Le Test de Concordance de Script comme outil d'évaluation du raisonnement clinique en Néonatologie

The script concordance test as a tool to evaluate clinical reasoning in Neonatology

Emira Ben Hamida[1], Imen Ayadi[1], Zahra Marrakchi[1], André Quinton[2]

1-Service de Néonatologie. Hôpital Charles Nicolle. Faculté de Médecine de Tunis. Université Tunis El Manar.
2-Centre de Recherches Appliquées en Méthodes éducatives (C.R.A.M.E). Université de Bordeaux.

RÉSUMÉ

Introduction : Le test de concordance de script vise à comparer l'organisation des connaissances, élément essentiel de la compétence clinique.
Objectifs : Construire un test de concordance de script et de démontrer sa pertinence dans l'évaluation des compétences en Néonatologie.
Méthodes : Un test de concordance de script comportant 20 vignettes a été administré à 52 étudiants de deuxième cycle d'études médicales et à 11 internes de médecine générale.
Résultats : Les experts ont eu un score moyen, rapporté sur 100, de 82.52 ± 7.35 IC95% [77.26-87.78], plus élevé que les externes de 58.52 ± 9.72 IC95% [55.82-61.23], et les internes de 63.17±11.36 IC95% [55.53-70.81] (p<0.0001).
Conclusion : Notre travail a démontré la faisabilité du TCS en stage de néonatologie. Il a permis de différencier les experts des internes et des étudiants.

Mots-clés
Test de concordance de script, Evaluation, Compétence, Raisonnement, Néonatologie.

SUMMARY

Background : Script concordance test aims to evaluate knowledge organization, which represents an essential component of the clinical competence.
Objective : To build a script concordance test and demonstrate its relevance in the evaluation of Neonatology skills.
Methods: A script concordance test including 20 vignettes and 20 items, was provided to 52 fourth year medical students and 11 family medicine interns.
Results : Script concordance test scores obtained by experts were higher then those obtained by students and family medicine interns. The scores (out of 100) were 82.52 ± 7.35 CI95% [77.26-87.78] for the experts, 58.52 ± 9.72 CI95% [55.82-61.23] for the students, and 63.17±11.36 IC95% [55.53-70.81] (p<0.0001) for the interns.
Conclusion: Our data suggest that script concordance tests could be used to assess the acquisition of clinical reasoning among fourth year medical students in neonatolgy.

Key-words
Script concordance test, Evaluation, Skills, Clinical reasoning, Neonatology.

Dans leur pratique de tous les jours, les professionnels de la santé rencontrent des problèmes plus au moins complexes qu'ils résolvent en faisant appel à leurs connaissances et à leur raisonnement clinique. Lorsque les données sont complètes et précises, le but à atteindre est clair et une solution est disponible. Mais il peut s'agir aussi de données incomplètes ou imprécises où le but à atteindre n'est pas toujours clair. Ce type de problème exige une capacité à raisonner en contexte d'incertitude [1, 2]. Le test de concordance de script (TCS) est un instrument d'évaluation du raisonnement en contexte d'incertitude.

La validité du TCS est aujourd'hui bien établie [3]. Une progression des scores selon qu'il est proposé respectivement à des étudiants en médecine, à des internes ou à des praticiens expérimentés, ses modalités techniques de construction et d'administration sont également bien connues [4-7]. Ce test n'a jamais fait partie des méthodes d'évaluation dans notre faculté de médecine et il n'a pas été testé en terrain de Néonatologie.

L'objectif de ce travail était de construire un TCS et de démontrer son intérêt dans l'évaluation des compétences en néonatologie des étudiants du deuxième cycle et des internes de notre faculté de médecine.

MÉTHODES

Population ayant effectué des TCS

Il s'agissait d'étudiants de quatrième année médecine (externes) et des internes de médecine générale. La durée de stage en néonatologie pour les externes était de deux à trois semaines en fonction des groupes et de leur répartition et celle des internes de deux mois. Chaque groupe d'externes était composé de 10 à 11 étudiants. Les groupes d'internes étaient composés de 5 à 6 étudiants. Les stages se sont déroulés au service de néonatologie, un service hospitalo-universitaire, de niveau III de soins néonataux offrant à côté des soins de réanimation néonatale, des soins de pédiatrie de maternité.

