Le test de Levene montre que les variances des scores avec les deux méthodes ne sont pas significativement différentes (valeur critique égale à 0,931, supérieure à 5%). La corrélation entre M2 et M3 est de 0,590 avec une valeur critique inférieure à 5%. La méthode selon une bonne réponse donne le score moyen le plus faible aussi en nous référant à la comparaison multiple (cf annexe 11).

Une comparaison multiple avait été aussi lancée avec le Tukey test pour les dimensions du test de concordance de script selon la méthode de détermination des scores. En effet avec M1, nous avons obtenu des valeurs statistiquement significatives (*p* inférieure à 0,05) pour la dimension diagnostique et la dimension investigation. Par contre la dimension thérapeutique a une différence de moyenne négative avec la dimension diagnostique et la dimension investigation ; les valeurs critiques sont supérieures à 0,05. Quant à la méthode des scores combinés avec pénalité de distance, la M2, la valeur critique est supérieure à 0,05 en comparant la dimension diagnostique et la dimension investigation avec une différence de moyenne négative. La dimension thérapeutique a des valeurs critiques inférieures à 0,05 quand on la compare aux deux autres dimensions ; la différence de moyenne est négative. En considérant la méthode 3, la méthode selon une bonne réponse, les valeurs critiques sont inférieures à 0,05 en comparant la dimension investigation aux dimensions diagnostique et thérapeutique.

Les scores obtenus avec les trois méthodes ont été comparés à l'aide d'une analyse de variance (ANOVA). Nous avons testé l'hypothèse nulle à savoir que les scores moyens sont égaux. Une valeur critique inférieure à 0,05 a été retenue. Il faut noter que l'hypothèse nulle est rejetée parce qu'une différence significative est retrouvée avec une valeur critique égale à 0,000. Le test Kruskal-Wallis H (test non paramétrique) a été aussi

réalisé et la valeur critique est égale à 0,000 avec 2 degrés de liberté. La comparaison entre deux méthodes soit M1-M2, soit M1-M3 soit M2-M3 avec Anova ou avec le test Kruskal-Wallis a donné des valeurs critiques inférieures à 5% dans l'ensemble. Avec la comparaison multiple, La méthode 2 a une différence de moyenne positive avec M1 et M3 mais avec des valeurs critiques inférieures à 0,05.

Nous avons diversifié nos analyses afin de trouver une réponse à notre question de recherche à savoir l'impact de la méthode de détermination des scores sur les propriétés métriques des scores. Ceci sera discuté dans le prochain chapitre. Dans la section 4.3 suivante nous présentons les résultats de la modélisation avec le logiciel Winsteps. Nous cherchions à voir jusqu'à quel point nos données s'ajustaient au modèle de crédit partiel retenu.

4.3 Modélisation de Rasch

En réalisant cette étude, nous nous étions fixés comme objectif de comparer les propriétés métriques des scores obtenus avec le test de concordance de script au regard de trois méthodes de détermination des scores. Mais nous avons voulu aussi aborder le TCS sous un autre angle en utilisant la modélisation Rash.

Le logiciel Winsteps version 3.90.0 (2015) a été utilisé et le modèle à crédit partiel a été retenu ; les valeurs manquantes ont été remplacées par le code 999. Nous avons analysé la qualité de l'ajustement des sujets d'abord puis celle de l'ajustement des items. Nous avons considéré l'indice d'ajustement carré moyen et l'indice d'ajustement standardisé. Toutefois Smith et *al.* (2003) recommandent de se baser sur la version standardisée des indices d'ajustement car elle est plus stable et permet de mieux détecter certains problèmes d'ajustement. La modélisation a été faite d'abord avec la méthode des scores combinés de Charlin, puis avec la méthode des scores combinés avec pénalité de distance et enfin la modélisation a été faite avec la méthode selon une bonne réponse.

4.3.1 La modélisation de Rasch et la méthode des scores combinés

La modélisation a été réalisée avec le modèle de crédit partiel comme précisé plus haut. L'échantillon total des 160 étudiants et des 135 items a été utilisé pour cette modélisation. Nous tenons à préciser que la modélisation n'a pas été réalisée pour chacune des dimensions de ce test de concordance de script ni pour les cinq cohortes d'étudiants séparément.

4.3.1.1 Qualité d'ajustement des sujets au modèle avec M1

Nous nous intéressons en tout premier lieu aux valeurs des différents indices pour les étudiants. Ces valeurs sont reportées dans le tableau 21 ci-dessous. Ainsi nous avons une moyenne de 1,02 ± 0,12 et 1,09 ± 0,57 respectivement pour le carré moyen *infit* et le carré moyen *outfit*. Pour la version standardisée de l'indice *infit*, la valeur moyenne est de 0,20 ± 1,10 tandis qu'elle est de 0,40 ± 1,50 pour la version standardisée de l'indice *outfit*.

Tableau 21 *Indices d'ajustement des 160 sujets évalués avec M1*

	Sujet *Infit*		Sujet *Outfit*	
	CM*	STD**	CM	STD
Moyenne	1,02	0,20	1,09	0,40
Écart type	0,12	1,10	0,57	1,50
Minimum	0,75	-2,00	0,62	**-2,20**
Maximum	1,52[a]	**5,00**	**7,53**	**9,90**
Indice de séparation = 1,69 ; indice de fidélité = 0,74				

Note. *CM : indice carré moyen ; **STD : version standardisée. [a] toute valeur supérieure aux critères d'évaluation est en surgras.

Des valeurs maximales extrêmes très importantes sont obtenues pour les deux types de statistiques étudiés. Ceci a justifié d'aller voir de plus près ces sujets. Nous avons trouvé des valeurs négatives pour la version standardisée à la fois pour le *infit* et le *outfit*. Selon les recommandations de Linacre (2015), nous nous sommes attardés sur ces

valeurs négatives. Nous avons ainsi procédé à des retraits d'items et/ou de sujets et de nouveau nous avons modélisé pour voir le comportement des données face au modèle. Aucune amélioration des valeurs des statistiques d'ajustement n'a été observée suite à ces retraits itératifs et nous avons jugé bon d'arrêter ce processus.

Dans le tableau 22 ci-dessous sont reportés les 10 premiers sujets ayant un mauvais ajustement en considérant la version standardisée de l'indice *outfit*. Quatorze sujets (soit 8,75%) ont une valeur de la version standardisée de l'indice *outfit* et/ou de l'indice *infit* non comprise dans l'intervalle [-2, 2].

Tableau 22 *Ajustement des sujets selon les valeurs extrêmes positives de la version STD de l'indice outfit avec M1*

Sujet	*Infit* CM	*Infit* STD	*Outfit* CM	*Outfit* STD
158	1,22	2,20	7,53	9,90
111	1,33	3,50	1,99	6,20
44	1,03	0,30	2,24	5,10
84	1,52	5,00	1,70	4,70
115	1,15	1,10	1,86	3,40
20	1,04	0,30	1,81	3,10
109	1,34	3,60	1,44	2,90
79	1,24	2,00	1,55	2,60
2	1,30	3,20	1,32	2,40
14	1,22	1,60	1,54	2,30

En considérant toujours la version standardisée de l'indice *outfit*, en référence au tableau *misfit order* produit par le logiciel Winsteps, nous n'avions qu'un seul sujet qui avait une valeur négative inférieure ou égale à -2,00.

4.3.1.2 Qualité d'ajustement des items au modèle avec M1

Nous présentons dans ce paragraphe les indices d'ajustement pour les items au nombre de 135. Des valeurs maximales extrêmes très importantes sont obtenues pour les statistiques étudiées. Le tableau 23 ci-dessous résume les valeurs des différents indices obtenues.

Tableau 23 *Indices d'ajustement des 135 items avec M1*

	Item *Infit*		Item *Outfit*	
	CM*	STD**	CM	STD
Moyenne	1,12	0,00	1,09	-0,10
Écart type	0,62	3,80	0,55	3,70
Minimum	0,34	**-9,90** [a]	0,34	**-9,90**
Maximum	**6,20**	**9,60**	**5,24**	**9,50**

Indice de séparation = 2,51 ; indice de fidélité = 0,86

Note. *CM : indice carré moyen ; **STD : version standardisée. [a] toute valeur supérieure aux critères d'évaluation est en surgras.

L'évaluation des statistiques *infit* pour les items révèle que la valeur moyenne pour le carré moyen est de 1,12 ± 0,62. En ce qui concerne les statistiques *outfit*, le carré moyen est en moyenne de 1,09 ± 0,55. Toutefois, on retrouve des valeurs maximales dépassant le seuil de 1,50 pour le carré moyen intéressant à la fois les statistiques *infit* et *outfit*. Quant à la valeur standardisée des indices d'ajustement, la valeur moyenne pour la statistique *infit* est de 0,00 ± 3,80 tandis qu'elle est de -0,10 ± 3,70 pour la statistique *outfit*. Les statistiques *infit* et *outfit* pour ces indices montrent aussi des valeurs maximales non comprises dans l'intervalle souhaité. Comme le dit Linacre (2015), une valeur élevée représente une plus grande menace à la validité qu'une valeur très faible. Nous avons aussi des valeurs minimales négatives. Ceci nous a poussés à regarder de près ces items de façon individuelle. Nous avons trouvé 72 items (soit 53,30%) dont les valeurs de la version standardisée de l'indice *outfit* ne sont pas comprises dans l'intervalle [-2, +2]. Dans les tableaux 24 et 25 nous avons reporté les dix premiers items à avoir soit une valeur inférieure ou égale à -2 soit supérieure ou égale à +2. Ces items ont été retirés et nous avons repris la modélisation. Les résultats n'étaient pas meilleurs avec ces retraits itératifs ainsi avons-nous mis fin à ce processus qui n'amenait aucune information utile et pertinente dans le cadre de nos analyses.

Tableau 24 *Ajustement des items selon les valeurs extrêmes positives de la version STD de l'outfit avec M1*

Item	*Infit* CM	*Infit* STD	*Outfit* CM	*Outfit* STD
93	1,69	9,60	1,70	9,50
37	1,64	8,90	1,63	8,60
49	1,63	8,20	1,64	8,10
107	1,59	7,40	1,59	7,20
123	1,55	6,30	1,56	6,10
28	1,49	6,20	1,49	6,10
34	1,61	6,00	1,61	5,70
32	1,74	5,70	1,77	5,50
88	1,38	5,70	1,38	5,50
89	1,36	5,50	1,37	5,50

Tableau 25 *Ajustement des items selon les valeurs extrêmes négatives de la version STD de l'outfit avec M1*

Item	*Infit* CM	*Infit* STD	*Outfit* CM	*Outfit* STD
133	0,46	-9,90	0,47	-9,90
62	0,41	-9,90	0,41	-9,90
14	0,37	-9,90	0,37	-9,90
74	0,34	-7,60	0,36	-6,80
90	0,39	-7,10	0,38	-6,70
1	0,49	-6,80	0,48	-6,50
54	0,56	-6,50	0,55	-6,40
106	0,57	-6,50	0,57	-6,30
30	0,52	-6,80	0,54	-6,20
118	0,44	-6,40	0,45	-5,80

Quant aux items avec un bon ajustement au modèle de crédit partiel, nous n'avons retrouvé que 63 items avec ajustement acceptable soit 46,67% des items. Par contre, 72 items (53,33%) ont un mauvais ajustement.

Nous nous penchons dans la prochaine section sur l'indépendance locale des items de cet instrument de mesure, un des postulats fondamentaux de la modélisation de Rasch.

4.3.1.3 Indépendance locale avec M1

L'analyse des résidus standardisés est capitale pour évaluer l'indépendance locale des items. Pour évoquer le problème de dépendance locale, nous rappelons qu'il faut avoir une corrélation positive entre deux items d'au moins 0,70. Selon Linacre (2015), on a un indice de faible dépendance inter-items pour toute corrélation de 0,40. Ceci veut dire qu'après extraction du facteur Rasch nous avons cherché s'il y avait des corrélations inter-items affichant une valeur d'au moins 0,40. Nous avons trouvé la paire d'items 110-135 qui affiche une corrélation de plus de 0,40 ; la paire d'items 5-82 a une corrélation de 0,40.

Quant aux corrélations négatives, nous notons qu'il y a la paire d'items 14-60 avec une valeur de -0,32. Les paires d'items 35-106 et 3-51 ont une valeur de -0,28. De même en nous référant au tableau des plus grandes corrélations standardisées utilisées pour identifier la dépendance entre les sujets, nous avons 0,50 comme la plus grande valeur ; ce sont les sujets 10 et 131 qui sont concernés. Nous n'avons pas eu de corrélations négatives.

Dans la prochaine section nous décrivons l'analyse de la dimensionnalité du test de concordance de script à l'étude au regard de la méthode 1.

4.3.1.4 L'unidimensionnalité avec M1

Nous avons procédé comme Dionne *et al*. (2017), Grondin *et al*. (2017) pour examiner ce postulat en nous basant sur la variance inexpliquée associée aux différents

contrastes de l'analyse en composantes principales effectuée sur les résidus standardisés, c'est-à-dire une fois le « facteur Rasch » extrait (Linacre, 2015).

Nous voulions savoir si les scores obtenus par les répondants étaient sous la seule influence d'un trait latent unique : le raisonnement clinique qui serait mesuré par un item ou bien endossé par un répondant. Pour ce, nous vous référons au tableau 26 dans lequel nous avons reporté les différentes valeurs obtenues pour la variance selon la méthode 1.

Tableau 26 *Résultats de l'analyse en composantes principales des résidus standardisés avec M1*

Variance expliquée et valeurs propres		Pourcentage de variance inexpliquée (variance propre associée)	
Mesure	16,30% (17,10)	Contraste 1	3,70% (4,40)
Répondants	1,40% (1,50)	Contraste 2	2,60% (3,10)
Items	14,90% (15,70)	Contraste 3	2,30% (2,80)
Inexpliquée	83,70% (82,90)	Contraste 4	2,00% (2,40)
		Contraste 5	2,00% (2,40)

Ainsi nous voyons que pour la méthode 1, la variance expliquée par la mesure est de 16,30% ; la variance pour les répondants est de 1,40% et pour les items, nous avons eu 14,90%. De plus la variance inexpliquée associée aux cinq contrastes est supérieure à 2. Les deux premiers contrastes ont d'ailleurs des valeurs propres supérieures à 3. Nous savons qu'un regroupement de 2 items peut être évoqué dès qu'il y a une valeur propre de 2 ainsi il y a potentiellement deux sous-dimensions au moins mesurées par les items. La qualité de l'échelle de mesure est abordée dans la section suivante.

4.3.1.5 Qualité de l'échelle de mesure avec M1

Avec la représentation schématique de Wright (Figure 4 ci-dessous), nous pouvons visualiser la distribution des items et des sujets au niveau de la même échelle de mesure. Ainsi nous apprécions la difficulté des items et l'habileté des sujets à savoir leur raisonnement clinique. Nous voyons que les sujets et les items sont agglutinés à un seul niveau; il n'y a pas d'alignement entre eux pour la méthode des scores combinés.

