RECHERCHE ET PERSPECTIVES

Pédagogie Médicale 2013; 14 (3): 157–168
DOI: 10.1051/pmed/2013047

Impact du panel de référence sur les qualités psychométriques d'un test de concordance de script développé en formation initiale des sages-femmes

Impact of the reference panel on the psychometric quality of a script concordance test developed for midwifery training

Madeleine GANTELET[1], Anne DEMEESTER[2], Vanessa PAULY[3], Robert GAGNON[4], Bernard CHARLIN[4]

[1] École de sages-femmes de Besançon, Centre hospitalier régional et universitaire de Besançon France
[2] École universitaire de maïeutique Marseille Méditerranée, Université d'Aix Marseille, France
[3] Centre d'investigation clinique, Assistance Publique – Hôpitaux de Marseille, France
[4] Centre de pédagogie appliquée aux sciences de la santé, Faculté de médecine, Université de Montréal, Québec, Canada

Manuscrit soumis le 26 juillet 2012 ; commentaires éditoriaux formulés aux auteurs le 14 octobre 2012 ; accepté pour publication le 2 juin 2013

Mots-clés
Raisonnement clinique ;
sages-femmes ;
test de concordance de script ;
clés de correction ;
cohérence interne ;
évaluation

Résumé – Contexte : Le test de concordance de script (TCS) a pour particularité de recourir à des professionnels expérimentés pour établir les scores du test. La composition de ce panel a été discutée dans des études. Une étude portant sur les étudiants sages-femmes a montré un coefficient α de Cronbach assez faible. **But :** Analyser l'impact de la composition spécialisée ou non d'un panel de professionnels sur les qualités psychométriques d'un TCS développé dans la filière maïeutique. **Matériel et Méthodes :** Une étude analytique prospective multicentrique a été menée auprès d'étudiants en dernière année dans six écoles de sages-femmes françaises. Deux panels de professionnels ont été constitués : Panel A de sages-femmes praticiennes polyvalentes, panel B de sages-femmes praticiennes spécialisées dans un des quatre secteurs d'activité de la profession. Les scores des étudiants sages-femmes ont été établis avec les panels A et B. **Résultats :** Les clés de correction et les scores ne varient pas avec le panel. Le coefficient de cohérence interne α de Cronbach n'est satisfaisant ni avec le panel A, ni avec le panel B. **Conclusion :** Les vignettes de ce test doivent être améliorées pour être utilisées par les écoles. Le nombre de vignettes doit être augmenté en créant une banque de vignettes pour chacun des domaines. L'hypothèse d'explication du faible coefficient α est la dispersion des vignettes entre les différents secteurs d'activité de la profession.

Keywords
Clinical reasoning; midwives; script concordance test; scoring key; coefficient of reliability; assessment

Abstract – Context: Script Concordance Tests (SCTs) are designed to measure the degree of concordance between examinees and a reference panel of experts on clinical decisions and actions under uncertainty. To create a scoring key, SCTs are submitted to a reference panel. The validity of such a panel is often challenged. Prior studies on SCTs in the training of midwives have shown a low reliability coefficient (Cronbach's α). **Objective:** To analyze the impact of a reference panel of generalist versus specialist midwives working in one of the four midwifery fields on the psychometric quality of SCTs developed for midwives. **Materials and Methods:** We conducted a prospective multicenter study among students in the final year of studies in 6 French midwifery schools. To create a scoring key, two panels of professionals were set up: panel A consisting of generalist midwives and panel B consisting of midwives specializing in one of the four midwifery fields. Students' scores were referenced with panels A and B. **Results:** The scoring key and the students' scores do not vary from the panels. The internal consistency estimate of reliability using Cronbach's α is unacceptable, whether with panel A or B. Some vignettes are unsatisfactory. **Conclusion:** The test should be improved before it is used. The wide range of vignettes covering various fields of practice (gynecology, obstetrics and pediatrics) can probably explain the low reliability coefficient. The test should be developed in all midwifery schools by creating a vignette bank for each midwifery field.