Rédaction du TCS

Les questions ont été élaborées en prenant en considération la liste des objectifs de stage de quatrième année des études médicales. Les objectifs de stages des internes ne comprennent pas une liste spécifique d'objectifs pour la néonatologie et rejoignent ceux des externes. Les objectifs de stage des externes concernent fondamentalement le recueil de l'information clinique par l'interrogatoire et l'examen physique, le choix et l'interprétation des examens complémentaires, l'initiation au diagnostic et au traitement des cas simples de pathologie courante et de situations d'urgence ainsi que les attitudes d'autoformation, de relation avec le patient et avec l'équipe sanitaire.

L'élaboration des TCS a été effectuée selon les recommandations de l'université de Montréal, disponibles sur le site *http://www.cpass.umontreal.ca/index.html.* Une liste de sujets a été définie, avec pour chacun le ou les domaines de connaissances à évaluer (diagnostique, d'investigation, thérapeutique et pronostique), conformément aux objectifs de stage.

Les vignettes composées chacune d'un seul item ont été rédigées par une seule personne, sénior en néonatologie, maitre de conférences agrégé, et ont été soumises à une relecture par un autre sénior.

Chaque vignette comprenait de façon explicite l'échelle de Likert, avec un score (ou ancrage) de -2,-1, 0, 1, et 2. Si l'information donnée rendait l'hypothèse, l'examen initialement prévu, la prescription ou le geste envisagé(e), plus probable ou plus justifié (e), selon le domaine de connaissance à évaluer, le score était de 1 ou 2 ; si elle était sans effet, il était de 0 ; si elle le (la)rendait moins probable ou moins justifiée, le score était de -1 ou -2. Figure 1

Validation des items et choix des experts

Une grille de contrôle de la qualité a priori a été consultée pour la validation du TCS. Cette grille est disponible sur le site de l'université de Montréal.

Le test a été soumis, de façon anonyme, à 10 médecins séniors en néonatologie qui ont été informés de son but

TCS Vignette Clinique N°3

Rayene, nouveau-né à terme eutrophique, est admis en unité de soins intensifs devant des convulsions à type de clonies du membre supérieur droit, notées à H10 de vie. Il est né à terme, l'accouchement était laborieux avec l'application de forceps pour défaut d'expulsion. Le score d'Apgar était à 2-7-8 à 1-3 et 5 mn de vie.

Si vous pensez faire :	Et que l'on vous fournit :	L 'examen initialement prévu est :
Une tomodensitométrie cérébrale	Une échographie transfontanellaire normale	-2 -1 0 1 2

-2 L'examen initialement prévu est absolument contre-indiqué
-1 L'examen initialement prévu est moins justifié
0 L'examen initialement prévu n'est pas influencé par la donnée supplémentaire
1 L'examen initialement prévu est plus justifié
2 L'examen initialement prévu est indispensable

Figure 1: Exemple de vignette de test de concordance de script à un item, constituée d'un scénario, d'une prescription d'une investigation et d'une information complémentaire. L'effet de la nouvelle information est évalué par l'échelle de Likert, par une note allant de -2 à +2.

et de son cadre. Ces dix séniors exerçaient dans quatre services hospitalo-universitaires différents : deux professeurs, deux maitres de conférences agrégés, quatre assistants hospitalo-universitaires et deux médecins spécialistes de la santé publique.
Au cours d'une analyse complémentaire, les ancrages situés à distance de la réponse modale ont été considérés comme erronés. Si l'erreur était liée à un seul expert, sa réponse n'a pas été prise en compte et si elle était faite par plusieurs experts, l'ensemble de l'item n'a pas été pris en compte dans la notation. Les items où la distribution des réponses n'était pas centrée sur une partie de l'échelle de Likert, n'ont pas été également pris en compte.