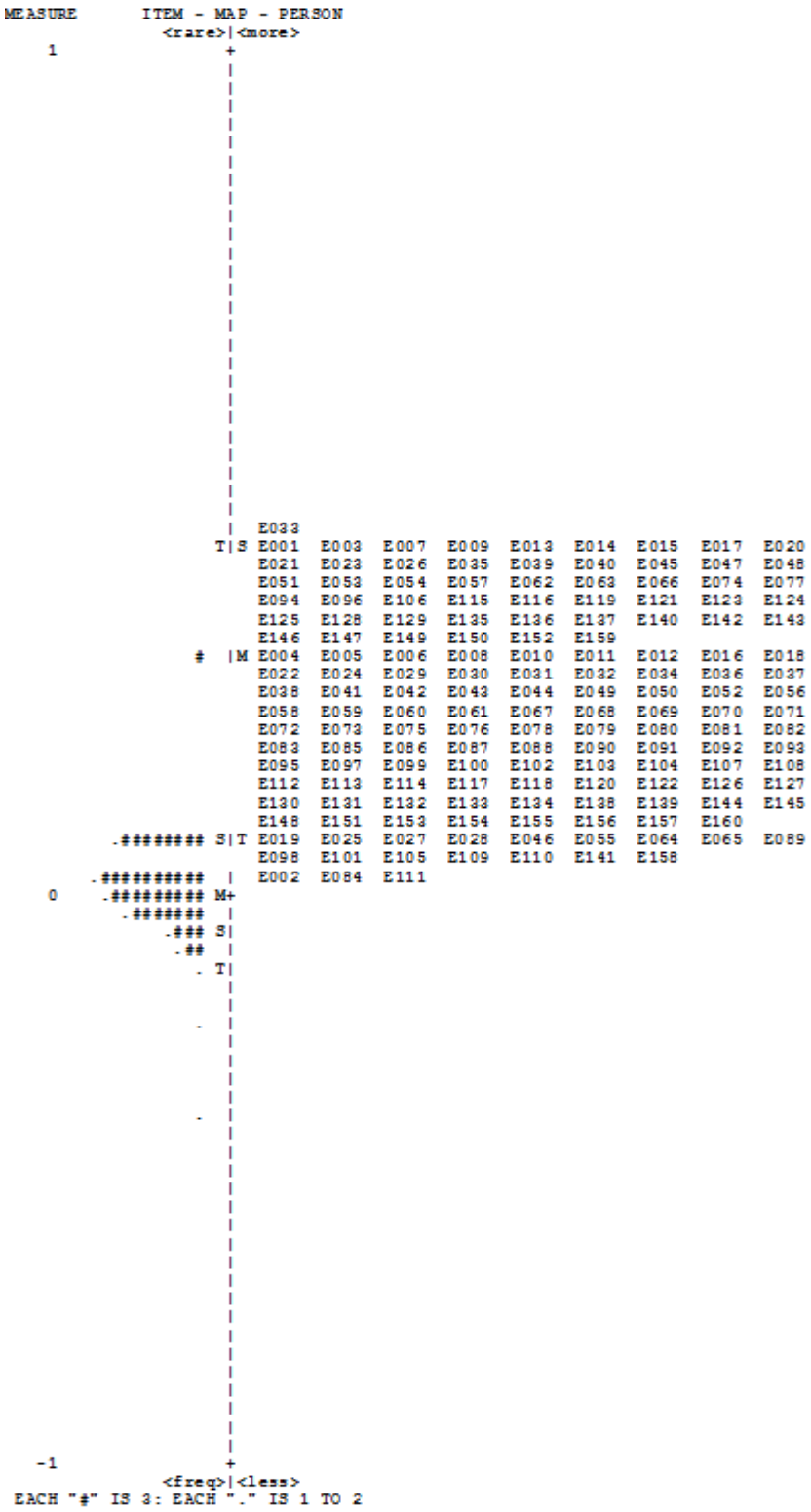

Figure 4. Position des sujets et des items sur l'échelle de mesure avec la méthode 1.

4.3.1.6 Fidélité avec M1

Nous nous intéressons d'abord aux valeurs de la fidélité pour les sujets ; l'indice de séparation est de 1,69. Pour la fidélité (*reliability*), la valeur est de 0,74. Quant aux items, l'indice de séparation est de 2,51 et la fidélité est de 0,86. Ainsi nous avons des valeurs faibles de fidélité relatives aux sujets et aux items.

Au vu des résultats obtenus après cette première modélisation, nous avons procédé au retrait itératif d'items et/ou au retrait de sujets en vue d'améliorer les résultats. Autrement dit nous étions en quête du meilleur ajustement des données au modèle de crédit partiel. Nous présentons les résultats dans la section suivante.

4.3.1.7 Processus de retrait et de modélisation Rasch avec M1

Nous avons consulté le tableau de mauvais ajustement des items et des sujets produit par le logiciel Winsteps et nous avons, au prime abord, retiré les sujets dont les valeurs excédaient l'intervalle d'acceptabilité retenu en considérant la version standardisée de l'indice *outfit*. Ceci dans le but d'améliorer l'ajustement des données du test de concordance de script au modèle choisi toujours en référence à la méthode des scores combinés. Ainsi avons-nous réalisé cinq retraits de façon itérative de sujets et/ou d'items à partir de la base de données Excel. Par la suite nous avons transféré comme précédemment décrit les scores bruts dans Winsteps. Nous résumons dans le tableau 27 ci-dessous les valeurs des indices d'ajustement des sujets obtenues après ces différents retraits. Par exemple, pour la première ligne et la première colonne, [160/135] indique que le nombre total des sujets est 160 et celui des items est 135. Les résultats obtenus avec la première modélisation sont reportés. Par contre à la deuxième ligne, nous avons procédé au retrait de 8 sujets soit 5% du nombre total des sujets et l'intégralité du nombre d'items a été respectée soit zéro item retiré. Nous avons eu comme au premier essai des valeurs extrêmes de la version standardisée et du carré moyen des indices *outfit* et *infit* toutefois moins élevées qu'avant. Nous pouvons voir qu'avec le retrait d'au moins 40% des items et d'au moins 13% des sujets [139/75] nous obtenons une étendue de la version standardisée de l'indice *outfit* de [-1,70 2,20]. Considérons maintenant l'indice de séparation. La valeur est comprise entre 1,51 et 1,56 à peine acceptable selon Boone, Staver et Yale (2014).

Tableau 27 *Étendues des indices d'ajustement des sujets obtenus après retrait itératif avec M1*

Sujets /Items	Étendue *Infit* CM	Étendue *Outfit* CM	Étendue *Infit* STD	Étendue *Outfit* STD	Indice de Séparation	Indice de Fidélité
160/135	[0,75 1,52]	[0,62 - 7,53]	[-2,00 5,00]	[-2,20 9,90]	1,69	0,74
			Après retrait			
152/135	[0,75 1,38]	[0,65 1,54]	[-2,70 4,10]	[-2,60 3,40]	1,56	0,71
152/98	[0,75 1,32]	[0,61 1,83]	[-1,9 3,00]	[-1,90 2,50]	1,57	0,71
152/81	[0,70 1,33]	[0,61 1,94]	[-2,6 2,90]	[-2,00 3,30]	1,51	0,69
142/75	[0,77 1,30]	[0,64 3,00]	[-1,60 1,90]	[-1,50 5,20]	1,51	0,69
139/75	[0,77 1,29]	[0,64 1,41]	[-1,60 1,90]	[-1,70 2,20]	1,51	0,70

En ce qui a trait aux indices d'ajustement des items suite au retrait itératif de sujets et/ou d'items, nous avons obtenu des valeurs hors de l'intervalle retenu pour la version standardisée de l'*outfit*. Par contre, les valeurs de l'indice de séparation pour chacun des retraits effectués sont supérieures à 2,5. Dans le tableau 28 ci-dessous, nous avons rapporté les indices d'ajustement des items relatifs aux 5 retraits itératifs précédents.

Tableau 28 *Étendues des indices d'ajustement des items obtenus après retrait itératif avec M1*

Sujets /Items	Étendue *Infit* CM	Étendue *Outfit* CM	Étendue *Infit* STD	Étendue *Outfit* STD	Indice de Séparation	Indice de Fidélité
160/135 [16]	[0,34 6,20]	[0,34 5,24]	[-9,90 9,60]	[-9,90 9,50]	2,51	0,86
			Après retrait			
152/135	[0,28 2,12]	[0,27 2,28]	[-9,90 9,90]	[-9,90 9,70]	4,35	0,95
152/98	[0,34 2,13]	[0,37 2,48]	[-8,00 4,10]	[-7,20 4,00]	3,86	0,94
152/81	[0,31 1,97]	[0,36 2,34]	[-3,80 2,90]	[-3,70 2,90]	3,76	0,93
142/75	[0,60 1,93]	[0,53 2,17]	[-2,80 2,80]	[-3,30 2,90]	3,30	0,92
139/75	[0,60 1,61]	[0,53 1,59]	[-2,90 2,70]	[-3,30 2,80]	3,54	0,93

Nous avons jugé que l'ajustement ne s'était point amélioré et nous avons décidé de poursuivre le retrait. Nous avons alors constaté au fil de ce processus que nous obtenions encore des valeurs aberrantes avec les indices d'ajustement. Les items et les sujets étaient toujours regroupés dans une partie de l'échelle ceci signifiait qu'il n'y avait pas du tout une bonne étendue. La discussion de ces résultats est faite dans le chapitre V. Le retrait itératif a été qualifié de non efficace et nous avons arrêté ce processus.

Nous allons maintenant présenter la modélisation Rasch des scores bruts obtenus avec la méthode des scores combinés avec pénalité de distance, la M2. Les mêmes démarches ont été faites pour évaluer l'ajustement des données au modèle, la distribution des items et des sujets avec la représentation graphique de Wright de même que la

[16] La description des lignes du tableau a été faite juste dans le paragraphe précédent.

vérification des deux postulats principaux de la modélisation à savoir l'indépendance locale et l'unidimensionnalité.

4.3.2 La modélisation de Rasch et la méthode des scores combinés avec pénalité de distance

Nous rappelons que la méthode des scores combinés avec pénalité de distance est dénommée M2. Nous avons procédé comme pour la méthode 1 en ce qui a trait aux données manquantes ; elles ont été codées 999. L'analyse a été lancée avec 160 sujets et 135 items donc l'ensemble du test. Nous présentons d'abord les statistiques des sujets puis celles des items.

4.3.2.1 Qualité d'ajustement des sujets au modèle avec M2

En considérant la statistique d'ajustement *infit*, la moyenne de l'indice carré moyen est de 1 ± 0,15. La moyenne de la version standardisée est de 0,00 ± 1,50. Pour la statistique d'ajustement *outfit*, la moyenne de l'indice carré moyen est de 0,99 ± 0,28 ; celle de la version standardisée est de -0,10 ± 1,40. Le tableau 29 ci-dessous regroupe ces résultats. Il faut aussi regarder les valeurs minimales et maximales des indices. Nous avons relevé des extrêmes que nous avons mis en surgras.

Tableau 29 *Indices d'ajustement des 160 sujets avec M2*

	Sujets *Infit*		Sujets *Outfit*	
	CM*	STD**	CM	STD
Moyenne	1,00	0,00	0,99	-0,10
Écart type	0,15	1,40	0,18	1,40
Minimum	0,74	**-2,60** [a]	0,68	**-2,70**
Maximum	1,75	**5,80**	**1,87**	**6,40**

Indice de séparation = 1,68 ; indice de fidélité = 0,74

*CM : indice carré moyen ; **STD : version standardisée. [a] la valeur supérieure aux critères d'évaluation est en surgras.

Nous pouvons voir qu'il n'y a pas de valeur minimale extrême pour le carré moyen. Les valeurs extrêmes concernent la version standardisée des indices d'ajustement : l'*infit* et l'*outfit* dont l'intervalle d'acceptabilité est de [-1,90 +1,90].

Le logiciel Winsteps produit un tableau de *misfit* pour les sujets et les items. Il est bien spécifié aussi dans ce tableau que les sujets ou les items ayant le meilleur ajustement sont omis ou encore non listés. En ce qui concerne la méthode 2 ou encore la méthode des scores combinés avec pénalité de distance, nous dénombrons 56 sujets (35%) pour lesquels il y a des valeurs hors de l'intervalle retenu ce qui permet de déduire qu'il y a 104 sujets à avoir eu le meilleur ajustement. Dans les deux tableaux 30 et 31 ci-dessous nous présentons les 10 sujets concernés ayant eu le plus grand *misfit* (valeurs négatives et positives).

Tableau 30 *Ajustement des sujets selon extrêmes positives de la version STD de l'outfit avec M2*

Sujet	Infit CM	Infit STD	Outfit CM	Outfit STD
84	1,75	5,80	1,87	6,40
111	1,36	3,30	1,40	3,40
22	1,39	3,50	1,39	3,10
25	1,38	3,50	1,36	3,00
14	1,19	1,60	1,41	2,60
89	1,29	2,80	1,28	2,50
16	1,20	1,80	1,31	2,40
101	1,26	2,40	1,26	2,30
145	1,07	0,70	1,31	2,30
36	1,34	3,10	1,26	2,10

Tableau 31 *Ajustement des sujets selon les valeurs extrêmes négatives de la version STD de l'outfit avec M2*

Sujet	Infit CM	Infit STD	Outfit CM	Outfit STD
5	0,75	-2,50	0,69	-2,70
149	0,74	-2,60	0,68	-2,70
122	0,80	-2,00	0,74	-2,20
88	0,81	-1,90	0,77	-2,10
99	0,76	-2,40	0,75	-2,10
129	0,81	-1,80	0,73	-2,10
142	0,80	-2,00	0,74	-2,10
26	0,77	-2,20	0,73	-2,10

L'ajustement des sujets a été évalué au préalable et les résultats sont exposés ci-dessus. Nous présentons les paramètres statistiques à la section 4.3.2.2.

4.3.2.2 Qualité d'ajustement des items au modèle avec M2

Nous commençons avec l'indice d'ajustement *infit*. Nous voyons que la moyenne de l'indice carré moyen est de 1,00 ± 0,25 ; celle pour la version standardisée est de -0,10 ± 2,70. Considérons maintenant l'indice d'ajustement *outfit* : la moyenne pour l'indice carré moyen est de 0,99 ± 0,24 ; par contre, elle est de -0,10 ± 2,70 pour la version standardisée. Nous remarquons que les valeurs minimales et maximales pour les deux statistiques d'ajustement à savoir *infit* et *outfit* sont également hors des intervalles d'acceptabilité. Nous vous référons au tableau 32 ci-dessous pour les différentes statistiques d'ajustement pour les items.

Tableau 32 *Indices d'ajustement des 135 items avec M2*

	Item *Infit*		Item *outfit*	
	CM*	STD**	CM	STD
Moyenne	1,00	-0,10	0,99	-0,10
Écart type	0,25	2,70	0,24	2,70
Minimum	**0,45**	**-8,50** [a]	**0,44**	**-8,30**
Maximum	**1,70**	**7,50**	**1,70**	**7,50**
Indice de séparation = 5,46 ; indice de fidélité = 0,97				

*CM : indice carré moyen ; **STD : version standardisé. [a] la valeur supérieure aux critères d'évaluation est en surgras.

Nous avons consulté le tableau de *misfit order* pour les items également produit par le logiciel Winsteps. Nous dénombrons 75 items avec misfit (55,60%). Ainsi nous déduisons que 60 items sur 135 (44,40%) ont eu le meilleur ajustement. Dans les deux tableaux 33 et 34 ci-dessous nous présentons les 10 items à avoir le plus grand *misfit* (valeurs positives) et les 10 autres ayant les valeurs extrêmes négatives.

Tableau 33 *Ajustement des items selon les valeurs extrêmes positives de la version STD de l'outfit avec M2*

Item	*Infit* CM	*Infit* STD	*Outfit* CM	*Outfit* STD
19	1,70	7,50	1,70	7,50
58	1,58	6,10	1,59	6,20
47	1,55	6,30	1,54	6,30
76	1,46	5,50	1,46	5,50
124	1,43	5,00	1,43	5,00
75	1,39	4,70	1,39	4,70
132	1,37	4,40	1,37	4,40
81	1,36	4,10	1,34	3,80
22	1,32	3,50	1,32	3,50
135	1,31	3,40	1,30	3,20

Et dans le tableau 34 suivant nous présentons les items ayant les valeurs les plus négatives en considérant la version standardisée des statistiques d'ajustement.

Tableau 34 *Ajustement des items selon les valeurs extrêmes négatives de la version STD de l'outfit avec M2*

Item	*Infit* CM	*Infit* ZSTD	*Outfit* CM	*Outfit* ZSTD
74	0,47	-8,50	0,48	-8,30
48	0,49	-8,10	0,50	-7,80
62	0,45	-6,70	0,44	-6,70
111	0,56	-6,20	0,58	-5,80
6	0,54	-5,30	0,52	-5,30
1	0,53	-4,10	0,50	-4,20
112	0,51	-5,10	0,54	-4,60
67	0,68	-4,10	0,68	-4,00
24	0,63	-4,00	0,63	-4,00
8	0,69	-3,60	0,68	-3,60

Dans la section suivante, nous traitons l'indépendance locale des items de cet instrument de mesure au regard de la méthode des scores combinés avec pénalité de distance.

4.3.2.3 Indépendance locale avec M2

L'indépendance locale est un postulat fondamental de la modélisation comme expliqué plus haut. Les réponses des individus doivent être indépendantes statistiquement les unes des autres. L'étude de la dépendance locale entre deux items se fait sur la matrice des résidus standardisés, c'est-à-dire après extraction du facteur Rasch.

