Université de Montréal

L'évaluation du raisonnement clinique des résidents en hématologie par l'approche de concordance de script

par
Alain Bestawros

Faculté des études supérieures

Mémoire présenté à la Faculté des Sciences de L'Éducation
en vue de l'obtention du grade de M.A.
en Sciences de l'Éducation
option Pédagogie universitaire des sciences médicales

Juillet 2013

Université de Montréal
Faculté des études supérieures et postdoctorales

Ce mémoire intitulé :

L'évaluation du raisonnement clinique des résidents en hématologie par l'approche de concordance de script

Présenté par :
Alain Bestawros

a été évalué par un jury composé des personnes suivantes :

Jean-Guy Blais, président-rapporteur
Dr. Bernard Charlin, directeur de recherche
Dr. Jeannine Kassis, membre du jury

Résumé

La pratique de l'hématologie, comme celle de toute profession, implique l'acquisition d'un raisonnement adéquat. Se basant sur une théorie de psychologie cognitive, le test de concordance de script (TCS) a été développé et validé comme un instrument permettant d'évaluer le raisonnement clinique dans diverses spécialités médicales. Le but de cette étude était d'examiner l'utilité et les paramètres psychométriques d'un TCS en hématologie. Nous avons construit un TCS composé de 60 questions que nous avons administré à 15 résidents juniors (R1 à R3 en médecine interne), 46 résidents séniors (R4, R5 et R6 en hématologie) et 17 hématologues à travers le Canada. Après optimisation, le TCS comptait 51 questions. Sa consistance interne mesurée par le coefficient de Cronbach alpha était 0.83. Le test était en mesure de discriminer entre les résidents selon leur niveau de formation. Les questions contenant des images (n=10) semblaient avoir un potentiel discriminatoire plus élevé. Les scores obtenus par les résidents séniors corrélaient modéremment avec ceux obtenus à un test conventionnel d'hématologie composé de questions à choix multiples et à courte réponse (r de Pearson = 0.42; p=0.02). Le TCS a été complété en 36 minutes en moyenne et a été bien reçu par les participants. Le TCS est un instrument d'évaluation utile et valide en hématologie. Il peut être utilisé à des fins formatives en aidant au suivi de la progression des résidents. Il pourrait aussi être combiné à d'autres instruments d'évaluation à des fins sanctionnelles, ou encore, en éducation médicale continue.

Mots-clés : raisonnement clinique, jugement clinique, évaluation, résidents, étudiants, hématologie, morphologie, test de concordance de script (TCS)

Abstract

The practice of hematology, like any other profession, requires the acquisition of adequate judgment. Based on cognitive psychology theory, the script concordance test (SCT) has been developed and validated as an instrument capable of evaluation clinical judgement in various medical specialties. The goal of this study was to examine the usefulness and the psychometric qualities of the SCT in hematology. We constructed a SCT composed of 60 questions and we administered it to 15 junior residents (R1 to R3 in internal medicine), 46 senior residents (R4, R5 and R6) and 17 hematologists from across Canada. After item optimization, the test comprised 51 questions. Its internal consistency measured by Cronbach alpha was 0.83. The test was able to discriminate between residents according to their year of training. Questions containing an image (n=10) seemed to offer a stronger discriminative potential. Scores obtained by the senior residents correlated moderately with those obtained on a conventional hematology exam made of multiple choice questions and short-answers (Pearson r: 0.42; p=0.02). The SCT was completed in an average of 36 minutes and was well received by participants. The SCT is a useful and valid evaluation instrument in hematology. It may be used during training to monitor resident progression. It may also be combined to other evaluation tools and used for summative purposes or in continuing medical education.

Keywords: clinical reasoning, clinical judgment, assessment, residents, medical students, hematology, morphology, script concordance test (SCT)

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et des abréviations

CRMCC (RCPSC): Collège Royal des Médecins et Chirurgiens du Canada (Royal College of Physicians and Surgeons of Canada)

EBM : evidence-based medicine (médecine fondée sur les preuves)

ECOS (OSCE) : Examen clinique objectif et structuré (Objective Standardized Clinical Examination)

QCM (MCQ) : Questions à choix multiples (multiple-choice qusetions)

SAMP: Short Answer Management Problems

SOO: Simulated Office Orals

Remerciements

En plus des co-auteurs de l'article ci-joint, je tiens à remercier plusieurs personnes sans lesquelles ce mémoire n'aurait pas vu voir le jour. En particulier, Dr. Driss Kazi-Tani (MD, maîtrise en sciences informatiques), Dr. Robert Gagnon (psychologue et méthodologiste), et Mme Émilie Noyeau (documentariste).

Ce travail a été supporté par une subvention du Département de Médecine et une subvention du Fonds des professionnels.

Introduction

La pédagogie médicale

La médecine et la pédagogie sont intimement liées. L'origine-même du terme *docteur* vient du verbe latin *docere* qui veut dire « enseigner »[1]. La pédagogie médicale est une science relativement jeune dans la francophonie, mais qui est devenue au cours des dernières décennies une préoccupation importante. L'évolution de cette science a été grandement influencée par divers facteurs, notamment certains courants sociaux, les progrès en psychologie cognitive et les avancées technologiques dans les sciences de la santé.

Des changements sociaux majeurs et des développements en psychologie débutant dans les années 1970 ont établi les fondements de l'école cognitive en pédagogie. Bien qu'une discussion approfondie de la nature de ces changements dépasse la portée de ce travail[†], il importe de comprendre que l'école cognitive s'intéresse surtout au *comment* de l'apprentissage [4]. Par opposition à l'école béhaviorale qui l'a précédée, l'école cognitive consacre un rôle central et actif à l'apprenant. La pensée critique et le raisonnement sont mis de l'avant, et les mécanismes cognitifs qui les sous-tendent sont étudiées et utilisées pour développer de nouvelles méthodes d'apprentissages et d'évaluation.

C'est en partie dans cette perspective que le Collège Royal des Médecins et Chirurgiens du Canada (CRMCC) a élaboré dans les années 1990 un cadre éducatif fondé sur l'acquisition de compétences [5]. Ce cadre décrit sept rôles menant à des soins de santé optimaux : expert médical, communicateur, collaborateur, gestionnaire, promoteur de la santé, érudit et professionnel. Tel qu'illustré dans le schéma CANMEDS, le rôle d'expert médical constitue le point de convergence de l'activité du médecin et s'appuie sur les autres

[†] Consulter ces ouvrages pour une révision détaillée de l'influence des courants sociaux sur les théories pédagogiques en sciences de la santé : 2. Torre DM, Daley BJ, Sebastian JL, Elnicki DM. Overview of current learning theories for medical educators. *Am J Med* 2006; 119: 903-7, 3. Vienneau R. *Apprentissage et enseignement*. Gaëtan Morin, 2005.

compétences. Ce cadre sera par la suite intégré aux normes d'agrément, aux objectifs de formation, aux évaluations en fin de formations, aux modèles d'examens et au programme d'éducation continue.

Des travaux antérieurs à ceux du CRMCC avaient tenté de mieux comprendre les étapes d'acquisition d'une compétence. En 1990, Miller avait proposé un modèle conceptuel qui illustre les étapes d'acquisition d'une compétence sous forme de pyramide (Figure 1) [6]. A la base de la pyramide se retrouve le *savoir (knows)* des connaissances factuelles, suivi du *savoir comment (knows how)* appliquer ces connaissances. Les deux étapes ultérieures sont la capacité de *montrer comment (shows how)* et *d'accomplir (does)* la tâche elle-même.

L'application du modèle proposé par Miller à la compétence centrale qu'est celle de l'expertise médicale nous permet de conclure ceci : après avoir acquis des connaissances médicales *(knows)*, le médecin doit être en mesure d'appliquer correctement ces connaissances *(knows how)* au contexte clinique. L'utilisation d'une structure pyramidale illustre que l'acquisition de chaque étape est pré-requise au passage à l'étape subséquente. Il est donc essentiel que tout curriculum médical s'attarde en tout premier lieu à l'acquisition et l'application des connaissances puisque ce sont ces capacités qui serviront de fondements aux autres.

Ces deux étapes à la base de la pyramide – le savoir et le savoir comment - sont généralement évaluées par des épreuves écrites de différents formats. Plusieurs instruments ont été développés dans ce but, dont notamment les questions à choix multiples (QCM) et les questions à courte et longue réponse. Avant de s'attarder aux forces et aux faiblesses de chacun de ces instruments d'évaluation, revoyons d'abord l'importance de l'évaluation elle-même dans un curriculum médical.

« Une bonne partie du travail scolaire est influencée
par la perspective de tels examens et …
il est même gravement déformé
par cette préoccupation devenue dominante. »
- Piaget [7]

L'importance de l'évaluation

En pédagogie médicale, l'évaluation sert d'abord et avant tout à quantifier ou qualifier de manière objective les compétences et connaissances acquises. Cela devrait, du moins en théorie, être prédictif de la compétence clinique future de l'étudiant. L'évaluation peut être utilisée à des fins formatives, sanctionnelles ou simplement comme source de motivation pour les apprenants [8]. Elle détermine en grande partie ce que les étudiants apprendront [9, 10]. Cela peut être attribué principalement à deux facteurs. D'une part, étant souvent surchargés de travail et d'étude, les étudiants mettent naturellement l'accent sur les parties du curriculum qui seront évaluées. D'autre part, les parties du curriculum qui ne sont pas évaluées peuvent être perçues comme étant moins importantes ou moins pertinentes à la pratique médicale. Malgré l'impact que l'évaluation peut avoir sur l'apprentissage, il s'avère qu'on préfère parfois évaluer ce qui facilement évaluable plutôt que ce qui est difficilement évaluable mais plus pertinent [11, 12]. On peut aisément s'imaginer qu'une telle approche, si appliquée systématiquement, pourrait avoir un impact délétère significatif sur l'apprentissage.

Les connaissances factuelles sont facilement évaluables et peuvent refléter, du moins jusqu'à un certain degré, les compétences cliniques. Il n'est donc pas étonnant de constater qu'elles sont évaluées par divers instruments largement utilisés par les facultés de médecine. Or, on peut se questionner à savoir si l'acquisition de connaissances factuelles en soi est gage d'un réel apprentissage et d'une compétence véritable.

L'apprentissage est défini comme étant « un changement relativement permanent dans le potentiel de comportement de l'apprenant dû à l'expérience. » [3] Pour qu'il y ait

apprentissage, donc, les connaissances ne suffisent pas. Elles doivent mener à un changement dans le comportement de l'apprenant. C'est donc dire que c'est la formation plutôt que l'information qui compte. C'est à travers l'usage et le raisonnement que l'apprenant fait avec ses connaissances que son comportement en est changé. Un bon instrument d'évaluation est donc celui qui mesure le comportement de l'apprenant, à savoir sa capacité à appliquer ses connaissances à des situations cliniques. Il s'agit en d'autres mots de mesurer sa capacité à raisonner. Est-ce que les méthodes d'évaluation conventionnelles le permettent?

Les instruments d'évaluation conventionnels

Étant donné les multiples dimensions de chaque compétence et la variété des domaines de la médecine, il n'existe pas de méthode d'évaluation parfaite. En ce qui concerne la compétence d'expertise du clinicien, plusieurs méthodes d'évaluations existent et ont été plus ou moins validés à certains égards. [8] Nous allons maintenant nous attarder brièvement aux caractéristiques, ainsi qu'aux forces et aux faiblesses, des méthodes d'évaluation les plus communes : les questions à choix multiples (QCM), les questions à réponse longue, les examens oraux et les examens cliniques objectifs structurés (ECOS).

Les questions à choix multiples (QCM)

Les principaux avantages des QCM sont : i) leur haut degré de fidélité qui tient au fait qu'elles permettent de mesurer de larges domaines de connaissances; ii) qu'elles peuvent être administrées à de grands groupes d'étudiants simultanément et avec peu de ressources; et iii) qu'elles sont corrigées de façon objective. De plus, les QCM possèdent une bonne validité prédictive en terme de performance des étudiants dans la pratique future [13]. Ces nombreux avantages expliquent sans doute leur usage ubiquitaire.

D'un autre côté, l'usage des QCM pose certains problèmes. D'abord, leur élaboration requiert un certain effort afin de s'assurer qu'il n'y ait pas d'ambiguïté ni d'indice clair

quant à la réponse recherchée. De plus, selon le contexte et l'envergure du contenu à évaluer, il a été estimé qu'il peut falloir jusqu'à 4 heures de temps d'examen pour assurer la fiabilité d'une épreuve de QCM. [14] Aussi, elles ont été critiquées pour l'emphase qu'elles mettent sur l'acquisition superficielle de simples connaissances factuelles mémorisées. Cette emphase se reflète sur l'apprentissage et peut mener à ce qu'on a appelé *l'incompétence dissimulée* (hidden incompetence) [15]. En d'autres mots, une bonne performance à un examen de QCM peut cacher une intégration déficiente et ainsi mal prédire la compétence réelle du candidat. Suivant la même logique, il n'est pas étonnant que les QCM offrent une faible capacité de discrimination entre les apprenants et les experts [16].

Dans le but de pallier à certaines de ces critiques, le QCM à contexte riche a été développé [17]. Le principe consistant à choisir une bonne réponse parmi un choix de réponses persiste, mais la vignette proposée est plus élaborée et contient plus de détails. Ainsi le QCM à contexte riche tend à évaluer les fonctions cognitives supérieures et non la simple mémorisation. Comme pour le QCM simple, une seule bonne réponse est possible.

Dans le même ordre d'idée et dans le but de pallier aux mêmes lacunes attribuées au QCM simple, Bordage et Page ont proposé un format de questions appelé "key features." Le principe est d'identifier une étape critique (ou "key feature") dans la résolution d'un problème difficile et d'y mettre l'emphase, augmentant ainsi la capacité discriminative du test et les domaines sondés en un temps donné. Ce format offre une approche flexible quant au nombre de réponses possibles et à l'attribution des scores. Bien que le candidat doive démontrer qu'il a su intégrer toutes les informations contenues dans la vignette, c'est toujours le résultat du processus et non le processus lui-même qui est évalué. Certaines lacunes associées à ces instruments d'évaluation persistent donc.

Les questions à réponse longue

L'avantage principal de cette méthode d'évaluation est de permettre de sonder jusqu'à un certain degré le processus de raisonnement du candidat. Cependant, étant donné les ressources nécessaires à sa correction (temps et correcteurs), un examen composé de questions à réponse longue peut difficilement couvrir suffisamment de domaines de connaissances pour être fiable. De plus, la fiabilité inter-juges est basse et l'objectivité est parfois manquante même avec l'usage de grille de correction [18].

Les examens oraux

L'examen oral a été souvent prisé pour sa capacité de mesurer plusieurs compétences simultanément. On peut évaluer les connaissances factuelles, sonder le raisonnement clinique du candidat et même des compétences et attributs personnels difficiles à mesurer autrement tel que la tolérance au stress, la confiance en soi, etc. Des études ont également démontré que l'examen oral possède une bonne validité prédictive des capacités cliniques futures [19].

Cependant, l'examen oral n'est pas sans défauts. D'abord, sa fiabilité peut être compromise par des examinateurs qui, étant simplement humains, peuvent parfois présenter des biais et donc manquer d'objectivité, même avec l'usage d'une grille de correction [18] [20]. D'autre part, le large déploiement d'un tel examen requiert énormément de ressources (correcteurs et temps). Enfin, de même que pour les questions à réponse longue, il est souvent difficile de sonder suffisamment de domaines de connaissances pour atteindre une fiabilité adéquate.

ECOS

Mise au point par Ronald Harden à la fin des années 1970, l'ECOS est un examen composé de plusieurs stations utilisant des patients, réels ou simulés, et visant à évaluer les habiletés, les attitudes, ainsi que les aspects cognitifs d'un candidat pour une discipline donnée [21].
L'ECOS permet de mesurer certaines habiletés que d'autres instruments d'évaluation ne peuvent pas sonder. Celles-ci incluent le recueil de données cliniques, l'examen physique et

les habiletés techniques. L'ECOS est cependant mal adapté à l'évaluation du raisonnement clinique puisque l'observation ne porte que sur des comportements observables. Les processus cognitifs qui prennent place dans l'esprit du candidat ne sont pas directement sondés par les examinateurs. De plus, afin d'être standardisé, le score est attribué en fonction d'une liste d'item ou de gestes que le candidat doit accomplir. Or, cette standardisation se fait à partir d'une certaine logique ou approche que les constructeurs du test pensent appropriée, mais qui n'est pas nécessairement celle utilisée par tous les experts. A ce titre, une étude effectuée à Toronto en 1999 a démontré que les cliniciens avec le plus d'expérience obtiennent des scores plus bas sur les ECOS que les résidents ou étudiants [22]. Des listes d'items ne constituent donc peut-être pas des mesures valides de la compétence clinique.

De plus, étant donné les coûts et ressources associés à ce genre d'évaluation, le nombre de cas reste limité et restreint ainsi la fiabilité de l'ECOS. Il a été démontré que la performance des médecins varie grandement d'un cas à un autre [23]. La fiabilité des ECOS peut atteindre 80% mais cela requiert un minimum de 4 heures d'examen, et au moins une vingtaine de stations [13].

Chaque méthode d'évaluation a donc des avantages et des inconvénients, et aucune ne semble parfaite. Une des caractéristiques principales d'un bon instrument d'évaluation est sa capacité à évaluer réellement ce qu'on désire évaluer ; c'est ce qu'on appelle la validité de construit. Or, plusieurs des méthodes d'évaluation classiques dont on a discutés plus haut ont démontré un effet surprenant : les scores obtenus par des experts sont inférieurs ou comparables à ceux obtenus par des résidents en formation [24, 25]. Cette observation évoque un doute sérieux quant à leur validité de construit. Cet effet paradoxal a été attribué à l'hypothèse voulant que plusieurs de ces instruments mesurent les connaissances factuelles sans évaluer la capacité des apprenants de les utiliser pour raisonner. Il est donc possible que ces méthodes d'évaluation ne reflètent pas réellement ce que les experts font

en pratique. Cette entité, qui semble échapper du moins en partie aux instruments d'évaluations conventionnels pourrait bien être le raisonnement clinique.

