Pédagogie Médicale
REVUE INTERNATIONALE FRANCOPHONE D'ÉDUCATION MÉDICALE

Test de concordance de script : un nouveau mode d'établissement des scores limitant l'effet du hasard

Script Concordance Test: a New Scoring Method to Limit Guessing Effects

Sophie VANBELLE[1], Valérie MASSART[2], Didier GIET[2], Adelin ALBERT[1]

Résumé ***Contexte :*** *La note obtenue au test de concordance de script (TCS) est actuellement calculée à partir de la méthode des scores combinés.* ***But :*** *Proposer une nouvelle mesure d'accord entre les réponses données au TCS par le candidat évalué et celles du panel d'experts.* ***Résultats :*** *Le système de notation actuel ne tient pas compte du fait qu'un certain nombre d'accords entre le candidat et le panel d'experts peut être fortuit. La mesure d'accord proposée permet de corriger ce fait.* ***Conclusion :*** *Cet article soulève le problème de scores obtenus fortuitement. La mesure d'accord développée permet d'améliorer le système de notation actuel.*

Mots clés Test de concordance de script ; degré d'accord ; système de notation ; effet du hasard ; kappa de Cohen.

Abstract ***Context:*** *The method of combined scores is currently used to determine the score obtained by the Script Concordance Test (SCT).* ***Goal:*** *To propose a novel measure of agreement between the SCT responses given by the candidate and those provided by the panel of experts.* ***Results:*** *The current scoring method of the SCT does not take into account the fact that agreement between the candidate and the panel of experts may be arbitrary. The proposed measure allows correcting this fact.* ***Conclusion:*** *This article addresses the problem of scores obtained by guessing. The new measure of agreement allows to improve the actual scoring method.*

Key words Script Concordance Test; agreement; scoring; guessing effect; Cohen kappa coefficient.

Pédagogie Médicale 2007;8:71-81

1- Informatique médicale et Biostatistique - Université de Liège – Liège - Belgique
2- Département de Médecine Générale - Université de Liège –Liège - Belgique

Correspondance : Sophie Vanbelle - Informatique médicale et Biostatistique - Université de Liège - CHU (B 23) - Sart Tilman. B-4000 Liège. Téléphone : 00 32 436 625 90 - Mailto:sophie.vanbelle@ulg.ac.be

Article disponible sur le site http://www.pedagogie-medicale.org ou http://dx.doi.org/10.1051/pmed:2007002

Introduction

Le test de concordance de script (TCS) est utilisé en éducation médicale afin d'évaluer la capacité d'étudiants en médecine ou de médecins en formation à résoudre des problèmes cliniques en situation d'incertitude[1]. Le test comprend une série d'items à évaluer sur une échelle de Likert à 5 points. La note des candidats est calculée à l'aide de la méthode des scores combinés par comparaison aux réponses données par un panel d'experts répondant aux mêmes questions dans les mêmes conditions. La note actuelle, reflétant la similitude entre les réponses du candidat et celles du panel d'experts, ne tient pas compte du fait qu'un certain nombre d'accords puissent être fortuits. Ainsi, des candidats donnant systématiquement la même réponse à chaque question, répondant de façon aléatoire ou interprétant de manière inadéquate l'échelle de Likert pourraient obtenir un score honorable. Une manière d'identifier ces candidats et de remédier au problème des accords dus au hasard est proposée dans cet article.

Méthodes

Le test de concordance de script

Le test de concordance de script est un outil d'évaluation du raisonnement clinique récemment décrit[1] et utilisé dans plusieurs domaines de spécialités médicales (chirurgie, gynécologie, radiologie, etc.). L'approche de ce test consiste à présenter au candidat une série de problèmes cliniques mal définis et à leur demander de poser un « micro-jugement » en matière diagnostique, d'investigation ou de traitement, lorsque des éléments d'information supplémentaires lui sont communiqués.

Chaque item est constitué d'une situation clinique susceptible d'être rencontrée par le candidat dans l'exercice de sa profession. Plusieurs options de diagnostic ou de prise en charge sont possibles et la description de la situation ne contient pas toutes les données permettant de 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, présentée comme hypothèse au candidat. Pour cela, le candidat doit choisir entre les différentes propositions d'une échelle de Likert à 5 points allant de « l'hypothèse est pratiquement éliminée » à « il ne peut s'agir pratiquement que de cette hypothèse ». Un exemple d'item est donné dans le *tableau 1*.