Établissement des scores : méthode des scores combinés
L'établissement des scores s'est basé sur la méthode des scores combinés qui utilise les réponses des experts recrutés pour construire un barème de scores, où chaque réponse des experts possède une valeur intrinsèque et reçoit une pondération en fonction du nombre d'experts qui l'ont choisie. Plus précisément, la réponse à un item la plus souvent choisie par les experts (la réponse modale) vaut un point ; les autres réponses des experts reflètent la variabilité dans l'interprétation des données et donnent droit à une fraction d'un point, en fonction du nombre d'experts qui les ont choisies. Par exemple à l'un des items, 7 experts répondent « -1 », 2 répondent « -2 », 1 répond « 0 » et aucune ne répond « 1 » ou « **+**2». Pour obtenir les scores, il s'agit de repérer la réponse qui a été la plus choisie par les experts (dans ce cas-ci, la réponse « -1 » a été choisie par 7 experts). Cette réponse donnera maintenant un point à toutes les personnes évaluées qui la choisiront et elle devient la norme. Ensuite, on divise toutes les réponses par le nombre d'experts le plus élevé à avoir choisi la réponse. Dans cet exemple, il faut diviser par 7. Les réponses non choisies par les experts donnent 0 point et les autres réponses donnent un crédit en fonction du nombre d'experts qui les ont choisies, soit 2/7 = 0,28 point pour la réponse « -2 » et 1/7 = 0,14 pour la réponse « 0 ». Tableau 1

Tableau 1 : Méthode des scores combinés. Exemple de réponse à un item par dix experts.

	-2	-1	1	2	Instructions
Nombre de réponses parmi les experts	2	7	0	0	Repérer la réponse modale la plus choisie)
Création de scores	2/7	7/7	0/7	0/7	Diviser par le nombre d'experts ayant donné la réponse la plus choisie
Crédit pour l'item	0.28	1	0	0	Donner les points aux étudiants examinés pour cet item et aux experts

Le score global a été obtenu en ajoutant les scores de chaque question, puis les transformant en un score combiné sur 100.

La réalisation du test
Les différents groupes d'externes et d'internes présents au cours de la période d'étude ont été évalués par le TCS au dernier jour de leur stage. Le déroulement du test a eu lieu dans des conditions identiques, en salle de staff pendant 20 à 30 minutes. L'évaluation était anonyme et les étudiants étaient informés de son but. Un paragraphe expliquant les principes du test était placé en préambule.

Analyse statistique
Les résultats ont été exprimés en moyennes ± écart-type. Les comparaisons de plusieurs moyennes sur séries indépendantes ont été effectuées au moyen du test F de Snedecor d'analyse de la variance paramétrique (ANOVA à un facteur). En cas de différence significative, les comparaisons 2 à 2 ont été faites par la méthode de Bonferroni. La normalité des distributions a été étudiée par le test de Kolmogorov et Smirnov et graphiquement. L'égalité des variances a été étudiée par le test de Levene.
La cohérence interne (fidélité) du TCS a été évaluée par le coefficient α de Cronbach à l'aide d'une feuille de calcul Excel téléchargeable sur le site de l'université de Montréal.

RÉSULTATS

Les réponses du groupe d'experts sont détaillées sur le tableau 2. Les experts ont eu un score moyen, rapporté sur 100, de 82,52 ± 7.35, $IC_{95\%}$ [77,26-87,78], les externes de 58,52 ± 9,72, $IC_{95\%}$ [55,82-61,23], (p<0,0001) ; et les internes de 63.17±11.36, $IC_{95\%}$ [55,53-70,81] (p<0.0001). Les coefficients α de Cronbach pour les réponses des externes et des internes respectifs étaient de 0,68 et de 0,61. La différence des scores entre étudiants et internes n'était pas statistiquement significative, p=0,46. Tableau 3, figure2.

