Pédagogie Médicale
REVUE INTERNATIONALE FRANCOPHONE D'ÉDUCATION MÉDICALE

Le test de concordance de script, un instrument d'évaluation du raisonnement clinique

Bernard CHARLIN, Robert GAGNON*, Louis SIBERT **, Cees Van der VLEUTEN****

Résumé *Contexte : La capacité à résoudre des problèmes cliniques mal définis caractérise les médecins expérimentés et compétents. En matière d'évaluation, les modalités actuelles exigent un consensus entre les membres du jury sur la bonne réponse, or les médecins expérimentés varient dans le processus de raisonnement pour ce type de problème qui est en conséquence généralement exclu des évaluations. Ceci explique peut-être pourquoi les médecins expérimentés n'obtiennent pas, dans les évaluations écrites de compétence, des scores plus élevés que les médecins en fin de formation.* ***But :*** *Le test de concordance de script (TCS) est conçu pour mesurer la capacité à résoudre des problèmes mal définis. L'objectif de l'article est de montrer l'intérêt du TCS aux enseignants en sciences de la santé en leur apportant les informations pratiques nécessaires à l'utilisation d'un tel test.* ***Méthode :*** *Les bases théoriques du test, les principes concernant la tâche soumise aux candidats, la façon d'enregistrer les réponses et le processus d'établissement des scores sont expliqués. Les étapes à suivre pour construire un TCS sont décrites. Les qualités psychométriques observées dans une série de travaux de recherche sont rapportées.* ***Résultats :*** *Plusieurs études menées dans différentes disciplines ont montré que le test permet de discriminer des niveaux d'expérience différents avec des caractéristiques psychométriques intéressantes (validité de construit, validité prédictive, fidélité). Ces résultats sont brièvement présentés et commentés.* ***Conclusion :*** *Le TCS permet une évaluation standardisée des processus de raisonnement sur des problèmes cliniques mal définis.*

Mots clés raisonnement clinique ; évaluation ; problèmes mal définis ; étudiants en médecine ; résidents ; internes ; médecins

Summary ***Context:*** *The capacity to solve ill-defined problems is a characteristic of experienced and competent physicians. The current methods used in standard evaluation of competence is to get a panel of experts to reach a consensus about the «good answer» whereas it is observed that experienced physicians vary in their reasoning process in the realm of ill-defined problems. Thus, these kinds of problems are usually excluded from the tests. This may explain why, on written assessments, experienced physicians frequently obtain lower scores than at the end-of-training physicians.* ***Goal:*** *The Script Concordance Test (SCT) is set to measure the capacity to resolve problems in the context of ill-defined clinical situations. The object of this article is to demonstrate the usefulness of the SCT to teachers in health sciences and to give all the practical information to use the test.* ***Method:*** *The theoretical foundations of the test, the principles surrounding the task given to candidates, the processes of answer recording and score setting are explained. The steps to follow in the construction of a SCT are described. The essential psychometric properties derived from a series of studied are also reported.* ***Results:*** *Studies in several medical fields have shown that the SCT do discriminate levels of clinical experience and possess satisfying psychometric properties (construct and predictive validity, reliability). Main results are presented and commented.****Conclusion:*** *The SCT permits a standardized assessment of reasoning processes in the context of ill-defined clinical situations.*

Keywords clinical reasoning evaluation; ill-defined problems; medical students; residents; interns; physicians.

Pédagogie Médicale 2002 ; 3 : 135-144

** Université of Montréal; Canada; ** Université de Rouen; France *** Université de Maastricht, Pays Bas.*
Correspondance : Bernard Charlin, URDESS, Faculté de Médecine-Direction, Université de Montréal, CP 6128, succursale Centre-ville, Montréal, Québec, H3C 3J7 Canada - mailto:bernard.charlin@umontreal.ca

Introduction

Les outils de mesure par écrit du raisonnement clinique existants à l'heure actuelle présentent un défaut majeur. Les médecins possédant plusieurs années d'expérience obtiennent des notes à peine meilleures et parfois inférieures à celles de médecins moins expérimentés (internes, résidents) dans les simulations de résolution de problèmes cliniques[1-2]. Ceci pourrait s'expliquer par la nature des problèmes habituellement soumis dans ce type d'évaluation. Il s'agit généralement de problèmes bien définis, alors que la capacité à résoudre des problèmes mal définis caractérise davantage les médecins expérimentés et compétents.

Les problèmes auxquels font face les professionnels sont, en effet, de deux sortes[3]. Certains découlent de ce que Schön nomme la rationalité technique. Il s'agit de problèmes que l'on peut résoudre en appliquant des solutions tirées des sciences fondamentales et des sciences appliquées, dans des contextes définis et stables, pour atteindre des buts clairs et prévisibles. Ce type de problème est dit bien défini. En pratique clinique cependant, chaque patient est particulier, les pathologies sont parfois multisystémiques et les buts à atteindre ne sont pas toujours évidents. L'information disponible au début d'une rencontre clinique est parfois imparfaite, inconsistante ou même inexacte et les données doivent être activement recherchées de façon à formuler le problème et à le résoudre. Les cliniciens sont fréquemment confrontés à ce que l'on nomme des problèmes mal définis. Selon Schön les connaissances nécessaires à la résolution de ce type de problème sont différentes des connaissances techniques utilisées pour résoudre les problèmes bien définis. Il les nomme connaissances professionnelles et affirme que les institutions de formation mettent trop d'emphase sur l'apprentissage et l'évaluation des connaissances techniques et négligent les connaissances professionnelles, alors que ce qui fait l'essence des professions, au-delà de la maîtrise des savoirs techniques, c'est la capacité à résoudre les problèmes mal définis.