Le raisonnement clinique

Les termes *raisonnement clinique* et *jugement clinique* sont souvent utilisés de façon interchangeable dans la littérature, ce qui ne facilite pas leurs définitions respectives [9, 26, 27]. Le concept sous-jacent cependant pourrait être défini comme étant « le processus cognitif utilisé par le clinicien afin de confirmer et/ou d'infirmer des hypothèses diagnostiques et thérapeutiques. » [28] (traduction libre de l'anglais). Ou encore, le raisonnement clinique a été défini comme étant « l'activité intellectuelle qui synthétise l'information obtenue à partir de la situation clinique, qui l'intègre aux connaissances et aux expériences antérieures et l'utilise pour prendre des décisions de diagnostic et de prise en charge du patient» [29, 30].

L'acquisition de connaissances est nécessaire mais non suffisante au raisonnement clinique. Bien raisonner c'est savoir appliquer les bonnes connaissances dans le bon contexte chez le bon patient et au bon moment. Le raisonnement clinique comporte donc notamment la capacité à intégrer les données, à générer des hypothèses pertinentes, et à décider du poids à attribuer à chaque donnée. Il fait partie de la compétence d'expertise et en constitue même la pierre angulaire [31].

Le raisonnement clinique implique bien plus que l'application d'un ensemble de règles ou de principes à une situation clinique donnée [32]. Il implique l'intégration des diverses connaissances et de l'expérience, et la pondération de toutes les données. Schön a proposé de faire la distinction entre deux types de situations cliniques auxquelles font face les cliniciens [33] :

i) Présentation clinique claire

Dans le premier cas, le clinicien confronté à un problème clinique possède toutes les données nécessaires afin d'établir le diagnostic et le plan de traitement. Le problème clinique est bien défini et les données sont complètes et non-équivoques. Le clinicien peut alors appliquer un ensemble de règles claires pour résoudre ce problème bien défini. Ce type de raisonnement, qui fait appel à ce que Schön nomme la *rationalité technique*, fait partie du raisonnement clinique du professionnel. Il peut être enseigné et évalué par des méthodes classiques d'évaluation telle que les QCM. Ce type de présentation clinique cependant ne représente que la minorité des cas rencontrés par les cliniciens dans leur pratique.

ii) Présentation clinique ambiguë

Ces problèmes représentent la majorité des cas rencontrés en clinique. Les données nécessaires pour résoudre le problème peuvent être manquantes, ambiguës, imprécises ou encore inconsistantes. En d'autres mots, elles sont marqués par de l'incertitude. Schön a proposé que la capacité de raisonner dans ces contextes d'incertitude constitue la pierre angulaire de la compétence clinique [33]. Ce deuxième genre de raisonnement est plus difficile à enseigner et à évaluer. Or il y a intérêt à suivre son développement non seulement parce qu'il est plus pertinent cliniquement, mais aussi parce que c'est l'incertitude qui amène les apprenants à abandonner un raisonnement automatique en faveur d'un raisonnement plus analytique [34].

Medicine is a science of uncertainty
and an art of probability
- Osler [35]

Évaluer le raisonnement clinique en contexte d'incertitude

Les progrès scientifiques ont certes permis une certaine standardisation de la médecine. La médecine moderne est marquée par de vastes études scientifiques qui ont mené à l'identification des meilleures lignes de conduite dans plusieurs situations cliniques. C'est ce qu'on appelle la médecine fondée sur la preuve (*evidence-based medicine)*. Bien que les conclusions qui émanent de recherches scientifiques puissent constituer des lignes de conduites, ces dernières demeurent insuffisantes pour deux raisons principales. D'abord, plusieurs questions cliniques, en raison de leur rareté ou de leur complexité, n'ont pas fait l'objet d'études, ou du moins d'études de bonne qualité. Deuxièmement, même en présence de ligne de conduites claires, il arrive que la situation clinique soit ambiguë ou que des informations soient manquantes ou encore que le patient ait un profil quelque peu différent de ceux étudiés. Les études scientifiques et les conseils d'experts n'adressent donc qu'une fraction de toutes les éventualités possibles en clinique. De l'adaptation, voire de l'improvisation à l'occasion, s'avère inévitable. Et on ne saurait le faire sans un bon raisonnement clinique. L'enseignement de cette gestion de l'incertitude serait déficient dans plusieurs curriculums médicaux [36].

Le raisonnement clinique doit donc faire partie d'une évaluation continue en cours de formation. Son évaluation a non seulement pour effet de promouvoir son apprentissage et son développement chez les étudiants et résidents mais aussi à permettre aux apprenants eux-mêmes ainsi qu'aux responsables du curriculum d'identifier les forces et les faiblesses du programme, et ainsi de mieux orienter les apprentissages.

Plusieurs problèmes se posent lorsqu'on tente d'évaluer le raisonnement clinique en contexte d'incertitude.

D'abord, l'élaboration d'examens contenant des problèmes cliniques mal définis amène des difficultés en matière de correction. D'une part, l'obtention du consensus des correcteurs pour l'inclusion d'une question dans un examen élimine de façon systématique les questions contenant un certain degré d'incertitude. D'autre part, la correction de questions avec des réponses variables conduit à des problématiques d'objectivité et de ressources dont on a discuté plus haut (ex : questions à réponse longue).

Deuxièmement, l'évaluation du jugement clinique en général, et d'autant plus en contexte d'incertitude, est compliquée par le fait que même dans des situations cliniques similaires, les experts ne collectent pas tous la même information et ne suivent pas nécessairement le même processus cognitif ; et ceci, même s'ils arrivent au même diagnostique [37, 38]. En contexte d'incertitude, il n'existe donc pas de parcours cognitif optimal que l'on peut enseigner et encore moins évaluer. De plus, même si l'on identifie les étapes essentielles du raisonnement clinique pour un problème donné, la difficulté de la pondération demeure. On doit attribuer le crédit approprié à chaque étape du raisonnement; raisonnement sur lequel les experts ne s'entendent pas nécessairement.

Troisièmement, tout comme dans la plupart des situations cliniques réelles, il n'existe pas une seule et unique « bonne réponse ». Non seulement les experts ne s'entendent pas toujours sur la meilleure conduite à suivre, mais plus d'une réponse peuvent être bonnes pour autant qu'elles émanent et soient supportées par un raisonnement adéquat. Il n'est donc pas pédagogique - ni équitable! - de demander aux étudiants de donner la « bonne réponse ». C'est pourtant le cas pour plusieurs instruments d'évaluation utilisés couramment, dont les QCM et les ECOS. Dans ce genre d'examens, les bonnes réponses sont établies par un consensus d'experts. Pour pallier à cette difficulté, des auteurs ont proposé une méthode d'agrégation des scores à laquelle nous reviendrons plus tard [39, 40].

L'évaluation du raisonnement clinique en contexte d'incertitude amène d'autres difficultés décrites dans la littérature. Par exemple, dans des questions où l'on présente au candidat une sélection limitée de réponses possibles, le candidat peut se retrouver à reconnaître la bonne réponse plutôt qu'à la générer. C'est ce que l'on a nommé *l'effet d'indice* (cueing effect) [38]. Une étude a d'ailleurs démontré que les étudiants obtiennent un score inférieur lorsque le même contenu est évalué par des questions qui proposent les réponses que par des questions où la réponse doit être générée [41].

L'évaluation du raisonnement clinique en contexte d'incertitude s'avère donc épineuse. D'une part, les méthodes d'évaluation conventionnelles semblent mal adaptées pour le faire, et d'autre part, la création d'un nouvel instrument d'évaluation doit surmonter plusieurs défis. L'instrument idéal doit permettre d'évaluer de façon objective le raisonnement clinique, en plus d'être facile à concevoir, administrer et corriger.

L'approche par concordance de script

C'est justement dans le but de concevoir un tel instrument que Charlin et al. ont développé une approche basée sur la concordance de script [32, 42]. L'idée de base était de développer un instrument capable d'évaluer objectivement le raisonnement clinique dans des contextes d'incertitude. Cet instrument devait posséder une bonne fiabilité et validité et devait être facile à construire, à administrer et à corriger. C'est en se basant sur la théorie des scripts que ces auteurs ont développé le test de concordance de script (TCS).

Concordance de script : la théorie

Plusieurs modèles théoriques ont été proposés pour expliquer le processus cognitif du raisonnement clinique, mais aucun n'a été unanimement accepté [30]. La théorie hypothético-déductive proposée par des chercheurs en psychologie cognitive stipule que le clinicien expérimenté génère des hypothèses dès qu'il est exposé à des données cliniques.

S'en suit alors un processus cognitif, que l'on pourrait appeler le raisonnement clinique, pendant lequel chacune des hypothèses générées est confirmée ou infirmée selon les données présentées. Chaque donnée est sujette à interprétation et le clinicien doit lui attribuer le poids approprié.

La qualité de ce processus dépend des représentations mentales que les experts se font des diverses maladies et de leurs présentations. Feltovitch et Barrows ont proposé la *théorie des scripts* pour expliquer la structure des connaissances médicales utilisée par l'expert médical afin d'arriver à un diagnostic clinique [43, 44]. Cette théorie suppose l'existence d'un réseau de connaissances pré-existantes et d'un spectre de valeurs qui sont jugées acceptables ou inacceptables pour chaque diagnostic [42]. Ainsi, lorsqu'un expert est confronté à une nouvelle présentation clinique, il collecte diverses données qui peuvent faire partie du questionnaire médical, de l'examen physique, ou des tests de laboratoire ou des épreuves radiologiques. Ces données activent des réseaux de connaissances préétablies qui permettent de comparer ces informations aux représentations mentales que le clinicien a de chacune des maladies. Ces représentations ont été appelées « illness scripts » [44].

La qualité et le raffinement de ces scripts augmentent avec l'expérience clinique et dépendent de la quantité, mais aussi et surtout de l'organisation de l'information détenue par le clinicien. L'activation de ces scripts dans l'esprit d'un expert est automatique et en grande partie inconsciente (*script triggering*) [45]. Charlin et al. ont proposé que les scripts fonctionnent par des associations de mémoire plutôt que des déductions causales, ce qui permet l'activation et l'application rapide des scripts [42, 46, 47]. L'expert se retrouve donc à comparer la présentation du présent patient à une banque de patients antérieurs et à établir les similitudes [26]. Le raisonnement se fait par séquence : une hypothèse est générée puis elle est testée en fonction des données disponibles. Si elle est rejetée, une autre hypothèse est générée et ainsi de suite. L'expérience du clinicien est donc directement reliée à la qualité et précision des scripts qu'il utilise. Les scripts sont dynamiques et peuvent se raffiner avec l'expérience du clinicien. Il serait donc possible de promouvoir leur

développement et de les évaluer en cours de formation, et voire même dans un contexte d'éducation continue.

Le script de chaque expert est particulier puisque chacun a son expérience propre. Cependant, les scripts des médecins expérimentés se ressemblent ; la preuve étant que confrontés au même patient, ils atteignent généralement le même diagnostic. Comme il sera détaillé dans la section Méthodologie, cette différence entre les scripts qu'on observe parmi les experts est capturée dans le processus de correction du TCS. Ainsi, il n'existe pas une seule bonne réponse. La méthode d'agrégation des scores permet plus de discrimination dans la détection du degré d'expertise [48].

La formation médicale serait donc une période au cours de laquelle les apprenants développent et raffinent leurs scripts jusqu'à ce qu'ils s'apparentent à ceux des experts dans le domaine. D'ailleurs, les apprenants rapportent que c'est en écoutant les cliniciens réfléchir au problème du patient – donc raisonner – qu'ils apprennent le plus [49, 50]. Les mentors qui rendent verbalement explicite leur raisonnement clinique sont les plus appréciés [51]. Les meilleurs enseignants exposent les ambiguïtés et les incertitudes entourant la question clinique ainsi que le cheminement cognitif utilisé pour en tenir compte tout en arrivant à une issue satisfaisante [36].

Test de concordance de script : les principes

La structure du TCS sera exposée plus en détails dans la section Méthodologie. Globalement, l'idée est de comparer le processus de raisonnement et l'organisation des connaissances (les scripts) des apprenants à celui d'un panel d'experts dans le domaine [52]. Les questions sont tirées de situations cliniques réelles. Elles débutent par une vignette clinique. Des hypothèses diagnostiques et/ou thérapeutiques sont émises. Une nouvelle information est ajoutée, puis on teste l'impact de cette information sur l'hypothèse. Étant donné que l'impact est qualitatif, une échelle apte à reproduire cette variabilité (échelle de Likert) est utilisée. Les résultats des apprenants sont comparés à ceux du panel de référence

en utilisant une méthode d'agrégation des scores. Ce test vise donc à évaluer le processus de raisonnement plutôt que l'issue du raisonnement [29].

Test de concordance de script : la littérature

Le TCS a déjà fait l'objet de nombreuses études dans plusieurs domaines des sciences de la santé, les premières études datant de 1998 [53, 54]. Dans la plupart de ces études, il a démontré plusieurs qualités, dont une bonne consistance interne tel que mesurée par l'alpha de Cronbach et a été en général très bien reçu par les participants (Tableau 1).

En médecine interne, un champ similaire à l'hématologie, le TCS a été évalué par Marie et al. [55] Un total de 95 items ont été soumis à 4 groupes composés d'étudiants en médecine, de résidents, d'internes et d'internistes. Le test s'est avéré très discriminatoire entre les groupes et avait un bon coefficient de consistance interne (alpha de Cronbach : 0,81). Le test avait été également bien reçu par les participants.

Le TCS a également été testé et validé en urologie par Sibert el al [56-58]. Une des études avait la particularité de comparer les scores d'étudiants canadiens et d'étudiants français en fonction de la provenance des experts. Les auteurs ont rapporté que les scores étaient plus élevés lorsque les réponses des étudiants étaient comparées à celles d'experts provenant du même pays. En d'autres mots, les étudiants apprennent à raisonner tel que leurs enseignants raisonnent, ce qui représente un argument de validité en faveur du TCS et qui aussi nous rassure à l'effet que le raisonnement peut être enseigné. Il est intéressant de noter que peu importe la provenance du panel d'experts auquel les réponses des étudiants sont comparées, leurs rangs demeurent identiques. Une telle observation renforce les fondements du TCS.

Le TCS a également été utilisé en radiologie, domaine où la perception et l'interprétation d'images cliniques sont essentielles. Cela peut s'apparenter en hématologie à l'interprétation des frottis sanguins et des aspirations médullaires. Une étude de Brazeau-Lamontagne et al. a testé le TCS sur des étudiants, des résidents juniors, des résidents

séniors et des experts radiologues de deux différents départements [59]. Les participants ont répondu à un total de 183 items, dont 145 évaluaient les capacités d'interprétation et 38 les capacités de perception. Les auteurs ont trouvé que les deux capacités, interprétative et perceptuelle, progressent pendant la formation en spécialité, mais l'habileté de perception se développe plus tôt que celle d'interprétation. L'habileté de perception était une meilleure prédicatrice du niveau de formation. Le test avait une bonne consistance interne (alpha de Cronbach : 0.79 et 0.81 pour l'interprétation et la perception respectivement), et il permettait de discriminer facilement entre chacun des groupes.

Le TCS a également été utilisé dans d'autres domaines médicaux, tel qu'en chirurgie où il s'est avéré utile pour l'évaluation des décisions intraopératoires [60, 61]. Des résultats similaires ont été rapportées en médecine d'urgence [16, 62], en neurochirurgie [63], en otorhinolaryngologie [64] et en dermatologie [65]. En radio-oncologie, Lambert et al. ont démontré qu'un TCS composé de 70 questions pouvait discriminer adéquatement entre les résidents et les étudiants et que la consistance interne étaient excellente (alpha de Cronbach : 0.90) [66]. Similairement, en neurologie, Lubarsky et al. ont démontré qu'un TCS composé de 94 items permettait de discriminer entre les résidents juniors et séniors, et qu'il était très bien reçu [67]. En chirurgie gynécologique, Park et al. ont étudié le TCS chez 75 résidents et ont confirmé sa capacité discriminative et une consistance interne acceptable (alpha de Cronbach : 0.73) [68].

Le TCS a trouvé application dans d'autres domaines des sciences de la santé. Une étude thaïlandaise a conclu qu'un TCS de 31 items pouvait discriminer entre des étudiants en pharmacie et des pharmaciens quant à la prise en charge pharmacologique du diabète [69]. Une étude similaire a également été menée auprès d'étudiantes en sciences infirmières [70]. Goulet et al. ont démontré que le TCS pouvait être utilisé comme un outil pour évaluer le raisonnement clinique chez les médecins en difficulté [71].

L'usage du TCS s'est aussi révélé pertinent à l'extérieur des champs conventionnels de l'expertise médicale. Llorca et al. ont utilisé le TCS pour examiner les opinions et attitudes

d'étudiants et de résidents en rhumatologie sur des questions d'éthique clinique [72]. Les experts choisis avaient des perspectives éthiques différentes. Le premier panel était constitué des membres du comité d'éthique du centre hospitalier; le deuxième de professeurs de différentes spécialités avec un intérêt en éthique; le troisième de médecins généralistes; et le quatrième d'enseignants de thérapeutique. Les auteurs ont trouvé que tous les professionnels émettent des décisions similaires. De plus, les scores obtenus par les résidents sont plus élevés que ceux obtenus par les étudiants, quel que soit le jury. Ils ont conclu que le TCS était un instrument valide pour évaluer la composante professionnelle des savoirs. Une autre étude utilisant le TCS en éthique a rapporté des résultats similaires [73].