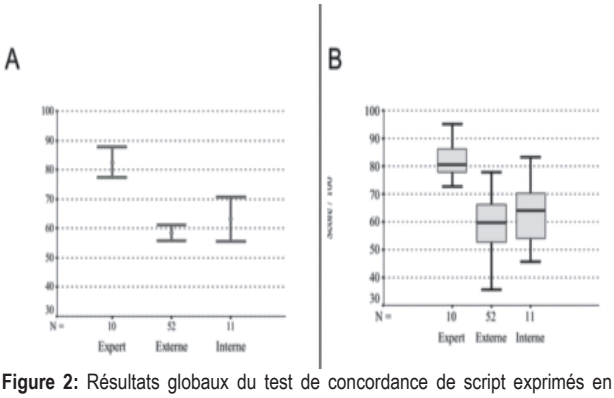

Figure 2: Résultats globaux du test de concordance de script exprimés en scores moyens avec barres d'erreurs (A) et en médianes et quartiles (B) pour les trois groupes : Expert, Externe et interne.

Tableau 2: Réponses des experts au test de concordance de script

Echelle Likert	Vignette 1	Vignette 2	Vignette* 3	Vignette* 4	Vignette 5	Vignette 6	Vignette 7	Vignette 8	Vignette 9	Vignette* 10
-2	0	0	0	1	2	9	0	8	6	0
-1	0	0	1	6	7	1	0	2	2	1
0	4	0	1	2	1	0	0	0	2	6
1	4	5	3	1	0	0	5	0	0	0
2	2	5	5	0	0	0	5	0	0	3
Echelle Likert	**Vignette* 11**	**Vignette 12**	**Vignette 13**	**Vignette 14**	**Vignette 15**	**Vignette 16**	**Vignette 17**	**Vignette 18**	**Vignette 19**	**Vignette 20**
-2	1	0	4	7	0	0	0	3	1	1**
-1	4	3	6	3	7	0	1	2	5	0
0	3	7	0	0	2	10	8	5	4	1
1	0	0	0	0	0	0	0	0	0	2
2	2	0	0	0	1**	0	1**	0	0	6

*Items pouvant être considérées comme erronées ou discutables
** ancrages pouvant être considérés comme erronés ou discutables

Tableau 3 : Répartition des scores Experts/Etudiants/Internes.

Scores	Experts N=10	Externes N=52	Internes N= 11
Moyenne	82.52	58.52	63.17
IC 95%	[77.26-87.78]	[55.82-61.23]	[55.53-70.81]
Médiane	80.67	59.72	64.15
Variance	54.11	94.55	129.25
Ecart type	7.35	9.72	11.36
Minimum	72.75	35.75	45.65
Maximum	95.05	77.8	83.15
Coefficient α de Cronbach		0.68	0.61

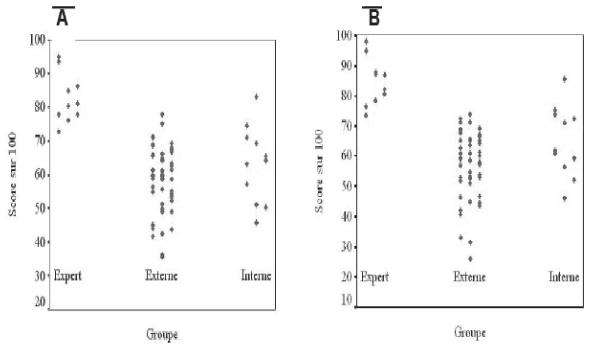

Figure 3 : Dispersion des scores individuels dans chacun des groupes avant (A) et après exclusion (B) des items erronés ou discutables

Trois ancrages et quatre items ont été considérés comme erronés ou discutables et ont été retirés dans l'analyse complémentaire. Tableau 2

La suppression des ancrages et des items erronés ou discutables a conduit aux résultats suivants : Experts 84,49± 7,88, $IC_{95\%}$ [78,91-90,19], externes 57,46±10,95, $IC_{95\%}$ [54,41-0,51] ($p<0.0001$), et internes 64,95±11,59, $IC_{95\%}$ [57,16-72,74] ($p<0,0001$). Les coefficients α de Cronbach respectifs pour les réponses des externes et des internes étaient de 0,96 et de 0,99. La différence entre étudiants et internes n'était pas statistiquement significative, p=0,11.