Pour que les examens puissent discriminer en fonction de la compétence, il faudrait qu'ils contiennent des tâches de résolution de problèmes mal définis. Cela amène des difficultés en matière de correction. En effet, l'usage veut que l'on demande aux membres de jurys qui valident les réponses aux questions d'examen de donner par consensus la bonne réponse à exiger des candidats. Or la recherche sur le raisonnement clinique[4-5] montre que les médecins expérimentés divergent considérablement dans leur cheminement vers la solution de problèmes mal définis, même s'ils parviennent généralement aux mêmes diagnostics et à des décisions similaires en matière de prise en charge des patients. De fait, dès que l'on essaie d'introduire dans des examens des questions correspondant à des problèmes complexes, les experts ne parviennent pas à se mettre d'accord sur la bonne réponse à exiger et les questions incriminées sont éliminées. Par ailleurs, le choix d'une bonne réponse, dans un QCM (questionnaire à choix multiples) par exemple, reflète mal les nuances qu'implique la nature du raisonnement dans les problèmes mal définis.

Le test de concordance de script (TCS) apporte une réponse à cette problématique[6]. Il est innovateur à trois niveaux : le stimulus soumis aux étudiants, le moyen d'enregistrer leur réponse et le processus d'établissement de score pour chaque question. L'article a pour objectif de montrer son intérêt en apportant les informations pratiques nécessaires à son utilisation. Il décrit ses bases théoriques, ses caractéristiques, ses principes de construction et d'élaboration du système de notation et expose les résultats issus des publications initiales. Le TCS est conçu pour évaluer les sciences cliniques. Il n'est pas adapté à l'évaluation des sciences fondamentales.

La base théorique du test

Le développement de l'expertise chez les professionnels est très lié à l'organisation des connaissances. La théorie des scripts[7-8-9], postule qu'en situation clinique les médecins mobilisent des réseaux de connaissance préétablis, des scripts, qui sont utilisés pour comprendre cette situation et agir en fonction de buts diagnostiques, d'investigation ou de traitement. Les scripts diagnostiques contiennent les associations que le clinicien a établies entre l'entité pathologique et ses différents attributs cliniques (les signes et symptômes notamment). Ce sont ces associations, ces liens, qui dans le contexte clinique permettent de prendre les décisions concernant la force ou la faiblesse d'une hypothèse ou de décider si un signe ou un symptôme n'est jamais associé à une maladie et donc que cette hypothèse doit être éliminée. Des liens similaires permettent de raisonner pour aboutir aux décisions d'investigation ou de traitement[9] Les scripts des médecins expérimentés varient dans leurs détails, parce que l'expérience personnelle de chacun diffère, mais ils sont proches pour l'essentiel. Si ce n'était pas le cas, les cliniciens seraient incapables de communiquer entre eux à propos

Tableau 1 : Exemple de test destiné à des internes en urologie, décrivant le format des items pour l'investigation diagnostique, pour la demande d'examens complémentaires, pour la prise en charge thérapeutique.

Questionnaire diagnostique

Scénario clinique :
Une patiente de 63 ans est hospitalisée en urgence pour une première crise de colique néphrétique droite. Les douleurs cèdent sous la prise d'antalgiques et il n'y a pas de fièvre. La bandelette urinaire montre 3 croix de sang, la radiographie d'abdomen sans préparation en urgence semble normale.

Si vous pensez à *(option diagnostique)*	Et qu'alors vous trouvez *(nouvelle information clinique)*	L'effet sur l'hypothèse diagnostique est le suivant
1- Tumeur urothéliale	Échographie rénale normale	-2 -1 0 +1 +2
2- Lithiase radio transparente	PH urinaire à 5,5	-2 -1 0 +1 +2

Entourez la proposition qui vous semble adéquate :
-2 l'hypothèse est pratiquement éliminée ; -1 l'hypothèse devient moins probable ;
0 l'information n'a aucun effet sur l'hypothèse ; +1 l'hypothèse devient plus probable.
+2 il ne peut s'agir pratiquement que de cette hypothèse.

Questionnaire d'investigation

Scénario clinique :
Vous prenez en charge le bilan étiologique de cette patiente

Si vous pensiez faire *(option d'examen paraclinique)*	Et qu'alors vous trouvez *(nouvelle information obtenue par l'anam - nèse ,l'examen clinique, l'imagerie ou un examen complémentaire)*	L'effet sur la nécessité de demander ce test est le suivant
Cystoscopie	Hématurie	-2 -1 0 +1 +2
Urétéropyélographie rétrograde	Cytologies urinaires négatives	-2 -1 0 +1 +2

Entourez la proposition qui vous semble adéquate :
-2 absolument contre-indiqué ; -1 peu utile ou plutôt néfaste ; 0 non pertinent dans cette situation.
+1 utile et souhaitable. +2 indispensable.