Nous rappelons qu'il a été établi que la corrélation inter-items doit être positive et d'une valeur d'au moins 0,70 afin d'évoquer une dépendance locale entre les items (Linacre, 2015). Toutes les valeurs de la corrélation inter-items avec la méthode 2 (M2) sont inférieures à 0,40. Quant aux corrélations négatives, nous notons qu'il y a la paire

43-62 à avoir une valeur de -0,32. Les paires 23-34, 23-113 et 4-128 ont une valeur de -0,31. Nous avons aussi des valeurs négatives de -0,29 et -0,28. En consultant le tableau des plus grandes corrélations standardisées produit par le logiciel Winsteps, nous avons 0,38 comme la plus grande valeur ; ce sont les sujets 3 et 77 qui sont concernés. Nous avons eu deux valeurs de corrélation négative : -0,34 pour les sujets 107 et 135 ; -0,33 pour les sujets 50 et 80.

Dans la prochaine section nous décrivons les résultats relatifs à l'unidimensionnalité du test de concordance de script toujours au regard de la méthode 2.

4.3.2.4 Unidimensionnalité avec M2

L'unidimensionnalité du test de concordance de script est appréciée à partir des valeurs de la variance inexpliquée associée aux différents contrastes de l'analyse en composantes principales effectuée sur les résidus standardisés, c'est-à-dire après l'extraction du « facteur Rasch » (Linacre, 2015). Les auteurs comme Grondin *et al.*, Dionne *et al.* en 2017 ont procédé ainsi en suivant les recommandations de Linacre. Nous vous référons au tableau 35 ci-dessous qui met en évidence ces valeurs. La variance expliquée par la mesure est de 22,50%. Si nous nous référons à Linacre (2015), la variance expliquée est très faible. Nous notons que les répondants ont la plus faible proportion de la variance 1,20% et pour les items la variance est de 21,20%.

Tableau 35 *Résultats de l'analyse en composantes principales des résidus standardisés avec M2*

Variance expliquée et valeurs propres		Pourcentage de variance inexpliquée (variance propre associée)	
Mesure	22,50% (22,70)	Contraste 1	3,20% (4,10)
Répondants	1,20% (1,20)	Contraste 2	2,80% (3,60)
Items	21,20% (21,40)	Contraste 3	2,30% (3,00)
Inexpliquée	77% (77.30)	Contraste 4	2,10% (2,70)
		Contraste 5	1,90% (2,40)

La variance inexpliquée associée aux cinq contrastes est supérieure à 2. Les trois premiers contrastes ont d'ailleurs des valeurs propres supérieures ou égales à 3. Nous savons qu'un regroupement de 2 items peut être évoqué dès qu'il y a une valeur propre de 2 ainsi il y a potentiellement deux sous-dimensions au moins mesurées par les items.

Qu'en est-il de l'échelle de mesure des items et des sujets au regard de la méthode des scores combinés avec pénalité de distance ?

4.3.2.5 Qualité de l'échelle de mesure avec M2

La représentation graphique de Wright permet de visualiser la distribution des items et des sujets au niveau de la même échelle de mesure au regard de la méthode de détermination des scores avec pénalité de distance, la M2. La Figure 5 ci-dessous illustre pour nous cette échelle. Il n'y a pas une bonne étendue des sujets. En effet, ces derniers restent agglutinés entre 0 et +1 logit.

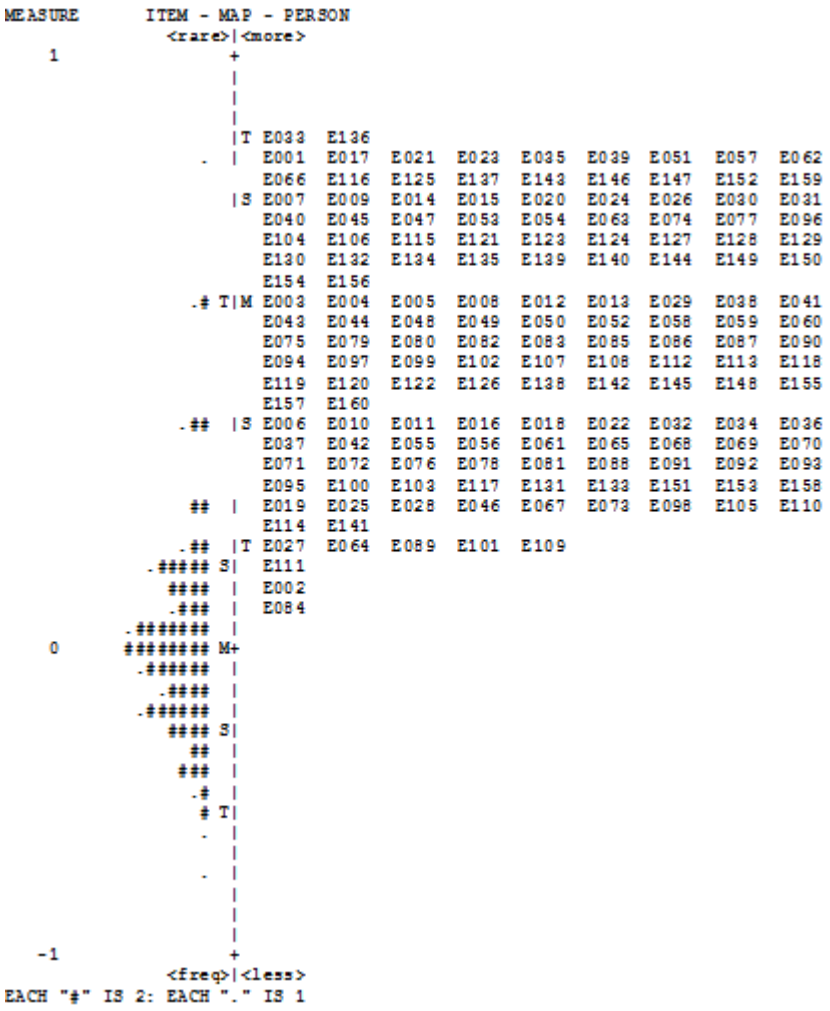

Figure 5. Position des répondants et des items sur l'échelle de mesure avec M2.

4.3.2.6 Fidélité avec M2

Pour les 160 étudiants ayant passé ce TCS, l'indice de séparation est de 1,68. Pour la fidélité (*reliability*), la valeur est de 0,74. Quant aux 135 items, nous avons trouvé un indice de séparation est de 5,46 et la fidélité est de 0,97. Nous reviendrons sur ces valeurs dans le prochain chapitre, la discussion.

Au vu des résultats obtenus avec la modélisation, nous avons procédé à un retrait itératif de sujets et/ou d'items et nous avons relancé la modélisation dans la section suivante. En effet, les valeurs extrêmes font beaucoup de bruit pour répéter Linacre (2004) et Tennant et Conaghan (2007). Leur importance influence la qualité de

l'ajustement du test au modèle de Rasch utilisé dans le cadre de cette étude, le modèle à crédit partiel.

4.3.2.7 Processus de retrait d'items et de sujets avec M2

En vue de chercher le meilleur ajustement des données au modèle retenu, nous avons procédé à des retraits itératifs d'items et/ou de sujets de la manière suivante. La liste de ces derniers est dans les annexes de notre travail. Les tableaux de mauvais ajustement ont été bien examinés et nous avons enlevé d'abord les valeurs excédant l'intervalle de valeurs normales à partir de la base de données Excel. Par la suite nous avons transféré les scores bruts dans Winsteps. Nous résumons dans le tableau 36 ci-dessous les valeurs des indices d'ajustement des sujets obtenues avec ces différents retraits. De plus aucun indice de séparation avec ces trois retraits itératifs n'a une valeur d'au moins 2,00, valeur considérée comme bonne (Boone, Staver et Yale, 2014). L'indice de fidélité reste inférieur à 0,90.

Tableau 36 *Étendues des indices d'ajustement des sujets obtenus après retrait itératif avec M2*

Sujets /Items	Étendue *Infit* CM	Étendue *Outfit* CM	Étendue *Infit* STD	Étendue *Outfit* STD	Indice de Séparation	Indice de Fidélité
160/135	[0,74 1,75]	[0,68 1,87]	[-2,60 5,80]	[-2,70 6,40]	1,68	0,74
			Après retrait			
150/112	[0,69 1,30]	[0,72 1,37]	[-2,30 2,30]	[-2,10 2,40]	1,40	0,66
139/135	[0,74 1,28]	[0,72 1,36]	[-2,20 2,20]	[-2,00 2,50]	1,25	0,61
139/74	[0,62 1,44	[0,61 1,64]	[-2,60 2,40]	[-2,00 1,70]	0,82	0,40

Attardons-nous maintenant sur les indices d'ajustement des items après retrait avec la méthode des scores combinés avec pénalité de distance. Pour tous ces retraits, l'indice de séparation a une valeur supérieure à 3 ce qui est souhaitable selon Boone *et al*. (2014). L'indice de fidélité (*reliability*) est supérieur à 0,90. Ces valeurs sont présentées dans le tableau 37 ci-dessous.

Tableau 37 *Étendues des indices d'ajustement des items obtenus après retrait itératif avec M2*

Sujets /Items	Étendue Infit CM	Étendue Outfit CM	Étendue Infit STD	Étendue Outfit STD	Indice de Séparation	Indice de Fidélité
160/135	[0,45 1,70]	[0,44 1,70]	[-8,50 7,50]	[-8,30 7,50]	5,46	0,97
			Après retrait			
150/112	[0,41 1,51]	[0,43 1,50]	[-7,30 4,60]	[-7,00 4,60]	5,36	0,97
139/135	[0,39 1,86]	[0,39 1,86]	[-7,70 6,90]	[-7,70 6,90]	5,15	0,96
139/74	[0,66 1,57]	[0,68 1,47]	[-2,90 2,30]	[-3,00 1,90]	5,10	0,96

Après avoir exposé les résultats de la modélisation avec la méthode de Charlin et la méthode des scores combinés avec pénalité de distance, il reste à traiter les résultats de la modélisation obtenus avec la méthode selon une bonne réponse, la M3. Ces derniers sont donc présentés dans la section suivante.

4.3.3 La modélisation de Rasch et la méthode selon une bonne réponse

Avec la méthode 3, les scores sont dichotomiques. Aucun codage particulier n'a été nécessaire. Les valeurs manquantes ont été remplacées comme pour les autres méthodes par 999 pour faciliter l'analyse avec le logiciel Winsteps. Voici donc les indices d'ajustement obtenus pour l'ensemble des sujets et des items du TCS.

4.3.3.1 Qualité d'ajustement des sujets avec M3

Pour la statistique *infit*, la moyenne de l'indice carré moyen est de 1,00 ± 0,15 tandis qu'elle est de 0,99 ± 0,28 pour la statistique *outfit*. Quant à la version standardisée, la moyenne est de 0,00 ± 1,5 pour la statistique *infi*t et de 0,00 ± 1,20 pour la statistique *outfit*. Ces différentes valeurs sont reportées dans le tableau 38 ci-dessous.

Tableau 38 *Indices d'ajustement des 160 sujets avec M3*

	Sujets *Infit*		Sujets *Outfit*	
	CM*	STD**	CM	STD
Moyenne	1,00	0,00	0,99	0,00
Écart type	0,15	1,50	0,28	1,20
Minimum	0,75	**-2,90** [a]	0,58	**-2,10**
Maximum	1,60	**5,30**	**2,06**	**4,40**
Indice de séparation = 0,66 ; indice de fidélité = 0,31				

*CM : indice carré moyen ; **STD : version standardisée. [a] la valeur supérieure aux critères d'évaluation est en surgras.

Nous avons consulté le tableau de *misfit order* des sujets et ceci a permis de relever les dix premiers sujets à avoir le plus grand *misfit*. A noter que pour la M3, la valeur la plus extrême est de 0,40. Nous vous référons au tableau 39 ci-dessous dans lequel sont reportées ces données.

Tableau 39 *Ajustement des sujets selon les valeurs extrêmes positives de la version STD de l'outfit avec M3*

Sujet	*Infit* CM	*Infit* STD	*Outfit* CM	*Outfit* STD
79	1,29	3,00	2,06	4,40
84	1,60	3,00	1,94	4,10
126	1,39	3,80	1,55	2,60
158	1,17	1,80	1,60	2,60
101	1,14	1,50	1,49	2,50
35	1,27	2,60	1,61	2,40
91	1,02	0,30	1,61	2,40
138	1,14	1,50	1,49	2,30
62	1,36	3,00	1,73	2,20
13	1,07	0,80	1,50	2,00

Nous n'avons retrouvé aucun sujet avec des valeurs extrêmes négatives de la version standardisée eu égard à la méthode 3. Passons à l'ajustement des items dans la prochaine section.

4.3.3.2 Qualité d'ajustement des items avec M3

En ce qui a trait aux items, pour la statistique *infit*, la moyenne de l'indice carré moyen est de 1,00 ± 0,02; cette moyenne est de 0,99 ± 0,07 pour la statistique *outfit*. Considérons maintenant la version standardisée, la moyenne est de 0,10 ± 0,40 pour la statistique *infi*t et elle est de 0,10 ± 0,50 pour la statistique *outfit*. Les valeurs positives extrêmes pour la version standardisée de l'indice *outfit* sont moins élevées avec la méthode M3 (en surgras dans le tableau) et concernent l'item 87 (2,20) ; l'item 122 a une valeur limite égale à 2,00. Aucune valeur négative extrême n'est retrouvée. Les résultats des indices d'ajustement des items sont reportés dans le tableau 40 ci-dessous.

Tableau 40 *Indices d'ajustement des 135 items avec M3*

	Item *Infit*		Item *Outfit*	
	CM*	STD**	CM	STD
Moyenne	1,00	0,10	0,99	0,10
Écart type	0,02	0,40	0,07	0,50
Minimum	0,95	-1,20	0,68	-1,20
Maximum	1,05	**2,10**	**1,23**	**2,20**
Indice de séparation = 5,67 ; indice de fidélité = 0,97				

*CM : indice carré moyen ; **STD : version standardisée. [a] la valeur supérieure aux critères d'évaluation est en surgras.

A la prochaine section nous abordons l'indépendance locale des items de cet instrument de mesure au regard de la méthode 3.

4.3.3.3 Indépendance locale avec M3

A partir de l'analyse des résidus standardisés, nous avons seulement la paire d'items 59-60 qui affiche une corrélation de 0,50 (supérieure à 0,40). Quant aux corrélations négatives, la paire [23-34] a une valeur de -0,42. Huit paires ont une corrélation positive inférieure ou égale à 0,40 et 10 autres ont une valeur négative inférieure à 0,40 [35-106, 43-62, 81-93, 39-95, 40-55, 30-61, 35-62, 23-113, 62-82, 4-128].

Quant aux sujets, en nous référant au tableau des plus grandes corrélations résiduelles standardisées utilisées pour identifier les personnes dépendantes produit par le logiciel Winsteps, la plus forte valeur positive de corrélation résiduelle standardisée (0,47) concerne les sujets 103 et 120. Il y a une seule valeur négative (-0,33) relative aux sujets [120-137, 74-120]. Qu'en est-il du postulat d'unidimensionnalité au regard de la M3 ?

4.3.3.4 Unidimensionnalité avec M3

Nous voulons savoir si les scores obtenus par les répondants étaient sous la seule influence de leur raisonnement clinique ou non. La variance inexpliquée associée aux différents contrastes de l'analyse en composantes principales effectuée sur les résidus

standardisés, c'est-à-dire une fois le « facteur Rasch » extrait (Linacre, 2015) va nous permettre de confirmer ce postulat.

Les différentes valeurs obtenues pour la variance des items sont reportées dans le tableau 41 ci-dessous. La variance expliquée par la mesure est de 27,80% dont 25,20% sont représentés par les items et 2,60% par les sujets. La variance non expliquée est de 72,20%.

Tableau 41 *Résultats de l'analyse en composantes principales des résidus standardisés avec M3*

Variance expliquée et valeurs propres		Pourcentage de variance inexpliquée (variance propre associée)	
Mesure	27,80% (27,90)	Contraste 1	3,50% (4,90)
Sujets	2,60% (2,60)	Contraste 2	2,20% (3,10)
Items	25,20% (25,30)	Contraste 3	2,00% (2,80)
Inexpliquée	72,20% (72,10)	Contraste 4	1,80% (2,50)
		Contraste 5	1,60% (2,20)

La variance inexpliquée associée aux cinq contrastes est supérieure à 2 et pour les deux premiers contrastes les valeurs propres sont supérieures à 3. Que pourra-t-on déduire de ces résultats ? Ceci va être discuté dans le prochain chapitre.