Le candidat est évalué sur base d'un score qui mesure la similitude entre ses réponses et celles d'un panel de référence constitué de médecins expérimentés dans le domaine envisagé. Le processus d'établissement des scores s'appuie sur les deux principes suivants[1] : a) la réponse de chaque membre du panel de référence reflète une opinion valide qui devrait être prise en compte et b) les réponses pour lesquelles il n'existe pas de consensus ne devraient pas être rejetées. La note finale du candidat est alors déterminée par la méthode des scores combinés en confrontant ses réponses à celles des experts.

Etablissement des scores

Pour chaque item, les réponses données par les experts du panel sont prises en compte pour établir le score du candidat. A cet effet, on définit un « crédit » pour chaque catégorie de l'échelle de Likert ; celui-ci représente le score qu'un candidat peut obtenir s'il choisit la catégorie

Tableau 1 :
Exemple d'item proposé dans un test de concordance de script

Raymond, 50 ans, se plaint de troubles érectiles depuis 6 mois. Il évoque la question pour la première fois avec vous et cherche à s'informer sur les origines de ces symptômes.

Si votre hypothèse est	Et que vous apprenez que	L'effet sur l'hypothèse diagnostique est (a)
« une origine vasculaire »	« il a souffert d'une prostatite aiguë il y a 4 mois »	-2 -1 0 +1 +2

(a) : (-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.

correspondante. Dans la méthode actuelle (M1), un crédit de 1 est accordé à la (les) catégorie(s) de l'échelle de Likert la (les) plus choisie(s) par les experts, les autres catégories recevant un crédit inférieur en fonction du nombre d'experts ayant choisi celles-ci. La note finale du candidat est alors la moyenne des scores qu'il a obtenus pour les différents items, reflétant ainsi la similitude entre les réponses du panel et celles du candidat.

Une alternative à ce système d'établissement des scores consiste à allouer à chaque catégorie de l'échelle de Likert un crédit qui correspond à la proportion d'experts du panel choisissant cette catégorie. La note obtenue par le candidat sur l'ensemble des items pourrait alors être interprétée comme un degré de concordance entre ses réponses et celles du panel d'experts, appelé « degré d'accord observé » (M2).

Le score du candidat calculé en utilisant ces deux premières définitions des crédits ne prend pas en compte le fait que l'accord obtenu entre les réponses du candidat et celles du panel d'experts puisse être fortuit. En effet, un candidat donnant systématiquement la même réponse à chaque item, répondant au hasard ou interprétant de manière inadéquate l'échelle de Likert, obtient un score de manière fortuite. De plus, chaque candidat a, *a priori*, tendance à choisir plus souvent certaines catégories de l'échelle que d'autres. Par exemple, certains pourraient avoir tendance à moins choisir les catégories extrêmes (-2 et +2) et concentrer leurs réponses sur la catégorie centrale de l'échelle (0) (« biais de la tendance centrale »), tandis que d'autres pourraient éviter les catégories intermédiaires (-1 et +1) et grouper leurs réponses sur les catégories extrêmes (« biais des extrêmes »). Une estimation de la tendance du candidat à choisir les différentes catégories peut être obtenue en déterminant la fréquence à laquelle celui-ci choisit les différentes catégories de l'échelle sur l'ensemble du TCS. Il en est de même pour le panel d'experts. Les distributions ainsi obtenues pour le candidat et le panel d'experts permettent de déterminer le degré d'accord dû au hasard. Ce degré d'accord reflète la proportion d'accords obtenue entre le candidat et le panel d'experts lorsqu'on connaît la fréquence à laquelle ceux-ci ont choisi chaque catégorie de l'échelle de Likert indépendamment des items.