Questionnaire de thérapeutique

Scénario clinique :
Un patient de 25 ans est admis aux urgences à la suite d'une chute à moto avec choc direct sur le pubis. L'état hémodynamique est stable. La radiographie du bassin révèle une fracture du bassin avec une disjonction de la symphyse pubienne. Vous prenez en charge la conduite thérapeutique de ce patient.

Si vous pensiez faire *(une option thérapeutique)*	Et qu'alors vous trouvez *(nouvelle information obtenue par l'anamnèse ,l'examen clinique, l'imagerie ou un examen complémentaire)*	L'effet sur la pertinence de réaliser ce geste est le suivant
Cathéter suspubien	Urétrorragie sans globe vésical	-2 -1 0 +1 +2
Laparotomie exploratrice	Rupture sous péritonéale de vessie	-2 -1 0 +1 +2
Cathéter suspubien	Rupture urétrale et globe vésical	-2 -1 0 +1 +2

Entourez la proposition qui vous semble adéquate :
-2 absolument contre-indiqué ; -1 peu utile ou plutôt néfaste ; 0 non pertinent dans cette situation.
+1 utile et souhaitable ; +2 indispensable.

des maladies de leurs patients et ils n'atteindraient pas les mêmes diagnostics dans des situations similaires.
L'approche du test consiste à présenter aux candidats une série de problèmes cliniques et à leur demander de prendre des décisions diagnostiques, d'investigation ou de traitement lorsque des éléments d'information clinique supplémentaire leur sont communiqués (Tableau 1). Le test permet de mesurer le degré d'organisation des connaissances, de vérifier si les connaissances sont élaborées[10], c'est-à-dire organisées pour agir efficacement dans le contexte clinique. Le test vise à mesurer l'adéquation des liens au sein des connaissances cliniques, plutôt que la simple présence des éléments de ces connaissances. Le système d'établissement des scores est conçu pour mesurer le degré de similitude qui existe entre le script du candidat et ceux des médecins expérimentés d'un panel de référence, d'où le nom de test de concordance. Le test est bâti à partir des questions posées ou des gestes effectués par les médecins expérimentés dans la situation décrite pour parvenir rapidement à une solution dans le problème clinique à résoudre et à partir de la signification apportée aux données qu'ils obtiennent par ces questions ou ces gestes.

Les principes du TCS

Le stimulus

Il consiste à décrire une situation clinique susceptible d'être rencontrée par l'étudiant (ou le médecin examiné) dans son exercice actuel ou futur. Cette situation doit être problématique, même pour un expert du domaine. Plusieurs options de diagnostic, ou de prise en charge, sont possibles et la description ne contient pas toutes les données nécessaires pour résoudre le problème.
La tâche consiste à envisager l'effet que produirait la découverte d'une nouvelle donnée sur le statut d'une des options pertinentes dans la situation. La question suivante investigue l'effet d'une autre donnée sur une autre option. Au sein d'une vignette (la description du cas clinique), chaque question est indépendante des autres. Le Tableau 1 décrit le format des questions. Les intitulés varient légèrement au sein des formats diagnostiques, d'investigation ou de traitement.

Le recueil des réponses

La recherche sur le raisonnement dans des situations complexes a montré qu'il n'est pas fait de combinaisons de probabilités liées à la présence ou l'absence des signes et symptômes, mais plutôt de jugements de nature qualitative[11] sur l'effet que ces attributs ont sur le statut (le degré d'activation) d'une hypothèse. En situation diagnostique, par exemple, il va s'agir de décider si telle ou telle donnée renforce le statut de l'hypothèse, le diminue, ou encore n'apporte rien par rapport à l'hypothèse. Le TCS reflète cela en utilisant une échelle de Likert pour enregistrer les réponses (voir Tableau 1). 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 qualitatif.

Tableau 2 : Méthode d'établissement des scores

	-2	-1	0	+1	+2	Commentaires
Nombre de réponses parmi les membres du panel	0	0	5	4	1	Identifier la réponse la plus choisie l (ici la valeur 0)
Mécanisme de création des scores	0	0	5/5	4/5	1/5	Division par le nombre de membres ayant donné la réponse la plus choisie (5 ici)
Crédit pour l'item	0	0	1	0,8	0,2	Points obtenus par l'étudiant pour cet item

Le processus d'établissement des scores

Ce processus est sous-tendu par un principe : la réponse de chaque membre du panel de référence reflète une opinion valide qui devrait être prise en compte et les réponses pour lesquelles il n'existe pas un consensus ne devraient pas être rejetées. La méthode des scores combinés[12-13] permet de prendre en compte la variabilité trouvée habituellement entre médecins expérimentés lorsqu'ils répondent à des questions complexes[5]. Les scores à chaque item (en docimologie, on appelle item chaque question qui reçoit un score) découlent des réponses données au sein du panel de référence.