4.3.3.5 Qualité de l'échelle de mesure avec M3

La représentation graphique de Wright est la Figure 6 ci-dessus et elle nous permet de voir la position des sujets et des items sur la même échelle. Les sujets sont regroupés entre -0,53 et +1,68, donc il y a une faible étendue. Quant à la position des items c'est-à-dire la position de leur difficulté, ils s'étalent entre -4,21 et +4,15. Un bon nombre d'items se retrouve dans une partie de l'échelle où nous ne retrouvons aucun sujet.

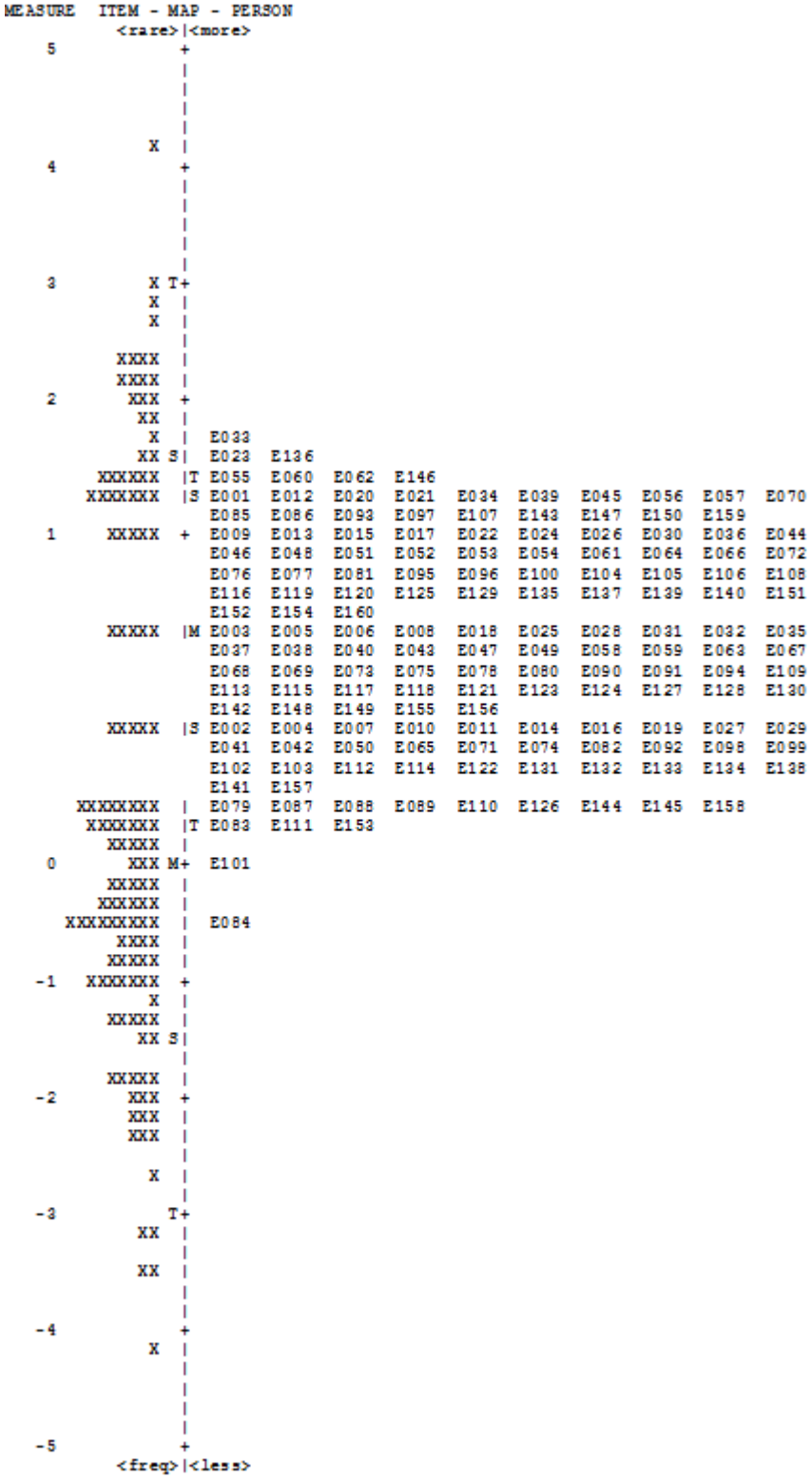

Figure 6. Position des étudiants et des items sur l'échelle de mesure avec M3.

En considérant le tableau de *misfit order* des sujets et des items produit par le logiciel Winsteps (voir annexe), nous avons procédé au retrait itératif d'items et/ou de sujets pour voir comment évolueraient les indices statistiques eu égard à la M3. Nous décrivons dans la section suivante les résultats après ces retraits.

4.3.3.6 Fidélité avec M3

L'indice de séparation des sujets est de 0,66. Pour la fidélité (*reliability*), la valeur est de 0,31. Pour les 135 items, nous avons trouvé un indice de séparation est de 5,67 et la fidélité est de 0,97. Nous reviendrons sur ces valeurs dans le prochain chapitre, la discussion.

Nous avons procédé à un retrait itératif de sujets et/ou d'items et nous avons relancé la modélisation dans la section suivante.

4.3.3.7 Processus de retrait et de modélisation de Rasch avec M3

En annexe, nous présentons les items et les sujets concernés par ces retraits. Voyons maintenant les valeurs des indices d'ajustement obtenues. Nous tenons à préciser d'entrée de jeu que le nombre d'itération est plus faible avec la méthode par consensus comparé aux deux autres méthodes, M1 et M2. Dans le tableau 42 ci-dessous, nous reportons les valeurs des indices d'ajustement des sujets après 2 retraits. Les valeurs de l'indice de séparation des sujets sont nettement inférieures à 1,50. Les valeurs de l'indice de fidélité (*reliability*) sont toutes aussi faibles pour les 2 retraits.

Tableau 42 *Étendues des indices d'ajustement des sujets obtenus après retrait itératif avec M3*

Sujets /Items	Étendues *Infit* CM	Étendues *Outfit* CM	Étendues *Infit* STD	Étendues *Outfit* STD	Indice de Séparation	Indice de Fidélité
160/135	[0,75 1,60]	[0,58 2,06]	[-2,90 5,30]	[-2,10 4,40]	0,66	0,31
			Après retrait			
160/126	[0,74 1,49]	[0,62 1,85]	[-2,90 4,40]	[-2,10 3,50]	0,80	0,39
136/131	[0,77 1,48]	[0,58 1,69	[-2,60 4,40]	[-2,00 2,70]	0,34	0,10

Vu les valeurs obtenues pour ces statistiques, nous avons jugé qu'il n'était pas nécessaire de pousser l'itération plus loin eu égard au nombre de sujets. En effet, le retrait supplémentaire de sujets a donné lieu à des valeurs non comprises dans l'intervalle retenu [-1,9, +1,9] donc les résultats étaient jugés encore plus inacceptables.

Qu'en est-il des indices d'ajustement des items? Nous n'avons pas eu de grande différence dans les valeurs des indices d'ajustement des items après le retrait de 24 sujets ou celui de 9 items. Toutefois la limite supérieure des valeurs obtenues pour la version standardisée est inférieure à 2,9. L'indice de séparation est excellent (supérieur à 3) comme avant le retrait. L'indice de fidélité (*reliability*) est supérieur à 0,90. Ces différentes valeurs sont résumées dans le tableau 43 ci-dessous en estimant les paramètres à partir des scores calculés avec la méthode 3.

Tableau 43 *Étendues des indices d'ajustement des items avec M3*

Sujets /Items	Étendues *Infit* CM	Étendues *Outfit* CM	Étendues *Infit* STD	Étendues *Outfit* STD	Indice de Séparation	Indice de Fidélité
160/135	[0,95 1,05]	[0,68 1,23]	[-1,20 2,10]	[-1,20 2,20]	5,67	0,97
			Après retrait			
160/126	[0,96 1,17]	[0,87 1,16]	[-0,80 2,30]	[-0,90 2,40]	5,41	0,97
136/131	[0,96 1,04]	[0,85 1,15]	[-1,50 1,90]	[-1,50 1,90]	5,43	0,97

La dernière section de ce chapitre est une synthèse des résultats obtenus après la modélisation de Rasch. Par la suite, la discussion est présentée dans le cinquième chapitre.

4.3.4 Synthèse des résultats selon la modélisation de Rasch

Le logiciel Winsteps 3.90.0 a été utilisé pour la modélisation de Rasch et le modèle *Partial Credit* a été ainsi mis à profit dans cette étude. Tout au long de la partie 4.3, nous avons décrit les résultats de cette modélisation obtenus avec les trois méthodes de détermination des scores. Les tableaux suivants reportent les valeurs moyennes des indices d'ajustement, celles des indices de fidélité et de séparation. Nous avons analysé également les postulats fondamentaux de la modélisation de Rasch à savoir l'indépendance locale et l'unidimensionnalité. Dans le tableau 44 ci-dessous, il y a un récapitulatif de ces données.

Tableau 44 *Récapitulatif des indices d'ajustement des items avec les trois méthodes*

Modélisation	Item *Infit*		Item *outfit*	
	CM*	STD**	CM	STD
Méthode 1	1,12 ±0 ,62	0,00 ± 3,80	1,09 ± 0,55	-0,10 ± 3,70
Méthode 2	1,00 ± 0,30	-0,20 ± 2,90	0,99 ± 0,29	-0,30 ± 2,80
Méthode 3	1,00 ± 0,02	0,10 ± 0,40	0,99 ± 0,07	0,10 ± 0,50

En ce qui a trait à la fidélité voici ci-dessous le tableau 45 avec le récapitulatif des valeurs de ces indices.

Tableau 45 *Indices de séparation et de fidélité pour les trois méthodes utilisées*

		Séparation	Fidélité
Sujets	M1	1,69	0,74
	M2	1,62	0,72
	M3	0,73	0,35
Items	M1	2,51	0,86
	M2	5,48	0,97
	M3	5,67	0,97

L'analyse des résidus standardisés a permis d'avoir les corrélations entre les items. Les valeurs sont dans l'ensemble inférieures ou égales à 0,50 et nous avons reporté dans le tableau 46 ci-dessous les paires d'items et leurs valeurs de corrélation pour les trois méthodes de détermination des scores utilisées dans le cadre de ce test de concordance de script.

Tableau 46 *Valeurs corrélationnelles des items basés sur les résidus standardisés*

Méthode 1		Méthode 2		Méthode 3	
Corrélation	Items	Corrélation	Items	Corrélation	Items
0,44	110-135	0,39	59-60	0,50	59-60
0,40	5-82	0,35	54-95	0,40	110-125
0,37	33-98	0,31	68-88 / 7-60 / 4-125	0,38	36-82
0,36	4-81	0,29	27-72 / 74-114 / 110-114	0,37	24-106
0,33	90-118 / 17-98	0,28	87-116	0,36	54-95
0,31	74-114 / 1-124	0,27	74-99	0,35	67-110
0,29	2-43 / 21-38 / 48-68	-	-	0,31	62-65
0,28	33-48 / 21-68 / 90-124 / 48-82	-	-	0,30	61-79 / 55-81
0,27	82-83	-	-	-	-
-0,32	14-60	-0,32	43-62	-0,42	23-34
-0,28	35-106 / 3-51	-0,31	23-113 / 23-34 / 4-128	-0,35	35-106 / 43-62
		-0,29	35-106 / 100-114	-0,32	81-93 / 39-95 /
		-0,28	39-95 / 102-126 /10-91 /14-135	-0,31	40-55 / 30-61
				-0,30	35-62 / 23-113 / 62-82 / 4-128

Le postulat d'unidimensionnalité a été testé et nous présentons le récapitulatif des différentes valeurs de variance expliquée et des valeurs propres pour les trois méthodes de détermination des scores utilisées dans le tableau 47 ci-dessous.

Tableau 47 *Récapitulatif des valeurs de variance pour les trois méthodes*

Variance expliquée et valeurs propres	Méthode 1	Méthode 2	Méthode 3
Mesure	16,30 % (17,10)	22,50 % (22,70)	27,80 % (27,90)
Répondants	1,40 % (1,50)	1,20 % (1,20)	2,60 % (2,60)
Items	14,90 % (15,70)	21,20 % (21,40)	25,20 % (25,30)
Inexpliquée	83,70 % (82,90)	77,00 % (77.30)	72,20 % (72,10)
Pourcentage de variance inexpliquée (variance propre associée)			
Contraste 1	3,70 % (4,40)	3,20 % (4,10)	3,60 % (5,00)
Contraste 2	2,60 % (3,10)	2,80 % (3,60)	2,40 % (3,40)
Contraste 3	2,30 % (2,80)	2,30 % (3,00)	2,00 % (2,80)
Contraste 4	2,00 % (2,40)	2,10 % (2,70)	1,80 % (2,40)
Contraste 5	2,00% (2,40)	1,90% (2,40)	1,70% (2,30)

En résumé nous avons exposé dans le chapitre IV les résultats relatifs à la détermination des scores du test de concordance de script selon la théorie classique des tests. Egalement une modélisation de ces scores a été menée avec le modèle Crédit Partiel en utilisant le logiciel Winsteps et les résultats ont également été exposés. La discussion de ces différents résultats est faite dans le prochain chapitre.

CHAPITRE V DISCUSSION

Nous avons conduit une étude exploratoire descriptive de l'évaluation du raisonnement clinique avec le test de concordance de script par comparaison de trois méthodes de détermination des scores. Nous allons d'abord centrer la discussion autour de notre question de recherche. Puis nous aborderons les limites et la contribution de notre étude. La fin de ce chapitre traitera de quelques recommandations pour les études futures.

5.1 Le TCS et la théorie classique des tests

Nous rappelons la question de recherche : En quoi le choix de la méthode de détermination des scores influence-t-il les propriétés métriques des scores obtenus avec un test de concordance de script au regard de trois méthodes de détermination des scores ? Dans cette section nous discutons des différents résultats obtenus avec la théorie classique des tests au regard des 3 méthodes. Nous avons procédé de la même manière que pour la présentation des résultats. Nous discutons des résultats pour l'ensemble des 160 étudiants puis ceux des différentes cohortes et pour finir les résultats selon les dimensions du test de concordance de script.

5.1.1 Échantillon total des 160 étudiants

Successivement sont discutées la validité du construit et la fidélité. Puis une comparaison entre les méthodes est présentée.

5.1.1.1 Validité de construit

À notre humble avis, les propriétés métriques des scores obtenus sont grandement influencées par la méthode de détermination des scores retenue en nous basant sur les différents résultats obtenus pour notre étude avec la théorie classique des tests. Comme il est rapporté dans les écrits recensés, ce TCS différencie les experts et les étudiants quelle que soit la méthode de détermination des scores considérée. En effet, le score moyen des experts était supérieur à celui des étudiants. Toutefois nous pensons que ceci ne constitue

pas une grande preuve de validité car c'est un résultat hautement prévisible. Aussi nous aurions pu certainement nous attendre à des valeurs plus élevées pour les scores des experts. Cependant il faut noter que le score moyen des experts dans l'ensemble des études recensées concernant le TCS n'atteint jamais les 90% (Lambert, Gagnon, Nguyen et Charlin, 2009 ; Carrière, Gagnon, Charlin, Downing et Bordage, 2009 ; Deschênes, Charlin, Gagnon et Goudreau, 2011 ; Subra *et al*. 2017, Kazour *et al*. 2016, Goos, Shubach, Seifer et Boeker *et al*. 2016 par exemple). Ce score varie en général entre 76-82% ce qui signifierait qu'il y a en moyenne 20% d'incertitude du côté des experts contrairement aux autres examens. Pour les étudiants, les valeurs des scores moyens diffèrent aussi selon la méthode considérée. Le score moyen pour les étudiants est, toutefois, plus élevé avec la méthode 2 (méthode des scores combinés avec pénalité de distance). Ceci a été aussi décrit par Goos *et al*. en 2016. Cent treize étudiants (70,62%) sur les 160 ont un score entre 70 et 79 ; 40 autres (25%) ont un score supérieur ou égal à 80%. En revanche avec la méthode 3 aucun étudiant n'a un score supérieur ou égal à 80% et la grande majorité (135/160) ont un score compris entre 60 et 69%. Ces données nous portent à avancer que la méthode 2 serait la méthode la plus appropriée dans ce contexte. De nombreux auteurs tels Ahmadi *et al.* (2014) ont conclu que le test de concordance de script est un test intéressant pour évaluer la prise de décision en médecine d'urgence. Cependant ils ont soulevé la question de la validité suffisante des jugements du panel d'experts pour constituer la référence standard du test. Les constructeurs du test de concordance de script recommandent que les experts soient des spécialistes dans le domaine étudié et ils doivent être aussi au nombre minimal de 15 pour un TCS à visée certificative ; ceci est retrouvé dans une seule étude. Dans le cadre du TCS utilisé, l'équipe de Liège, Belgique avait dû retirer un expert comme expliquée dans la section 3.2. Avec la M1, au moins un des 12 experts a un score moyen inférieur au score moyen des étudiants. Quant à la méthode des scores combinés avec pénalité de distance, la M2, nous avons fait le même constat qu'un expert a aussi un score moyen plus petit que le score moyen des étudiants. Par contre, l'écart le plus important entre les experts et les étudiants s'observe avec la méthode 3. Les résultats sur le score moyen de notre étude corroborent ceux des écrits. En effet dans toutes ces études, le score moyen des experts était supérieur à celui des apprenants de tout ordre avec des différences statistiquement

significatives (Brailovsky *et al*., 2001 ; Charlin, Brailovsky, Brazeau-Lamontagne *et al*., 1998 ; Dawson *et al*., 2014 ; Marie *et al*., 2005 ; Sibert *et al*., 2002).