Introduction

Le référentiel de métier et de compétences des sages-femmes françaises a été élaboré par le Collectif des associations et de syndicats de sages-femmes (avec la participation du Conseil national de l'Ordre des sages-femmes)[1]. Dans ce document, le raisonnement clinique, qui vise, dans une situation donnée, à élaborer un diagnostic, une conduite à tenir et un pronostic, est reconnu comme une compétence de la sage-femme. Il est donc important de pouvoir évaluer et valider cette compétence de raisonnement clinique dans les 35 écoles de sages-femmes françaises. Les sages-femmes françaises ont la particularité de posséder un statut médical qui légitime cette compétence. Actuellement, à l'école de sages-femmes de Besançon, comme dans d'autres écoles de sages-femmes françaises, l'évaluation du raisonnement clinique est réalisée de façon transversale dans le cadre de contrôles continus qui comportent des analyses de cas cliniques, et lors d'évaluations cliniques pratiquées sur les lieux de stages. Des grilles d'observation inspirées de la pédagogie par objectifs sont utilisées. Le test de concordance de script (TCS) est un outil d'évaluation spécifique du raisonnement clinique élaboré par Charlin[2,3]. Ce test a pour particularité de mesurer des indicateurs du raisonnement clinique en contexte d'incertitude ; pour cela il recourt à un panel de professionnels pour élaborer la grille de correction. La composition quantitative et qualitative du panel est l'une des difficultés de l'élaboration d'un TCS. Une étude menée à l'école de sages-femmes de Marseille[4] a permis de montrer qu'il discrimine le niveau de performance entre novice et expert. Cette étude a permis de documenter la validité de construit partielle de ce test en formation initiale des sages-femmes mais a révélé un faible coefficient de cohérence interne. Il était dès lors légitime de conduire une recherche complémentaire pour améliorer les qualités psychométriques du TCS lorsqu'il est utilisé dans le domaine de la maïeutique, notamment la cohérence interne dont le coefficient α de Cronbach est un indicateur. D'une façon plus générale, il paraissait opportun d'examiner l'influence d'une spécialisation des membres du panel sur les qualités psychométriques du test.

Le but du travail rapporté dans cet article est d'analyser l'impact de la spécialisation du panel de professionnels sur les qualités psychométriques du TCS. Plus précisément, nous souhaitons analyser la variabilité du coefficient de cohérence interne du test en utilisant deux panels différents : un panel de sages-femmes polyvalentes et un panel de sages-femmes spécialisées par secteur d'activité. L'étude analyse l'évolution des clés de correction, des scores obtenus par des étudiants et du cœfficient α de Cronbach d'un

TCS, en fonction des clés de correction obtenues par l'un ou l'autre des panels.

Cadre conceptuel du test de concordance de script

Le TCS explore les processus mis en œuvre par l'étudiant au cours du raisonnement clinique (et non le résultat). L'approche du TCS se fonde sur la théorie des scripts[5]. En 1988 et 1989, Nelson et Fayol, cités par Nendaz[6], ont décrit cette architecture de connaissances. Les scripts sont des connaissances élaborées ou compilées, spécifiquement organisées pour être efficaces dans une tâche clinique. Le novice commence par un raisonnement causal ou conditionnel qui est exigeant cognitivement. Il se construit ensuite des réseaux de connaissances pour agir plus vite : les scripts.

L'efficacité de l'architecture scripturale des connaissances est due à l'automatisme et à la conscience de son utilisation qui permet d'argumenter le raisonnement. Le script guide la recherche et la sélection efficace d'informations complémentaires. Il n'est pas exclusif des savoirs biomédicaux qui peuvent être exploités comme éléments de contrôle et comme ressource complémentaire quand la genèse d'hypothèse fait défaut.