Les membres du panel complètent le test individuellement. Leurs réponses sont utilisées pour bâtir la grille de correction (Tableau 2). Pour chaque item la réponse donne droit à un crédit qui correspond au nombre de panélistes qui l'ont choisie. Prenons, par exemple, un panel qui serait composé de 10 membres. Pour un item 5 membres ont choisi la réponse 0 ; quatre ont choisi la réponse + 1 et un a choisi la réponse + 2. Les scores bruts sont 5/10 pour la réponse 0, 4/10 pour la réponse + 1 et 1/10 pour la réponse + 2. Les réponses non choisies par les membres du panel reçoivent 0.

Tous les items doivent offrir le même crédit maximum. Pour obtenir ceci, les scores bruts sont transformés proportionnellement pour obtenir un crédit d'un point pour la réponse qui a été la plus choisie par les membres du panel. Les autres choix reçoivent un crédit partiel. Pour transformer les scores, tous sont divisés par le nombre de membres qui ont donné la réponse la plus choisie. Dans notre exemple, tout étudiant qui répond 0 reçoit 1 point, + 1 reçoit 4/5^{e} de point et + 2 reçoit 1/5^{e} de point. Les autres réponses sont cotées 0. Le score total pour le test est la somme de tous les crédits obtenus à chaque item. Ce score total est ensuite transformé par une simple règle de trois pour obtenir une note exprimée sur 100. Un candidat qui aurait systématiquement choisi comme la majorité des membres aurait une note de 100.

Étapes de construction d'un TCS

La première étape, comme pour tout examen, consiste à bien définir la population et le domaine médical à évaluer. La deuxième consiste à sélectionner des situations représentatives du domaine, en terme de fréquence, mode de présentation, gravité et possibilité de traitement. Cette étape nécessite la collaboration d'un petit groupe d'experts (2 experts sont généralement suffisants à ce stade d'élaboration du test) chargés de sélectionner les situations cliniques. Pour chaque situation, ils doivent spécifier : 1- les hypothèses pertinentes de diagnostic, d'investigation ou de traitement ; 2- les principaux signes à rechercher à l'anamnèse et à l'examen physique, les principaux examens complémentaires à demander pour résoudre le problème ; 3- quelles informations cliniques, positives ou négatives, ils chercheraient pour vérifier les hypothèses. Les items du test sont rédigés à partir du matériel obtenu à cette étape. Il est inutile de chercher des situations ou données cliniques inhabituelles pour que le test puisse être discriminant. Le test discrimine très bien en utilisant les situations communes de chaque domaine.

La troisième étape consiste à rédiger les vignettes cliniques et les items du test à partir du matériel obtenu à l'étape 2. Chaque situation clinique est ensuite présentée sous forme de vignette de quelques lignes. La vignette contient suffisamment d'information pour décrire un problème crédible, mais pas assez pour que le problème puisse être résolu tel quel, même par un expert du domaine. La vignette est suivie par une série d'items. Le format d'un item dépend de l'objectif d'évaluation (diagnostic, investigation, traitement). Chaque item est constitué de 3 parties (Tableau 1). La première partie comprend une hypothèse diagnostique, une investigation paraclinique ou une option thérapeutique, la deuxième partie présente une information nouvelle (par exemple, un signe clinique, le résultat d'un examen complémentaire). La troisième partie est une échelle de type Likert. Chaque item est bâti de telle manière qu'une réflexion est nécessaire pour y répondre et chaque item est indépendant des autres (ceci est clairement précisé dans les instructions données aux candidats). Le but est non pas de déterminer l'effet cumulatif d'une série d'informations cliniques mais de déterminer l'effet d'une information clinique ou para-clinique isolée sur une hypothèse diagnostique, une proposition d'investigation ou une option thérapeutique.

L'étape de validation du test consiste à demander à quelques experts et à quelques personnes appartenant à la population qui doit être évaluée (médecins en pratique, internes ou étudiants) de passer le test afin de corriger ou d'éliminer les items qui apparaissent confus ou non pertinents. La dernière étape est celle de la construction des grilles de réponse. Elle a été décrite

plus haut dans le texte et dans le Tableau 2. La détermination de la composition du panel de référence est une étape importante. Il n'y a pas de règle absolue ici. Cela dépend de l'objet de l'évaluation. Pour évaluer, par exemple, le raisonnement clinique de résidents en médecine familiale sur les problèmes courants d'infection gynécologique, il convient de se demander si ce sont des gynécologues ou des médecins de famille qui voient le plus ce type de pathologie et de décider ensuite de la composition du panel. Le nombre de membres sur le panel doit pouvoir refléter les variations d'opinion entre médecins expérimentés. Un nombre de 7 à 10 permet d'obtenir une bonne précision de correction. Le nombre total d'items suffisant pour un test de Concordance de Script dépend de l'objet de l'évaluation. Pour une activité de formation continue, le nombre d'items requis n'est pas très élevé (20 à 30). Pour une évaluation sommative, où les préoccupations de fidélité sont importantes et où il est nécessaire de bien sonder l'ensemble du domaine (validité de contenu), le nombre d'items est plus important (60 ou plus).