5.1.1.2 Fidélité

Pour la cohérence interne de cet instrument de mesure, le coefficient alpha de Cronbach avec la méthode 1 se rapproche des valeurs du coefficient dans les écrits recensés. Par exemple le tableau 2 résumant une dizaine d'écrits recensés montre que le coefficient alpha de Cronbach varie entre 0,61 et 0,90. En effet Lambert et al. (2016) ont eu une valeur de 0,90 pour un TCS constitué de 90 items. Des auteurs avec un nombre d'items se rapprochant de celui du TCS à l'étude (132 ou 137 items) ont obtenu un coefficient de cohérence interne à 0,80 ou 0,85. En ce qui a trait à notre étude, cette valeur signifie qu'il y a une bonne homogénéité à l'intérieur du test de concordance de script à l'étude. Le coefficient alpha de Cronbach est plus faible pour la méthode 2 que celui obtenu avec la M1. Par contre, il est extrêmement faible avec la méthode 3 selon une bonne réponse (méthode dichotomique). Les valeurs du coefficient alpha calculé pour l'ensemble des sujets et des items étaient aberrantes (près de 0 ou carrément négatives ce qui indiquait que plusieurs dimensions étaient présentes.

Aussi nous avons vu qu'avec la méthode M2, la méthode des scores combinés avec pénalité de distance, le score moyen des étudiants était le plus élevé en comparaison à celui des deux autres. Par contre, le coefficient alpha de Cronbach est plus faible que celui obtenu avec la M1.

La méthode selon une bonne réponse, méthode dichotomique (en 3 échelons), est la méthode la plus fréquemment utilisée dans les évaluations classiques telles les questions à choix multiples. Mais il faut bien noter que les résultats sont les plus faibles peu importe le paramètre statistique qu'on voudrait considérer par exemple la robustesse du test, la corrélation item-total ou le score moyen. Le coefficient alpha de Cronbach est très faible et doit pousser à une réflexion plus profonde. Nous ne pouvons pas prétendre que la M3 différencie correctement les participants et nous ne pouvons pas la recommander dans le contexte d'évaluation du raisonnement clinique, selon nous, avec le test de concordance de script. Nous sommes ainsi en désaccord avec Bland *et al*. (2005).

Ces derniers avaient conclu que la méthode 3 avait une bonne fidélité et permettait de bien différencier les participants. Wilson *et al*. (2014) ont, comme nous, trouvé des résultats discordants quant à la supériorité de la méthode 3. Wilson *et al.* ont aussi avancé qu'avec la méthode des scores combinés (M1), il est difficile de différencier les étudiants proches de la réponse modale et ceux qui y étaient éloignés. D'où l'intérêt de la méthode 2, selon ces mêmes auteurs, car elle permet d'évaluer la direction de la réponse donnée par l'étudiant et aussi son effet.

Les analyses faites avec l'échantillon total des étudiants de même que le TCS dans son ensemble nous permettent d'avancer que les méthodes de détermination des scores ont un impact certain sur les propriétés métriques des scores obtenus avec le TCS. En effet nous avons vu que la fidélité diffère selon la méthode utilisée de même que les scores moyens. De plus en considérant les scores selon les dimensions avec chacune des méthodes considérées nous avons obtenu des différences statistiquement significatives au niveau des scores.

Pour bien étayer la réponse à notre question de recherche, nous avons considéré dans la discussion les analyses faites par cohortes pour chacune des méthodes et également selon les dimensions du TCS.

5.1.2. Discussion des résultats obtenus par cohortes d'étudiants et selon les dimensions

Les analyses ont été faites par cohorte c'est-à-dire par année de passation de l'examen. Il a été jugé important de les faire en vue d'éliminer tout effet de cohorte vu que nous avions procédé dès le début au regroupement de tous les participants en un seul échantillon de 160 étudiants. Mais les caractéristiques des étudiants n'étaient pas disponibles ce qui a gêné la recherche de cet effet de cohorte dans le vrai sens du terme. Le seul point connu était qu'ils étaient tous au même niveau dans leur cursus. A notre connaissance, cette étude fait partie des rares études ayant eu des participants répartis sur cinq années. Les analyses descriptives par cohorte ont révélé que la cohorte 2013 est la cohorte à avoir les plus faibles scores moyens selon la M1 et la M2. Par contre pour la M3, la cohorte 2012 a les plus faibles valeurs. Le coefficient alpha de Cronbach est très

variable avec des valeurs quasi nulles et même négatives pour la M3. Ceci nous a interpellés et nous avons formulé les hypothèses suivantes soit le codage initial est entaché d'erreurs soit on est en présence d'une association d'items mesurant différentes dimensions. Nous avions voulu éliminer toute erreur et nous avions repris les analyses avec M3 en commençant par la transformation des scores de l'échelle de Likert (-2 →+2) à l'échelle (1, 2, 3, 4, 5). Nous avons déterminé les scores en fonction de la réponse modale comme étant la bonne réponse. L'analyse a été lancée de nouveau et nous avons obtenu les mêmes résultats. A notre humble avis, ces résultats sont cohérents d'autant que nous avons un fort pourcentage d'items à avoir une corrélation item-total faible c'est-à-dire inférieure à 0,10.

En considérant les dimensions du test de concordance de script, il est noté que les valeurs du coefficient alpha de Cronbach sont supérieures aux valeurs obtenues avec l'échantillon total des 160 étudiants. Et ceci même pour la méthode selon la bonne réponse, la M3. De plus les valeurs du coefficient alpha de Cronbach sont plus élevées pour les dimensions investigation et thérapeutique. Ces deux dimensions ont moins d'items que la dimension diagnostique (36 items chacune contre 63 items pour la dimension diagnostique). Par contre les résultats de Latreille (2012) sont différents des nôtres car Latreille avait obtenu des valeurs du coefficient alpha de Cronbach très faibles, inférieures à 0,50 en considérant chaque dimension de son TCS. Nous pouvons avancer que la grande différence entre ces deux travaux réside déjà dans le nombre d'items. En comparant les valeurs du coefficient alpha de Cronbach total pour l'ensemble du test aux valeurs obtenues avec chaque dimension, nous voyons que les items de ces 3 dimensions ne sont pas parallèles ni additifs car le coefficient total diminue. Donc l'addition des dimensions ne contribue pas à améliorer la fidélité du score total au test (Laveault et Grégoire, 2014). Ceci nous porte à suggérer d'établir des sous-scores par dimension au lieu d'établir un score global. Nous notons aussi que la méthode 2 a permis d'obtenir de meilleurs scores comparée à M1 et M3 eu égard à la dimension investigation. De plus les meilleures corrélations pour échantillons appariés sont obtenues avec la paire M1-M2 et ce quelle que soit la dimension considérée.

Selon la théorie classique des tests, nous pouvons dire que la méthode 2 nous parait la plus appropriée pour évaluer le raisonnement clinique en contexte d'incertitude. La méthode des scores combinés de Charlin vient en deuxième position. Et la méthode 3 nous parait être la moins appropriée pour évaluer le raisonnement clinique car elle joue beaucoup sur les propriétés métriques des scores. Nous remettons en question l'optimisation préalable du test de concordance de script car d'elle dépend toutes les conclusions qu'on pourrait tirer de l'instrument de mesure. Il aurait fallu avoir le texte du TCS pour analyser le libellé des vignettes ou des catégories de réponse par exemple. Grondin *et al*. (2017) avaient identifié des items problématiques rien que dans leur libellé.

En résumé en regard de la théorie classique des tests, les meilleurs résultats relatifs aux propriétés métriques du test de concordance de script ont été notés en considérant séparément les résultats des scores de chaque dimension du TCS. Comme expliqué précédemment un des objectifs en réalisant cette étude exploratoire était d'analyser le TCS et ses méthodes de détermination des scores avec la modélisation Rasch. Ainsi dans la section suivante, allons-nous présenter la discussion des résultats des trois méthodes avec cette modélisation. Nous sommes restés fidèles à la question de recherche c'est-à-dire vérifier l'impact de chacune des méthodes de détermination des scores sur la qualité de la mesure même avec le modèle de Rasch.

5.2 Discussion des analyses avec la modélisation de Rasch

Dans l'ensemble avec la modélisation de Rasch nous avons eu des résultats quelque peu déroutants quelle que soit la méthode de détermination des scores utilisée. Il faut noter que la modélisation a été faite uniquement avec l'ensemble des items c'est-à dire contrairement à la théorie classique des tests nous n'avons pas pu modéliser selon les trois dimensions du TCS. Avec les premiers tests d'ajustement pour les sujets et les items, les indices de fidélité et de séparation obtenus diffèrent des valeurs seuils retenues dans la littérature.

5.2.1 Indices d'ajustement pour les étudiants et les items

Nous tenons à rappeler, préalablement à la discussion des valeurs des statistiques d'ajustement obtenues, que les auteurs Linacre et Wright (1994) ont défini plusieurs intervalles de valeurs jugées comme acceptables selon le contexte du test. En effet dans un contexte à enjeux critiques, ils recommandent l'intervalle [0,80 – 1,20] pour la statistique carré moyen (*infit* et *outfit*). Et on juge que l'ajustement est parfait avec une valeur de 1. Comme Dionne *et al*. (2017), nous avons considéré que le TCS est un instrument ayant des points communs autant avec le test à choix multiple administré dans un contexte à enjeux critiques qu'avec un instrument qui collige des données sur des observations cliniques. Avons-nous ainsi retenu l'intervalle compris entre 0,50 et 1,70 qui correspond, selon Linacre et Wright, à une étendue acceptable pour un instrument évaluant des données basées sur des observations cliniques. Qu'en est-il pour les statistiques standardisées? Pour mieux interpréter les statistiques standardisées, Linacre (2002) suggère, en effet, que ces valeurs soient comprises entre -1,90 et +1,90. De telles valeurs témoigneraient d'un ajustement adéquat des données obtenues au modèle choisi.

Dans l'ensemble les scores modélisés des étudiants sont tous concentrés au même endroit sur l'échelle. Ceci fait supposer que les items de ce TCS sont probablement trop semblables et ils se comportent donc de la même façon. Avec la méthode des scores combinés (méthode 1), nous avons des valeurs moyennes de l'indice carré moyen comprises entre 0,50 et 1,70 (Linacre et Wright, 1994) ce qui pourrait signifier que nous avons une mesure optimale du raisonnement clinique. Cependant les valeurs extrêmes obtenues nous ont interpellés et nous ont fait évoquer qu'il y avait beaucoup de bruit dans les données. Nous avons ainsi procédé à des retraits itératifs de sujets et/ou d'items à la recherche d'un meilleur ajustement du TCS au modèle à crédit partiel de Rasch. Ce processus n'a pas bonifié l'ajustement de nos données au modèle.

Qu'en est-il de la méthode 2, la méthode des scores combinés avec pénalité de distance ? Les moyennes des indices d'ajustement sont aussi comprises dans l'intervalle décrit par Linacre et Wright (1994). Nous avons noté aussi beaucoup de valeurs aberrantes et comme pour la méthode 1, le retrait itératif simultané ou non de sujets et d'items n'a pas amélioré la qualité de la mesure.

Nous avons répété la même démarche avec la méthode 3. Pour rappel, cette méthode est basée sur une bonne réponse et elle est donc dichotomique. Nous voyons que les valeurs moyennes sont aussi retrouvées dans l'intervalle de Linacre et Wright. Mais nous sommes confrontés aussi aux valeurs extrêmes très aberrantes. Nous tenons à signaler à l'attention du lecteur que les items sont mieux étendus sur l'échelle en considérant la carte Wright contrairement aux méthodes 1 et 2.

5.2.2 La fidélité

Nous avons apprécié la fidélité du test de concordance de script en nous appuyant sur les deux théories à savoir d'une part sur la théorie classique des tests avec le coefficient alpha de Cronbach et d'autre part sur la modélisation de Rasch. Nous allons discuter respectivement des indices de séparation *(separation*) et de l'indice de fidélité (*reliability*) obtenus avec chacune de nos trois méthodes pour la modélisation de Rasch.

L'indice de séparation des items est faible pour la méthode 1. Par contre, il est meilleur avec la méthode 2 et la méthode 3. Ceci suggèrerait que l'échantillon des 160 étudiants est trop faible en ce qui a trait à la méthode des scores combinés de Charlin ; donc la hiérarchie des items ne peut pas être assurée.

Quant aux répondants, l'indice de séparation est très faible en considérant la référence de Boone et *al*. quelle que soit la méthode utilisée. Nous pouvons donc dire que le modèle utilisé ne nous permet pas de bien différencier les répondants en ce qui a trait à leur raisonnement clinique. Même en nous référant à Duncan, Bode, Lai et Perera (2003) qui suggèrent une limite inférieure acceptable de 1,50, les valeurs obtenues restent en deçà. Nous rappelons qu'avec la représentation de Wright nous avons vu que les sujets et les items étaient positionnés sur un même point ; il n'y avait pas du tout d'étendue surtout avec les méthodes 1 et 2. Par contre avec la méthode 3, il y avait une meilleure étendue des items sur l'échelle. Ceci est en accord avec les valeurs de l'indice de séparation. Les auteurs Dionne *et al*. (2017) et Grondin *et al*. (2017) ont aussi eu dans le contexte de modélisation d'un TCS les mêmes observations à savoir de faibles valeurs de l'indice de séparation. Si nous nous référons aux valeurs du coefficient alpha de Cronbach, nous

pouvons dire que ces deux paramètres statistiques sont cohérents dans le sens qu'ils sont les deux faibles.

Nous allons nous intéresser maintenant aux indices de fidélité des sujets et des items.

Nous avons précédemment apprécié la fidélité de ce test de concordance de script en nous appuyant sur l'indice de fidélité de Cronbach selon la théorie classique des tests. Ce dernier est critiqué par certains car il y aurait un risque de surestimation de la fidélité quand on a affaire à des scores bruts de nature non-linéaire dans la modélisation de Rasch (Dionne *et al*., 2017). L'indice de fidélité des sujets avec la modélisation est faible quel que soit la méthode considérée. Dionne *et al*. ont rapporté aussi des valeurs faibles pour l'indice de fidélité des sujets et nous rappelons que ces auteurs avaient utilisé la méthode de Charlin. Il n'y a pas beaucoup d'études à date pour nous permettre de faire une comparaison plus large.

Quant aux items, les indices de fidélité sont bons (> 0,80) et ceci peut suggérer que l'échantillon de répondants est adéquat pour l'estimation des paramètres des modèles. Nous remarquons qu'avec la méthode 3 nous avons l'indice de fidélité et l'indice de séparation les plus faibles en ce qui a trait aux sujets. En nous attardant sur les items, la méthode 1, la méthode de Charlin ou encore la méthode des scores combinés, obtient les plus faibles valeurs.

5.2.3 Indépendance locale

En considérant les résidus standardisés obtenus après extraction du facteur de Rasch, aucune valeur de corrélation supérieure à 0,70 n'est retrouvée peu importe la méthode de détermination des scores utilisée. Ainsi nous pouvons dire que les items ont montré une indépendance locale, caractéristique très importante dans la modélisation de Rasch.

5.2.4 Unidimensionnalité

Une analyse des composantes principales sur les résidus standardisés a été faite pour vérifier ce postulat après extraction du facteur Rasch. En consultant les écrits peu

nombreux à date à ce sujet, nous n'avions pas trouvé de démarche très explicite pour la réaliser. Les travaux de Grondin *et al*. (2017) nous ont servi de guide. Nous voulions savoir est-ce que les scores obtenus par nos répondants seraient sous la seule influence de leur raisonnement clinique ou non.