Application à l'hématologie

L'hématologie adulte est une des branches de la médecine interne. Pour y être certifiés au Canada, les étudiants qui terminent leurs études en médecine doivent compléter une résidence de 3 ans en médecine interne avant d'y appliquer. Le programme d'hématologie lui-même est d'une durée de 2 ans et couvre les deux grandes branches de l'hématologie : l'hématologie bénigne qui traite des désordres des diverses composantes du sang (ex : anémie, défaut de coagulation, etc.); et l'hématologie maligne qui traite des divers cancers qui émanent des cellules hématopoïétiques, tel que les leucémies et les lymphomes.

Tant en hématologie bénigne que maligne, les tests de laboratoire et d'imagerie sont omniprésents. La principal analyse de laboratoire est la formule sanguine complète (FSC), qui si anormale, sera accompagnée d'un frottis sanguin qu'il faut examiner sous le microscope. Il existe une multitude d'autres analyses sanguines qu'un hématologue doit apprendre à interpréter, comme par exemple des données de biochimie tel que l'électrophorèse des protéines sériques ou le bilan martial. Lorsque certaines maladies hématologiques sont suspectées, l'hématologue doit procéder à une aspiration médullaire (biopsie de moelle osseuse) qui sera interprétée également sous le microscope. Au cours des dernières décennies, des analyses génétiques et moléculaires se sont ajoutées aux autres

données. Bien qu'elles soient censées être objectives et précises, les données de laboratoire ne concordent pas toujours entre elles. Elles ne sont pas non plus toujours consistantes avec les données cliniques. L'interface entre les deux types de données – cliniques et laboratoire - a pris beaucoup d'importance au cours des dernières décennies, à tel point qu'est née une nouvelle branche de la recherche appelée *recherche translationnelle*. Cette dernière est définie comme étant le domaine de la recherche qui s'intéresse à l'application des connaissances et progrès en sciences fondamentales à la clinique.

De plus, l'analyse des frottis sanguins et des aspirations médullaires requière des capacités de perception et d'interprétation d'images sous le microscope. L'hématologue doit non seulement apprendre à reconnaître les anomalies morphologiques, mais également établir leur signification, et ce, en tenant compte des autres données.

Le raisonnement clinique d'un hématologue doit tenir compte d'une multitude de données de natures différentes. Ainsi, il doit concilier entre i) des éléments de la présentation clinique tel que des symptômes ou trouvailles physiques particulières, ii) des résultats de divers tests de laboratoire incluant des analyses aussi banales qu'une FSC et aussi complexes qu'une étude chromosomique, iii) la morphologie du frottis sanguin et de l'aspiration médullaire, et finalement, iv) un vaste éventail de maladies, allant de la simple anémie ferriprive à la leucémie aiguë. Les étapes menant au diagnostic et la prise en charge diffèrent donc considérablement. Le bon hématologue doit non seulement jouir d'un vaste domaine de connaissances variées, mais aussi et surtout de beaucoup de discrimination et jugement clinique. Toutes ces caractéristiques font du TCS l'instrument d'évaluation tout désigné pour mesurer objectivement le raisonnement clinique en hématologie.

L'applicabilité dans le curriculum

L'importance de l'acquisition du raisonnement clinique est maintenant bien établie. La question devient : quel est le moment le plus propice durant le curriculum médical où l'enseignement et l'évaluation du raisonnement clinique devrait commencer. A ce sujet, Konner a fait remarquer que les curriculums classiques des études médicales sont construits sur le principe que les étudiants doivent acquérir beaucoup de connaissances durant leurs études médicales et que, une fois en résidence, du jour au lendemain, ils doivent apprendre à les appliquer dans la vraie vie [74]. La théorie des scripts suggère que le développement et le raffinement des scripts commencent dès les premières rencontres cliniques [43, 44]. Il est donc essentiel de s'attarder au raisonnement clinique tôt dans le curriculum médical, bien qu'un minimum de connaissances soit requis.

Le TCS a été testé en préclinique [75-77]. Dans un but formatif, Hoff et al. ont employé un TCS pour amener des étudiants en première année de médecine à utiliser en contexte clinique les connaissances qu'ils venaient d'acquérir [76]. L'activité a été administrée en petit groupe, et a été appréciée tant par les étudiants que par leurs tuteurs. Les réflexions et les discussions induites par ces questions semblaient favoriser l'intégration des connaissances. Duggan et al. ont employé le TCS dans un but sommatif chez des étudiants en 5^{e} année de médecine [77].

Que ce soit dans un contexte formatif ou sanctionnel, l'évaluation du raisonnement clinique doit occuper une place importante dans le curriculum médical. La théorie de psychologie cognitive qui sous-tend le TCS, ainsi que quelques études préliminaires, suggèrent que l'introduction précoce de l'apprentissage et l'évaluation du raisonnement clinique soit souhaitable.

Le but principal de cette étude était d'examiner l'utilité et les paramètres psychométriques d'un TCS en hématologie. Nous avons également voulu évaluer la capacité discriminatoire de questions contenant des images de frottis sanguins ou d'aspirations médullaires. Enfin, dans le but de mieux définir sa validité, nous avons comparé les résultats obtenus à un TCS à ceux obtenus à un test conventionnel d'hématologie.

Méthodologie

Le format

Chaque cas du TCS débute par une vignette clinique simple, mais incomplète, pour laquelle il existe plusieurs hypothèses pertinentes. Ces cas sont inspirés le plus possible de la pratique quotidienne de l'hématologie et doivent être quelque peu problématiques même pour un expert du domaine. Plusieurs options de diagnostic, ou de prise en charge sont possibles et la vignette clinique ne contient pas suffisamment d'information pour résoudre le problème. Une hypothèse diagnostique est ensuite proposée. Puis une nouvelle donnée est apportée. La tâche du participant consiste à évaluer l'impact (négatif, neutre ou positif) de cette nouvelle donnée clinique sur l'hypothèse initiale. Les réponses sont recueillies sur une échelle de Likert. Il s'agit d'un format de questionnaire dans lequel il est demandé de faire un choix parmi une série de brèves propositions d'ordre quantitatif. Des exemples de TCS sont contenus dans l'article joint à ce mémoire.

Construction, administration, correction et optimisation.

D'abord, il a fallu définir la population et le domaine médical visé. Nous avons choisi d'évaluation le raisonnement clinique en hématologie clinique adulte chez les résidents juniors et séniors. Nous nous sommes référés aux objectifs de formation du CRMCC [78] pour établir une grille de spécification (Tableau 2) qu'on a utilisée pour sélectionner des vignettes cliniques pertinentes.

La construction du TCS

On a ensuite construit un questionnaire informatique composé d'environ 25 cas cliniques en suivant un guide de pratique écrit par Fournier et al. [79]. Chaque vignette devait contenir

suffisamment d'information pour générer des hypothèses, mais pas assez d'information pour arriver à une réponse claire. Pour chaque vignette clinique, une moyenne de trois questions ou items (total de 60 questions) ont été développés tel que proposé par Gagnon et al. [80]. Tel que décrit plus haut, chaque item comprend trois parties : la première partie énonce une hypothèse diagnostique ou thérapeutique; la deuxième présente une information nouvelle; et la troisième partie est constituée d'une échelle de Likert qui est utilisée pour mesurer l'impact de cette nouvelle information sur l'hypothèse proposée. Chaque item est indépendant des autres.

Dix (10) des 75 questions contenaient une image d'un frottis sanguin ou d'une aspiration médullaire. Les images ont été capturées au laboratoire de morphologie de l'Hôpital Maisonneuve-Rosemont avec une caméra de 5-mégapixel conçue à cet effet. Les images reflétaient des cas cliniques usuels, mais n'étaient pas assez spécifiques pour mener en elles-mêmes à un diagnostic.

La révision des questions

On a ensuite demandé à 3 hématologues académiques de réviser le TCS pour s'assurer qu'il rencontre les objectifs du CRMCC et que les situations cliniques et hypothèses suggérées sont réalistes et pertinentes. Certaines questions faisaient référence à des situations cliniques inhabituelles ou exceptionnelles, et elles ont été éliminées afin de ne pas nuire au potentiel discriminant du test [79].

Le panel d'expert

Un panel d'expert a ensuite été constitué. Selon la littérature, le nombre requis d'experts sur le panel de référence peut varier selon l'étendue et la complexité du champs médical examiné ainsi que le but de l'évaluation, formative ou sanctionnelle. Les règles statistiques usuelles prédisent une meilleure fiabilité du test avec un plus grand nombre d'experts sur le panel. Une étude par Gagnon et al. s'est spécifiquement penchée sur la question [81]. Les auteurs ont comparé les propriétés psychométriques d'un TCS de 73 items administré à 80 résidents et un panel constitué de 38 médecins de famille. La fiabilité des scores des résidents a été calculée en prenant un nombre variable d'expert, soit 5, 10, 15, 20, 25, 30,

ou l'ensemble des 38 experts. Il s'est avéré que la fiabilité augmentait substantiellement entre un panel de 5 et 10 membres, mais que les gains devenaient négligeables au delà de 20 membres. Les auteurs ont conclu en recommandant que le panel d'expert soit composé d'un moins 10 membres. Si l'évaluation est utilisée à des fins sanctionnelles, un panel de 20 membres est préférable.

Le choix des membres qui composent le panel peut avoir un impact important sur la capacité de discrimination du TCS. L'idée principale est de choisir un groupe de spécialistes qui représente la profession de façon globalitaire et légitime. Il est donc préférable de choisir des experts avec des connaissances cliniques et un champ de pratique représentatif de toute la spécialité, plutôt que des sous-spécialistes ou des chercheurs. En étudiant trois différents panels, Charlin et al. ont démontré que plus la variabilité était importante au sein du panel, plus la capacité discriminatoire du TCS augmentait [82].

Nous avons donc choisi un panel de 17 hématologues, dont 10 travaillaient dans un milieu académique et 7 dans un milieu communautaire repartis sur trois provinces canadiennes : Québec, Ontario et Colombie-Britannique. Tous les hématologues étaient des spécialistes certifiés par le CRMCC. Les hématologues académiques devaient consacrer plus de 70% de leur temps au travail clinique et être impliqués dans le service de consultation et de garde dans leurs établissements respectifs. Les hématologues communautaires n'avaient pas d'affiliation universitaire.

Les participants

Nous avons ensuite obtenu l'approbation des comités d'éthique des Universités de Montréal, de Toronto et de McGill. Nous avons également rédigé un consentement que les participants devaient lire et accepter avant de procéder au questionnaire informatique. En plus du panel de référence, deux groupes de participants ont écrit l'examen :

a. 17 résidents juniors en 1^{ière}, 2^{e} ou 3^{e} année de médecine interne (R1 à R3) à l'Université de Montréal et McGill ont été approchés durant leur rotation d'hématologie. 15 des 17 ont accepté d'écrire l'examen dans leur temps libre à l'intérieur d'un intervalle d'une semaine.

b. Les résidents séniors en hématologie (R4 à R6) ont été approchés durant la retraite annuelle d'hématologie qui a lieu une fois par année à Toronto. Des 57 résidents éligibles, 46 ont complété l'examen (15 R4, 21 R5 et 10 R6). Le TCS a été administré après un examen conventionnel d'hématologie que les participants écrivent à chaque année. Cet examen est développé chaque année par des hématologues de l'Université de Toronto. Il contient des QCM et des questions à réponse courte. Il est conçu avec la perspective de couvrir tous les domaines de l'hématologie et ainsi aider les participants à se préparer pour l'examen final du CRMCC.

Les scores

L'établissement des scores repose sur le principe que la réponse de chaque membre du panel d'expert reflète une opinion valide et doit ainsi être prise en compte. Il s'agit donc d'évaluer le degré de similitude entre la réponse de l'étudiant et celle d'un groupe d'expert. La réponse choisie par le plus grand nombre d'expert constitue la réponse modale et donne au candidat un crédit de 1 point. Les autres réponses donne au candidat un pointage qui correspond à la proportion d'experts qui l'ont choisi. Par exemple, si sur un panel de 10 experts, 6 ont choisi « +1 » et 4 ont choisi « +2 », alors les candidats qui ont répondu « +1 » se verront attribuer un point (6 sur 6) et ceux qui ont choisi « +2 » mériteront 0.66 point (4 sur 6). Le score total est calculé en additionnant les points obtenus à tous les items. Ce score est ensuite divisé par le nombre d'items et multiplié par 100 pour obtenir un score en pourcentage.

Cette méthode des scores combinées ou de l'agrégation des scores prend donc en compte la variabilité qui existe normalement entre des experts. Les avantages reliées à l'utilisation de cette méthode de correction plutôt qu'une méthode basée sur le consensus a été bien démontrée par Charlin et al. [48]

Optimisation du test en 2 étapes

La première étape d'optimisation consistait à éliminer les questions dont les réponses étaient trop variables. Ces dernières étaient définies comme celles qui ont généré des

réponses qui traversaient le « 0 ». L'idée sous-jacente est que de telles questions menaient probablement à une certaine confusion ou manquaient de clarté. Ainsi, 2 questions ont été éliminées. Inversement, nous avions prévu d'éliminer les questions qui généraient des réponses unanimes, puisqu'elles seraient au fond très comparables à des QCM. Cependant, ce n'était le cas pour aucune question. Théoriquement, si des membres du panel d'experts ont une ou plusieurs réponses déviantes, la fiabilité du TCS pourrait s'en trouver affaiblie par l'introduction d'une erreur de mesure. Gagnon et al. se sont penchés sur la question et ont conclu que l'élimination des réponses déviantes n'avait pas d'impact majeur sur l'erreur de mesure pourvu que le panel contienne suffisamment d'experts (> 15) [83].

La deuxième étape d'optimisation consistait à éliminer les questions qui avaient une faible corrélation avec l'ensemble des autres (total). Ainsi les questions avec une corrélation item-total inférieure ou égale à 0.05 ont été éliminées (7 questions). Aucune des questions contenant une image n'a été éliminée. Au total, l'instrument optimisé contenait 18 scénarios et 51 questions. La consistance interne sous forme du coefficient alpha de Cronbach a été calculée en utilisant le programme Microsoft Excel. Avant optimisation, l'alpha de Cronbach était 0.71, et après optimisation, 0.83.

Statistiques

La consistance interne du test a été évaluée avec le coefficient d'alpha de Cronbach. La méthode d'ANOVA a été utilisée pour comparer les scores des différents groupes de participants. Le test r de Pearson a été utilisé pour corréler les résultats des résidents séniors au TCS et à l'examen conventionnel d'hématologie. Tous les p étaient considérés significatifs si $\alpha \leq 0.05$.

Article

Voir document ci-joint.

L'article intitulé « Evaluation of clinical judgment of hematology trainees by the script concordance approach » est l'article principal rapportant les résultats de ma recherche. Le manuscrit est présentement en préparation pour soumission à une revue scientifique.

J'ai obtenu l'autorisation des autres auteurs en vue d'inclure cet article dans mon mémoire.

Discussion

Le raisonnement clinque est au cœur de l'expertise médicale. Son acquisition débute dès les premières rencontres cliniques et il se raffine avec le temps et l'expérience clinique. Son développement repose sur une structure et une organisation des connaissances médicales qui permettent leur utilisation efficiente dans des contextes cliniques variées. L'évaluation du raisonnement clinique en contexte d'incertitude pose certaines difficultés que des instruments d'évaluation conventionnels ont peine à résoudre. Le TCS, basé sur la théorie cognitive des scripts, est conçu dans le but d'évaluer objectivement le raisonnement clinique dans un contexte d'incertitude tel que représenté par des situations cliniques inspirées de la pratique. Nous avons construit et testé un TCS en hématologie adulte avec des résidents juniors, séniors et un panel d'hématologues. Nos résultats suggèrent que le TCS est un instrument d'évaluation utile et valide en hématologie.

Validité

La validité constitue l'attribut central de tout instrument d'évaluation. Démontrer qu'un test est valide, c'est démontrer que le sens et l'interprétation des résultats qu'on y donne correspondent réellement à ce que l'on veut mesurer dans une population donnée et à un moment donné [84, 85]. En évaluant la validité d'un instrument d'évaluation, il faut garder en tête quelques points. D'abord, il s'avère difficile, voire impossible, de *prouver* la validité d'un instrument d'évaluation. La validité d'un instrument dans une sphère donnée peut être établie par l'accumulation de données et d'évidences. Deuxièmement, et en conséquence de ce premier point, la validité d'un instrument est relative, et n'est jamais parfaite. Finalement, la validité de tout instrument est situationnelle. Elle dépend spécifiquement du contexte, de la population, et du moment où elle a été établie. On peut parfois se permettre d'extrapoler, mais il faut rester prudent.

La conception contemporaine de la validité établit un cadre global appelé *validité de construit* et identifie 5 sources qui peuvent supporter cette validité : le contenu, le processus de réponse, la structure interne, les relations aux autres variables et les conséquences [84-86]. Ces sources de validité sont complémentaires et contribuent individuellement et collectivement à l'établissement de la validité d'un instrument d'évaluation. Examinons une à une chacune de ces sources de validité tel qu'appliquée à notre TCS.