TCS et questions à choix multiples

Le tableau 3 illustre la différence dans la nature des stimulus utilisés dans des questions à choix multiples (QCM) et dans des TCS. Le QCM présenté est un exemple de QCM de qualité[14], qui vise à évaluer les fonctions cognitives supérieures et non la simple mémorisation. Ce type de QCM est appelé QCM à contexte riche. Il est important de souligner la différence importante existant entre un QCM et un TCS, bien que tous deux soient des examens standardisés (tous les étudiants sont soumis aux mêmes tâches cognitives) à correction objective. Dans un QCM, toutes les données nécessaires à la résolution du problème posé sont présentes dans la

Tableau 3 : Différences entre TCS et QCM

QCM[14]

Un jeune homme de 20 ans est poignardé dans le bras avec un couteau.
La face dorsale de l'avant-bras est insensible, ainsi que la face dorsale de la main entre le pouce et l'index.
Les extenseurs du poignet sont paralysés et il ne peut étendre le pouce entre les articulations métacarpo-phalangienne et interphalangienne. Quel nerf a été atteint?

A- le nerf médian
B- le nerf radial
C- le nerf cubital
D- le nerf inter-osseux postérieur
E- le nerf brachial postérieur

Dans un QCM, l'ensemble des données décrites permet de répondre au problème posé et il n'existe qu'une seule bonne réponse qui est soumise avec d'autres réponses qui sont des leurres.

TCS

Un jeune homme de 20 ans est poignardé dans le bras avec un couteau.
La face dorsale de l'avant bras est insensible.

Si vous pensez à	Et qu'alors vous trouvez	L'effet sur l'hypothèse diagnostique est
Une lésion du nerf médian	une paralysie des extenseurs du poignet	2 -1 0 +1 +2

Dans un TC, même un expert ne peut résoudre le problème avec les données décrites dans la vignette. D'autres informations cliniques sont nécessaires. Chaque item mesure l'effet provoqué par une nouvelle information sur le statut d'une des hypothèses pertinentes à la situation.

vignette. On demande au candidat d'intégrer toutes ces données et de prendre une décision, qui est enregistrée par un choix au sein d'une liste. La tâche est donc globale et on enregistre le résultat du processus de raisonnement, pour lequel une seule bonne réponse est possible. Dans un TCS, la vignette ne contient pas toutes les données nécessaires, plusieurs options sont possibles et pour pouvoir prendre une décision définitive d'autres données sont nécessaires. La tâche consiste donc à évaluer l'effet qu'une de ces nouvelles données aurait sur une des options pertinentes à la situation. La réponse enregistrée concerne donc le processus du raisonnement et non son résultat. Il est logique de demander un consensus entre experts pour la réponse à obtenir à un QCM, car cela représente l'issue d'un raisonnement. On ne peut par contre s'attendre à un consensus pendant le processus de raisonnement (ce qui est le cas dans un TCS), car pendant ce processus, les médecins ne suivent pas le même cheminement[4-5], même s'ils arrivent en fin de compte généralement au même diagnostic.

Si, en accord avec la distinction de Schön, on accepte l'idée que le savoir dans une profession comporte un savoir technique et un savoir professionnel, il est logique de concevoir une évaluation de ces deux composantes du savoir par deux instruments distincts. Les QCM sont utilisés pour évaluer la connaissance des règles, des lois et des procédures bien établies, tandis que des TCS sont utilisés pour évaluer la compétence à raisonner dans des problèmes complexes qui ne peuvent se résoudre par une simple application de connaissances. Le champ de la médecine basée sur les preuves fournit un exemple caractéristique[15]. La profession médicale ne peut se résumer à l'application de règles établies dans des contextes précis, sur des populations de patients sélectionnés par des critères d'inclusion stricts. Dans la réalité clinique, il est nécessaire de tenir compte des particularités de chaque patient et de données parfois contradictoires. Dans cette perspective la connaissance de règles de pratique basée sur les données probantes peut donc être évaluée par des QCM si le niveau de preuve est élevé (par exemple niveaux 1 et 2 de l'échelle de la Canadian Task Force on Periodic Examination)[16]. Lorsque le niveau de preuve est plus bas pour des raisons d'impossibilité technique ou d'insuffisance d'études, l'analyse contextuelle devient déterminante et explique, pour une bonne part, les divergences d'experts. L'utilisation de TCS devient alors tout à fait adaptée[17].

Les qualités psychométriques du TCS

Les premières études portant sur le test ont été publiées en 1998[18-19]. Elles vérifiaient sa validité de construit. On appelle ainsi l'adéquation entre un test et le cadre théorique dans lequel il est construit. La démarche consiste à poser une hypothèse qui découle des intentions avec lesquelles le test à été élaboré, puis à faire des mesures empiriques pour vérifier cette hypothèse. Ces deux études, menées en radiologie et en gynécologie-obstétrique montraient que les médecins expérimentés obtiennent des scores supérieurs à ceux des internes et résidents, qui sont eux même supérieurs à ceux des externes apportant ainsi des arguments en faveur de la validité de construit. Cet effet a été trouvé dans toutes les études publiées ultérieurement.