Afin de corriger le degré d'accord observé pour l'effet du hasard, on retranche simplement le degré d'accord lié au hasard du degré d'accord observé. Le résultat est alors standardisé (M3) afin d'obtenir la valeur 1 lorsque l'accord est parfait (le candidat choisit systématiquement la (les) proposition(s) de l'échelle la (les) plus choisie(s) par le panel d'experts à chaque item). Le degré d'accord corrigé est négatif si le degré d'accord observé est inférieur au degré d'accord dû au hasard. Il est nul si le degré d'accord observé est égal au degré d'accord dû au hasard. Ces propriétés sont identiques à celles du coefficient kappa de Cohen[2] utilisé dans de nombreux domaines pour quantifier le degré d'accord entre deux observateurs sur un critère qualitatif. Landis et Koch[3] ont proposé une classification pour apprécier ce degré d'accord *(Tableau 2)*. Cette classification pourrait aussi être utilisée pour apprécier le degré d'accord entre les réponses au TCS du candidat (un observateur) et celles du panel d'experts (un groupe d'observateurs). Les formules mathématiques permettant de déterminer les notes des candidats par les différentes méthodes sont explicitées en *annexe*.

Tableau 2 :
Classification du degré d'accord corrigé entre deux observateurs (cœfficient kappa de Cohen) selon Landis et Koch[3]

Kappa de Cohen	Qualification de l'accord
> 0,81	Très bon
0,61-0,80	Bon
0,41-0,60	Modéré
0,21-0,40	Médiocre
0-0,20	Mauvais
<0	Très mauvais

L'étude du degré d'accord corrigé entre les candidats et le panel d'experts pourrait mettre en évidence : a) les candidats faibles (degré d'accord corrigé < 0). Ces candidats ont en effet un degré d'accord observé inférieur à celui qu'ils obtiendraient fortuitement ; b) les candidats susceptibles d'avoir répondu au hasard ou de présenter des problèmes de compréhension du test (degré d'accord corrigé = 0). Ces candidats ont un degré d'accord observé similaire à celui qu'ils obtiendraient fortuitement. De plus, elle permet d'identifier les candidats qui répondent au hasard ou qui donnent systématiquement la même réponse car ces candidats obtiendraient tous un degré d'accord corrigé nul (M3), tandis qu'ils pourraient obtenir un score honorable selon la méthode actuelle (M1).

Résultats

Les étudiants en médecine (2003-2005) de l'Université de Liège (Liège, Belgique) souhaitant se spécialiser en médecine générale ont eu la possibilité de se soumettre à un TCS comportant 34 items et portant sur des situations rencontrées dans ce domaine. Un panel de 11 experts auxquels le TCS a été soumis a permis de déterminer la distribution des crédits dans les catégories de l'échelle de Likert pour chaque item. Parmi les 54 étudiants ayant répondu au TCS, seuls ceux ayant répondu à tous les items (n = 39) ont été inclus dans la présente étude.

Un résumé des résultats obtenus par les 39 étudiants ainsi que le cœfficient alpha de Cronbach[4] sont donnés au *tableau 3* pour chacune des 3 méthodes exposées. Le taux d'échec au TCS a été déterminé en fixant le seuil de réussite à 0,5 pour le score obtenu par la méthode actuelle (M1) et à 0,4 pour les degrés d'accord (M2 et M3). Le seuil 0,5 correspond au critère de réussite classique et 0,4 au seuil d'un accord médiocre selon la classification de Landis et Koch[3].

Lorsqu'on considère les scores standardisés, le coefficient de corrélation intraclasse vaut 0,97 entre la méthode actuelle et le degré d'accord observé ; 0,95 entre la méthode actuelle et le degré d'accord corrigé et 0,98 entre le degré d'accord observé et le degré d'accord corrigé. Les notes obtenues selon la méthode actuelle (M1) et le degré

Tableau 3 :
Résultats obtenus au Test de concordance de script de médecine générale par les 39 étudiants à l'aide des 3 différentes méthodes.

Le tableau reprend, pour chaque méthode, le coefficient alpha de Cronbach, le nombre et le pourcentage d'étudiants en échec, la moyenne ± l'écart-type des notes obtenues par les étudiants.