Le contenu

Il y a validité de contenu lorsque les questions contenues dans un instrument d'évaluation sont représentatives, en tant d'échantillons, des sphères de questions possibles en lien avec un domaine particulier. La validité de contenu revêt une importance capitale puisque la littérature a démontré que la performance dans un problème clinique prédit mal la performance dans un autre problème clinique [87]. Que l'instrument soit utilisé à des fins formatives ou sanctionnelles, il importe de construire des questions qui rencontrent spécifiquement les objectifs établis. Dans ce cas, le but de l'instrument était d'évaluer le raisonnement clinique des résidents en hématologie adulte. Tel qu'établi par la théorie cognitive qui sous-tend les scripts et le raisonnement clinique, les cas doivent être inspirés de la pratique réelle d'un hématologue et doivent contenir un élément d'incertitude ou d'ambiguïté. Pour ce faire, nous avons suivi les lignes de conduites énoncées par Fournier et al. pour construire les questions du TCS [79]. De plus, nous avons optimisé le test en éliminant les questions qui génèrent un consensus parmi les experts (aucune dans notre TCS), puisque de telles questions s'apparenteraient à des QCM. De même, nous avons éliminé les questions qui ont généré une distribution trop importante de réponses (2 questions) et les questions avec une corrélation item-total trop faible (7 questions).

Pour le contenu lui-même, nous nous sommes référés à une grille de spécifications établie à partir des objectifs du CRMCC pour le programme d'hématologie adulte et avons inclus des images de frottis sanguins et d'aspirations médullaires dans 10 questions. Les questions ont été révisées par 3 hématologues académiques afin de s'assurer qu'elles rencontrent les objectifs du CRMCC et que les situations cliniques et hypothèses suggérées sont réalistes et

pertinentes. Ainsi, nous croyons que l'instrument final contenait des questions de bonne qualité consistantes avec le construit et supportant donc une validité de contenu. Dans le futur, d'autres questions pourraient être générées en suivant le même processus.

Le processus de réponse

L'élément de validité ayant trait au processus de réponse repose sur la relation entre le construit théorique soutenant l'instrument et le processus cognitif que les questions contenues dans cet instrument génèrent chez les candidats [84]. Pour appliquer ceci au TCS, il faut se rappeler des bases théoriques sur lequel la théorie des scripts repose. Comme nous avons exposé plus haut, le TCS prend source dans le modèle de raisonnement clinique appelé hypothético-déductif en psychologie cognitive [32]. Ce modèle suggère que les cliniciens, lorsque confrontés à des données cliniques activent inconsciemment leurs « illness scripts » et génèrent automatiquement des hypothèses diagnostiques [44]. Ils tendent ensuite de confirmer ou d'infirmer chacune de ces hypothèses à la lumière de données complémentaires. La structure du TCS vise à reproduire en trois étapes cette séquence exacte telle que proposée par le modèle théorique. Pour chaque vignette clinique, on émet d'abord une hypothèse (*Si vous pensiez à…*), puis on introduit une nouvelle donnée (*Et que vous trouvez…*). On évalue enfin l'impact de cette nouvelle donnée sur l'hypothèse initiale (*Cette hypothèse devient…*) Il est à noter que les deux premières étapes ne requièrent pas l'intervention du candidat puisqu'elles sont générées par le TCS lui-même. Seule la dernière étape, soit celle de l'interprétation, est évaluée. Bien que le format du TCS soit fidèle au modèle cognitif qui le sous-tend, il ne permet de sonder que la dernière étape du processus cognitif impliqué. Il s'agit ici d'une des faiblesses possibles du TCS. Pour l'illustrer, on pourrait à titre d'exemple imaginer un candidat qui possède de bonnes capacités d'interprétation mais qui éprouve des difficultés à générer des hypothèses ou à déterminer quelles informations complémentaires rechercher. Un tel candidat, bien qu'ayant un raisonnement clinique limité, pourra tout-de-même répondre adéquatement à une question du TCS.

Cette faiblesse du TCS, cependant, est relative et théorique. Puisque le modèle tout entier repose sur l'existence de représentations mentales structurées des connaissances, on peut supposer - sans le prouver cependant - que les candidats possédant des scripts plus raffinés sont plus aptes à générer des hypothèses, chercher des données complémentaires et les interpréter. La dernière étape, celle évaluée par le TCS, serait donc représentative de tout le processus cognitif et pourrait suffire pour le sonder. Bien qu'aucune donnée de notre étude, ou à notre connaissance de la littérature traitant du TCS, ne puisse directement prouver ceci, certains arguments peuvent être amenés en support au lien entre le modèle cognitif sous-jacent et le format du TCS. Ces arguments se retrouvent parallèlement à supporter la validité de construit du TCS.

D'abord, dans notre étude, tel qu'on s'y attendrait et tel que démontré dans d'autres études utilisant le TCS (Tableau 2), les scores augmentaient avec le niveau d'expertise. Ainsi, notre TCS a pu discriminer entre les résidents juniors, séniors et le panel d'experts. Les scores ont pu même discriminer entre les résidents séniors selon leur niveau de formation (R4 vs R5 vs R6). Étonnamment, ce n'est pas le cas de plusieurs instruments d'évaluation conventionnels incluant même les ECOS [22, 24, 25]. Il est intéressant de noter que les résidents en dernière année de formation (R6) ont obtenus des scores similaires à ceux des experts. C'est une trouvaille qui a aussi été notée dans d'autres études [60, 61, 88]. Cette observation est rassurante puisqu'elle confirme justement que ces résidents sont prêts à agir à titre d'experts dans cette spécialité. La plupart de ces résidents s'apprêtent justement à assumer une pratique indépendante.

On retrouve dans la littérature d'autres arguments en faveur de la validité de construit du TCS. La démonstration que la méthode de pondération utilisant l'agrégation des scores est supérieure à celle du consensus milite en faveur de la validité de construit [48]. En urologie, l'utilisation du TCS à travers deux pays a permis de démontrer la stabilité des rangs des étudiants peu importe la provenance du panel d'experts [58]. Enfin, une étude par Gagnon et al. a sondé le processus de réponse des candidats [47]. Les auteurs ont soumis un TCS de 64 items à des étudiants en médecine et des gériatres. Les nouvelles

informations proposées dans chaque item étaient choisies pour être soit typiques, atypiques ou incompatibles. Tel que prédit par le modèle cognitif qui sous-tend le TCS, les items présentant une information typique était traités beaucoup plus rapidement que les items amenant une information atypique ou incompatible.

Il existe donc plusieurs arguments en faveur de la validité de construit du TCS. Ces arguments démontrent et établissent de solides liens entre le modèle théorique qui sous-tend le TCS et le format des questions utilisées. Cependant, à ce stade, il existe peu d'arguments soutenant directement le processus de réponse au TCS comme source de validité. Les preuves amenées sont indirectes et devront être supportées dans le futur par d'autres études. C'est également à cette même conclusion que sont arrivés Lubarsky et al. dans leur revue des éléments de validité du TCS [89].

La structure interne

En plus d'être valide du point de vue du construit, du contenu et du processus de réponse, un instrument d'évaluation doit être en mesure de dupliquer cette validité de façon consistante et reproductible. La structure interne d'un instrument d'évaluation constitue donc la 3^{e} source de validité qu'on examinera.

Un instrument d'évaluation est jugé fiable s'il donne avec constance le même résultat. Un des paramètres utilisés pour mesurer la fiabilité d'un test est le coefficient alpha de Cronbach. Ce dernier mesure la consistance interne du test, c'est-à-dire jusqu'à quel point les items du questionnaire contribuent à la mesure de la même dimension. Un test est généralement considéré fiable si l'alpha de Cronbach est égal ou supérieur à 0.80 [84, 85]. La fiabilité d'un instrument d'évaluation dépend de plusieurs facteurs dont le nombre d'items qu'il contient. Un des avantages du format TCS c'est que chaque item requière peu de temps pour y répondre, ce qui permet que l'administration de tests contenant un grand nombre de questions en un temps raisonnable. Tout en augmentant la fiabilité du TCS, un nombre élevé de questions permet aussi de tester plusieurs domaines de connaissances du

candidat. Ceci, comme mentionné plus haut, s'avère essentiel puisque la performance d'un candidat dans un domaine prédit mal sa performance dans un autre [87, 90]

Les études traitant du TCS dans diverses spécialités médicales ont démontré de façon consistante de bons coefficients alpha de consistance interne, généralement entre 0.70 et 0.90 (voir Tableau 2). Dans notre étude, le coefficient alpha était de 0.83 après optimisation. Le processus d'optimisation détaillé plus haut a permis de ne conserver dans le questionnaire final que les questions qui contribuent positivement à la corrélation item-total. Le TCS final contenait 51 items, ce qui est légèrement inférieur au nombre qu'on retrouve dans d'autres TCS, soit généralement entre 60 et 90 items. Étant relativement court, notre TCS a pu être complété en une moyenne de 36 minutes, tout en maintenant une bonne fiabilité. C'est là l'un des avantages du format TCS : permettre l'administration de tests ayant un haut coefficient de fiabilité en peu de temps comparativement à d'autres instruments d'évaluation.

Dix (10) des 51 questions contenaient une image d'un frottis sanguin ou d'une aspiration médullaire. Lorsqu'analysés séparément des 41 autres, le score obtenu à ces questions pouvait, à lui seul, discriminer entre les R4 et les R5. Par contre, il ne pouvait pas distinguer entre les R5 et les R6 ni entre les R5 et le panel d'experts. Ceci n'est pas surprenant. Les résidents séniors effectuent leur stage de morphologie durant leur R4. Puisque la retraite annuelle a lieu en début d'année académique, les R5 ont l'avantage d'avoir déjà fait leur stage. Nos résultats suggèrent que le facteur principal influençant la performance dans les items contenant des images soit davantage le stage de morphologie plutôt que l'année de formation. Cependant, il n'est pas exclu qu'un test comprenant plus d'images ou utilisant des stratégies de questionnement différentes puisse s'avérer plus discriminatif. En radiologie, Brazeau-Lamontagne et al. ont trouvé qu'en matière d'image radiologique, l'habileté de perception se développe plus rapidement que celle d'interprétation, la première étant une meilleure prédicatrice du niveau de formation [59]. De nouvelles plateformes informatiques permettent notamment de questionner le candidat sur l'endroit ou l'aspect précis d'une image qu'il aperçoit et sa signification. Ces réponses

peuvent ensuite être comparées à celles d'experts. Un TCS en morphologie hématologique utilisant cette technologie serait tout-à-fait envisageable.

La composition du panel d'expert a également une incidence sur la fiabilité et la structure interne d'un instrument d'évaluation. Plusieurs études se sont penchées sur la composition optimale du panel d'expert. Une étude de Gagnon et al. supportée par une littérature substantielle traitant du TCS, a conclu que le panel d'expert devrait être composé d'au moins 15 membres [81]. Nous avons donc inclus dans le nôtre 17 hématologues provenant de milieux différents. Cette variabilité contribue positivement à la capacité discriminative du test tel que démontré par Charlin et al. [82].

On peut conclure que la structure interne de notre TCS contribue positivement à sa validité. Dans le futur, l'optimisation systématique des nouvelles questions ainsi que le maintien d'une composition optimale du panel d'expert permettra de conserver cette source de validité.

La relation avec d'autres variables

Théoriquement, si notre TCS mesure seulement et spécifiquement ce qu'il est censé mesurer, c'est-à-dire le raisonnement clinique, il ne devrait pas y avoir de corrélation entre les scores obtenus au TCS et ceux obtenus par le biais de d'autres instruments d'évaluation. Cela reste théorique cependant puisque d'une part les sphères de compétences se recoupent et d'autre part, aucun instrument d'évaluation n'est pur. En d'autres mots, bien que les connaissances et le raisonnement clinique soient deux entités conceptuellement distinctes, ils dépendent en partie l'un de l'autre et il s'avère difficile, voire impossible, de mesurer l'un sans sonder l'autre.

Avant d'analyser la relation entre les scores obtenus au TCS et ceux obtenus à d'autres épreuves, il importe de distinguer deux sous-types de validité en relation aux autres

variables : la validité concomitante et la validité prédictive. La validité concomitante mesure au moyen d'un coefficient de corrélation jusqu'à quel point les scores à un test donné peuvent être utilisés pour estimer les scores qui sont effectivement obtenus par les mêmes sujets à une autre épreuve administrée au même moment. La validité prédictive est semblable, mais vise à explorer la corrélation avec des variables mesurées dans le futur. Nous reviendrons sur ce dernier aspect lorsque nous traiterons des conséquences d'une épreuve comme source de sa validité.

Validité concomitante

Dans notre étude, 46 résidents séniors ont complété le TCS ainsi qu'un examen conventionnel écrit d'hématologie composé de QCM et de questions à courtes réponses. La corrélation r de Pearson entre les scores obtenus aux deux épreuves était de 0.42 ($p=0.02$), ce qui indique une corrélation modérée. Ceci n'est pas étonnant. Le modèle cognitif qui sous-tend le TCS implique l'existence de scripts basés sur des connaissances. Bien que le raisonnement clinique découle de l'organisation de ces scripts, il n'en demeure pas moins qu'il dépend des connaissances factuelles sous-jacentes. Selon le modèle proposé, l'examen écrit testerait davantage les connaissances factuelles des candidats, alors que le TCS mesurerait davantage leur raisonnement clinique. Les deux domaines sont fondamentalement reliés, ce qui expliquerait aisément la corrélation positive mais modérée retrouvée.

D'autres études ont également trouvé une corrélation entre le TCS et d'autres instruments d'évaluation. Park et al. ont trouvé une corrélation ($r=0.38$, $p=0.001$) entre les scores obtenus par les résidents en chirurgie gynécologique à un TCS à ceux obtenus à un examen écrit préparé localement par le service [68]. Utilisant un TCS en médecine d'urgence, Humbert et al. ont démontré que les scores obtenus par les résidents au TCS corrélaient très bien avec ceux obtenus dans un examen préparé localement ($r=0.69$, $p<0.001$) ainsi qu'avec les scores obtenus à un examen national sanctionnel (USMLE - United States Medical Licensing Examination Step 2 - Clinical Knowledge) ($r=0.56$ $p<0.001$) [62]. En médecine interne, Kelly et al. ont trouvé une corrélation entre la performance d'étudiants et de résidents sur un TCS comparés aux évaluations cliniques obtenues ($r=0.22$, $p=0.005$) ainsi

qu'aux résultats obtenus à un examen national (NBME - National Board of Medical Examiners) ($r=0.35$, $p=0<0.001$) [91]. Par contre, les scores obtenus au TCS ne corrélaient pas avec ceux obtenus à un examen composé de QCM ($r=0.11$, $p=0.159$). Fournier et al. n'ont pas trouvé de corrélation entre les résultats obtenus à un TCS en médecine d'urgence et ceux obtenus à un QCM ($r^2=0.0164$, $p=0.59$) [16].

La corrélation entre le TCS et d'autres épreuves n'est donc pas consistante. A ce sujet, une étude intéressante a été menée par Collard et al. [92]. Les auteurs ont comparé les scores obtenus à un TCS en sciences biomédicales à ceux obtenus à un examen factuel composé de vrais ou faux. Une corrélation positive a été trouvé chez les étudiants en début de formation (3^e et 4^e année; $r=0.53$, $p<0.0001$), mais pas chez ceux en fin de formation (5^e et 6 année; $r=0.07$, $p=0.64$). Les auteurs ont conclu qu'en début de formation, les connaissances factuelles et le raisonnement clinique sont dépendants l'un de l'autre, mais qu'avec l'expérience, une indépendance relative s'installe, affaiblissant ainsi la corrélation entre les deux épreuves. Bien qu'hypothétique, cette conclusion supporterait la validité de construit du TCS.

Validité prédictive

Les deux épreuves faisant partie de notre étude, soit le TCS et l'examen écrit, ont eu lieu au même moment. Nous ne sommes donc pas en mesure d'analyser la valeur prédictive du TCS en hématologie.

D'autres études se sont par contre penchées sur la question dans d'autres domaines. Brailovsky et al. ont examiné la validité prédictive du TCS chez une cohorte de 24 étudiants dans leur dernière année de médecine [93]. Les résultats obtenus à ce TCS ont été comparés à ceux obtenus à trois autres épreuves administrées à la fin de la résidence en médecine familiale, soit 2 ans plus tard. Deux de ces trois tests (SAMP et SOO, Short Answer Management Problems et Simulated Office Orals) mesuraient principalement le raisonnement clinique et le troisième (ECOS) évaluait surtout le comportement et la capacité de prise en charge des patients. Les auteurs ont confirmé leur hypothèse de départ,

à savoir que les scores obtenus au TCS corréleraient de façon significative avec les scores obtenus aux SAMP (r=0.45, p=0.013) et SOO (r=0.45, p=0.015) mais pas avec ceux obtenus aux ECOS (r=0.35, p=0.052). Ils ont donc conclu que le TCS, en permettant de prédire les résultats à d'autres épreuves administrées ultérieurement, offrait une bonne validité prédictive.

Ces résultats supportent aussi indirectement deux autres aspects du TCS. Premièrement, la validité de construit du TCS s'en retrouve renforcée puisque les trois instruments d'évaluation (TCS, SAMP et SOO) prétendent évaluer principalement le même concept, soit le raisonnement clinique. Il est donc rassurant que les scores qu'on y obtient corrèlent. Deuxièmement, on peut en conclure indirectement que les candidats ayant un bon raisonnement clinique en début de formation risquent de le conserver plus tard en cours de formation. Ceci peut être utilisé comme un argument en faveur de l'enseignement et de l'évaluation du raisonnement clinique tôt en cours de formation.

Les études examinant la relation entre le TCS et d'autres épreuves demeurent contradictoires, et il est difficile de tirer des conclusions franches. Cependant, dans notre étude il y avait une corrélation modérée entre les scores obtenus au TCS et ceux obtenus à un examen conventionnel d'hématologie. Cela support le construit sous-jacent du TCS ; à savoir, il évalue le raisonnement clinique bien que celui-ci ne soit pas indépendant de d'autres sphères tel que les connaissances factuelles. Dans le futur, il serait intéressant que d'autres études se penchent sur la relation entre le TCS et d'autres épreuves.