La dispersion des scores individuels dans chacun des groupes avant et après exclusion des items erronés ou discutables est représentée sur la figure 3.

DISCUSSION

Dans ce travail, nous avons retrouvé des différences statistiquement significatives des scores moyens entre experts et externes et entre experts et internes. Les scores moyens étaient plus élevés chez les internes que chez les externes mais cette différence n'était pas statistiquement significative. L'absence de différence peut être expliquée par le fait qu'une seule année d'étude les sépare et que l'encadrement prodigués en terrain de stage est plus régulier pour les externes.

L'organisation des connaissances adaptées aux tâches cliniques se développe progressivement avec l'expérience acquise au cours de la formation et donc très tôt au cours des stages hospitaliers. Cette progression des scores de TCS a été validée dans d'autres spécialités à différents niveaux de formation dont de deuxième cycle des études médicales [8-10].

Des scores significativement plus bas chez les externes pourraient faire discuter la pertinence même de l'utilisation du TCS chez des étudiants de deuxième cycle, dans la mesure où à leur niveau, il faut commencer par leur transmettre des certitudes, leur permettant d'acquérir des représentations mentales correspondant à la fois à des réalités, et à des repères. Certes, au cours de leurs stages hospitaliers, ils vont être confrontés inéluctablement à des situations d'incertitudes, mais encore faut-il qu'ils apprennent à raisonner devant des situations typiques imposant un diagnostic de certitude.

La principale limite de notre travail était le nombre réduit d'items. Le caractère nouveau du test, aussi bien pour les étudiants que pour les enseignants était la principale raison ayant conduit à un nombre limité de vignettes, comprenant un seul item pour chaque vignette. Malgré ce nombre limité d'items, le coefficient de Cronbach était déjà satisfaisant à l'analyse globale, 0,68 pour les

réponses d'externes et 0,61 pour celles des internes. Ces coefficients ont augmenté après exclusion des items jugés erronés ou discutables à variabilité très élevée entre experts. L'exclusion des items à haute variabilité enlèverait du test une part importante d'incertitude. Une variabilité élevée témoigne de la difficulté liée à l'incertitude, et la variabilité du test augmente la discrimination, effet recherché dans un TCS, mais quand la variabilité devient extrême ceci pourrait témoigner plutôt d'erreur [11]. Cette erreur peut émaner du panel d'experts ou alors être au niveau de l'élaboration même de la vignette clinique. Dans un TCS destiné à des internes exerçant dans un service d'accueil des urgences pédiatriques, on avait fait passer un test comprenant 60 questions réparties en 38 vignettes mais les auteurs n'avaient retenu que 50 questions dans 30 vignettes, optimisant ainsi le score alpha de Cronbach à 0,77[12].

Si on crée un seul item par vignette l'effort de réflexion est considérable pour les étudiants qui doivent se plonger dans une grande série de cas clinique [4]. La durée de notre test (30 mn) était suffisante pour répondre à 20 questions dans 20 vignettes cliniques. Les différents coefficients de Cronbach calculés étaient satisfaisants même après élimination de quatre vignettes jugées erronées, suggérant que c'est la qualité de la vignette elle-même et la variabilité des réponses qu'elle génère qui interviennent plutôt que leur nombre seul.

L'élaboration des vignettes dans un contexte d'incertitude est une tâche difficile surtout quand on n'est pas formé et habitué à cet outil. Dans notre test, les vignettes étaient rédigées par un seul expert et validé par un autre, tous les deux sans expérience antérieure.

Il est préférable que trois experts participent à l'élaboration d'un TCS afin de multiplier les expériences et les jugements pour obtenir des vignettes et des questions optimales pour ce test, sans ancrage anormal et avec un certain degré de dispersion mais limité autour de la réponse modale, donc à variabilité correcte [6].