La validité prédictive du test a fait l'objet d'une investigation publiée en 2001[20]. Le test a été administré à tous les étudiants d'une faculté immédiatement avant leurs examens de fin de formation. Le test portait sur les connaissances cliniques en chirurgie. La cohorte d'étudiants qui s'est ensuite engagée en formation de médecine familiale a été suivie jusqu'à l'examen d'obtention de la spécialité, deux ans plus tard. Compte tenu de l'objet de mesure du test, l'organisation des connaissances adaptée aux tâches cliniques, nous faisions l'hypothèse que le test prédirait bien les résultats des tests de raisonnement clinique et moins bien ceux concernant les habiletés cliniques, mesurées par un examen clinique objectif et structuré (ECOS). Les résultats ont confirmé cette hypothèse, suggérant que les étudiants qui organisent bien leurs connaissances à un moment de leur formation continuent à le faire aux stades ultérieurs de formation.

Les études décrites ci-dessus ont donné des indications relatives à la fidélité du test. Un test est dit fiable s'il donne avec constance un même résultat et s'il mesure de façon précise avec le moins d'erreur d'estimation possible. Le coefficient Alpha de Cronbach, un des indices de fidélité, mesure la cohérence interne du test, c'est-à-dire jusqu'à quel point les items contribuent à la mesure d'une même dimension dans le test. Il est communément admis en pratique évaluative qu'un test est fiable lorsque ce coefficient a une valeur supérieure à 0,80. Essentiellement, la fidélité d'un test dépend du nombre d'items contenus dans le test. Naturellement si un test a besoin d'un grand nombre d'items pour obtenir un degré de fidélité satisfaisante, cela pose des problèmes lorsque le temps d'examen devient excessif. À partir des données recueillies dans les

trois études citées ci-dessus il a été montré, par des études de généralisabilité, que l'on atteint systématiquement une valeur de 0.80 avec 60 items, ce qui permet d'administrer les tests en une heure ou moins. Les études ultérieures ont ici encore confirmé cet effet.

La stabilité des scores en fonction des cultures a été étudiée en urologie[21]. Un TCS a été élaboré par des enseignants français. Il a ensuite été traduit en langue anglaise. Le test a été administré d'une part aux étudiants d'une université et aux internes d'une inter-région en France et d'autre part aux étudiants et résidents d'une université canadienne anglophone. Deux panels de référence ont été constitués, l'un avec des urologues français, l'autre avec des urologues anglophones. Les tests des étudiants et résidents français ainsi que ceux des étudiants et résidents canadiens anglophones ont fait l'objet d'une double correction, l'une avec la grille de réponse obtenue par le panel de référence français, l'autre par le panel de référence canadien. Nous faisions l'hypothèse que le test permettrait de discriminer en fonction de la compétence quel que soit le pays, mais qu'il existerait un biais en faveur de la culture médicale dans un même pays. Les résultats ont confirmé la capacité de l'instrument à distinguer le niveau de formation des participants quelle que soit la grille utilisée et les sujets français obtenaient des scores supérieurs quand ils étaient jugés par le panel français tandis que les sujets canadiens obtenaient des scores supérieurs quand ils étaient jugés par le panel canadien. L'instrument garde donc ses capacités de discrimination lorsqu'il est utilisé dans une autre culture, mais avec une meilleure concordance dans la culture d'origine.

Dans les examens habituels, la méthode de construction des grilles de réponse par le jury consiste à demander aux experts de déterminer par consensus la bonne réponse qui doit être exigée des étudiants pour chaque item. Nous avons examiné[22] l'effet produit par deux méthodes d'établissement des scores, l'une commune pour le TCS, la méthode des scores combinés, l'autre basée sur la méthode habituelle de recherche de consensus. Cent cinquante étudiants ont passé le test à la fin de leur stage d'externat en gynécologie et sept gynécologues enseignants ont accepté de passer le test, comme le font les étudiants. Deux grilles de réponses ont été établies par un autre groupe de gynécologues enseignants. L'une par la méthode des scores combinés (les membres du panel de référence remplissent le test individuellement), l'autre réalisée un an plus tard, dans laquelle la bonne réponse pour chaque item est déterminée par consensus. L'étude a révélé que dans 59 % des cas, les membres du panel de référence donnent une réponse différente dans les deux contextes. Le raisonnement des experts à propos d'un même cas présenté dans une vignette diffère donc selon le contexte dans lequel sont placés les experts. Lorsqu'ils sont seuls, ils résolvent le cas en utilisant leurs connaissances personnelles et les souvenirs de patients présentant des pathologies similaires, tandis qu'ils partagent d'autres données lorsqu'ils raisonnent en groupe. Ceci conduit à se demander s'il est légitime de donner des scores à des étudiants à partir de raisonnements d'experts qui sont placés dans un contexte différent de celui dans lequel les étudiants sont placés lorsqu'ils passent leurs examens. L'étude a également révélé que la méthode des scores combinés permet de distinguer les experts au sein de l'échantillonnage de personnes examinées, alors que la méthode par consensus ne le permet pas, confirmant ainsi la validité de construit de la méthode des scores combinés.