Les conséquences

Dans notre étude, le TCS a été administré dans le cadre d'une retraite annuelle à laquelle sont invités l'ensemble des résidents en hématologie du Canada. Le but de cette retraite est

purement formatif, visant à aider les résidents à se préparer à l'examen sanctionnel du CRMCC. Il n'y avait donc aucune conséquence tangible à la performance des résidents au TCS. A notre connaissance, toutes les études qui se sont penchées sur l'usage du TCS l'ont fait dans un cadre formatif.

Pour la discussion, nous allons diviser l'examen de cette 5^{e} source de validité en deux parties. D'abord, nous allons nous attarder aux conséquences qui peuvent être engendrées par la passation du TCS dans un contexte formatif et ensuite dans un contexte sanctionnel.

Contexte formatif

Les conséquences de la passation d'un TCS dans un contexte formatif pourraient être étudiées formellement si l'on menait par exemple une étude comparant la performance de participants ayant écrit ou pas un TCS à leur performance à une autre épreuve. A notre connaissance, aucune étude de la sorte n'a été publiée. En revanche, certaines études dont on a discuté plus haut ont étudié la validité prédictive du TCS [93]. En l'absence de randomisation et d'un groupe contrôle par contre, de telles études ne peuvent pas établir un lien de causalité entre l'exercice du TCS et une performance future.

Dans le contexte donc, l'évaluation des conséquences du TCS repose surtout sur les commentaires émis par les participants. Dans notre étude, comme dans la plupart des autres, l'exercice a été bien reçu par les participants. Dans un sondage administré à la fin du TCS, la plupart ont rapporté que la longueur du test ainsi que son niveau de difficulté étaient adéquats. Vingt-deux pourcents (22%) des résidents ont indiqué que l'expérience dans son ensemble était « neutre » alors que 78% la qualifiait de « très bien » ou « bien ». Bien que la plupart des commentaires fussent positifs, deux problématiques ressortaient. La première était l'impossibilité de retourner en arrière durant le test pour revoir ses réponses. En théorie, une telle option ne devrait pas nuire au processus de réponse ou la validité du test, chaque question étant indépendante des autres. Cet ajustement dans le programme informatique pourrait être effectué dans le futur. Deuxièmement, les résidents ont exprimé leur fatigue durant le TCS puisqu'ils venaient de terminer l'examen écrit qui était d'une durée de 2.5 heures. Nous avions pressenti ce problème avant l'administration et avons

tenté d'y pallier à travers des changements dans l'horaire de la retraite. Malheureusement, cela ne fut pas possible. Dans le futur, c'est un aspect dont il faudra tenir compte.

Chez les étudiants en préclinique, Hoff et al. ont rapporté que l'activité a été appréciée à la fois par les étudiants et par les tuteurs. Les points forts identifiés par les étudiants incluent l'occasion de discuter et réviser les notions apprises, d'intégrer leurs connaissances et de développer le raisonnement clinique [76].

Le TCS peut donc jouer un rôle important dans un contexte formatif tant en préclinique qu'au niveau de la résidence. On pourrait également lui imaginer un rôle dans un contexte d'éducation continue, bien qu'il n'ait pas encore été formellement étudié dans cette situation.

Contexte sanctionnel

L'idée d'utiliser le TCS dans un cadre sanctionnel a été mise de l'avant, et cette idée gagne de la crédibilité à mesure que des études supportant sa validité dans diverses spécialités médicales sont publiées. Duggan et al. ont utilisé le TCS dans un contexte sommatif chez des étudiants en 5^{e} année de médecine [77]. A notre connaissance, aucune étude publiée ne relate l'usage du TCS à des fins sanctionnelles au niveau de la résidence ou des autorités médicales régissant la pratique de la profession.

Pour que le TCS en hématologie (ou dans d'autres spécialités) soit utilisé à des fins sanctionnelles, il faudrait satisfaire deux critères principaux. D'une part, à défaut de prouver, il faudra démontrer sa validité à la satisfaction de l'autorité sanctionnelle qui compte l'utiliser (ex : Université ou CRMCC). Or, on se rappellera que la validité d'un instrument est démontrée par une accumulation de données, et qu'elle est situationnelle. Il faudrait donc accumuler des données propres à chaque spécialité, population, et niveau de formation. Deuxièmement, la méthode de correction et d'établissement des scores revêt une importance capitale lorsque l'instrument est utilisé dans un contexte sanctionnel. La

méthode des scores combinés ou d'agrégation des scores utilisée dans le TCS a démontré des avantages comparativement à la méthode du consensus et permet une plus grande discrimination [48, 82]. Comme pour les examens conventionnels, il faudra établir un seuil de succès. Dans ce but, Charlin et al. ont proposé une méthode statistique de standardisation des scores obtenus au TCS [94]. Selon cette méthode, the scores brutes sont convertis à l'aide d'une échelle utilisant comme points de référence le score modal et la déviation standard des scores du panel d'expert. Les auteurs ont proposé d'utiliser cette méthode pour standardiser les scores obtenus à différents TCS et même à d'autres instruments d'évaluation qui comparent les réponses des participants à ceux d'un panel d'experts.

L'accumulation d'évidences en faveur de la validité du TCS ainsi qu'une standardisation des scores permettraient éventuellement son usage à des fins sanctionnelles. N'étant pas un instrument d'évaluation complet en soi, le TCS pourrait être combiné à d'autres méthodes d'évaluation.

Enfin, nous nous permettons ici d'aborder brièvement un sujet qui dépasse l'usage du TCS mais qui est à notre avis de grand intérêt. Tel qu'exposé plus haut, le but de l'évaluation dans le curriculum médical est de favoriser les apprentissages, qui nous pensons, à leur tour contribuent à la formation de meilleurs médecins qui délivreront de meilleurs soins aux patients. Il est intéressant de noter que, bien qu'en apparence logique, cette séquence de raisonnement repose sur de maigres données. Le lien entre le curriculum médical et ultimement les soins prodigués aux patients et leur devenir reste théorique et abstrait. Malgré son importance capitale, cette question n'a reçu que très peu d'attention dans la littérature [36, 95-97]. Mourad et al n'ont trouvé aucun lien entre la qualité de l'enseignement aux résidents tel que perçu par eux-mêmes et la durée d'hospitalisation des patients avec des pathologies communes tel qu'une pneumonie, un épisode d'insuffisance cardiaque ou de maladie pulmonaire obstructive. Il n'y avait pas non plus de relation entre la qualité de l'enseignement et les taux de réadmission des patients avec les mêmes pathologies à l'intérieur d'un an [95].

De telles études ont certes d'importantes limitations, notamment les courtes périodes de suivi. La conception et la réalisation de telles études soulèvent plusieurs défis, et c'est sans doute pourquoi elles sont rares. Bien qu'une revue détaillée de ce sujet dépasse la portée de ce texte, nous croyons qu'il est important de souligner que l'évaluation des conséquences réelles – les seules qui comptent vraiment! – sur le bienêtre et le devenir des patients n'occupe pas la place qu'elle mérite dans la littérature en pédagogie médicale. Pourtant, n'est-ce pas là le but premier de tout curriculum médical?

Les forces de notre étude

En examinant les forces de notre étude, on se retrouve indirectement à explorer les forces de la méthode TCS. Quelques autres forces sont propres à la nôtre et nous en discuterons.

Nous avons démontré que le format du TCS comportait d'importants avantages. D'abord, en examinant les 5 sources de validités généralement acceptées, on a conclu qu'il y a avait des arguments de poids en faveur de la validité du TCS en hématologie mais aussi dans d'autres domaines médicaux. En suivant des critères bien établis, on peut construire un examen dont le contenu respecte la théorie cognitive sous-jacente – c'est-à-dire l'évaluation du raisonnement clinique en contexte d'incertitude – tout en démontrant d'excellentes propriétés psychométriques. Le coefficient de consistance interne de notre TCS optimisé était de 0.83, ce qui se situe dans la moyenne des coefficients rapportés dans d'autres études. La facilité d'administration du TCS s'avère également un avantage majeur. En effet, une plateforme informatique existe et permet non seulement la création rapide et aisée de questions, mais également l'administration et la correction des réponses de manière automatisée. Des images et même des vidéos pourraient être incorporées facilement. L'examen peut être administré à un grand nombre d'individus simultanément, tout comme il peut être écrit à différents moments et à des lieux différents. Il suffit d'avoir accès à un ordinateur. De plus, la confidentialité des résultats est protégée par le logiciel informatique. Et enfin, le tout se fait à un coût très faible comparativement à d'autres instruments

d'évaluation. En ce sens, le TCS rassemble plusieurs caractéristiques d'un test informatique optimal [98].

Plus spécifiquement, notre étude avait certaines particularités et points forts. D'abord, c'est la première à se pencher sur l'usage du TCS en hématologie. Comme nous l'avons exposé ci-haut, il s'agit d'une discipline qui se prête particulièrement bien à ce format puisque le raisonnement clinique, notamment en contexte d'incertitude, en fait intégralement partie. De plus, le format TCS en privilégiant des questions courtes a permis de couvrir un ensemble représentatif de sujets auxquels est confronté l'hématologue. Nous avons également incorporé des questions contenant des images à notre TCS. Les analyses, bien qu'embryonnaires, ont suggéré que les items contenant des images avaient un plus grand potentiel de discrimination. Cet aspect doit encore être exploré davantage.

Notre étude est multicentrique. Elle a été menée dans trois provinces canadiennes et a recruté au moment de l'administration du test la plupart des résidents en hématologie du Canada. Le panel d'experts était hétérogène et contenait des hématologues académiques et communautaires. Bien que les nombres n'aient pas été suffisamment grands pour permettre des comparaisons interprovinciales, le caractère multicentrique de l'étude ajoute à sa force. Nous avons également pu profiter de la retraite annuelle en hématologie pour comparer les résultats du TCS à ceux d'un examen conventionnel. La corrélation modérée qu'on a trouvée supporte le construit du TCS et contribue à sa validité. Enfin, notre TCS a été bien reçu par les participants.

Les faiblesses et limites de notre étude

Comme pour les avantages, plusieurs des faiblesses de notre étude sont propres au format du TCS lui-même. D'autres sont en lien avec notre étude et nous en discuterons.

D'abord, même si plusieurs arguments militent en faveur de la validité du TCS comme instrument d'évaluation du raisonnement clinique, certaines limites et questions demeurent. Lubarsky et al. ont d'ailleurs fait un excellent résumé de la question [89]. Notamment, comme nous l'avons exposé plus haut, l'évidence de validité ayant trait au processus de réponse au TCS repose sur une preuve indirecte. D'autres études dans ce sens seront nécessaires.

Étant un test écrit, le TCS ne permet pas d'évaluer d'autres domaines importants de la pratique clinique, tel que des habiletés de recueil de données, d'examen physique, et des compétences tel que la collaboration, le professionnalisme, etc. L'utilisation de plusieurs méthodes d'évaluation complémentaires s'avère nécessaire [9, 24] . Que ce soit pour des fins formatives ou sanctionnelles, le TCS devrait donc faire partie d'un ensemble d'instruments d'évaluation visant à évaluer les diverses facettes de la compétence clinique [29].

Le format du TCS permet l'évaluation du raisonnement clinique en ce qui a trait aux sciences cliniques. Il est peu adapté à l'évaluation des acquis dans les sciences fondamentales [52]. Bien que cela puisse être une limite à son utilisation, le TCS peut s'avérer utile dans l'interprétation ou l'intégration d'une donnée de science fondamentale à une situation clinique.

Notre TCS, plus spécifiquement, comportait certaines faiblesses. D'abord, le nombre de questions contenant des images aurait pu être plus grand. Cela nous aurait peut-être permis de tirer des conclusions plus robustes quant à l'usage de telles questions en hématologie.

Aussi, nous aurions aimé faire une analyse de validité prédictive en comparant les scores à ceux obtenus par les résidents séniors aux examens du CRMCC. Pour des raisons de confidentialité cependant, cela n'a pas été possible. Enfin, le moment d'administration de notre TCS aurait pu être mieux choisi afin que les candidats soient moins fatigués.

Place du TCS dans le curriculum médical

En raison de la théorie sur laquelle il repose, ainsi que ses forces démontrées par cette étude et bien d'autres, le TCS mérite certes une place dans le curriculum médical, mais celle-ci doit encore être mieux définie. Pour les fins de cette discussion, imaginons le parcours d'un étudiant en médecine qui aspire à être hématologue ! On pourrait diviser – même si quelque peu arbitrairement - sa formation médicale en quatre étapes chronologiques : i) étudiant en médecine; ii) résident junior ; iii) résident sénior et enfin iv) hématologue. Quelle pourrait être la place du TCS à chacune de ces étapes de formation ? Étant donné l'absence de données scientifiques probantes quant à l'usage du TCS à ces diverses étapes de formation, je me permets ici d'émettre un certain nombre d'hypothèses et de propositions me basant le plus possible sur ce qu'on sait du TCS. L'usage du TCS dans chacun des ces contextes devra éventuellement faire l'objet d'étude.

i) étudiant en médecine

Durant les premières années en médecine, l'accent est mis sur l'acquisition des connaissances factuelles, ou le savoir (*knows* de Miller). L'acquisition de telles connaissances est évaluée par des épreuves telles que les QCM et les questions à réponses courtes ou longues qui s'y prêtent bien ; le TCS étant mal adapté à l'évaluation des connaissances factuelles et des sciences de base tel que discuté précédemment. Cependant, se basant sur la théorie des scripts, on pourrait proposer que l'acquisition de ces connaissances serait facilitée par le développement concomitant de scripts. Ces derniers offriraient une structure autour de laquelle les connaissances peuvent s'arrimer, facilitant ainsi leur mémorisation et leur extraction en temps opportun. Plus tôt semées, plus tôt

l'étudiant pourra commencer à enrichir et raffiner ces scripts, et ainsi à apprendre comment utiliser (le *know how* de Miller) ses connaissances dans un contexte clinique.

Durant les études médicales, il serait donc concevable que le TCS soit utilisé systématiquement à des fins formatives comme a été démontré par Hoff et al. en hématologie. Durant ces activités, les étudiants ont rapporté que le TCS favorisait entre autres l'intégration des connaissances.

L'usage du TCS à des fins sommatives dans ce contexte devrait faire l'objet de plus de recherche avant d'être proposé.

ii) résident junior

Durant cette étape de formation, l'acquisition de connaissances factuelles se poursuit, mais c'est surtout l'intégration, l'organisation et l'application de ces connaissances à la clinique qui devient le focus. Dans ce contexte, le TCS a beaucoup à offrir. D'abord, à des fins formatives, il pourrait favoriser l'acquisition du raisonnement clinique en contexte d'incertitude. Tel que présenté plus haut, les instruments d'évaluation conventionnels offrent peu dans ce domaine. On pourrait s'imaginer par exemple offrir un TCS bi-annuellement durant les 3 années de formation en médecine interne. Une telle épreuve permettrait au programme d'identifier les résidents en difficultés et identifier les domaines mal enseignés dans le curriculum.

Le TCS a cependant ses limites et il ne serait pas suffisant dans l'évaluation des résidents juniors. Des ECOS et\ou des examens oraux seraient nécessaires afin d'évaluer d'autres sphères de l'apprentissage tel quel la collecte de données, l'examen physique, les habiletés techniques, l'esprit de synthèse, les capacités interpersonnelles , etc. Faisant partie d'un ensemble comprenant d'autres instruments d'évaluation, on pourrait aisément concevoir un

rôle formatif, et peut-être même sommatif, du TCS durant les premières années de résidence.

iii) résident sénior

Durant ses dernières années de formation, le résident doit désormais faire preuve d'un raisonnement clinique qui s'approche de celui d'un hématologue. Dans ce contexte, il serait aisé d'imaginer l'usage du TCS dans un contexte sommatif. Après tout, le résident finissant doit être en mesure de démontrer que son raisonnement clinique lui permet désormais d'entrer dans le groupe de spécialistes auquel il aspire appartenir. Encore une fois, le TCS serait utilisé en combinaison avec d'autres instruments d'évaluation. Cela pourrait faire l'objet d'une proposition imminente au CRMCC.

Étant donné les enjeux – devenir ou pas hématologue! – l'intégration du TCS dans les examens du CRMCC devrait d'abord faire l'objet d'un projet pilote. On pourrait envisager par exemple, que le comité de spécialité qui développe annuellement les questions d'examen soit invité à agir à titre de panel d'expert sur un questionnaire TCS. Pour la première année, celui-ci serait administré aux candidats en plus des autres épreuves prévues. Si sa validité par rapport aux autres variables reste supportée et qu'il est bien reçu par tous, le TCS pourrait éventuellement remplacer, du moins en partie, la composante écrite du présent examen sanctionnel du CRMCC composé exclusivement de questions à réponses courtes et longues. En plus du TCS, en hématologie, un examen oral et un examen de morphologie seraient nécessaires pour évaluer quelques unes des autres sphères de compétences. L'examen de morphologie se faisait traditionnellement sur lame, mais le CRMCC a récemment adopté un logiciel informatique qui permet la visualisation de larges champs morphologiques. Si on réussit à intégrer un tel logiciel au TCS, on pourrait s'imaginer éventuellement l'usage exclusif du TCS même pour évaluer les aptitudes en morphologie.