Les scripts de maladie ont été décrits par Feltovitch et Barrow[7] en 1984. Ils font référence à un réseau de connaissances relatives à une maladie donnée : connaissances quantitatives (physiopathologies, sémiologie et expériences relatives à une maladie donnée) et connaissances qualitatives (organisation par mise en lien et hiérarchisation de toutes ces informations)[8].

Le TCS consiste à présenter au candidat une série de problèmes cliniques. Une hypothèse initiale de diagnostic, d'investigation, de traitement ou de pronostic est faite Le candidat est invité à évaluer l'impact d'une nouvelle information sur cette hypothèse initiale par le choix d'un des ancrages d'une échelle de Likert. Autrement dit, le candidat est invité à évaluer la liaison entre ces deux informations (l'hypothèse initiale et la nouvelle donnée). Le TCS permet d'observer et mesurer une concordance de réponse (résultat), à partir de laquelle on infère une concordance de scripts et une concordance de processus. Un guide de construction du TCS a été élaboré[9]. La base du test est une vignette qui se présente sous forme de cas clinique, suivie de plusieurs questions portant sur le cas décrit. Le fait que la vignette soit courte participe à l'incertitude.

Dix à vingt professionnels constituent le panel de référence et passent le test dans des conditions identiques à celle de l'examen[10]. Pour chaque question, les résultats de ce panel sont analysés comme suit pour élaborer les clés de correction. À chaque option de réponse, le nombre de membres du panel qui l'ont choisie est transformé proportionnellement : la réponse qui a été la plus choisie (la valeur modale) est créditée d'un point ; les autres choix reçoivent un crédit partiel. Ainsi, pour transformer les scores, tous sont divisés par le nombre de membres qui avaient choisis la valeur modale. On obtient ainsi les clés de correction de chaque proposition.

Une série de plusieurs études a permis d'évaluer les qualités psychométriques du TCS. La fidélité du TCS c'est à dire la capacité à donner le même résultat en situation identique a été évaluée[11]. Les études montrent qu'avec 60 questions administrées en une heure, le test est presque toujours fidèle en conservant le coefficient α de Cronbach satisfaisant (valeur égale ou supérieure à 0,75). Sibert a mesuré la stabilité du test entre deux cultures[12]. Dans le domaine de l'urologie, il a noté une stabilité dans les résultats d'étudiants français et canadiens si on réalise la correction avec un panel de professionnels français ou canadiens. Le test est facilement réalisable car il suffit en général d'une heure pour l'administrer lorsque le champ de la mesure est délimité. Un TCS peut cependant durer plus d'une heure en fonction de l'ampleur du champ de connaissance à évaluer et de l'enjeu de l'examen. De plus, le test peut être complètement ou partiellement informatisé. Il est réutilisable s'il n'est pas divulgué aux étudiants en dehors de l'administration du test. Il est bien accepté par les étudiants sages-femmes[13].

Matériel et méthodes

L'étude est analytique, prospective et multicentrique. Elle a été réalisée au sein de six écoles de

Tableau I. Critères d'expérience et de formation continue des sages-femmes des panels A et B.

Critère d'inclusion des sages-femmes pour le panel A	**Critère d'inclusion des sages-femmes pour le panel B = mêmes critères que A et en plus :**
2 sages-femmes (polyvalentes) - Ayant au moins 5 ans d'expérience (en équivalent temps plein) - Travaillant dans un centre hospitalier universitaire depuis au moins 2 ans - En activité à ce jour. - Ayant réalisé au moins une formation continue* dans l'année écoulée.	2 sages-femmes - Dont au moins 2 ans en gynécologie, consultations prénatales ou consultations d'urgence sur ces 5 ans d'expérience - Sage-femme ayant réalisé au moins une formation continue* dans ce domaine
	2 sages-femmes - Dont au moins 2 ans en pathologies maternelles et fœtales - Sage-femme ayant réalisé au moins une formation continue* dans ce domaine
	2 sages-femmes - Dont au moins 2 ans en salle de naissance - Sage-femme ayant réalisé au moins une formation continue* dans ce domaine*
	2 sages-femmes - Dont au moins 2 ans en suites de couches post-natal mère – enfant - Sage-femme ayant réalisé au moins une formation continue* dans ce domaine