En nous référant aux valeurs définies par Linacre (2015), la variance expliquée par la mesure est très faible pour les trois méthodes et les répondants ont la plus faible proportion de la variance. Que pourrait-on avancer au vu de ces résultats? En effet, ces derniers nous font penser que des facteurs autres que le raisonnement clinique des répondants influencent probablement les scores obtenus. Il existerait probablement une sous-dimension. Contrairement à Grondin *et al*., nous n'avons pas pu réaliser d'étude qualitative des items du TCS non plus pour étudier la force de la corrélation atténuée entre les paramètres de sujets modélisés et deux groupes d'items à savoir ceux qui corrèlent le plus fortement et ceux qui corrèlent le plus faiblement avec le facteur.

Comment expliquer de tels résultats? Que nous dit Linacre là-dessus? En effet selon Linacre (2003), il y a trois hypothèses. a- soit les scores des répondants sont aléatoires mais ils restent alignés sur les prédictions du modèle; b- soit les scores ne sont pas du tout prédits par le modèle, c- soit encore l'estimation de la difficulté des items et l'estimation de l'habileté des répondants ne sont pas précises. Laquelle de ces hypothèses convient le mieux à notre contexte ? En nous basant sur les valeurs des statistiques d'ajustement, nous pensons que la troisième hypothèse nous conviendrait le plus. D'autant qu'en référence avec la représentation de Wright nous n'avions pas du tout observé d'étendue des sujets ni des items. De plus il y a un lien étroit entre le pourcentage de variance expliqué par le facteur Rasch d'une part et, d'autre part, avec l'étendue de la distribution de l'estimation de l'habileté des répondants et l'estimation de la difficulté des items selon Linacre (2008). Toutefois, il avait considéré dans cet écrit des items dichotomiques. Dans notre situation nous avons seulement la méthode 3 qui est dichotomique et avec cette dernière, la variance est faible mais relativement moins que les deux autres méthodes.

Toujours en nous référant à Linacre (2015) une valeur propre d'au moins 2 correspond à un regroupement de 2 items au moins. Ceci représente le nombre minimal

d'items pour constituer une sous-dimension. Qu'en est-il de nos résultats ? Pour chacune des méthodes considérées, la valeur propre associée est supérieure à 2 pour tous les contrastes (1 à 5). Ainsi nous pensons que nous sommes probablement en présence de regroupements d'items formant des sous-dimensions comme mentionné plus haut. L'estimation des paramètres fournie par le modèle risque d'être déformée.

Attardons-nous sur la valeur de la corrélation atténuée entre les items. La corrélation atténuée est la corrélation obtenue après extraction de l'erreur standard liée à chaque mesure d'un répondant pour chaque groupe d'items. Toujours selon Linacre (201), si la valeur est éloignée de l'unité 1, les groupes d'items mesurent 2 dimensions différentes. Ainsi l'hypothèse nulle que ces groupes d'items mesurent la même chose peut être rejetée. Voyons d'abord la méthode 1 et les valeurs de la corrélation atténuée. Les groupes 1-3 et 1-2 ont des valeurs absolues inférieures à 0,57. Ceci signifie que les items auraient autant la moitié de leur variance en commun ; C'est la valeur-seuil, la probabilité qu'on a affaire à différents traits latents variables. Pour les autres contrastes il y a autant de variation.

En ce qui a trait à la méthode 2, pour les premiers contrastes, on a une corrélation atténuée supérieure à 0,57 mais moindre que 0,71 pour le groupe 1-2. Par contre pour les groupes 1-3 et 2-3, la valeur est moindre que 0,57 faisant évoquer qu'il y a probablement des traits latents différents.

Pour la méthode 3, la valeur de la corrélation atténuée des deux premiers contrastes varie beaucoup. Certaines sont supérieures à 0,71 respectivement les groupes 1-3 de ces 2 premiers contrastes. Ceci indique que les mesures des répondants aux deux groupes d'items ont plus que la moitié de leur variance en commun donc ils sont plus dépendants qu'indépendants. Il y a autant de variation pour les autres contrastes. Donc les deux postulats fondamentaux de la modélisation de Rasch ne sont pas vérifiés. Tout ceci est en cohérence avec les statistiques d'ajustement non retrouvées dans les intervalles recommandés et aussi avec la mauvaise qualité de la mesure.

5.3 Contribution de l'étude

Notre étude fait partie des rares études ayant abordé la question de la détermination des scores dans le contexte du test de concordance de script. Des auteurs tels Bland *et al*. (2005) et Wilson *et al*. (2014) ont été les premiers à avoir initié la réflexion sur cette variable de cet instrument de mesure. De plus il faut aussi souligner que les études relatives au TCS ont, dans leur grande majorité, utilisé la théorie classique des tests. Le point original de l'étude exploratoire descriptive que nous avions menée est la modélisation de Rasch qui a permis de mieux comprendre les propriétés métriques des scores. Nous pouvons détecter les items problématiques et ainsi nous pouvons améliorer l'optimisation du TCS.

Et cette étude a contribué à faire ressortir le caractère limité de la théorie classique des tests et de la modélisation de Rasch avec les méthodes de détermination des scores dans leur forme actuelle. Notre étude espère avoir ouvert la porte pour d'autres recherches dans ce domaine.

Vu les enjeux critiques (formatif, sommatif, certificatif) dans l'utilisation du TCS dans différents domaines de formation en santé, il est judicieux de pouvoir analyser les résultats des études et de pouvoir émettre un jugement critique pertinent. Il s'en suivra alors une plus grande utilisation du TCS.

L'auteure a pu grâce à cette étude se familiariser avec les concepts sous-jacents à l'élaboration des outils dans le but d'évaluer les compétences des apprenants en médecine, particulièrement le raisonnement clinique.

5.4 Limites de l'étude

Plusieurs limites ont été identifiées dans la réalisation de cette étude. Nous allons les aborder sur le plan général, mais aussi selon les 2 modèles théoriques.

Sur le plan général

- La conception même de l'étude à partir des données secondaires constitue une grande limite parce que l'utilisation de données recueillies par des tiers présente des avantages et des inconvénients (Dionne et Fleuret, 2016). Le risque potentiel d'inhiber la créativité de l'auteure est l'un des inconvénients. En effet, l'auteure n'a pas pu contrôler l'élaboration du questionnaire et encore moins la collecte des données. Au vu des résultats non satisfaisants avec la TCT et la modélisation de Rasch, nous avons pensé qu'une analyse qualitative de l'énoncé des items serait essentielle. Mais cette analyse n'a pas pu se faire car nous ne disposions pas du texte.

- Une limite souvent rapportée par les auteurs dans le cadre de l'analyse des données secondaires est le respect de l'intégrité des données surtout lors du transfert de ces dernières d'un logiciel à un autre. Pour écarter cette limite, nous avons examiné un échantillon aléatoire à chaque fois correspondant à 10% des variables pour voir si les données étaient intègres.

- Une autre limite est la gestion des données manquantes pour lesquelles nous avons évalué l'ampleur. Pour faciliter les calculs et donc l'analyse, les données manquantes ont été codifiées 0 pour la TCT et 999 pour la modélisation. Mais elles étaient de moins de 7%. Les données ont été collectées dans le même but que nous à savoir évaluation du raisonnement clinique avec un TCS

- Malgré l'optimisation dont a été l'objet ce TCS par ses constructeurs de l'Université de Liège en Belgique, nous avons pensé que cette démarche n'a pas été optimale vu les résultats obtenus que ce soit avec la théorie classique des tests que ce soit avec la modélisation de Rasch. Certains paramètres relatifs aux participants n'étaient pas disponibles ce qui a empêché d'étudier, par exemple, le fonctionnement différentiel des items. Toutefois, l'auteure a réalisé quelques vérifications sur les propriétés statistiques de cette base de données, notamment l'effet de cohorte sur les résultats, l'effet des trois dimensions du TCS sur la cohérence interne.

- Les étudiants appartenaient à cinq cohortes annuelles de 2010 à 2014. Ils sont tous des étudiants finissants qui vont s'orienter vers la médecine générale. Il faut aussi

noter que la taille de la population en fonction de l'année considérée est très inhomogène. Nous avons tenté de faire une analyse comparative de ces cohortes et il a été remarqué des différences dans la moyenne des scores. Mais nous ne pouvons pas savoir si ces différences sont dues à l'échantillonnage ou si certaines cohortes ont pu être meilleures que d'autres.

- De plus les experts constituent un groupe pour lequel nous n'avons aucune information relative à leur nombre d'années d'expérience par exemple. Ceci constitue aussi une limite à notre sens.

Théorie classique des tests et Modélisation de Rasch

L'analyse des résultats obtenus est déroutante que ce soit selon la théorie classique des tests que ce soit selon la modélisation de Rasch. Ceci nous amène à questionner le choix des experts depuis la conception de la clé des scores.

Nous avons été confrontés au problème de carence d'écrits relatifs à l'utilisation de la modélisation de Rasch dans le contexte du TCS. La majorité de ces écrits sont en anglais et non réalisés dans le domaine de l'éducation médicale. Toutefois en 2017, l'auteure a pu trouver un papier de Grondin *et al*. expliquant la démarche de la modélisation de Rasch a tout débutant. L'auteure a du se familiariser avec le logiciel Winsteps et ainsi découvrir le modèle à crédit partiel.

L'ajustement des scores calculés est mauvais ceci est probablement lié à la mauvaise qualité des items. Il serait pertinent de répéter la modélisation des scores répartis selon les trois dimensions du TCS. En effet la TCT réalisée en considérant chacune des dimensions séparément nous a permis d'avoir de meilleurs résultats. Ceci laisse présumer qu'il en serait de même avec la modélisation. Malheureusement nous avons été confronté à des problèmes techniques ayant empêché de la faire.

5.5 Recommandations pour le futur

Il est essentiel d'avoir un instrument de mesure dont l'interprétation des scores obtenus pourra être considérée comme valide et fidèle pour évaluer le raisonnement clinique des apprenants en médecine. Nous pouvons émettre quelques recommandations aux constructeurs du test de concordance de script.

Il serait important de construire le test de concordance de script avec des items optimisés bien rédigés reflétant les situations complexes en contexte d'incertitude auxquelles les apprenants seront confrontés dans leur pratique. Et il faudrait augmenter également le nombre des vignettes avec une répartition homogène en fonction des dimensions. Il nous parait essentiel aussi d'augmenter le nombre d'experts avec un bon profil pour le domaine d'étude considérée et aussi de les classer en fonction de leur année d'expérience.

Quant à la question principale de notre étude, malgré toutes les limites évoquées, il est clair que nous ne pouvons recommander la méthode selon une bonne réponse pour attribuer des notes aux étudiants. En revanche, la méthode des scores combinés de Charlin et la méthode des scores combinés avec pénalité de distance restent les deux méthodes à approfondir dans ce contexte-là. La modélisation Rasch peut être d'une grande aide pour la consécration de l'une de ces deux méthodes et pourrait être complémentaire à la théorie classique des tests. Elle reste une avenue à explorer dans le futur. De plus nous pensons qu'on doit considérer chacune des trois dimensions séparément en déterminant la cotation de chaque vignette et ainsi calculer le score composite total.

L'auteure recommande de réaliser une étude avec des données primaires relatives au test de concordance de script en utilisant les deux modèles théoriques. Si les données s'ajustent bien au modèle et les conditions d'utilisation sont bien respectées, ceci permettra certainement de mieux maitriser les méthodes d'analyse et d'affiner l'outil de mesure et ainsi obtenir des propriétés métriques des scores de haute valeur.

CONCLUSION

Le raisonnement clinique est une composante importante de la compétence clinique et se construit au fil des situations cliniques quotidiennes. Son évaluation est donc essentielle avec des outils valides. Cette étude exploratoire a apporté un éclairage utile et original sur les méthodes de détermination des scores par l'utilisation de deux théories. Les résultats des étudiants ne sont pas justes vu qu'ils sont attribués selon la réponse modale des experts lesquels d'ailleurs n'avaient pas un score moyen trop grand par rapport à celui des étudiants et dans l'ensemble sans une grande étendue. Les méthodes actuelles ne sont pas satisfaisantes dans leur forme actuelle et la méthode selon une bonne réponse est la moins appropriée dans le contexte du test de concordance de script. Avec la théorie classique des tests, nous aurions émis la conclusion à savoir que les propriétés métriques des scores sont bonnes avec la méthode de Charlin à coup sûr et avec la méthode des scores combines avec pénalité de distance. Cependant la modélisation de Rasch est complémentaire à la théorie classique des tests car elle a permis de mieux déceler les items problématiques. Également elle a permis de comprendre que la performance des étudiants était probablement sous l'influence de plusieurs sous-dimensions et que pour le TCS à l'étude les items ne sont pas, par conséquent, unidimensionnels ni indépendants. Nous ne pouvons pas ainsi situer les sujets de façon fiable sur le continuum de mesure. Nous recommandons ainsi aux constructeurs de ce TCS d'ajouter des items dans les zones de distribution des sujets peu ou pas couverts par les items, de mieux catégoriser les items, et aussi de valider le contenu des items en augmentant le nombre d'experts du domaine choisi.

Nous avons compris la grande variabilité des conclusions des travaux précédents sur la méthode de détermination des scores. Le consensus n'est donc pas encore trouvé ce qui explique qu'il reste encore des points à élucider. Cette étude exploratoire espère avoir ouvert la porte à d'autres recherches plus poussées dans le but de faciliter la compréhension et surtout la généralisation de l'utilisation de la modélisation de Rasch dans le contexte du test de concordance de script. Nous pensons qu'elle permettra aux constructeurs de TCS de mieux élaborer les tâches et de prendre de meilleures décisions dans le processus d'évaluation du raisonnement clinique des apprenants en sciences de la

santé. Il est tentant d'avancer que ce serait mieux de considérer des sous-scores pour chacune des trois dimensions du test de concordance de script et de calculer le score composite total. Ainsi les évaluateurs ayant recours au test de concordance de script pour évaluer le raisonnement clinique des apprenants en médecine pourraient avoir une appréciation plus large de cette compétence tant attendue et ainsi apporter une remédiation plus adaptée aussi à leurs faiblesses en fonction de la dimension en difficulté.

Références

Ahmadi, S-F., Khoshkish, S., Soltani-Arabshahi, K., Hafezi-Moghadam, P., Zahmatkesh, G., Heidari, P., …Lotfipou, S. (2014). Challenging script concordance test reference standard by evidence: do judgments by emergency medicine consultants agree with likelihood ratios*? International Journal of Emergency Medicine* 7:34.

Alinier, G. (2003). Nursing students'and lectures'perpesctives of objective structured clinical examination incorporating simulation. *Nurse Education Tearoday*, 23 (6), 4190426.

Amini M, Moghadami M, Kojuri J, Abbasi H, Abadi AA, Molaee NA, … Charlin B. (2011). An innovative method to assess clinical reasoning skills: clinical reasoning tests in the second national medical science Olympiad in Iran. *BMC Research Notes*. 4(1), 418. doi: 10.1186/1756-0500-4-418.

Audétat, M., Dory, V., Nendaz, M., Vanpee, D., Pestiaux, D., Junod Perron, N. et Charlin, B. (2012). What is so difficult about managing clinical reasoning difficulties? *Medical Education*, 46, 216-227.

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

Blais, J.G. (1987). Effets de la violation du postulat d'unidimensionnalité dans la théorie des réponses aux items (thèse de doctorat, Université de Montréal à Montréal, Canada, consultée le 8 octobre 2014.

Bland, A. C., Kreiter, C. D. et Gordon, J.A. (2005). The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine*, 80, 395-399.

Bond, T.G. et Fox, C. M. (2007). Applying the Rasch model : fundamental measurement in the human sciences, 2nd ed. Mahwah: Lawrence Erlbaum Associates Publishers

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

Boulouffe, C., Doucet, B., Muschart, X., Charlin, B. et Vanpee, D. (2013). Assessing clinical reasoning using a script concordance test with electrocardiogram in an emergency medicine clerkship rotation. *Emergency Medicine Journal*, 1-4. doi:10.1136/emermed-2012-201737.

Brailovsky, C., Charlin, B., Beausoleil, S., Coté, S. et Van der Vleuten, C. (2001). Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. *Medical Education*, 35, 430-436.