Méthode	Description	Alpha de Cronbach	Echec[(a)] n (%)	Moyenne ± écart-type
M1	Méthode actuelle	0,14	0 (0,0)	0,74 ± 0,06
M2	Degré d'accord observé	0,14	8 (21)	0,43 ± 0,10
M3	Degré d'accord corrigé	-0,23	2 (5,1)	0,57 ± 0,11

(a) seuil fixé à 0,5 pour M1 et à 0,4 pour M2 et M3

d'accord corrigé pour l'effet du hasard (M3) sont représentées pour chacun des 39 étudiants sur *la figure 1*. Le degré d'accord observé (M2) et celui dû au hasard y sont aussi repris. Les résultats obtenus à l'aide de l'accord corrigé (M3) sont systématiquement plus bas que ceux obtenus par la méthode actuelle (M1) car la méthode de l'accord corrigé pénalise pour l'effet du hasard. On constate sur cette figure que deux étudiants sont en échec (n^{os} 20 et 26) lorsqu'on utilise le degré d'accord corrigé (M3), alors qu'ils ont réussi selon la méthode actuelle (M1). Ces deux étudiants sont classés derniers par la méthode actuelle.

Les différences obtenues dans le classement des candidats lorsque ceux-ci sont classés du plus fort au plus faible en fonction du degré d'accord corrigé ou de la note qu'ils ont obtenue par la méthode actuelle sont illustrées sur la *figure 2.* On constate que l'écart entre les rangs obtenus par la méthode actuelle et le degré d'accord corrigé varie de -7 à +9. Si on applique un principe de « concours classant » visant à ne retenir que les « x premiers candidats » (x = 1, 2,…, 38 ou 39), on s'aperçoit que les candidats retenus par les deux méthodes diffèrent dans 29 (74 %) cas sur 39.

Afin d'illustrer le fait que le degré d'accord corrigé permet de mettre en évidence les candidats choisissant systématiquement la même catégorie de l'échelle ou répondant au hasard, considérons les deux cas suivants : a) un étudiant donnant systématiquement la même réponse aux items du TCS de médecine générale obtiendrait un degré d'accord corrigé nul (M3). Par contre, il obtiendrait une note (M1) de 0,28 ; 0,40 ; 0,64 ; 0,43 ou 0,14 s'il répond systématiquement (-2), (-1), (0), (1) ou (2), respectivement ; b) un étudiant répondant au TCS de médecine générale au hasard (c'est-à-dire qu'il a 1 chance sur 5 de choisir chaque catégorie de l'échelle de Likert) obtiendrait une note (M1) de 0,28/5 + 0,40/5 + 0,64/5 + 0,43/5 + 0,14/5 = 0,38 ; tandis que son degré d'accord corrigé avec le panel d'experts (M3) serait nul.

Discussion

La méthode développée dans cet article propose de corriger le score actuel du TCS pour l'effet du hasard. Les résultats obtenus par la méthode actuelle et le degré d'accord corrigé sont hautement corrélés (r = 0,96) car le degré d'accord corrigé contient le degré d'accord observé, qui est une variante du système de notation actuel. Néanmoins, l'utilisation de la correction pour l'effet du hasard permet une vision plus objective des notes actuelles qui sont surestimées car elles comprennent une part pouvant être attribuée au hasard. Il existe aussi une corrélation hautement significative entre les classements des étudiants, bien que les variations entre ceux-ci ne soient pas négligeables (de -7 à +10 places) et impliquent une sélection différente des candidats dans le cadre d'un concours classant.

La question de la correction pour l'effet du hasard a déjà été posée pour d'autres systèmes de notation. Chevalier[5] a conclu, dans le cadre des questionnaires à choix multiples (QCM), que la correction n'améliorait pas le système de notation actuel de manière significative. Notons que la correction apportée ici est différente de celle exposée par Chevalier[5]. De plus, une des différences majeures entre les QCM et le TCS réside dans le fait que, contrairement aux QCM, il n'existe pas une réponse unique et correcte à chaque question. Le candidat est en effet évalué en confrontant ses réponses à celles d'un panel d'experts. Il est donc important de tenir compte de la variabilité des réponses données par le panel d'experts, comme le fait déjà la méthode actuelle, mais aussi des biais qui peuvent exister lorsque l'on utilise une échelle de Likert (par exemple, « biais de la valeur centrale »). En effet, si les étudiants et le panel d'experts répondaient au hasard au sens strict du terme, le degré d'accord dû au hasard entre un candidat et le panel d'experts serait de 0,20 (= 5x0.20x0.20) dans le cadre d'une échelle à cinq points, et ce pour tous les candidats. La correction pour l'effet du hasard serait alors inutile. Cependant, comme illustré dans la *figure 1*, le degré d'accord dû au hasard n'est pas le même pour tous les candidats car le panel d'experts et les candidats ont, *a priori*, tendance à choisir plus souvent certaines catégories de l'échelle plutôt que d'autres, ce qui soutient l'utilité de la correction.