*Au moins une journée de formation (colloque, congrès sur ce thème) mais si possible formation plus longue ou diplômante.

sages-femmes françaises, dispersées sur le territoire, afin d'éviter « l'effet d'école » supposé non négligeable en obstétrique. Le choix de ces écoles s'est fait sur la base du volontariat. L'étude s'est déroulée de décembre 2007 à mars 2008 dans les six écoles de sages-femmes et les six maternités-écoles de Besançon, Brest, Clermont-Ferrand, Marseille, Paris Saint-Antoine et Toulouse.

Sélection des panels

La population de sages-femmes expertes des deux panels répondait à des critères définis d'expérience et de formation continue (Tableau I). La sélection des sages-femmes a été faite par les sages-femmes directrices et enseignantes des écoles, en concertation avec les sages-femmes cadres des maternités.

Le panel A était classiquement constitué de 12 professionnels polyvalents. Les sages femmes de ce panel ont permis le calcul des clés de correction de l'ensemble des vignettes du test.

Le panel B était constitué de quatre groupes de chacun 12 professionnels experts par domaine d'activité des sages-femmes. La spécialisation du panel de référence est définie au regard de quatre domaines d'activité de la profession sage-femme, chacune étant « référente » pour l'un des domaines suivants : 1) gynécologie, urgences 1er trimestre, consultations prénatales ; 2) pathologies maternelles et fœtales ; 3) salle de naissance ; 4) post-natal mère et enfant. Le panel B a permis de calculer les clés de correction des vignettes par domaine.

Étudiants participants

La population étudiée était constituée de tous les étudiants sages-femmes inscrits en dernière année des études de sages-femmes dans ces six écoles, soit

Tableau II. Répartition des vignettes en fonction des domaines et dimensions explorés.

Domaine	Dimension explorée	Nombre de vignettes
Gynécologie – Urgences 1er trimestre – Consultations prénatales	Diagnostic	2
	Thérapeutique ou Exploration	2
	Pronostic	1
Pathologies maternelles et fœtales	Diagnostic	2
	Thérapeutique ou Exploration	2
	Pronostic	1
Salle de naissance	Diagnostic	2
	Thérapeutique ou Exploration	2
	Pronostic	1
Post-natal : mère – enfant	Diagnostic	2
	Thérapeutique ou Exploration	2
	Pronostic	1

160 participants. La participation des étudiants au TCS était obligatoire. En fonction des modalités de contrôle des connaissances de chacune des écoles, la note obtenue par l'étudiant au TCS était incluse dans la moyenne des évaluations.

Construction du test de concordance de script

Les 20 cas cliniques du test ont été classés en quatre groupes correspondants à quatre secteurs d'activité de la profession sage-femme : 1) gynécologie, urgences 1er trimestre, consultations prénatales ; 2) pathologies maternelles et fœtales ; 3) salle de naissance ; 4) post-natal mère et enfant. Les cas ont été répartis régulièrement dans ces quatre domaines et dans les trois dimensions du raisonnement clinique (Tableau II).

Le guide de construction du TCS[8] recommande de construire le test « par cas ». Ainsi la construction de chacune des vignettes du TCS a été déclinée en trois questions pour un cas. Aussi, cette correction permet d'établir un score par cas en faisant une moyenne des scores des trois questions du cas.

Pour ce même test, chaque étudiant avait donc quatre scores : un score A par question avec le panel A (addition des scores des trois questions du cas), un score Ac par cas avec le panel A (moyenne des scores des trois questions du cas), un score B par question avec le panel B et un score Bc par cas avec le panel B.

Analyse statistique

Les tests statistiques utilisés ont été le test Z (loi normale centrée réduite) pour la comparaison de moyennes. Le risque α de première espèce de conclure à tort à une différence significative était fixé à 5 %.