Bursztein, A.-C., Cuny, J.-F., Adam, J.-L., Sido, L., Schmutz, J.L., de Korwin, J.-D., …Barbaud, A. (2011). Usefulness of the script concordance test in dermatology. Journal of the European Academy of Dermatology and Venereology, 25, 1471-1475. Doi:10.1111/j.1468-3083.2011.04008.x.

Caire, F., Sol, J.C., Charlin, B., Isodiri, P. et Moreau, J.J. (2004). Le test de concordance de script (TCS) comme outil d'évaluation formative des internes en neurochirurgie : implantation du test sur internet à l'échelle nationale. *Pédagogie Médicale*, 5(2), 87-94.

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

Cavanagh, R. F. et Waugh. R. F. (2011). Applications of Rasch measurement in learning environments research. Rotterdam, The Netherlands: Sense publishers.

Centre de pédagogie appliquée aux sciences de la santé. Image de vignette de test de concordance de script. Récupéré du site internet de la Faculté de Médecine de l'Université de Montréal : http://www.cpass.umontreal.ca/documents/images/TCS/Tcs1.gif.

Chang, T. P., Kessler, D., McAnninch, B., Fein, D. M., Scherzer, D. J., Seelbach, E., …Pusic, M.V. (2014). Script concordance testing : assessing residents' clinical decision-making skills for infant lumbar punctures. *Academic Medicine*, 89, 1-8. doi : 10.1097/ACM.0000000000000059.

Charlin, B. (2006). Évaluer la dimension d'incertitude du raisonnement clinique. *Pédagogie Médicale*, 7(1), 5-6.

Charlin, B., Bordage, G. et van der Vleuten, C. (2003). L'évaluation du raisonnement clinique. *Pédagogie Médicale*, 4, 42-52.

Charlin, B., Boshuizen, H. P. A., Custers, E. J. F. M. et Feltovich, P. J. (2007). Scripts and clinical reasoning. *Medical Education*, 41, 1178-1184.

Charlin, B., Brailovsky C. A., Brazeau-Lamontagne L., Samson L. et Leduc, C. (1998) Script Questionnaires : Their Use for Assessment of Diagnostic Knowledge in Radiology. *Medical Teacher*, 20: 567-571.

Charlin, B., Brailovsky, C., Leduc, C. et Blouin, D. (1998). The diagnosis script questionnaire: A new tool to assess a specific dimension of clinical competence. *Advances in Health Sciences Education*, 3(1), 51-58. doi: 10.1023/A:1009741430850.

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

Charlin, B., Gagnon, R., Kazi-Tani, D. et Thivierge, R. (2006). Le test de concordance de script comme outil d'évaluation en ligne du raisonnement des professionnels en situation d'incertitude. *Revue Internationale des Technologies en Pédagogie*, 2(1).

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

Charlin, B., Gagnon, R., Lubarsky, S., Lambert, C., Meterissian, S. Chalk, C., ...van der Vleuten, C. (2010). Assessment in the context of uncertainty using the script concordance test: more meaning for scores. *Teaching and Learning Medicine*, 22, 180-186. doi:10.1080/10401334.2010.488197.

Charlin, B., Roy, L., Brailovsky, C. A. et van der Vleuten, C. (2000). The Script Concordance Test : A tool to assess the reflective clinician. *Teaching and Learning in Medical Education*, 12, 189-195.

Charlin, B. et St-Jean, M. (2002). Le test de concordance de script : un outil pour évaluer le jugement en médecine. *Bulletin du CÉFES*, 6, 4-5, Université de Montréal.

Charlin, B., Tardif, J. et Boshuizen, H. P. A. (2000). Scripts and medical diagnostic knowledge : Theory and applications for clinical reasoning instruction and research. *Academic Medicine*, 75, 182-190.

Charlin, B. et van der Vleuten, C. (2004). Standardized assessment of reasoning in context of uncertainty. The script concordance test approach. *Evaluation and the Health Professions*, 27, 304-319.

Cobb, K. A., Brown, G., Hammond, R. et H. Mossop, L. H. (2015). Students' Perceptions of the Script Concordance Test and Its Impact on Their Learning Behavior: A Mixed Methods Study. *Journal of Veterinary Medical Education, 42(1) 8.* doi: 10.3138/jvme.0514-057R1.

Cohen, L. J., Fitzgerald, S. G., Lane, S. et Boninger, M. L. (2005). Development of the seating and mobility script concordance test for spinal cord injury: Obtaining content validity evidence. *Assistive Technology*, 17(2), 122-132. doi:10.1080/10400435.2005.10132102.

Cooke, S., Lemay, J.-F., Beran, T., Sandhu et Amin, H. (2016). Developpment of a method to measure clinical reasoning of pediatric residents: The pediatric script concordance test. *Creative Education*, 7, 814-823. http://dx.doi.org/10.4236/ce.2016.76084

Compere V., Moriceau, J., Gouin, A., Guitard, P-G., Damm, C., Provost, D., …Dureuil, B. (2015). Residents in tutored practice exchange groups have better medical raisoning as measured by the script concordance test: a pilot study. *Anaesthesia Critical Care Pain & Medicine*, 34 (1), 17-21. doi.org/10.1016/j.accpm.2014.12.001.

Dawson, T., Comer, L., Kossick, M.A.et Neubrander, J. (2014). Can script concordance testing be used in nursing education to accurately assess clinical reasoning skills? *Journal of Nursing Education*, 54, 281-286.

Deschênes, M.F., Charlin, B., Gagnon, R. et Goudreau, J. (2011). Use of a Script Concordance Test to Assess Development of Clinical Reasoning in Nursing Students. *Journal of Nursing Education*, 50 (7), 381-387.

Dionne, E. et Fleuret C. (2016). L'analyse de données secondaires dans le cadre d'évaluation de programme, regard théorique et expérientiel. La Revue canadienne d'évaluation de programme 31.2, 253–261 doi: 10.3138/cjpe.142.000.

Dionne, E., Grondin, J. et Latreille, M-E. (2017). Exploration des scores à un test de concordance de script sous la loupe du modèle de Rasch. Dans E. Dionne et I. Raîche (dir.), (p.77-110). Mesure et évaluation en éducation médicale. Regards actuels et prospectifs. Presse des Universités du Québec

Dory, V., Gagnon, R., Vanpee, D. et Charlin, B. (2012). How to construct and implement script concordance tests: insights from a systematic review. *Medical Education*, 46 (6), 552-563.

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

Drolet, P. (2015). Assessing clinical reasoning in anesthesiology: Making the case for the Script Concordance Test. *Anaesthesia Critical Care Pain & Medicine* 34 (1), 5-7. Doi:10.1016/j.accpm.2015.01.003.

Ducos, G., Lejus, C., Sztark, F., Nathan, N., Fourcade, O., Tack, I., …Minville, V. (2015). The Script Concordance Test in anesthesiology: Validation of a new tool for assessing clinical reasoning. *Anaesthesia Critical Care Pain & Medicine*, 34(1), 11-15.

Dumont, K., Loye, N., et Goudreau, J. (2015). Le potentiel diagnostique des questions d'un test de concordance de script pour évaluer le raisonnement clinique infirmier. *Pédagogie médicale*, 16(1), 49-64. doi: 10.1051/pmed/2015012.

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

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

Eva, K. W. (2005). What every teacher needs to know about clinical reasoning. *Medical Education*, 39 (1), 98-106. doi:10.1111/j.1365-2929.2004.01972.x.

Faucher, C., Dufour-Guindon, M.-P., Lapointe, G., Gagnon, R. et Charlin, B. (2016). Assessing clinical reasoning in optometry using the script concordance test. *Clinical and Experimental Optometry*, 99 (3), 280–286. doi:10.1111/cxo.12354.

Feltovich, P. J. et Barrows, H. S. (1984). Issues of generality in medical problem solving. In : Schmidt HG, De Volder ML, eds. Tutorials in Problem-Based Learning (p.128-142). Assen/Maastrich, The Netherlands: VanGorcum 128–142.

Fournier, J-P., Demeester, A. et Charlin, B. (2008). Script concordance tests: guidelines for construction. *BMC Medical Informatics and Decision Making*, 8 (18). doi: 10.1186/1472-6947-8-18

Fournier, J-P., Thiercelin, D., Pulcini, C., Alunnil-Perret, V., Gilbert, E., Minguet, J-M. et Bertrand, F. (2006). Clinical reasoning assessment in emergency medicine: Script concordance tests are more efficient to detect clinical experience than rich-context multiple choice questions. *Pédagogie Médicale*, 7, 20-30.

Gagnon, R. et Charlin, B. (2007). Qui gagne? Faut-il tenir compte des réponses faites au hasard au cours des examens? *Pédagogie Médicale*, 8, 69-70.

Gagnon, R., Charlin, B., Coletti, M., Sauvé, E. et van der Vleuten, C. (2005). Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test? *Medical Education,* 39 (3), 284-291. doi:10.1111/j.1365-2929.2005.02092.x.

Giet, D., Massart, V., Gagnon, R. et Charlin, B. (2013). Le test de concordance de script en 20 questions. *Pédagogie Médicale*, 14 (1) 39–48.

Goos, M., Schubach, F., Seifert, G. et Boeker, M. (2016). Validation of undergraduate medical student concordance script test scores on the clinical assessment of the acute abdomen. *BMC Surgery*, 16,57 doi 10.1186/s12893-016-0173-y.

Goulet, F., Jacques, A., Gagnon, R., Charlin, B. et Shabah, A. (2010). Poorly performing physicians: Does the script concordance test detect bad clinical reasoning? *Journal of Continuing Education in the Health Professions*, 30 (3)161-166. doi:10.1002/chp.20076.

Grondin, J., Dionne, E., Savard, J. et Casimiro, L. (2017). Démonstration d'une méthodologie mettant à profit les modèles de Rasch : l'exemple d'une échelle de la mesure de l'offre active de services de santé en français. Dans E. Dionne et I. Raîche (dir.), (p.11-52). *Mesure et évaluation en éducation médicale*. Regards actuels et prospectifs. Presse des Universités du Québec.

Haladyna, T. M. et Rodriguez, M.C. (2013). Developing and validating test items

Harden, R.M. (1983). Preparation and presentation of patient-management problem (PMPs). *Medical Education*, 17(4), 256-276.

Harden, R.M. et Gleeson, F.A. (1979). Assessment of clinical competence using an objective structural clinical examination (OCSE). *Medical Education*, 13 (1) 41-54.

Hayward, J., Cheung, A., Velji, A., Altarejos, J., Gill, P., Scarfe, A. et Lewis, M. (2016). Script-theory virtual case: A novel tool for education and research. *Medical Teacher*, doi: 10.3109/0142159X.2016.1170776.

Hornos, E. H., Pleguezuelos, E. M., Brailovsky, C. A., Harillo, L. D., Dory, V. et Charlin, B. (2013). The practicum script concordance: an online continuing professional development format to foster reflection on clinical practice. *Journal of Continuing Education in the Health Professions*, 33 (1) 59- 66.c.

Humbert, A. J., Besinger, B. et Miech, E. J. (2011). Assessing clinical reasoning skills in scenarios of uncertainty: Convergent validity for a script concordance test in an emergency medicine clerkship and residency. *Academic Emergency Medicine*, 18 (6), 627-634. Récupéré du site : http://dx.doi.org.proxy.bib.uottawa.ca/10.1111/j.1553-2712.2011.01084.x le 16 juillet 2015.

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

Iravani, K., Amini, M., Doostkam, A., et Dehbozorgian, M. (2016). The validity and reliability of script concordance test in otolaryngology residency training. *Journal of Advances in Medical Education & Professionalism*, 4(2):93-96.

Jolly, B. et Grant, J. (1997). The Good Assessment Guide. A practical Guide to Assessment and Appraisal for Higher Specialist Training. Joint Center for Education in Medicine. London: UK.

Jouquan, J. (2002). L'évaluation des apprentissages des étudiants en formation médicale initiale. *Pédagogie médicale*, 3, 38-52.

Kazour, F., Richa, S., Zoghbi, M., El-Hage, W. et Haddad, F. G. (2016). Using the script concordance test to evaluate clinical reasoning skills in psychiatry. *Academic Psychiatry*. doi 10.1007/s40596-016-0539-6.

Labelle, M., Gagnon, R., Thivierge, R.L., Laprise, R., Ste-Marie, L-G. et Charlin. B. (2003). Formation continue en petits groupes sur l'ostéoporose : comparaison d'un atelier basé sur le test de concordance de scripts (TCS) et d'un atelier classique. *Pédagogie Médicale*, 4, 145-53.

Lambert, C., Gagnon, R., Nguyen, D. et Charlin, B. (2009). The script concordance test in radiation oncology: validation study of a new tool to assess clinical reasoning. *Radiation Oncology*, 9 (4):7.

Latreille, M. E. (2012). Évaluation du raisonnement clinique d'étudiantes et d'infirmières dans le domaine de la pédiatrie, à l'aide d'un test de concordance de script. (thèse de maîtrise, Université d'Ottawa, Canada). Récupéré du site : https://www.ruor.uottawa.ca/bitstream/10393/22698/3/Latreille_Marie_Eve_2012_these.pdf .

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

Lebeau, J. et Pagonis, D. (2006). Le test de concordance de script (TCS) : comment évaluer le raisonnement médical en situation d'incertitude? *Revue de Stomatologie et Chirurgie Maxillo-Faciale*, 107, 327-329.

Lemay, J-F., Donnon, T. et Charlin, B. (2010). The reliability and validity of a paediatric script concordance test with medical students, paediatric residents and experienced paediatricians. *Canadian Medical Education Journal*, 1, e89-e95.

Linacre, J. M. (2002). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3* (1), 85–106.

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

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

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

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

Linacre, J.M. (2015). Winsteps Rasch measurement computer program User's Guide (Version 3.90.2). Beaverton, Oregon. Récupéré du site http://www.winsteps.com/index.htm le 15 Juin 2017.

Lineberry, M., Kreiter, C.D. et Bordage, G. (2013). Threats to validity in the use and interpretation of script concordance test scores. *Medical Education*, 47, 1175–1183.

Lubarsky, S., Charlin, B., Cook, D. A., Chalk, C. et van der Vleuten, C. (2011). Script concordance testing: A review of published validity evidence. *Medical Education*, 45(4), 329-338. doi:10.1111/j.1365-2923.2010.03863.x.

Lubarsky, S., Dory, V., Duggan, P., Gagnon, R. et Charlin, B. (2013). Script concordance testing: From theory to practice: AMEE guide no. 75. *Medical Teacher*, 35 (3), 184-193. doi:10.3109/0142159X.2013.760036.

Lubarsky, S., Gagnon, R. et Charlin, B. (2013). Scoring the script concordance test: not a white and black issue. *Medical Education*, 47, 1152-1161.

Lubarsky, S., Dory, V., Audétat, M.-C., Custers, E. et Charlin, B. (2015). Using script theory to cultivate illness script formation and clinical reasoning in health professions education. *Canadian Medical Education Journal*, 6(2), e61-e70.

Marie, I., Sibert, L., Roussel, F., Hellot, M., Lechevallier, J. et Weber, J. (2005). Le test de concordance de script : un nouvel outil d'évaluation du raisonnement et de la compétence clinique en médecine interne? *La Revue de Médecine Interne,* 26 (6), 501-507.

Mathieu, S., Couderc, M., Glace, B., Tournadre A., Malochet-Guinamand, S., Pereira, B., …Soubrier, M. (2013). Construction and utilization of a script concordance test as an assessment tool for DCEM3 (5th year) medical students in rheumatology. *BMC Medical Education*. 13:166. https:/doi.org/10.1186/1472-6920-13-166.

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

Meterissian, S.H. (2006). A novel method of assessing clinical reasoning in surgical residents. *Surgical Innovation*, 13, 115–119.

Norcini, J. J., Shea, J. A. et Day, S. C. (1990). The use of the aggregate scoring for a recertification examination. *Evaluation and Health Professions*, 13, 241–251.

Norman, G.R. (1985). Objective measurement of clinical performance. *Medical Education*, 19, 43–47.

Norman, G.R. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education*, 39 (4), 418-427. doi: 10.1111/j.1365-2929.2005.02127.x.

Nouh, T., Boutros, M., Gagnon, R., Reid, S., Leslie, K., Pace, D., …Meterissian, S. (2010). The script concordance test as a measure of clinical reasoning: a national validation study. *American Journal of Surgery*, 203 (4) 530-534. doi 10.1016/j.amjsurg.2011.11.006.

Page, G. et Bordage, G. (1995). The medical coucnil of Canada's key features project : a more valid written examination of clinical decision-making skills. *Academic Med*icine, 70 (2), 104-110.