Par contre, un désavantage de l'utilisation du degré d'accord corrigé réside dans l'interprétation de celui-ci. En effet, les scores obtenus ne peuvent plus être interprétés de manière classique car ils peuvent être négatifs. Les critères absolus de réussite doivent donc être adaptés en conséquence. La valeur 0,4, définie comme seuil pour un accord « médiocre » par Landis et Koch[3], a été choisie arbitrairement dans ce travail. Il faut encore noter que le TCS de médecine générale, sur lequel sont basés les résultats, ne montre pas une validité interne d'un niveau suffisant quelle que soit la méthode utilisée. Une révision du TCS pourrait permettre d'augmenter la validité interne du questionnaire et une meilleure distribution des crédits dans le sens où un étudiant répondant de manière systématique n'obtiendrait plus un score honorable.

Figure 1 :
Notes obtenues par les 39 étudiants au Test de concordance de script de médecine générale à l'aide de la méthode actuelle (M1) et du degré d'accord corrigé pour l'effet du hasard (M3).

Le degré d'accord observé (M2) et dû au hasard sont aussi représentés. La ligne horizontale continue représente la limite du critère de réussite pour la méthode actuelle (0,5) et la pointillée pour le degré d'accord corrigé (0,4)

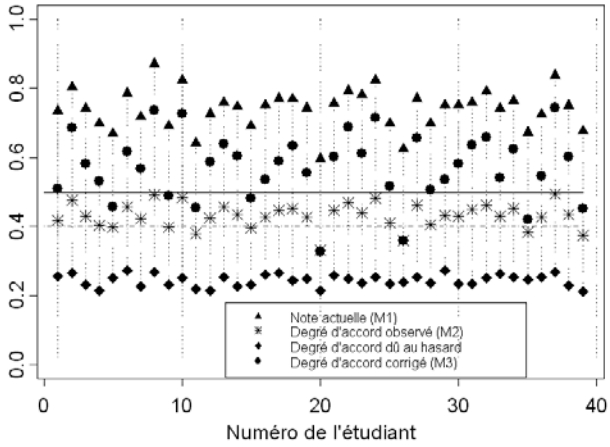
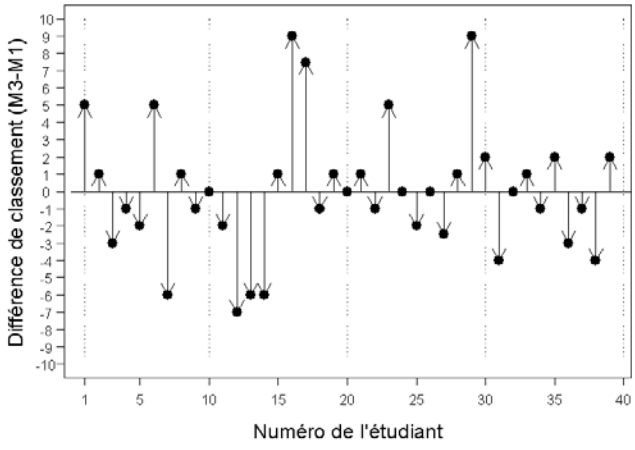


Figure 2 :
Différence de position de classement lorsqu'on classe les 39 étudiants ayant répondu au Test de concordance de script de médecine générale du plus fort au plus faible en fonction du degré d'accord corrigé (M3) et de la note obtenue par la méthode actuelle (M1)



Les auteurs jugent qu'il est important d'appliquer la méthode définie dans cet article à d'autres TCS, à plus grande validité interne, afin de conforter les résultats obtenus.