Nous avons récemment étudié[23] la progression chez les internes (résidents) de deux habiletés essentielles pour les radiologistes, la perception des signes sur les films (mesurée par un test de perception) et l'interprétation de ces signes (mesurée par un TCS). Les procédures d'établissement des scores ont confirmé la variabilité des réponses des radiologues expérimentés, même dans les tâches de perception, où l'on aurait pu s'attendre à un consensus à propos de la présence ou de l'absence de signes sur un film. La méthode des scores compilés a donc été utilisée dans les deux tests. Les deux habiletés progressent pendant la formation en spécialité, l'habileté de perception progressant plus vite que celle d'interprétation. Cette étude apporte des arguments supplémentaires en faveur de la validité de construit du test car elle indique que perception et interprétation sont deux compétences distinctes qui doivent être évaluées par des instruments distincts pendant la formation de spécialité en radiologie.

Les études publiées, ou en voie de publication, révèlent donc un certain nombre de qualités pour le TCS, en terme de validité et de fidélité notamment. En suivant les règles énumérées ci-dessus, la construction d'un TCS s'avère par ailleurs relativement aisée surtout si on la compare avec la construction d'autres types d'examens (QCM par exemple). Le TCS peut être administré aussi bien sous forme papier-crayon que sous forme informatisée. Il est enfin bien accepté par les personnes examinées, quel que soit leur niveau, étudiants, internes ou médecins en exercice. Les uns comme les autres commentent souvent favorablement le test en soulignant sa similarité avec la

démarche clinique réelle. En effet, le candidat qui se soumet à un test de concordance de script doit utiliser son expérience pour moduler son raisonnement (diagnostique ou thérapeutique) en fonction du contexte du patient tel que décrit. Il fait d'abord appel à la mobilisation de connaissances préenregistrées afin de saisir les données spécifiques du problème clinique et décider ensuite de la meilleure conduite à tenir. En pratique, c'est ainsi que, spontanément, le clinicien chevronné s'ajuste devant un patient.

La place du TCS dans une approche intégrée d'évaluation de la compétence clinique

Le jugement, alimenté par un bagage de connaissances factuelles, est au cœur même de la compétence professionnelle dans les domaines de la santé, comme dans les autres professions. Il est basé sur l'existence de liens au sein des connaissances cliniques. Le TCS, en évaluant dans un contexte de simulation de situation clinique l'existence et la fonctionnalité de ces liens permet d'évaluer le jugement clinique. Il permet de tester de grands groupes d'étudiants de façon standardisée (toutes les personnes évaluées sont soumises aux mêmes tâches), avec une méthode de correction objective.

Ceci amène à discuter de la place que peut occuper le TCS dans une stratégie d'évaluation de la compétence clinique. Cette dernière est multidimensionnelle. Aucun instrument d'évaluation ne permet de l'évaluer globalement. L'utilisation de plusieurs instruments est nécessaire si l'on veut avoir un portrait valide de cette compétence[24]. L'examen clinique objectif et structuré (ECOS), permet par exemple de mesurer avec fidélité les habiletés de recueil des données cliniques et les habiletés techniques[25]. Il permet mal cependant de mesurer le raisonnement clinique puisque l'observation ne porte que sur des comportements. La compétence clinique repose également sur les connaissances de règles établies, de lois et de procédures, bien évaluées par des examens de type QCM. Une stratégie d'évaluation crédible pourrait donc reposer sur l'utilisation concomitante de QCM, de TCS et d'ECOS, qui sont tous trois des examens standardisés à correction objective. Ces deux qualités sont d'autant plus essentielles que la compétence est mesurée dans un contexte à enjeu important, tel que celui des examens de certification (attribution du diplôme certifiant la compétence en fin de formation) ou d'épreuves de classement (le concours de l'internat par exemple).

Le TCS est un instrument en cours de développement. Il fait l'objet de plusieurs travaux de recherche subventionnée, par des organismes canadiens ou américains. Étant donné l'importance du nombre et des caractéristiques des participants au panel de référence, un projet de recherche en cours vise à déterminer le nombre de membre nécessaire sur les panels de référence pour obtenir une stabilité des résultats au sein du groupe testé. Un autre cherche à préciser quel est l'effet de la variabilité des caractéristiques professionnelles (médecins hospitaliers, médecins en pratique rurale, médecins en pratique privée, etc) des membres du panel sur cette même stabilité des résultats. Un troisième s'intéresse à la nature des items qui permettent le mieux d'identifier les personnes les plus compétentes dans un groupe de personnes examinées. S'agit-il des items pour lesquels il existe un grand consensus entre membres du panel ou ceux pour lesquels il existe certaines divergences ? Nous faisons l'hypothèse que ce sont ces derniers. Nous avons enfin mis en route un projet d'évaluation de tous les résidents d'une spécialité médicale à travers le Canada afin de comparer les résultats obtenus au TCS avec ceux obtenus dans les autres mesures de compétence de l'examen de certification.

Le TCS s'inscrit résolument dans le paradigme docimologique de l'évaluation, qui favorise la sélection d'examens standardisés comportant un recueil d'informations fondé sur une mesure. Il convient de rappeler qu'il existe en matière d'évaluation des alternatives conceptuelles et notamment celle de l'évaluation authentique[26-27]. Au sein de son propre paradigme le test de concordance de script apparaît cependant comme un instrument qui vient combler une lacune parmi les instruments d'évaluation de la compétence clinique actuellement disponibles. Les résultats de recherche obtenus en font un outil prometteur, mais nous sommes loin d'avoir répondu à toutes les questions qu'il soulève. Le test, développé en médecine est tout à fait transposable aux autres domaines professionnels dans lesquels la compétence implique la prise de décision dans des contextes comportant un certain degré d'incertitude.