Peden, N.R., Cairncross, R.G., Harden, R.M. et Crooks, J. (1985). Assessment of clinical competence in therapeutics: the use of the objective structured clinical examination. *Medical Teacher*, 7 (2), 217-223.

Psiuk, T. (2012). Le processus d'apprentissage du raisonnement clinique. Dans L'apprentissage du raisonnement clinique - Concepts fondamentaux, Contexte et processus d'apprentissage. (1e éd., p.111-157). Bruxelles, Belgique : De Boeck Université.

Ramaekers, S., Kremer, W., Pilot, A., Van Beukelen, P. et Van Keulen, H. (2010). Assessment of competence in clinical reasoning and decision-making under uncertainty: the script concordance test method. *Assessment and Evaluation in Higher Education*, 35 (6), 661-673.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Achievement Tests* (Expanded edition, 1980. Chicago: University of Chicago Press ed.). Copenhagen: Danish Institute for Educational Research.

Sibert, L., Charlin, B., Gagnon, R., Corcos, J. et Khalaf, A. (2001). Evaluation du raisonnement clinique en urologie : l'apport du test de concordance de script. *Prog. Urol.*, 11, 1213-1219.

Sibert, L., Charlin, B., Gagnon, R., Corcos, J, Lechevallier J. et Grise P. (2002). Assessment of clinical reasoning competence in urology with the script concordance test: an exploratory study across two sites from different countries. *Eur Urol* 41, 227-233.

Sibert, L., Darmoni, S. J., Dahamna, B., Weber, J. et Charlin, B. (2005). Online clinical reasoning assessment with the script concordance test: A feasibility study. *Biomedical Central Medical Informatics and Decision Making*, 5, 18-28.

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

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

Subra, J., Chicoulaa, B., Stillmunkès, A., Mesthé, P., Oustric, S. et Bugat, M-E. (2017). Reliability and validity of the script concordance test for postgraduate students of general practice. *European Journal of General Practice*, 23, 209-214. doi:10.1080/13814788.2017.1358709.

Tennant, A. et Conaghan, P. G. (2007). The Rasch Measurement Model in Rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis and Rheumatism*, 57 (8), 1358-1362.

Turgeon, J. et Bernatchez, J. (2003). Les données secondaires. Dans Gauthier, B. (dir.), Recherche sociale : de la problématique à la collecte de données (4e éd., p.431-468). Sainte-Foy, Canada : Presses de l'université du Québec.

Vanbelle, S., Massart, V., Giet, D. et Albert, A. (2007). Test de concordance de script : un nouveau mode d'établissement des scores limitant l'effet du hasard. *Pédagogie Médicale*, 8, 71–80.

Vialle, R. (2009). Application du test de concordance de script à une évaluation du raisonnement clinique en orthopédie pédiatrique. Mémoire de diplôme interuniversitaire, Université Pierre et Marie Curie. Récupéré du site : http://www.edu.upmc.fr/medecine/pedagogie/memoire/Memoires_2009_PDF/Memoire_Vialle_2009.pdf.

Wan, S H. (2015). Using the script concordance test to assess clinical reasoning skills in undergraduate and postgraduate medicine. *Hong Kong Med J*, 21, Epub doi: 10.12809/hkmj154572.

Wilson, A. B., Pike, G. R. et Humbert, A. J. (2014). Analyzing script concordance test scoring methods and items by difficulty and type. *Teaching and Learning in Medicine*, 26 (2), 135-145. doi:10.1080/10401334.2014.884464.

Wilson, M., Allen, D. D. et Corser Li, J. (2006). Improving measurement in health education and health behavior research using item response modeling: comparison with the classical test theory approach. *Health Education Research*, 21, i19-i32. doi:10.1093/her/cyl053.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis: Rasch measurement. Chicago: Mesa Press.

Wright, B. D. et Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8* (3), 370-371.

Wright, B. D. et Mok, M. M. (2004). An overview of the family of rasch measurement models, in Introduction to Rasch measurement theory, Models and Applications, J. Everitt and R. M. Smith, Eds.

Yang, S.-C., M.-Y. Tsou, E.-T. Chen, K.-H. Chan et K.-Y. Chang (2011). « Statistical item analysis of the examination in anesthesiology for medical students using the Rasch model », Journal of the Chinese Medical Association, 74(3), p. 125-129, <http://dx.doi.org/10.1016/j.jcma.2011.01.027>, consulté le 16 janvier 2018.

Annexes

Annexe 1 Patron de réponses des experts

Numéro d'items	Réponse modale	Variance	Réponse à 2 cat	Numéro d'items	Réponse modale	Variance	Réponse à 2 cat
1	1	0,63	0	**18**	3	0,63	0
2	4	0,55	0	**19**	2	*1,79	1
3	4	0,42	0	**20**	3	0,27	0
4	1	0,57	0	**21**	4	0,45	0
5	4	0,33	0	**22**	4	*1,15	0
6	1	0,27	0	**23**	3	*1,60	0
7	2	0,27	0	**24**	3	0,39	0
8	3	0,79	0	**25**	4	0,64	0
9	4	0,33	0	**26**	2	0,93	0
10	3	0,61	0	**27**	4	0,15	0
11	4	0,27	0	**28**	4	0,52	0
12	4	0,39	0	**29**	4	0,63	0
13	5	0,88	1	**30**	3	0,52	0
14	3	0,61	0	**31**	3	0,45	0

15	4	0,57	0	**32**	3	0,08	0
16	3	0,24	0	**33**	4	0,81	0
17	2	0,64	0	**34**	2	0,08	0
Numéro d'items	Réponse modale	Variance	Réponse à 2 cat	Numéro d'items	Réponse modale	Variance	Réponse à 2 cat
35	2	0,20	0	**55**	4	0,61	0
36	3	0,81	0	**56**	3	0,42	0
37	3	0,08	0	**57**	2	0,45	0
38	2	0,52	0	**58**	4	*1,24	0
39	2	0,24	0	**59**	4	0,77	0
40	3	0,45	0	**60**	2	*1,25	0
41	4	0,70	0	**61**	4	0,45	0
42	3	0,00	0	**62**	3	0,61	0
43	4	0,24	0	**63**	3	0,52	0
44	2	0,39	0	**64**	2	0,79	0
45	3	0,08	0	**65**	3	*1,33	0
46	3	0,36	0	**66**	3	0,63	0
47	1	*1,48	1	**67**	4	0,24	0
48	2	0,27	0	**68**	3	0,64	0
49	3	0,00	0	**69**	4	0,24	0
50	3	0,08	0	**70**	3	0,33	0

51	3	1,15	0	**71**	2	0,93	0
52	4	0,88	1	**72**	4	0,33	0
53	3	0,75	0	**73**	3	0,24	0
Numéro d'items	Réponse modale	Variance	Réponse à 2 cat	Numéro d'items	Réponse modale	Variance	Réponse à 2 cat
54	3	0,56	0	**74**	4	0,24	0
75	4	0,63	0	**95**	3	0,15	0
76	4	0,75	0	**96**	4	*1,06	0
77	4	0,27	0	**97**	3	0,73	0
78	5	0,27	0	**98**	4	0,45	0
79	4	0,99	3	**99**	4	0,24	0
80	3	0,63	0	**100**	5	0,61	0
81	3	0,70	0	**101**	3	0,45	0
82	2	0,42	0	**102**	2	0,63	0
83	3	0,27	0	**103**	4	0,70	0
84	4	0,52	0	**104**	2	0,27	0
85	2	0,52	0	**105**	3	0,75	0
86	2	0,45	0	**106**	3	0,82	0
87	4	0,39	0	**107**	1	0,42	0
88	3	0,27	0	**108**	4	*1,30	1
89	4	0,27	0	**109**	3	0,63	0

90	5	*1,54	1	**110**	4	0,55	0
91	2	*1,17	0	**111**	2	0,39	0
92	4	0,99	3	**112**	3	0,42	0
Numéro d'items	Réponse modale	Variance	Réponse à 2 cat	Numéro d'items	Réponse modale	Variance	Réponse à 2 cat
93	3	0,45	0	**113**	1	0,20	0
94	3	0,27	0	**114**	4	0,39	0
115	2	0,39	0	**126**	4	*1,27	0
116	4	0,45	0	**127**	2	*1,70	1
117	3	0,36	0	**128**	4	0,73	0
118	1	0,61	0	**129**	2	0,87	0
119	2	0,15	0	**130**	4	0,33	0
120	4	0,45	0	**131**	2	0,20	0
121	2	0,93	0	**132**	4	*1,55	0
122	4	0,75	0	**133**	3	*1,09	0
123	3	0,52	0	**134**	2	0,45	0
124	1	*1,17	1	**135**	4	*1,30	1
125	1	0,45	0	*item avec variance élevée			

Annexe 2 *Corrélation item-total pour les 135 items (Méthode 1)*

	$r \leq 0,10$	$0,10 < r \leq 0,20$	$r > 0,20$
Items	1, 11, 13, 14, 19, 20, 23, 25, 29, 30, 32, 48, 51, 54, 55, 65, 66, 78, 87, 90, 93, 95, 96, 100, 106, 107, 113, 118, 123, 124, 125	4, 6, 8,10, 15, 16, 21, 22, 28, 33, 35, 36, 37, 42, 45, 46, 47, 52, 58, 61, 62, 69, 70, 75, 76, 81, 83, 84, 89, 91, 97, 101, 109, 115, 116, 119, 127, 133	2, 3, 5, 7, 9, 12, 17, 18, 24, 26, 27, 31, 34, 38, 39, 40, 41, 43, 44, 49, 50, 53, 56, 57, 59, 60, 63, 64, 67, 68, 72, 73, 74, 75, 77, 79, 80, 82, 85, 86, 88, 92, 94, 98, 99, 102, 103, 104, 105, 108, 110, 111, 112, 114, 117, 120, 121, 122, 126, 128, 129, 130, 131, 132, 134, 135
Total	31	38	66

Annexe 3 *Corrélation item-total pour les 135 items (M2)*

	$r \leq 0,10$	$0,10 < r \leq 0,20$	$r > 0,20$
	1, 4, 7, 11, 13, 20, 21, 25, 29, 33, 47, **51**, **52**, 54, 55, 57, 58, 66, 78, 87, 90, 95, 96, 100, 107, 113, **115**, 118, 121, **123**, 124, 125, 127, 128	3, 6, **8**, 14, 15, 16, 17, 19, 22, 26, 27, 28, **32**, 35, 36, 39, **43**, 44, 59, 61, 71, **75**, 76, 77, **79**, 86, 91, 97, 101, 102, 103, 106, 108, 116, 120, 122, 126, **130**, 132, 135	2, 5, 9, 10, 12, 18, 23, 24, 30, 31, 34, 37, 38, 40, 41, 42, 45, 46, 48, 49, 50, 53, 56, 60, 62, 63, 64, 65, 67, 68, 69, 70, 72, 73, 74, 80, 81, 82, 83, 84, 85, 88, 89, 92, 93, 94, 98, 99, 104, 105, 109, 110, 111, 112, 114, 117, 119, 129, 131, 133, 134
Total	34	40	61

Annexe 4 *Corrélation item-total des 135 items (M3)*

	$r \leq 0,10$	$0,10 < r \leq 0,20$	$r > 0,20$
Items	2, 3, 13, 14, 16, 20, 21, 22, 23, 24, 25, 26, 28, 31, 41, 43, 48, 51, 54, 55, 58, 59, **61**, 62, 66, 68, 69, 70, 72, 74, 76, 78, 81, 84, 87, **89**, **90**, **93**, 95, 96, 97, 99, 100, 102, 103, 106, 112, 113, 114, 120, 121, 122, 126, 128, 130, 133, 135	1, 6, 7, **8**, 9, 10, 11, 15, 17, 18, 19, 29, 30, 33, **34**, 38, 39, 40, 46, 47, 49, 52, 53, 56, 57, 63, 64, 65, 67, 71, 73, 75, 77, 79, 85, 86, 88, 91, 94, 101, 105, **107**, 108, 109, 115, 116, **118**, 119, **123**, 124, 127, 129, 131, 132, 134	4, 5, 12, 27, 32, 35, 36, 37, 42, 44, 45, 50, 60, 80, 82, 83, 92, 98, 104, 110, 111, 117, 125
Total	57	55	23

Retrait M1

Annexe 5. *Récapitulatif des 8 sujets et des 37 items retirés*

Sujets	**Items**
14, 20, 44, 79, 84, 111, 115, 158	1, 14, 20, 22, 23, 24, 25, 28, 30, 32, 34, 36, 37, 39, 48, 49, 53, 54, 55, 62, 63, 73, 74, 88, 89, 90, 91, 92, 93, 106, 107, 118, 119, 123, 124, 129, 133

Annexe 6. *Récapitulatif des 8 sujets et des 54 items retirés*

Sujets	**Items**
14, 20, 44, 79, 84, 111, 115, 158	1, 8, 11, 14, 20, 22, 23, 24, 25, 28, 30, 31, 32, 34, 36, 37, 39, 42, 48, 49, 50, 53, 54, 55, 56, 58, 62, 63, 65, 67, 73, 74, 81, 85, 88, 89, 90, 91, 92, 93, 94, 95, 100, 106, 107, 112, 118, 119, 122, 123, 124, 127, 129, 133

Annexe 7. *Récapitulatif des 18 sujets et des 60 items retirés*

Sujets	**Items**
8, 10, 14, 16, 20, 44, 68, 79, 80, 84, 101, 111, 115, 131, 133, 158	1, 8, 11, 14, 20, 21, 22, 23, 24, 25, 28, 30, 31, 32, 34, 36, 37, 39, 42, 45, 48, 49, 50, 53, 54, 55, 56, 58, 62, 63, 65, 67, 73, 74, 78, 81, 83, 85, 88, 89, 90, 91, 92, 93, 94, 95, 100, 106, 107, 109, 111, 112, 118, 119, 122, 123, 124, 127, 129, 133

Annexe 8. *Récapitulatif des 21 sujets et des 60 items retirés*

Sujets	**Items**
8, 10, 14, 16, 17, 20, 23, 44, 68, 79, 80, 84, 88, 109, 101, 111, 115, 131, 133, 154, 158	1, 8, 11, 14, 20, 21, 22, 23, 24, 25, 28, 30, 31, 32, 34, 36, 37, 39, 42, 45, 48, 49, 50, 53, 54, 55, 56, 58, 62, 63, 65, 67, 73, 74, 78, 81, 83, 85, 88, 89, 90, 91, 92, 93, 94, 95, 100, 106, 107, 109, 111, 112, 118, 119, 122, 123, 124, 127, 129, 133

M2 Retrait

Retrait de 21 sujets

Nous avons retiré 21 sujets [2, 3, 5, 16, 22, 25, 26, 27, 36, 83, 84, 88, 89, 101, 109, 111, 118, 122, 124, 142, 149].

Annexe 9. *Récapitulatif des 10 sujets et des 23 items*

Sujets	Items
22, 25, 36, 59, 84, 89, 118, 122, 124, 149	1, 6, 14, 15, 19, 30, 39, 41, 56, 57, 58, 62, 74, 75, 78, 80, 111, 112, 117, 119, 122, 123, 125

Annexe 10. *Récapitulatif des 21 sujets et 61 items retirés*

Sujets	Items
2, 3, 5, 16, 22, 25, 26, 27, 36, 83, 84, 88, 89, 101, 109, 111, 118, 122, 124, 142, 149	1, 4, 6, 8, 13, 14, 15, 17, 18, 19, 20, 22, 24, 30, 37, 38, 39, 41, 46, 47, 48, 52, 54, 56, 57, 58, 61, 62, 64, 65, 66, 67, 70, 71, 72, 73, 74, 75, 76, 78, 80, 81, 86, 92, 99, 104, 105, 106, 110, 111, 112, 115, 118, 119, 122, 123, 125, 126, 131, 132, 135

Annexe 11. *Comparaison multiples entre M1, M2 et M3*

(I) Methode		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
M1	M2	-4,7490*	.56031	.000	-6.0663	-3.4316
	M3	7,7786*	.56031	.000	6.4613	9.0959
M2	M1	4,7490*	.56031	.000	3.4316	6.0663
	M3	12,5276*	.56031	.000	11.2103	13.8449
M3	M1	-7,7786*	.56031	.000	-9.0959	-6.4613
	M2	-12,5276*	.56031	.000	-13.8449	-11.2103

M3 Retrait

Nous avons fait un retrait de 9 items [5, 12, 20, 54, 69, 81, 98, 112 et 114].