Pour analyser la cohérence interne du test, le coefficient α de Cronbach a été calculé. Plus la cohérence interne d'un test est satisfaisante, plus le coefficient α est proche de 1.

Le test étant nouvellement construit, nous avons souhaité analyser la qualité des questions. Le coefficient alpha a été calculé en supprimant alternativement chacune des questions. Ceci nous a permis de repérer les questions qui diminuaient ou augmentaient la cohérence interne globale du test.

Pour compléter cette analyse nous avons voulu analyser l'impact de la dispersion des domaines dans la cohérence interne du test. Pour chaque question, nous avons mesuré deux coefficients de corrélation : un coefficient de corrélation multiple sur les autres

Tableau III. Variabilité des clés de correction entre les panels A et B.

Clé de correction	A-B	A GYN-B GYN	A PMF-B PMF	A SDN-B SDN	A SDC-B SDC
Moyenne	**0**	0,02	**–0,04**	0,05	**–0,03**
Ecart type	**0,32**	**0,37**	**0,34**	**0,28**	**0,30**

GYN : Gynécologie – Urgences 1er trimestre – Consultations prénatales.
PMF : Pathologies maternelles et fœtales.
SDN : Salle de naissance.
SDC : Post-natal : mère – enfant.

Tableau IVa. Variabilité des scores entre les panels A et B et selon le mode de correction (méthode traditionnelle ou méthodes des cas).

SCORE	A	B	A Cas	B Cas	A	A Cas	B	B Cas
Moyenne	13,71	13,71	13,71	13,71	13,71	13,71	13,71	**13,71**
Ecart type	1,31	1,09	1,31	1,08	1,31	1,31	1,09	**1,08**
Loi normale	0,04		–0,016		0		**–0,06**	
p	***p* > 0,48**		***p* > 0,505**		***p* = 0,5**		***p* > 0,52**	

Tableau IVb. Variabilité du coefficient α de Cronbach selon les panels A et B et selon le mode de correction (méthode traditionnelle ou méthodes des cas).

Panel A		Panel B	
Méthode traditionnelle **A**	Méthode cas **A Cas**	Méthode traditionnelle **B**	Méthode des cas **B Cas**
α_A= 0,416	**$\alpha_{A\ cas}$ = 0,395**	**α_B = 0,312**	**$\alpha_{B\ cas}$ = 0,300**

questions du même domaine (coefficient de corrélation intra-domaine) et un autre coefficient de corrélation multiple sur les questions des autres domaines (coefficient de corrélation extra-domaine). Ces coefficients de corrélations ont été calculés via deux régressions linéaires distinctes de la question d'intérêt sur les autres questions (intra-domaine puis extra-domaine). Ainsi ce coefficient analyse la force du lien d'une question avec les questions de son domaine et les questions des autres domaines.

Résultats

Sur l'ensemble du test, la moyenne des différences de clés de correction entre le panel A et le panel B est nulle. Dans les différents domaines spécifiques, la moyenne des différences de clés de correction varie de –0,04 à 0,05 (Tableau III).

Aucune différence significative de moyenne entre les scores issus du panel A et les scores issus du panel B n'est mise en évidence (Tableau IVa).

Que ce soit avec le panel A ou le panel B, il n'y a pas de différence de moyenne entre les scores émanant de la correction classique ou par cas.

Le coefficient α de Cronbach a été calculé dans chacun des modes de correction utilisés. Le coefficient α de Cronbach est inférieur à 0,5 dans les quatre situations (α_A= 0,416 ; $\alpha_{A\ cas}$ = 0,395 ; α_B = 0,312 ; $\alpha_{B\ cas}$ = 0,300). Que ce soit dans le panel A ou dans le panel B, le coefficient α de Cronbach est plus faible quand la correction est réalisée par la méthode des cas. Il est également plus bas quand les clés de correction sont issues du panel B par rapport au panel A (Tableau n° IVb).