Afin de promouvoir le raffinement des scripts des résidents séniors et les familiariser avec ce format d'examen, un certain nombre de TCS formatifs pourraient être offerts durant la résidence en hématologie. Ceux-ci pourront inclure des questions contenant des images de frottis sanguins et d'aspirations médullaires. Comme pour le résident junior, le TCS permettrait d'identifier les faiblesses à adresser par le résident et le programme en vue de l'examen sommatif.

iv) hématologue

Une fois devenu hématologue, la priorité devient la formation continue. En effet, il s'agit alors de conserver et mettre à jour ses connaissances et ses scripts, tout en poursuivant le raffinement de ceux-ci. Le TCS pourrait jouer un rôle important dans ce contexte. Par exemple, on pourrait envisager que diverses associations médicales offrent à leurs membres une ou deux questions TCS à chaque semaine par le biais d'internet. De telles initiatives ont d'ailleurs déjà vues le jour à différents endroits dans le monde. L'expérience déjà acquise par l'Université de Montréal et le CPASS dans ce domaine faciliterait énormément la mise ne œuvre d'un tel projet au niveau provincial et même national. Il serait également envisageable que le CRMCC incorpore un tel projet à son programme de maintien de la certification.

Le TCS a donc beaucoup de potentiel. Il pourrait être utilisé à chacune des étapes de formation d'un médecin, tantôt pour promouvoir l'intégration des connaissances, tantôt pour faciliter la mise en place et le raffinement des scripts, tantôt à des fins sommatives ou encore d'éducation continue. La prochaine étape consiste à concevoir un ou des projet(s) pilote(s) dans l'un ou l'autre de ces contextes.

Conclusion et retombées potentielles

La compétence clinique est multidimensionnelle et aucun instrument d'évaluation à lui-seul ne saurait la saisir globalement. Le meilleur instrument d'évaluation peut dépendre de plusieurs facteurs dont les objectifs à atteindre, la compétence évaluée, le moment dans le curriculum, et l'enjeu de l'évaluation (formatif versus sanctionnel). Les méthodes d'évaluation conventionnelles ont toutes des avantages et des inconvénients. En évaluant plus spécifiquement le raisonnement clinique en contexte d'incertitude, le TCS vient combler une des lacunes des autres instruments d'évaluation. Il a été proposé qu'une stratégie d'évaluation crédible pourrait consister en l'utilisation concomitante de QCM, TCS et ECOS puisque tous trois sont des examens standardisés à correction objective [52].

Nous avons exposé plusieurs arguments en faveur de la validité du TCS en général, et plus spécifiquement du nôtre en hématologie. Certains points faibles persistent, et d'autres études seront nécessaires avant d'étendre l'usage du TCS à des fins sanctionnelles. Il demeure par contre un excellent outil d'apprentissage qui est facile à concevoir, à administrer et à corriger. Il démontre également de bonnes propriétés psychométriques. Son utilisation en préclinique ainsi qu'en résidence a été étudiée, et ce, dans plusieurs domaines de la médecine. Son usage pourrait également s'étendre dans un contexte de formation continue.

Le TCS a démontré des résultats prometteurs dans tous les domaines où il a été évalué. Bien que certaines questions devront faire l'objet d'études supplémentaires, il demeure un instrument prometteur dans la formation des médecins.

Bibliographie

1. Origine du terme "docteur". http://www.fr.wikipedia.org/wiki/Docteur_(titre).

2. Torre DM, Daley BJ, Sebastian JL, Elnicki DM. Overview of current learning theories for medical educators. *Am J Med* 2006; 119: 903-7.

3. Vienneau R. *Apprentissage et enseignement*. Gaëtan Morin, 2005.

4. Cooke M, Irby DM, Sullivan W, Ludmerer KM. American medical education 100 years after the Flexner report. *N Engl J Med* 2006; 355: 1339-44.

5. Compétences CanMEDS. http://www.royalcollege.ca/portal/page/portal/rc/canmeds: Collège Royal des Médecins et Chirurgiens du Canada (CRMCC).

6. Miller GE. The assessment of clinical skills/competence/performance. *Academic medicine : journal of the Association of American Medical Colleges* 1990; 65: S63-7.

7. Piaget J. *Psychologie et pédagogie*. Paris: Gallimard, 1988.

8. Epstein RM. Assessment in medical education. *N Engl J Med* 2007; 356: 387-96.

9. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet* 2001; 357: 945-9.

10. Friedman Ben-David M. The role of assessment in expanding professional horizons. *Med Teach* 2000; 22: 472-77.

11. Frank JR, Snell LS, Cate OT*, et al.* Competency-based medical education: theory to practice. *Medical teacher* 2010; 32: 638-45.

12. Frank JR, Mungroo R, Ahmad Y*, et al.* Toward a definition of competency-based education in medicine: a systematic review of published definitions. *Medical teacher* 2010; 32: 631-7.

13. GR. N. *Theoretical and psychometric considerations. In: Report on the evaluation system for specialist certification.* Ottawa1993.

14. Newble DI SD. Psychometric characteristics of the obejctive structured clinical examination. *Medical education* 1996; 22: 325-34.

15. Hodges B. Medical education and the maintenance of incompetence. *Medical teacher* 2006; 28: 690-6.

16. Fournier JP. Thiercelin D PC, Alunni-Perret V, Gilbert E, Minguet JM, Bertrand F. . Clinical reasoning assessment in emergency medicine: script concordance tests are more efficient to detect clinical experience that rich-context multiple-choice questions. . *Pédagogie Médicale* 2006; 7: 20-30.

17. Jolly B GJ. *The Good Assessment Guide. A practical guide to assessment and appraisal for higher specialist training*. . London, UK.1997.

18. Norcini JJ, Diserens D, Day SC*, et al.* The scoring and reproducibility of an essay test of clinical judgment. *Academic medicine : journal of the Association of American Medical Colleges* 1990; 65: S41-2.

19. Solomon DJ, Reinhart MA, Bridgham RG*, et al.* An assessment of an oral examination format for evaluating clinical competence in emergency medicine. *Academic medicine : journal of the Association of American Medical Colleges* 1990; 65: S43-4.

20. Levine HG MC. The validity and reliability of oral examinations in assessing cognitive skills in medicine. *J Educ Meas* 1970; 7: 63-73.

21. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical education* 1979; 13: 41-54.

22. Hodges B, Regehr G, McNaughton N*, et al.* OSCE checklists do not capture increasing levels of expertise. *Academic medicine : journal of the Association of American Medical Colleges* 1999; 74: 1129-34.

23. Swanson DB NG, Linn RL. . Performance-based assessment: lessons learnt from the health professions. *Educ Res* 1995; 24: 5-11.

24. Van der Vleuten C. The assessment of professional competence: development, research and practical implications. *Advances in Health Sciences Education* 1996; 1: 41-67.

25. Schmidt HG, Boshuizen HP. On the origin of intermediate effects in clinical case recall. *Mem Cognit* 1993; 21: 338-51.

26. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. *Academic medicine : journal of the Association of American Medical Colleges* 1990; 65: 611-21.

27. Norman G. Research in clinical reasoning: past history and current trends. *Medical education* 2005; 39: 418-27.

28. Meterissian SH. A novel method of assessing clinical reasoning in surgical residents. *Surgical innovation* 2006; 13: 115-9.

29. Charlin BB, G. Van Der Vleuten, C. L'évaluation du raisonnement clinique. *Pédagogie médicale* 2003; 4: 42-52.

30. Higgs J, Jones MA. *Clinical reasoning in the health professions*. Oxford ; Boston: Butterworth-Heinemann, 2000: xiv, 322 p.p.

31. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. *N Engl J Med* 2006; 355: 2217-25.

32. Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. *Evaluation & the health professions* 2004; 27: 304-19.

33. Schön DA. *The reflective practitioner: How professionals think in action.* New York: Basic Books., 1983.

34. Mamede S, Schmidt HG, Rikers RM*, et al.* Breaking down automaticity: case ambiguity and the shift to reflective approaches in clinical reasoning. *Medical education* 2007; 41: 1185-92.

35. WB. B. *Sir William Osler: aphorisms from his bedside teachings and writings*. Springfield: Charles C Thomas, 1968.

36. Reilly BM. Inconvenient truths about effective clinical teaching. *Lancet* 2007; 370: 705-11.

37. Grant J, Marsden P. Primary knowledge, medical education and consultant expertise. *Medical education* 1988; 22: 173-9.

38. Norman GR, Feightner JW. A comparison of behaviour on simulated patients and patient management problems. *Medical education* 1981; 15: 26-32.

39. Norman GR. Objective measurement of clinical performance. *Medical education* 1985; 19: 43-7.

40. Norcini JJ SJ, Say SC. The use of the aggregate scoring for a recertification examination. *Eval Health Prof* 1990; 13: 241-51.

41. Page G, Bordage G. The Medical Council of Canada's key features project: a more valid written examination of clinical decision-making skills. *Academic medicine : journal of the Association of American Medical Colleges* 1995; 70: 104-10.

42. Charlin B, Roy L, Brailovsky C*, et al.* The Script Concordance test: a tool to assess the reflective clinician. *Teaching and learning in medicine* 2000; 12: 189-95.

43. Feltovitch PJ BH. Issues of generality in medical problem solving. In: Assen, ed. *Tutorials in Problem-based Learning: A New Direction in Teaching the Health Professions* The Netherlands: Van Gorcum, 1984.

44. Barrows HS, Feltovich PJ. The clinical reasoning process. *Medical education* 1987; 21: 86-91.

45. Barrows HS TR. *Problem-based Learning: An Approach to Medical Education*. New York: Spring Publishing Co., 1980.

46. Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Academic medicine : journal of the Association of American Medical Colleges* 2000; 75: 182-90.

47. Gagnon R, Charlin B, Roy L*, et al.* The cognitive validity of the Script Concordance Test: a processing time study. *Teaching and learning in medicine* 2006; 18: 22-7.

48. Charlin B, Desaulniers M, Gagnon R*, et al.* Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. *Teaching and learning in medicine* 2002; 14: 150-6.

49. Mattern WD, Weinholtz D, Friedman CP. The attending physician as teacher. *N Engl J Med* 1983; 308: 1129-32.

50. Irby DM. What clinical teachers in medicine need to know. *Academic medicine : journal of the Association of American Medical Colleges* 1994; 69: 333-42.

51. Smith CA, Varkey AB, Evans AT, Reilly BM. Evaluating the performance of inpatient attending physicians: a new instrument for today's teaching hospitals. *J Gen Intern Med* 2004; 19: 766-71.

52. Charlin BGRS, L., Vleuten C. Le test de concordance de script, un instrument d'évaluation du raisonnement clinique. *Pédagogie médicale* 2002; 3: 135-44.

53. Charlin B BC, Brazeau-Lamontage L, Samson L, Leduc C. Script questionnaires: their use for assessment of diagnostic knowledge in radiology. *Med Teach* 1998; 20: 567-71.

54. Charlin B, Brailovsky C, Leduc C, Blouin D. The Diagnosis Script Questionnaire: A New Tool to Assess a Specific Dimension of Clinical Competence. *Advances in health sciences education : theory and practice* 1998; 3: 51-8.

55. Marie I, Sibert L, Roussel F*, et al.* [The script concordance test: a new evaluation method of both clinical reasoning and skills in internal medicine]. *La Revue de medecine interne / fondee par la Societe nationale francaise de medecine interne* 2005; 26: 501-7.

56. Sibert L, Darmoni SJ, Dahamna B*, et al.* Online clinical reasoning assessment with the Script Concordance test: a feasibility study. *BMC medical informatics and decision making* 2005; 5: 18.

57. Sibert L, Darmoni SJ, Dahamna B*, et al.* On line clinical reasoning assessment with Script Concordance test in urology: results of a French pilot study. *BMC medical education* 2006; 6: 45.

58. Sibert L, Charlin B, Corcos J*, et al.* Stability of clinical reasoning assessment results with the Script Concordance test across two different linguistic, cultural and learning environments. *Medical teacher* 2002; 24: 522-7.

59. Brazeau-Lamontagne L, Charlin B, Gagnon R*, et al.* Measurement of perception and interpretation skills during radiology training: utility of the script concordance approach. *Medical teacher* 2004; 26: 326-32.

60. Meterissian S, Zabolotny B, Gagnon R, Charlin B. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *American journal of surgery* 2007; 193: 248-51.

61. Nouh T, Boutros M, Gagnon R*, et al.* The script concordance test as a measure of clinical reasoning: a national validation study. *American journal of surgery* 2012; 203: 530-4.

62. Humbert AJ, Besinger B, Miech EJ. 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 : official journal of the Society for Academic Emergency Medicine* 2011; 18: 627-34.

63. Caire F, Sol JC, Moreau JJ*, et al.* [Self-assessment for neurosurgery residents by script concordance test (SCT). The

process of test elaboration]. *Neuro-Chirurgie* 2004; 50: 66-72.

64. Kania RE, Verillaud B, Tran H*, et al.* Online script concordance test for clinical reasoning assessment in otorhinolaryngology: the association between performance and clinical experience. *Archives of otolaryngology--head & neck surgery* 2011; 137: 751-5.

65. Bursztejn AC, Cuny JF, Adam JL*, et al.* Usefulness of the script concordance test in dermatology. *Journal of the European Academy of Dermatology and Venereology : JEADV* 2011; 25: 1471-5.

66. Lambert C, Gagnon R, Nguyen D, Charlin B. The script concordance test in radiation oncology: validation study of a new tool to assess clinical reasoning. *Radiat Oncol* 2009; 4: 7.

67. Lubarsky S, Chalk C, Kazitani D*, et al.* The Script Concordance Test: a new tool assessing clinical judgement in neurology. *The Canadian journal of neurological sciences Le journal canadien des sciences neurologiques* 2009; 36: 326-31.

68. Park AJ, Barber MD, Bent AE*, et al.* Assessment of intraoperative judgment during gynecologic surgery using the Script Concordance Test. *American journal of obstetrics and gynecology* 2010; 203: 240 e1-6.

69. Khonputsa P, Besinque K, Fisher D, Gong WC. Use of script concordance test to assess pharmaceutical diabetic care: a pilot study in Thailand. *Medical teacher* 2006; 28: 570-3.

70. Deschenes MF, Charlin B, Gagnon R, Goudreau J. Use of a script concordance test to assess development of clinical reasoning in nursing students. *The Journal of nursing education* 2011; 50: 381-7.

71. Goulet F, Jacques A, Gagnon R*, et al.* Poorly performing physicians: does the Script Concordance Test detect bad clinical reasoning? *The Journal of continuing education in the health professions* 2010; 30: 161-6.

72. Llorca G RP, Riche B. Évaluation de résolution de problèmes mal définies en éthique clinique: variation des scores selon les méthodes de correction et selon les caractéristiques des jurys. *Pédagogie médicale* 2003; 4: 80-8.

73. Tsai TC, Chen DF, Lei SM. The ethics script concordance test in assessing ethical reasoning. *Medical education* 2012; 46: 527.

74. M. K. *Becoming a doctor: a journey of initiation in medical school*. New York: Penguin Books, 1988.

75. Humbert AJ, Johnson MT, Miech E*, et al.* Assessment of clinical reasoning: A Script Concordance test designed for pre-clinical medical students. *Medical teacher* 2011; 33: 472-7.

76. Hoff LB, A. Kassis, J., Charlin, B. . Le test de concordance de script comme outil d'enseignement et d'apprentissage: un projet-pilote pour les étudiants de première année de médecine. *Pédagogie médicale* 2010; 11: 51-6.

77. Duggan P, Charlin B. Summative assessment of 5th year medical students' clinical reasoning by script concordance test: requirements and challenges. *BMC medical education* 2012; 12: 29.

78. CRMCC. Objectifs de formation pour le programme d'hématologie adulte. In: http://crmcc.medical.org/residency/certification/objectives/hematology_f.pdf, ed.2008.

79. Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. *BMC medical informatics and decision making* 2008; 8: 18.

80. Gagnon R, Charlin B, Lambert C*, et al.* Script concordance testing: more cases or more questions? *Advances in health sciences education : theory and practice* 2009; 14: 367-75.

81. Gagnon R, Charlin B, Coletti M*, et al.* Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? *Medical education* 2005; 39: 284-91.

82. Charlin B, Gagnon R, Pelletier J*, et al.* Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Medical education* 2006; 40: 848-54.

83. Gagnon R, Lubarsky S, Lambert C, Charlin B. Optimization of answer keys for script concordance testing: should we exclude deviant panelists, deviant responses, or neither? *Advances in health sciences education : theory and practice* 2011; 16: 601-8.

84. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006; 119: 166 e7-16.

85. Downing SM. Validity: on meaningful interpretation of assessment data. *Medical education* 2003; 37: 830-7.

86. S. M. *Validity. In: Linn RL, ed. Educational Measurement*. New York, NY.: Macmillan, 1989.

87. Gale J MP. *Medical diagnosis: from student to clinician.* Oxford: Oxford University Press, 1983.

88. Ruiz JG, Tunuguntla R, Charlin B*, et al.* The script concordance test as a measure of clinical reasoning skills in geriatric urinary incontinence. *Journal of the American Geriatrics Society* 2010; 58: 2178-84.

89. Lubarsky S, Charlin B, Cook DA*, et al.* Script concordance testing: a review of published validity evidence. *Medical education* 2011; 45: 329-38.

90. Norcini JJ SD, Grosso LJ, Shea Ja, Webster GD. Reliability, validity and efficiency of multiple choice questions and patient management problem items formats in the assessement of physician competence. *Medical education* 1985; 19: 238-47.

91. Kelly W, Durning S, Denton G. Comparing a script concordance examination to a multiple-choice examination on a core internal medicine clerkship. *Teaching and learning in medicine* 2012; 24: 187-93.

92. Collard A, Gelaes S, Vanbelle S*, et al.* Reasoning versus knowledge retention and ascertainment throughout a problem-based learning curriculum. *Medical education* 2009; 43: 854-65.

93. Brailovsky C, Charlin B, Beausoleil S*, et al.* 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* 2001; 35: 430-6.

94. Charlin B, Gagnon R, Lubarsky S*, et al.* Assessment in the context of uncertainty using the script concordance test: more meaning for scores. *Teaching and learning in medicine* 2010; 22: 180-6.