Les tests psychométriques montrent des résultats meilleurs à l'échelle individuelle pour les items à variabilité modérée, tandis que les items à grande variabilité discriminent mieux l'expérience clinique à l'échelle collective [11]. Ces constatations expliquent en partie l'absence d'amélioration des scores de nos externes et internes après élimination de certains items considérés erronés ou discutables, mais d'un autre côté l'amélioration des coefficients α de Cronbach. En effet, les items à faible variabilité explorent les connaissances cliniques formelles plutôt que le raisonnement face à des données cliniques complexes et ambigües. Le panel d'experts peut posséder plus de facilité à raisonner dans un contexte d'incertitude que les juniors, mais pas forcément devant des connaissances où il n'y a pas d'incertitude [11]. Charlin et al, en étudiant l'effet de la variabilité au sein des réponses du panel de référence, ont démontré que la variabilité des réponses était un élément clé du pouvoir discriminant du TCS [11]. Il recommandait ainsi de construire délibérément des items pouvant induire une variabilité au sein du panel d'experts. Cependant, l'existence d'items à faible variabilité est nécessaire. Ces items évaluent les connaissances de problèmes bien définis et bien établis. Or ce type d'items s'apparente à une QCM à contexte riche sauf que le format et la tâche requise aux étudiants diffèrent.

CONCLUSION

Notre travail a démontré la faisabilité du TCS en stage de néonatologie. Cependant, elle a fait ressortir des difficultés que nous n'avions pas anticipées, au niveau de la rédaction des vignettes et des items et de leur validation et au niveau du choix des experts. L'élaboration du TCS impose de constituer un groupe d'enseignants formés et habitués progressivement à ce test pour respecter les recommandations actuelles. Le recours au TCS serait plus pertinent pour l'évaluation des internes et des résidents de spécialité confrontés à des situations parfois complexes et incertaines. Il l'est moins à un niveau d'études où l'étudiant doit encore acquérir des repères avant de manier l'incertitude.

Références

1. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C.The Script Concordance test: a tool to assess the reflective clinician. Teach Learn Med 2000; 12:189-95.
2. Charlin B, van der Vleuten C. Standardized assessment of reasoning in context of uncertainty. The script concordance test approach. Eval Health Prof. 2004; 27(3):304-19.
3. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CP. Script concordance method: a review of published validity evidence. Med Educ. 2011; 45(4):329-38.
4. Fournier JP, Demeester A, Charlin B. Script Concordance Tests: Guidelines for Construction. BMC Med Inform Decis Mak 2008; 8:18.
5. Gagnon R, Charlin B, Coletti M, Sauvé E,Van der Vleuten C. Assessment in context of uncertainty: How many members are needed on the panel of reference of a script concordance test? Med Educ 2005; 39:204-91.
6. Gagnon R, Charlin B, Lambert C, Carrière B, der Vleuten V. Script Concordance Testing: More Cases or More Questions? Adv Health Sci Educ 2009; 14:1382-96.
7. Charlin B, Gagnon R, Lubarsky S and al. Assessment in the context of uncertainty using the script concordance test: More meaning for scores. Teach Learn Med 2010; 22:180-6.
8. Gibot S, Boulaert PE. Le test de concordance de script outil d'évaluation formative en reanimation médicale. Pédagogie médicale 2008 ; 9:7-18.
9. Sibert L, Charlin B, Gagnon R,Corcos J, Khalaf A, Grise P. Evaluation du raisonnement clinique en urologie : l'apport du Test de Concordance de Script. Prog Urol 2001 ;11 :1213-1219.
10. Caire F, MarinB, Cuny E. Utilisation du test de concordance de script au cours du 2ème cycle des études médicales. Expérience dans l'enseignement de la neuro-chirurgie. Pédagogie médicale 2011 ; 12(1):29-35.
11. Charlin B, Gagnon R, Pelletier J and al. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. Med Educ 2006; 40: 848-854.
12. Carriere B, Gagnon R, Charlin B, et al. Assessing clinical reasoning in pediatric emergency medicine: validity evidence for a Script Concordance Test. Ann Emerg Med 2009;53:647-52.