L'étape suivante dans la construction d'un système d'appréciation des candidats serait de tenir compte du caractère ordinal de l'échelle de Likert. Bland *et al.*[6] ont proposé de prendre une mesure de distance entre la réponse de l'étudiant et la valeur modale ou la valeur moyenne du panel d'experts pour établir le score du candidat. Cette méthode est inadéquate si l'on souhaite conserver le fait qu'il n'existe pas une réponse unique et correcte à chaque question car les réponses des experts sont résumées dans une quantité unique. On ne tient donc plus compte de la variabilité des réponses des experts. Une solution plus correcte consisterait à étendre le coefficient de kappa pondéré[7], qui mesure le degré d'accord entre deux observateurs sur une échelle ordinale, au cas d'un observateur (le candidat) et d'un groupe d'observateurs (le panel d'experts). Communément, dans le cadre du coefficient d'accord pondéré, on utilise les poids introduits par Chiccetti et Allison[8] ou Fleiss et Cohen[9]. Cependant, l'introduction de ces poids soulève un autre problème car, même si ceux-ci sont utilisés de manière usuelle, leur définition est arbitraire. Ainsi, un système de pondération pourrait être choisi par un groupe de travail sur le test de concordance de script. L'application de modèles de Rasch[10] *(Item Response Theory)* pourrait aussi être envisagée.

Conclusion

Le Test de concordance de script permet d'évaluer le raisonnement clinique de médecins en spécialisation ou d'étudiants en médecine soumis à des problèmes cliniques mal définis. Actuellement, le système de cotation ne permet pas de prendre en compte le fait qu'un certain nombre d'accords peut être fortuit. Le présent travail ouvre des pistes de réflexion au sujet d'une modification du système actuel de notation du TCS. L'avantage en termes d'objectivité de cette méthode doit être confirmé par son application à d'autres tests à forte validité interne.

Contributions

Didier Giet et Valérie Massart se sont attachés aux aspects plus pédagogiques de la nouvelle méthode et le test de concordance de script a été développé dans leur service tandis que Sophie Vanbelle et Adelin Albert se sont focalisés sur les aspects statistiques et mathématiques et ont développé la méthode d'accord corrigé.

Annexe
Principes des méthodes exposées

Exemple : Considérons un exemple hypothétique de test de concordance de script (TCS) comportant 3 items et supposons qu'un panel de 12 experts soit désigné pour déterminer les crédits alloués à chaque catégorie de l'échelle pour chaque item. Les réponses données par les experts sont résumées dans le *tableau A*. Les réponses de 2 candidats (notés I et II) aux 3 items du TCS sont reproduites dans le *tableau B*.

Tableau A :
Réponses données à un TCS hypothétique comportant 3 items par 12 experts

	Catégorie de l'échelle de Likert				
Item du TCS	(-2)	(-1)	(0)	(1)	(2)
1	0	0	1	7	4
2	0	6	5	1	0
3	3	4	0	5	0

Tableau B :
Réponses hypothétiques de 2 candidats au TCS comportant 3 items

	Item du TCS		
Candidat	1	2	3
I	(1)	(0)	(-2)
II	(2)	(1)	(-1)

Méthode actuelle (M1)
Le crédit accordé à chaque catégorie de l'échelle correspond au nombre d'experts du panel ayant choisi cette catégorie divisé par le nombre maximum d'experts ayant choisi la même catégorie. Le score du candidat à un item correspond au crédit de la proposition qu'il a choisie. Sa note est la moyenne des scores qu'il a obtenus pour l'ensemble du TCS. Par exemple, pour l'item 1, le nombre maximum d'experts ayant donné la même réponse est 7. Ainsi, la catégorie (1) de l'échelle de Likert reçoit un crédit de 1, la catégorie (2) reçoit un crédit de 4/7 = 0,57, la catégorie (0) reçoit 1/7 = 0,14 et les autres catégories un crédit nul. En procédant de même avec les items 2 et 3 du TCS, on obtient le *tableau C*.

Tableau C :
Crédits accordés à chacune des catégories de l'échelle pour chaque item par la méthode actuelle (M1)

	Catégorie de l'échelle de Likert				
Item du TCS	(-2)	(-1)	(0)	(1)	(2)
1	0	0	0,14	1	0,57
2	0	1	0,83	0,17	0
3	0,26	0,8	0	1	0

Sur base des tableaux B et C, le candidat I obtient donc (1+0,83+0,6)/3=0,81 et le candidat II obtient (0,57+0,17+0,8)/3=0,51. Rappelons que le score maximal que l'on peut obtenir vaut 3/3 = 1 !

Méthode d'accord observé (M2)
Le crédit accordé à chaque catégorie de l'échelle correspond à la proportion d'experts qui ont choisi cette proposition. Le score du candidat à un item ainsi que sa note sont déterminés de la même manière que pour la méthode actuelle (M1). Les crédits sont repris au *tableau D*.