Tableau Va. Calcul du coefficient alpha de Cronbach en supprimant alternativement chaque vignette et question dans le panel A.

Vignette	Question	α_A	$\alpha_{A\ cas}$	Vignette	Question	α_A	$\alpha_{A\ cas}$	Vignette	Question	α_A	$\alpha_{A\ cas}$
1	1.1	0,414	**0,403***	8	8.1	0,389	0,370	15	15.1	0,393	0,375
	1.2	0,410			8.2	0,420			15.2	0,414	
	1.3	**0,429**			8.3	0,402			15.3	0,417	
2	2.1	0,414	**0,407**	9	9.1	**0,427**	**0,411**	16	16.1	0,400	0,386
	2.2	0,411			9.2	0,417			16.2	0,414	
	2.3	**0,429**			9.3	0,416			16.3	**0,424**	
3	3.1	0,392	0,382	10	10.1	0,420	0,390	17	17.1	0,412	0,338
	3.2	0,413			10.2	0,396			17.2	0,400	
	3.3	**0,424**			10.3	**0,425**			17.3	0,386	
4	4.1	**0,422**	0,379	11	11.1	0,403	0,359	18	18.1	0,409	0,384
	4.2	0,410			11.2	0,414			18.2	0,413	
	4.3	0,416			11.3	0,395			18.3	**0,422**	
5	5.1	0,386	0,343	12	12.1	0,417	0,381	19	19.1	**0,425**	**0,412**
	5.2	**0,422**			12.2	0,418			19.2	0,418	
	5.3	0,397			12.3	0,395			19.3	**0,425**	
6	6.1	0,392	0,374	13	13.1	**0,427**	0,392	20	20.1	**0,432**	**0,441**
	6.2	0,414			13.2	0,415			20.2	**0,431**	
	6.3	**0,421**			13.3	0,401			20.3	**0,422**	
7	7.1	0,370	0,339	14	14.1	0,400	0,370				
	7.2	0,410			14.2	0,410					
	7.3	0,412			14.3	0,403					

* Les résultats en gras soulignés montrent qu'en supprimant cette question ou cette vignette, le coefficient alpha de Cronbach augmente d'au moins 0,05.

Une analyse par question a montré que la suppression de certaines questions améliorait le coefficient α de Cronbach. Dans la correction avec le panel A, c'est le cas de 16 questions et cinq vignettes (Tableau Va). Dans la correction avec le panel B, c'est le cas de 17 questions et cinq vignettes (Tableau Vb).

La corrélation multiple intra-domaine est comprise entre 0,168 et 0,478. La corrélation multiple extra domaine est comprise entre 0,456 et 0,670. Pour chaque question la corrélation intra domaine est inférieure à la corrélation extra domaine (Tableau VI).

Discussion

Biais et limites de l'étude

L'échantillonnage des écoles a été de convenance. Les écoles volontaires étaient probablement sensibilisées à la formation au raisonnement clinique des étudiants sages-femmes. Ceci nous a permis de mener à bien l'étude par leur participation active mais a peut être sélectionné un type d'étudiant particulier.

Tableau Vb. Calcul du coefficient alpha de Cronbach, en supprimant alternativement chaque vignette et question dans le panel B.