95. Mourad O, Redelmeier DA. Clinical teaching and clinical outcomes: teaching capability and its association with patient outcomes. *Medical education* 2006; 40: 637-44.

96. Dimitroff A, Davis WK. Content analysis of research in undergraduate medical education. *Academic medicine : journal of the Association of American Medical Colleges* 1996; 71: 60-7.

97. Prystowsky JB, Bordage G. An outcomes research perspective on medical education: the predominance of trainee assessment and satisfaction. *Medical education* 2001; 35: 331-6.

98. Hols-Elders W, Bloemendaal P, Bos N*, et al.* Twelve tips for computer-based assessment in medical education. *Medical teacher* 2008; 30: 673-8.

Tableaux

Tableau 1. Grille de spécification

SUJET	# QUESTIONS (total de 60)
HÉMATOLOGIE MALIGNE	**36**
Leucémies aiguës : présentation, diagnostic, thérapie	6
Syndromes lymphoprolifératifs (indolents et agressifs)	6
Syndromes myélodysplasiques et myéloprolifératifs	6
Désordres plasmocytaires	6
Maladies congénitales et acquises de la moelle osseuse	3
Chimiothérapie : principes d'usage et complications	6
Greffe de cellules souches : indications et complications	3
HÉMATOLOGIE BÉNIGNE	**24**
Désordres des globules rouges	6
Désordres des globules blancs	6
Désordre de la coagulation	6
Banque de sang	6
TOTAL	**60**

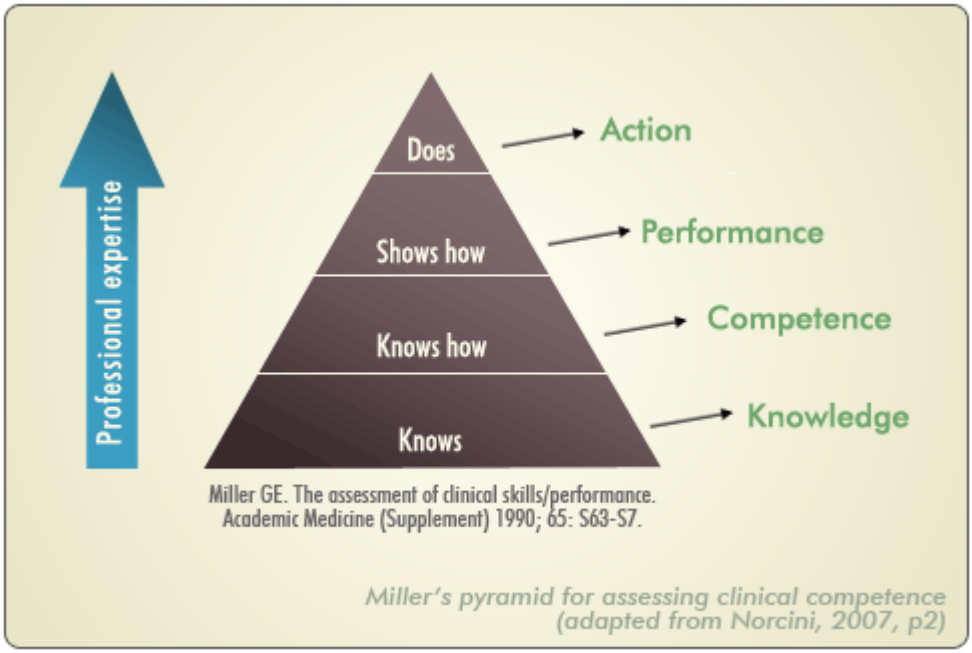
* basée sur les objectifs académiques établis dans le programme d'Hématologie adulte du Collège Royal des Médecins et Chirurgiens du Canada (CRMCC)

Tableau 2. Liste des études publiées traitant du TCS

Auteurs	Domaine	Nb d'items	Nb de candidats	Nb d'experts	Discrimination entre les groupes	Coefficient alpha de Cronbach
Sibert et al. [58] (2002)	Urologie	80	48	22		0.79
Llorca et al. [72] (2003)	Éthique et attitude	1	61	139 de différents groupes	Oui	Non applicable
Brazeau-Lamontagne et al. [59] (2004)	Radiologie	183	60	11	Oui	0.79 pour la perception et 0.81 pour l'interprétation
Marie et al. [55] (2005)	Médecine interne	95	31	7	Oui	0.81
Sibert et al. [57] (2006)	Urologie	97	207	26	Oui	0.73
Fournier et al. [16] (2006)	Médecine d'urgence	30	26	9	Oui	0.92-0.96 selon les groupes étudiés
Khonputsa et al. [69] (2006)	Pharmacie et diabète	31	60	16	Oui	Non disponible
Meteressian et al. [60] (2007)	Chirurgie	62	36	10	Oui	0.85
Lambert et al. [66]	Radio-oncologie	70	155	47	Oui	0.90

(2009)						
Lubarsky et al. [67] (2009)	Neurologie	94	49	16	Oui	0.79
Park et al. [68] (2010)	Chirurgie gynécologique	98	78	17	Oui	0.73
Ruiz et al. [88] (2010)	Gériatrie	70	88	8	Oui	0.72
Humbert et al. [62] (2011)	Médecine d'urgence	59	354	13	Oui	0.78
Bursztejn et al. [65] (2011)	Dermatologie	132	35	16	Oui	0.80
Kania et al. [64] (2011)	Otorhino-laryngologie	94	65	22	Oui	0.95
Nouh et al. [61] (2012)	Chirurgie	152	202	22	Oui	0.85

Figure 1. La Pyramide de Miller



Evaluation of clinical judgment of hematology trainees by the script concordance approach

Alain Bestawros[1,4], Jeannine Kassis[2,4], Christine Chen[3], Eugenia Piliotis[3], Robert Gagnon[4], Bernard Charlin[4]

Target : General hematology or medical education journal

[1] Division of Hematology and Medical Oncology, Centre Hospitalier de l'Université de Montréal, Montreal, Quebec, Canada
[2] Division of Hematology, Maisonneuve-Rosemont Hospital, Montreal, Quebec, Canada
[3] Division of Hematology, University of Toronto, Ontario, Canada
[4] Centre for Applied Teaching in Health Sciences (Centre de Pédagogie Appliquée aux Sciences de la Santé [CPASS]), University of Montreal, Montreal, Quebec, Canada

Corresponding author:

Alain Bestawros, MD
Centre Hospitalier de l'Université de Montréal
Hôpital Notre-Dame – Pavillon Deschamps (G-6149)
1560 Sherbrooke Street East
Montreal (Quebec) H2L 4M1
Phone : 514-890-8000 (27012)
Fax : 514-412-7803

Keywords: clinical judgment, resident evaluation, student evaluation, hematology, script concordance

Abstract

The practice of hematology, like any other profession, requires the acquisition of adequate clinical judgement. Based on a cognitive psychology theory, the script concordance test (SCT) has been developed and validated as an instrument capable of evaluation clinical judgement in various medical specialties. The goal of this study was to examine the usefulness and the psychometric qualities of the SCT in hematology. We constructed a SCT composed of 60 questions and we administered it to 15 junior residents (R1 to R3 in internal medicine), 46 senior residents (R4, R5 and R6) and 17 hematologists from across Canada. After item optimization, the test comprised 51 questions. Its internal consistency measured by Cronbach alpha was 0.83. The test was able to discriminate between junior and senior residents as well as between senior residents and staff. Questions containing an image (n=10) seemed to offer a stronger discriminative potential. Scores obtained by the senior residents correlated to some extent with those obtained on a conventional hematology exam made of multiple choice questions and short-answers (Pearson r : 0.42, p=0.02). The SCT was completed in an average of 36 minutes and was well received by participants. The SCT is a useful and valid evaluation instrument in hematology. It may be used during training to monitor resident progression. It may also be combined to other evaluation methods and used for summative purposes or in continuing education.

Introduction

Most real-life clinical presentations in hematology, and more generally in medicine, carry an element of uncertainty or ambiguity. Clinical judgment has been defined as the ability to make appropriate decisions in uncertain situations[1-2-3]. Throughout medical training, the acquisition of knowledge but most importantly its integration is essential to the development of clinical judgment. This is particularly true in hematology where clinicians are often confronted and must reconcile clinical, laboratory and radiological data, all of which may be subject to interpretation and may even occasionally be inconsistent.

In an effort to objectively evaluate clinical judgment, Charlin et al. developed a test based on the script concordance approach[4-5]. The script concordance test (SCT) is an evaluation tool that is built on the "illness script" theory advanced by Barrows and Feltovich[6]. Stated simply, this theory proposes that networks of knowledge, termed "illness scripts" begin to form during the very first clinical encounter and become more refined as experience is gained[7]. Each time a physician is faced with a new patient, incoming information, such as symptoms, signs or laboratory data, activate the relevant "illness script" and lead to early diagnostic hypotheses. As more information is added, expert clinicians are thought to use mental probability matrices[8], which in conjunction with refined illness scripts help them arrive at the right diagnosis[9].

The SCT aims to evaluate the development of illness scripts in trainees by comparing them to those of a panel of expert clinicians. The SCT has been tested in several medical specialties, such as radiology[10-11], surgery[12-13], urology[14], neurology,[15] internal medicine[16], radiation oncology[17] and emergency medicine[18-19]. In these disciplines, SCTs have consistently shown that scores increase with increasing levels of training, which supports its construct validity[20]. As well, SCTs were shown to have high measures of internal consistency as evidenced by Cronbach's alpha coefficient values generally above 0.70. Finally, SCTs were well received and their practical web-based adm.inistration

allows for easy incorporation of images, such as blood smears or bone marrow aspirates, which is particularly pertinent in hematology.

The primary goal of this study was to develop a valid and reliable tool to assess clinical judgment of junior and senior trainees in hematology using the script concordance methodology. We hereby report on the development and optimization of the test as well as on its psychometric qualities.

METHODS

Test construction

Using the objectives and training requirements for the Hematology Program published by the Royal College of Physicians and Surgeons of Canada (RCPSC)[21], we constructed 20 online clinical scenarios using the SCT format and following guidelines published by Fournier et al[22]. Accordingly, scenarios were inspired from real-life cases and test items contained a degree of uncertainty, imprecision or incompleteness. Each scenario contained an average of 3 questions, for a total of 60 questions. Ten (10) questions included a picture of a blood smear or bone marrow aspiration taken with a 5-megapixel camera in the morphology laboratory at Maisonneuve-Rosemont Hospital. The answer to each question was collected on a 5-point Likert scale (-2, -1, 0, +1, +2). Two examples of SCT scenarios and questions are shown in Figure 1. To ensure that the questions were clear and met the RCPSC objectives, they were reviewed by three academic hematologists, two of which were hematology program directors.

The reference panel

Previous studies have suggested that the reference panel should consist of at least 10 to 15 experts in order to achieve optimal reliability[23-24]. The reference panel comprised 17 hematologists (10 academic and 7 from a community practice). Experts were defined as attending hematologists certified by the RCPSC. Academic hematologists needed to attend regularly on ward and consultation services in university-affiliated hospitals and devote more than 70% of their time to clinical activities. Community hematologists practiced outside university-affiliated hospitals.

Participants

After obtaining ethical approval from the institutional review boards of the Universities of McGill, Montreal and Toronto, junior and senior residents were asked to write the

online test. Fifteen internal medicine residents (R1 through R3) rotating through hematology at University of Montreal and McGill University wrote the test. Senior residents (R4 through R6) were asked to write the test during the annual Hematology retreat organized by the University of Toronto. Of the 57 eligible senior residents, 46 completed the test (15 R4s, 21 R5s and 10 R6s). During the same retreat and just before completing the SCT, all residents wrote a 2.5 hour conventional hematology exam composed of multiple-choice questions (MCQ) and short-answers. All test responses were collected and analyzed anonymously; only information regarding their level of training, experience, and institutional affiliation was collected.

Scoring

Scoring of the SCT is based on the principal that even experts within a field may use slightly different reasoning or cognitive pathways to reach the answer. The scoring system is therefore designed to take into account this variability. Thus, any answer given by an expert has an intrinsic value. As such, there is no one single right answer. The response of examinees to each question is compared to the aggregate responses of the reference panel. The credit attributed to each answer is determined by the frequency with which the answer was chosen by the reference panel. Answers not chosen by experts receive a score of 0. For example, if 10 out of the 17 panel members answered "+2" to a given question, "+2" is established as the modal response and thus participants choosing this answer get 1 point (10/10). If the remaining 7 panel members chose "+1" as the answer, participants with this answer are credited 0.70 points (7/10). Answers not chosen by any experts receive a score of 0.

Ideally, the distribution of answers should be clustered around the mean. A distribution that is too wide likely indicates that the question was misleading or unclear[10]. Alternatively, a question to which all experts give the same answer is acting akin a multiple-choice question and is either assessing solely knowledge or very simple judgment. Such questions were removed from the test during the optimization process.

A conventional scoring system established by members of the Department of Hematology of the University of Toronto was used to evaluate residents' performance on the MCQ and short-answer exam. We were only provided with the final scores for the residents who had written the SCT.

Statistical analysis

Internal consistency was measured by computing Cronbach's alpha coefficient. Optimization of the test comprised two steps. First, questions generating a wide distribution (ie: crossing the ''0'' on the Likert scale) of answers were removed as they were assumed to be unclear or lead to confusion. Second, the test was optimized by eliminating questions with an item-total correlation lower than 0.05. The iterative application of this process ensured that only questions contributing positively to consistency were left. An analysis of variance was used to test the differences between the groups and determine the impact of the level of training on the scores obtained. The linear trend among the groups was calculated using the ANOVA method. The groups were compared to one another. The correlation between the SCT results and those obtained on the conventional hematology exam was calculated using Pearson's r.

RESULTS

Test development and optimization

The initial test contained 20 scenarios and a total of 60 questions. As part of the optimization process, 2 questions were removed because they generated a wide range of answers by the reference panel. Further item reduction was accomplished by iteratively eliminating questions with an item-total correlation lower than 0.05 (n=7 questions). The final instrument contained 18 scenarios with a total of 51 questions. Of note, all 10 items containing an image of a blood smear or bone marrow aspirate remained in the optimized test.

Internal consistency

The Cronbach's alpha coefficient of internal consistency of the full 60-question unoptimized exam was 0.71. The optimization process described above led to the removal of a total of 9 questions. The final optimized test containing 51 questions yielded a Cronbach's alpha coefficient of 0.83.

Differences between participant groups

The test was written by a total of 78 participants (15 R1-R3s, 15 R4s, 21 R5s, 10 R6s and 17 expert hematologists). Mean scores and standard deviations are illustrated in Figure 2. There were statistically significant differences between junior (R1 to R3) and senior residents (R4 to R6) (56.0%±10.2 vs. 70.4%±10.9, p=0.02). Moreover, even within the group of senior residents, the mean score differed significantly between R4s and R5s (61.7%±8.2 vs. 71.0%±8.9, p=0.03) as well as between R5s and R6s (71.9%±8.9 vs. 82.0%±6.6, p=0.03). There were no statistically significant differences between junior residents and R4s (56.0%±10.2 vs. 61.7%±8.2, p=0.33) nor between R6s and experts (82.0%±6.6 vs. 80.0±6.3, p=0.54).

Items containing images

When analyzed separately, the scores obtained on the 10 questions containing images were able to discriminate between R4s and R5s (60.5%±15 vs. 80.0%±9.4, p=0.01) but not between R5s and R6s (80.0%±9.4 vs 87.4%±8.5, p=0.42) nor between R6s and experts (87.4%±8.5 vs. 87.6%±9.3, p=0.55). Moreover, when compared to questions without images, questions containing an image yielded on average a higher item-total correlation (0.45±0.10 vs. 0.29±0.13 respectively).

Panel of experts

The mean score obtained by academic hematologists was not significantly different for the one obtained by community hematologists (80.8%±7.5 vs. 78.9%±4.4, p=0.37). There were also no significant differences or trends in expert scores with regards to the number of years they had been in practice (p=0.24) (Figure 4), nor the province where they practiced (p=0.44).

Relationship with between the SCT and the conventional test

The 46 senior residents who attended the annual Toronto hematology retreat wrote the conventional exam comprised of MCQ and short-answer questions. Junior residents and experts were not invited to write this exam. The scores for the R4s, R5s and R6s were 50.1%±12.1, 72.1%±10.1 and 82.1%±7.1 respectively. These differences were all statistically significant (p=0.03 for R4s vs R5s; p=0.02 for R5s vs R6s). There was a moderate correlation between the scores obtained on the SCT and the conventional exam correlated using the Pearson's r test (0.42, p=0.02).

Perception by participants

All participants completed the test within the allotted 45 minutes. The average time-to-completion of the test was 36 minutes (ranging between 19-51 minutes) with no significant differences between junior, senior residents and experts. Regarding the length of the test, 76% of responding participants found the test "just about right" and 24% found it "too long". With regards to the level of the questions, 91% found them "just about right". Their overall experience was judged "very good" by 50% of the residents,

“good” by 28% and “neutral” by the remaining. Feedback was generally positive. Two issues were raised however: first, the lack of possibility to go back and change the answers; and second, the timing of the exam, being administered after the 2.5 hour conventional hematology exam.

DISCUSSION

The SCT is designed to evaluate clinical judgment by measuring the degree of concordance between examinees' answers and those of a reference panel of experts. In our study, we developed and tested a SCT composed of 60 questions designed to evaluate clinical judgment in hematology trainees. The optimized SCT contained 51 questions.

The test displayed good psychometric properties. Internal consistency measured by the Cronbach alpha coefficient was 0.83. Most agree that coefficient above 0.80 are adequate not only for formative but also summative examinations. In comparison to other evaluation methods, the TCS compares very favorable especially when the time to task completion is taken into consideration[19-25].