Cet instrument d'évaluation a été développé grâce à des fonds de recherche obtenus du Conseil de la Recherche Médicale du Canada, de l'Association des Facultés de Médecine du Canada, du Conseil Médical du Canada et du Collège Royal des Médecins et Chirurgiens du Canada.

Références

1. Newble DI, Hoare J. & Baxter. Patient Mangement Problems. Issues of validity. Med Educ 1982, 16: 137-42
2. Van der Vleuten CPM, Newble D, Case S, Holsgrove G, McCann B, McRae C, Saunders N. Methods of assessment in certification, In Newble D, Jolly B. Wakeford, R (Eds) The certification and recertification of doctors: Issues in the assessment of clinical competence. Cambridge: Cambridge University Press. 1994.
3. Schön D A (1983) The reflective Practitioner: How Professionals Think in Action. New York: Basic Books.
4. Elstein, AS, Shulman, LS, Sprafka, S.A. Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press, 1978.
5. Grant, J. & Marsden, P. Primary knowledge, medical education and consultant expertise. Med Educ 1988 ; .22 : 173-79.
6. Charlin, B., Roy, L., Brailovsky, C.A., Van der Vleuten, C.P.M. The Script Concordance Test: A Tool to Assess the Reflective Clinician. Teach Learn Med Educ 2000: 12: 189-95.
7. Feltovich PJ, Barrows HS. Issues of generality in medical problem solving. In : Schmidt HG, De Volder ML (eds). Tutorials in problem-based learning: A new direction in teaching the health professions. Assen (Holland) :Van Gorcum, 1984 : 128-142.
8. Schmidt, H.G., Norman, G.R. and Boshuizen, H.P.A. A Cognitive Perspective on Medical Expertise: Theory and Implications. Acad Med 1990; 65: 611-621.
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 : 182-90.
10. Bordage, G. Elaborated Knowledge: A key to Successful Diagnostic Thinking. Academic Medicine, 1994 ; 69 : 883-85.
11. Smith, E.E. Concepts and Induction. In Posner M.I.(Ed.). Foundations of Cognitive Science. Cambridge, MA: MIT Press, 1989.
12. Norman GR Objective Measurement of Clinical Performance. Medical Education, 1985 : 19 : 43-47.
13. Norcini JJ, Shea JA, Day SC The Use of the Aggregate Scoring for a Recertification Examination. Evaluation and the Health Professions, 1990 : 13 : 241-51.
14. Norman G, Swanson DB, and Case SM. Conceptual and methodological issues in studies comparing assessment formats. Teaching and Learning in Medicine, 1996 : 8 : 208-16.
15. Colin R. Médecine basée sur les preuves et éducation médicale. Pédagogie Médicale, 2001, 2 : 69-70.
16. Woolf SH. Practice guidelines, a new reality in medicine: II. Methods of developing guidelines. Arch Intern Med 1992,152:946-952.
17. Guy Llorca. Communication personnelle, mai 2002.
18. 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.
19. Charlin B, Brailovsky CA., Leduc C, Blouin D. The Diagnostic Script Questionnaire: A New Tool to Assess a Specific Dimension of Clinical Competence. Advances in Heath Sciences Education, 1998; 3: 51-58.
20. Brailovsky C, Charlin B, Beausoleil S, Coté S, van der Vleuten, C. Measurement of Clinical Reflective Capacity Early in Training as a Predictor of Clinical Reasoning Performance at the End of Residency: An Exploratory Study on the Script Concordance Test. Med Educ 2001; 35: 430-36.
21. Sibert L, Charlin B, Corcos J, Gagnon R, Lechevallier J, Grise P. Assessment of clinical reasoning competence in urology with the Script Concordance test: an exploratory study across two sites form different countries. European Urology 2001, sous presse.
22. Charlin B, Desaulniers M, Gagnon R, Blouin D, van der Vleuten C. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. Teaching and Learning in Medicine, In press, July 2002.
23. Brazeau-Lamontagne L, Charlin B, Gagnon R, Samson L, Van der Vleuten C. Measurement of competence in radiology: Evolution of knowledge elaboration and perception skill scores along training. Radiology 2002, submitted.
24. Van Der Vleuten CPM. The assessment of professional competence: development, research and practical implications. Adv Health Sci Educ 1996 ; 1 : 41-67.
25. Sibert L, Grand'Maison P, Charlin B, Grise P. Développement d'un Examen Clinique Objectif Structuré pour évaluer les compétences des internes en urologie. Pédagogie médicale 2000, 1: 33-39.
26. Wiggins G. Curricular coherence and assessment: making sure that the effects match the intent. In: Beane JA (Ed) Toward a coherent curriculum. The ASCD yearbook. Alexandria (VA): Association for Curriculum Development, 1995 : 101-19.
27. Jouquan J. L'évaluation des apprentissages des étudiants en formation médicale initiale. Pédagogie médicale 2002, 3 :38-52.