Tableau D :
Crédits accordés à chacune des catégories de l'échelle pour chaque item par la méthode du degré d'accord observé (M2)

	Catégorie de l'échelle de Likert				
Item du TCS	(-2)	(-1)	(0)	(1)	(2)
1	0	0	0,08	0,58	0,33
2	0	0,50	0,42	0,08	0
3	0,25	0,33	0	0,42	0

Le candidat I obtient (0,58+0,42+0,25)/3=0,42 et le candidat II (0,33+0,08+0,33)/3=0,25. La valeur maximale que peut prendre l'accord observé, vu la distribution des réponses des membres du panel d'experts, est égale à (0,58+0,50+0,42)/3=0,50

Annexe (suite)

Accord dû au hasard

On détermine le nombre de fois que le candidat choisit chaque catégorie de l'échelle de Likert pour l'ensemble du TCS et on divise ce nombre par le nombre d'items composant le TCS. On fait de même pour chaque membre du panel d'experts. On calcule ensuite la moyenne pour le panel d'experts. Ensuite, pour chaque catégorie de l'échelle, on multiplie la quantité obtenue par le candidat par celle du panel d'experts. Lorsqu'on somme ces dernières quantités, on obtient le degré d'accord dû au hasard. Les distributions obtenues pour les candidats I, II et le panel d'experts sont exposées au *tableau E.*

Tableau E :
Fréquence à laquelle les candidats et le panel d'experts choisissent les différentes catégories de l'échelle de Likert

	Catégorie de l'échelle de Likert				
	(-2)	(-1)	(0)	(1)	(2)
Candidat I	1/3=0,33	0	1/3=0,33	1/3 = 0,33	0
Candidat II	0	1/3=0,33	0	1/3 = 0,33	1/3=0,33
Panel	0,25/3=0,08	(0,5+0,33)/3 = 0,28	(0,08+0,42)/3 = 0,17	(0.58+0,08+0,42)/3 = 0,36	0,33/3=0,11

A partir de ces distributions, on obtient un degré d'accord dû au hasard égal à
0,33x0,08+0x0,28+0,33x0.17+0,33x0,36+0x0,11 = 0,20 pour le candidat I
et à 0x0,08+0x0,28+0x0,17+0,33x0,36+0,33x0,11 = 0,25 pour le candidat II.

Méthode d'accord corrigé (M3)

Premièrement, on retranche le degré d'accord dû au hasard du degré d'accord observé. On divise alors cette quantité par la différence entre le degré d'accord observé maximal et le degré d'accord dû au hasard. On obtient ainsi le degré d'accord corrigé (M3).

Tableau F :
Notes obtenues par les candidats I et II pour les différentes méthodes

	M1	M2	Accord dû au hasard	M3
Candidat I	0,81	0,42	0,20	0,73
Candidat II	0,51	0,25	0,25	0,00

On obtient (0,42-0,20)/(0,50-0,20)=0,73 pour le candidat I et (0,25-0,25)/(0,50-0,25) = 0 pour le candidat II. Les différentes valeurs obtenues par les candidats sont résumées dans le *tableau F.*

Références

1. Charlin B, Gagnon R, Sibert L, Van der Vleuten C. Le test de concordance de script : un instrument d'évaluation du raisonnement clinique. Pédagogie Médicale 2002;3:135-44.

2. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37-46.

3. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

4. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-334.

5. Chevalier SA. A review of scoring algorithms for Ability and Aptitude Tests. paper presented at the Annual Meeting of the Southwestern Psychological Association, New Orleans (LA), 1998.

6. Bland AC, Kreiter CD, Gordon JA. The psychometric properties of five scoring methods applied to the Script Concordance Test. Acad Med 2005;80:395-9.

7. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70:213-20.

8. Chiccheti DV, Allison T. A new procedure for assessing reliability of scoring EEG sleep recordings. Amer J EEG Technol 1971;11:101-9.

9. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measure of reliability. Educational and Psychological Measurement 1973;33:613-9.

10. Van der Linden WJ & Hambleton RK (Eds.) Handbook of modern item response theory. New York: Springer, 1997.

Manuscrit reçu le 11 octobre 2006 ; commentaires éditoriaux formulés aux auteurs le 12 février 2007 ; accepté pour publication le 30 mars 2007.