The test was also able to discriminate between residents of different training levels, thus confirming its construct validity. Such findings are consistent with those previously reported in the literature in other areas of medicine[10-19]. There were no significant differences in the scores obtained by R6s and the panel of experts. At first glance, this may appear somewhat striking as experts on the panel have more clinical experience than R6s. On the other hand, it is rather reassuring that R6s are performing similarly to experts, as they are expected to enter independent clinical practice within months. Other evaluation tools, such as MCQs, have shown what is known as the ''intermediate effect'' whereby trainees towards the end of their schooling perform better than experts[26]. What such evaluation tools are truly measuring may be questionable.

The SCT was able to distinguish not only between junior and senior residents, but also most interestingly between senior residents themselves (R4s, R5s and R6s). It may therefore constitute a worthy assessment instrument able to monitor resident progress and help identify those for whom extra help or remedial action is required. Similarly, scores obtained on particular questions may help identify areas of strengths and weaknesses of the curriculum.

The validity of the SCT in hematology may be assessed by examining each of the 5 generally accepted sources of validity[27]. The *content* was constructed while keeping in mind the script concordance theory; that is, the assessment of clinical judgment in the context of uncertainty. We followed guidelines previously published[22] and optimized the test by removing all questions that did not contribute positively to the construct. The *response process* examines the relationship between the intended construct and the thought process utilized by participants. The SCT format uses a sequence of three steps (If you were considering…, and you find..., your hypothesis becomes…) that are designed to reproduce the cognitive process of the hypothetico-deductive model of reasoning. While the first two steps are provided by the question itself, the candidate is left with the last step that involves data interpretation. The SCT method is therefore relying on one step of the process to probe into the entire pathway. While this may be a valid assumption, more proof is required[20]. The *internal structure* of the SCT is robust as evidenced by high internal consistency measures (Cronbach alpha). The composition of the expert panel also contributes to that validity. In *relation to other variables*, the SCT has demonstrated moderate correlation with results obtained on the conventional written hematology exam. On the one hand, this indicates that it evaluates a construct – ie clinical judgment – that is sufficiently distinct from factual knowledge. On the other hand, it supports the cognitive script concordance model that relies on the existence of a structure of necessary knowledge. The *consequences* of writing a SCT test depend on whether it is used in formative or sanctionnel intent. In the former context, participants gave a generally positive feedback after completing the test, although the long-term educational impact of the SCT method is unknown. As evidence pertaining to its validity continue to accumulate, it is possible that the SCT be used in a sanctionnel setting, probably in combination with other evaluation tools.

Images

11 résidents

Our study has several strengths, some of which have to do the SCT methodology itself. First, several arguments are in favor of the validity of the SCT method. The test format allows computerized building, administration and correction of the questions. Vast areas of knowledge may be tested in a relatively short time and the whole process may be accomplished at very low cost compared to other assessment tools.

Our study was multicentric and recruited participants and experts from across the country. A novelty brought by our study was the introduction of images of blood smear and bone marrow aspirates into the SCT. Questions containing an image were able to distinguish R4s from the rest of the participants but were not discriminative between the other groups. There are two possible reasons for this finding. For one, the number of questions may have been insufficient. Secondly, in most training programs, the morphology rotation occurs in the R4 year (after the time of the annual retreat), which explains why there is a substantial difference in the scores between R4s and R5s. It would have been interesting to include more images in the SCT as previous studies have found image perception and interpretation quite discriminative of the training level[11]. We correlated the SCT results with those obtained on a conventional written exam. Finally, the feedback from participants was overall positive.

Our study had also some limitations, some of which apply to the SCT methodology itself. First, although empirical arguments support its validity, the basis of some areas remains theoretical. As one of the sources of validity, the process of response requires more investigation. As well, the SCT is designed for the evaluation of clinical judgment pertaining to clinical problems and is less adapted to the basic sciences. Translational interpretation of the latter may however be probed by the SCT. We could have incorporated more questions containing images in our study. This would have perhaps allowed us to make clearer observations. We could have also chosen a better timing to administer the exam so as to minimize participant fatigue.

Au total

CONCLUSION

Our study shows that the SCT is a valid and reliable tool for assessing clinical judgment in hematology. It may play a bigger role during training and help identify areas of strengths of weaknesses for trainees and curricula. It may also play a role in sanctionnel settings such as the RCPCP certification examinations, likely in combination with other assessment tools. It may also be of value in continuing medical education. More studies may consolidate its validity and better define its place within medical education worldwide.

Une image

FIGURE 1

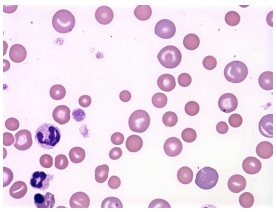
Case 1: You are seeing a 64 year-old man with diffuse adenopathy. He has been previously healthy.						
If you were considering:	**And you find:**	**Your hypothesis becomes:**				
1. Follicular lymphoma	The LDH is 800 U\L (normal <270 U\L)	-2	-1	0	+1	+2
2. Hairy cell leukemia	CT scan reveals extensive retroperitoneal adenopathy measuring up to 10 cm	-2	-1	0	+1	+2
3. Marginal zone lymphoma	A PET scan shows no FDG uptake	-2	-1	0	+1	+2
Case 2: You are seeing a 50 year-old female with anemia and biochemical evidence of hemolysis						
If you were considering:	**And you find:**	**Your hypothesis becomes:**				
1. Paroxysmal nocturnal hemoglobinuria	Serum ferritin is low	-2	-1	0	+1	+2
2. Autoimmune hemolytic anemia	Urine is positive for hemoglobinuria	-2	-1	0	+1	+2
3. Hereditary spherocytosis	Blood smear (click to enlarge) 	-2	-1	0	+1	+2

Figure 1*: Two script concordance test scenarios with 3 questions each. -2=ruled out or almost ruled out; -1=less probable; 0=neither less nor more probable; +1=more probable; +2=certain or almost certain*

FIGURE 2

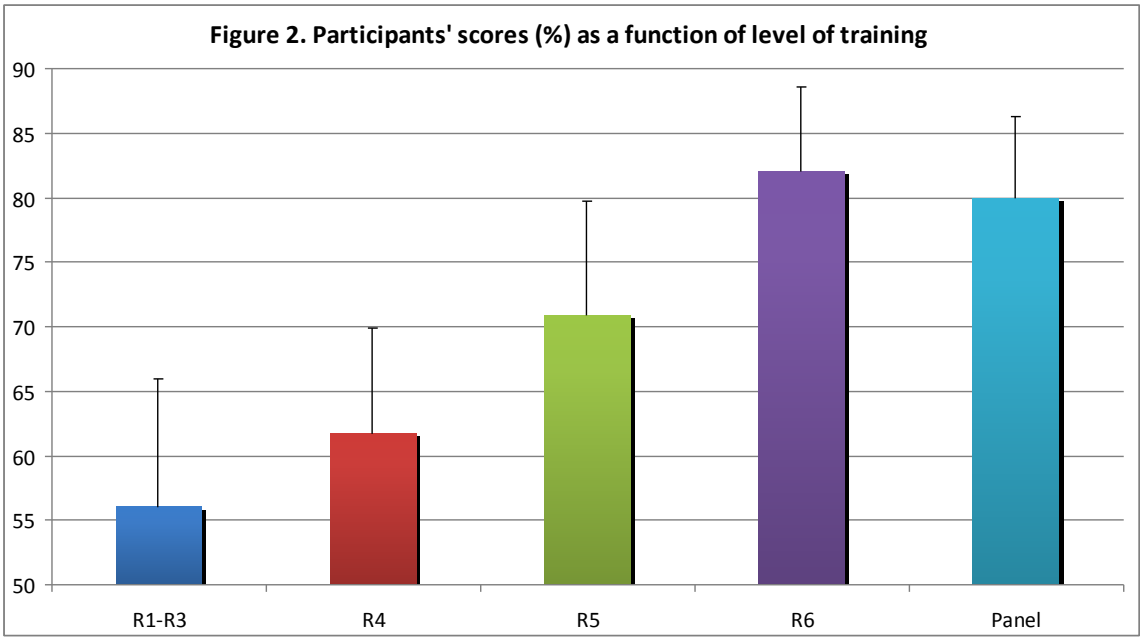


Figure 2. Participants' scores (%) as a function of level of training
90
85
80
75
70
65
60
55
50
R1-R3
R4
R5
R6
Panel

FIGURE 3

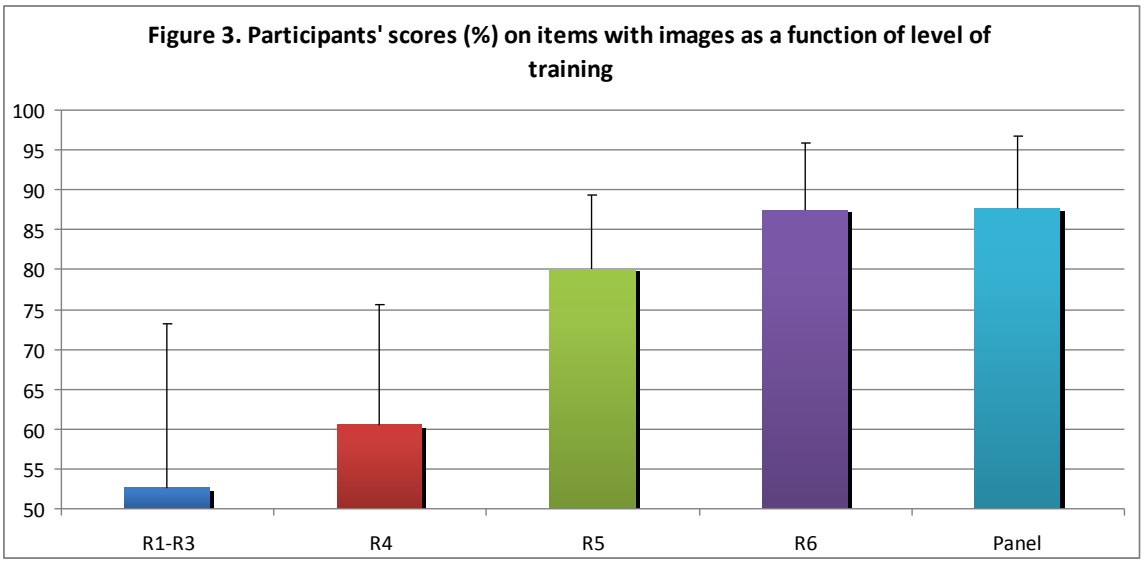


Figure 3. Participants' scores (%) on items with images as a function of level of training
100
95
90
85
80
75
70
65
60
55
50
R1-R3
R4
R5
R6
Panel

FIGURE 4

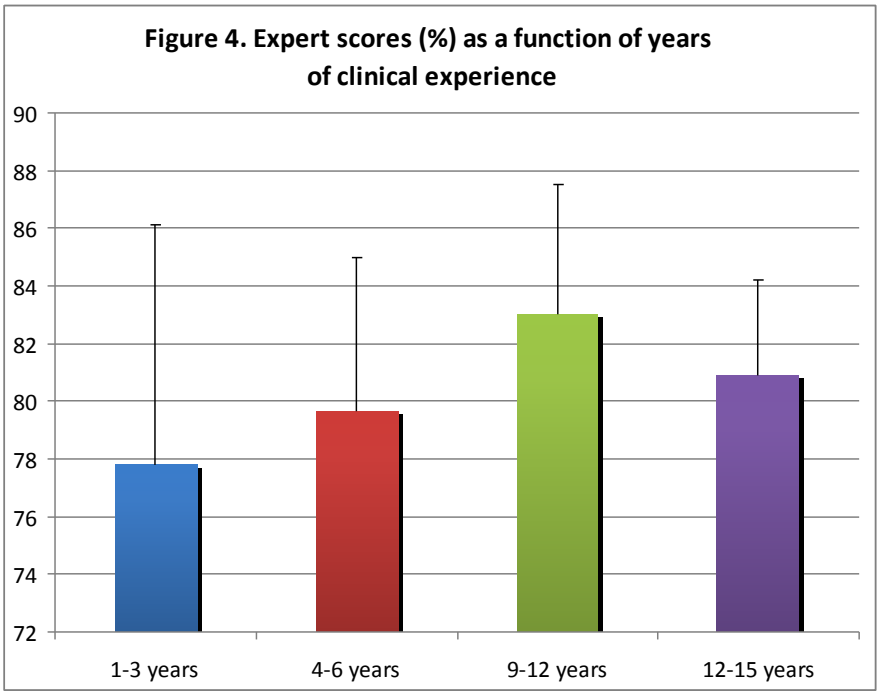


Figure 4. Expert scores (%) as a function of years of clinical experience
90
88
86
84
82
80
78
76
74
72
1-3 years
4-6 years
9-12 years
12-15 years

[1] Schön DA. The reflective practitioner: how professionals think in action. New York: Basic Books: 1983

[2] Montgomery K. How doctors think: clinical judgment and the practice of medicine. Oxford: Oxford University Press; 2006.

[3] Denig P, Wahlstrom R, Caput de Saintonge M, et al. The value of clinical judgement analysis for improving the quality of doctors' prescribing decisions. *Medical Education*. 2002; 36:770-780.

[4] Charlin B, Roy L, Brailovsky CA, Van der Vleuten CPM. The Script Concordance Test: A Tool to Assess the Reflective Clinician. *Teaching and Learning in Medical Education*, 2000; 12:189-195.

[5] Charlin B, Van der Vleuten C. Standardized assessment of ill-defined clinical problems: The script concordance approach. *Evaluation in the Health Professions*, in Press, 2004.

[6] Barrows HS, Feltovitch PJ. The clinical reasoning process. *Med Educ*. 1987;21:86-91.

[7] Charlin B, Boshuizen HPA, Custers EJFM, Feltovitch PJ. Scripts and clinical reasoning. *Med Ed*. 2007;31:1178-84.

[8] Papa FJ, Shores JH, Meyer S. Effects of pattern matching, pattern discrimination and experience in the development of diagnostic expertise. *Acad Med.* 1990;5:521-22.

[9] Charlin B, Tardif J, Boshuizen HPA. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Acad Med.* 2000;75:182-190.

[10] Charlin B, Brailovsky CA, Brazeau-Lamontagne L, Samson L, Leduc C. Script Questionnaires: Their Use for Assessment of Diagnostic Knowledge in Radiology. *Medical Teacher*, 1998; 20: 567-571.

[11] Brazeau-Lamontagne L, Charlin B, Gagnon R, Samson L, Van der Vleuten C. Measurement of perception and interpretation skills along radiology training: Utility of the script concordance approach. *Medical Teacher*, 2004; 26:326-32.

[12] Meterissian S, Zabolotny B, Gagnon R, Charlin B. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills ? *Amer Jour of Surg*, 2007; 193: 248-251.

[13] Meteressian SH. A novel method of assessing clinical reasoning in surgical residents. *Surg Innov*. 2006;13:115-9.

[14] Sibert L, Darmoni SJ, Dahamna B. On line clinical reasoning assessment with Script Concordance test in urology : results of a French pilot study. *BMC Medical Education*. 2006; 6:45.

[15] Lubarsky S, Chalk C, Kazitani D, Gagnon R, Charlin B. The script concordance test: a new tool assessing clinical judgment in neurology. Can J Neurol. Sci 2009; 36:326-331.

[16] Maire I, Sibert L, Roussel F, et al. Le test de concordance de script : un nouvel outil d'évaluation du raisonnement et de la compétence clinique en médecine interne ? *La revue de med int.* 2005; 26 :501-07.

Je crois que le 1er auteur est Marie

[17] Lambert C, Gagnon R, Nguyen D, Charlin B. The script concordance test in radiation oncology: validation study of a new tool to assess clinical reasoning. *Radiat Oncol* 2009; 4:7.

[18] Carriere B, Gagnon R, Charlin B, Downing S, Bordage G. Assessing clinical reasoning in paediactric emergency medicine: validity evidence for a script concordance test. *Ann Emerg Med* 2009;53(5):647-52.

[19] Fournier JP. Thiercelin D, Pulcini C, Alunni-Perret V, Gilbert E, Minguet JM, Bertrand F. Clinical reasoning assessment in emergency medicine: script concordance tests are more efficient to detect clinical experience that rich-context multiple-choice questions. Pédagogie Médicale 2006; 7:20-30.

[20] Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten C. Script concordance testing: a review of published validity evidence. *Med Educ* 2011; 45: 329-338.
[21] http://crmcc.medical.org/residency/certification/objectives/hematology_f.pdf
[22] Fournier JP, Demeester A, Charlin B. Script concordance tests : guidelines for construction. *BMC Med Inform Decis Mark* 2008 ;8 :18.
[23] Gagnon R, Charlin B, Coletti M, Sauvé E, Van der Vleuten C. Assessment in context of uncertainty: how many experts are needed on the Script Concordance Test panel of reference? M*ed Educ*. 2005; 39:284-91.
[24] Charlin B, Gagnon R, Pelletier J, Coletti., Abi-Rizk G, Nasr C, Sauvé E, Van der Vleuten C. Assessment of clinical reasoning in the context of uncertainty; the effect of variability within the reference panel. *Med Educ*. 2006; 40:848-54.
[25] Kelly W, Durning S, Denton G. Comparing a script concordance examination to a multiple-choice examination on a core internal medicine clerkship. Teaching and learning in medicine. 2012;24(3):187-93.
[26] Schmidt HG, Boshuisen HPA. On the origin of intermediate effects in clinical case recall. Mem Cogn. 1993; 21:328-51.
[27] Downing SM. Validity: on meaningful interpretation of assessment data. *Medical education* 2003; 37: 830-7.