Vignette	Question	α_B	$\alpha_{B\ cas}$	Vignette	Question	α_B	$\alpha_{B\ cas}$	Vignette	Question	α_B	$\alpha_{B\ cas}$
1	1.1	0,303	0,271	8	8.1	0,292	0,257	15	15.1	0,295	0,303
	1.2	0,292			8.2	0,289			15.2	**0,326**	
	1.3	0,309			8.3	0,313			15.3	0,313	
2	2.1	0,306	**0,341***	9	9.1	**0,334**	0,300	16	16.1	0,308	**0,306**
	2.2	**0,326**			9.2	0,297			16.2	0,310	
	2.3	**0,339**			9.3	0,309			16.3	**0,322**	
3	3.1	0,298	0,282	10	10.1	**0,334**	**0,312**	17	17.1	0,293	0,285
	3.2	**0,321**			10.2	0,300			17.2	0,306	
	3.3	0,308			10.3	0,310			17.3	**0,318**	
4	4.1	0,297	0,266	11	11.1	0,282	0,260	18	18.1	0,314	0,293
	4.2	0,311			11.2	**0,317**			18.2	0,311	
	4.3	0,302			11.3	0,296			18.3	0,303	
5	5.1	0,289	0,264	12	12.1	0,292	0,266	19	19.1	0,310	0,299
	5.2	**0,318**			12.2	**0,317**			19.2	0,315	
	5.3	0,297			12.3	0,288			19.3	**0,321**	
6	6.1	0,286	0,286	13	13.1	0,316	0,292	20	20.1	**0,334**	**0,328**
	6.2	**0,328**			13.2	**0,334**			20.2	0,315	
	6.3	0,309			13.3	0,287			20.3	0,312	
7	7.1	0,294	0,248	14	14.1	0,272	**0,315**				
	7.2	0,283			14.2	**0,321**					
	7.3	0,315			14.3	0,352					

* Les résultats en gras soulignés montrent qu'en supprimant cette question ou cette vignette, le Coefficient alpha de Cronbach augmente d'au moins 0,05.

Évolution des clés de corrections et scores

Les résultats montrent une stabilité des clés de correction et des scores obtenus si les clés de correction sont respectivement issues d'un seul panel de professionnels polyvalents ou d'un panel multiple de professionnels experts.

Évolution de la cohérence interne du test

Etant éloigné de la valeur 1, le coefficient α de Cronbach du TCS n'est pas satisfaisant dans cette étude. Il semble d'autant plus faible que les clés de correction sont issues d'un panel de multiples professionnels experts (α_B = 0,312 ; $\alpha_{B\ cas}$ = 0,300) par rapport à un seul panel de professionnels polyvalents (α_A = 0,416 ; $\alpha_{A\ cas}$ = 0,395). L'objectif d'amélioration de la cohérence interne par la spécialisation du panel n'est pas atteint mais les résultats de l'étude montrent que la spécialisation du panel ne modifie pas les qualités psychométriques du test.

Nous ne remettons pas en cause les principes d'élaboration du TCS basé sur le concept de script mais le test que nous avons construit manque de cohérence interne et ne peut être utilisé en l'état pour

Tableau VI. Corrélation des scores obtenus respectivement : pour chaque question et pour les autres questions du domaine (corrélation intra – domaine) ; pour chaque question et pour les questions des autres domaines (corrélation extra – domaine) avec le panel A.

Question	*R* intra domaine	*R* extra domaine	Questions	*R* intra domaine	*R* extra domaine	Questions	*R* intra domaine	*R* extra domaine
1.1	0,460	0,610	**8.1**	0,383	0,535	**15.1**	0,352	0,597
1.2	0,288	0,510	**8.2**	0,472	0,584	**15.2**	0,168	0,552
1.3	0,264	0,571	**8.3**	0,460	0,531	**15.3**	0,295	0,544
2.1	0,295	0,573	**9.1**	0,272	0,504	**16.1**	0,258	0,562
2.2	0,441	0,538	**9.2**	0,478	0,512	**16.2**	0,317	0,466
2.3	0,245	0,554	**9.3**	0,357	0,573	**16.3**	0,232	0,552
3.1	0,389	0,589	**10.1**	0,297	0,528	**17.1**	0,379	0,545
3.2	0,177	0,568	**10.2**	0,307	0,664	**17.2**	0,324	0,575
3.3	0,354	0,589	**10.3**	0,257	0,516	**17.3**	0,365	0,666
4.1	0,395	0,620	**11.1**	0,333	0,670	**18.1**	0,355	0,622
4.2	0,313	0,498	**11.2**	0,342	0,503	**18.2**	0,324	0,608
4.3	0,343	0,567	**11.3**	0,328	0,631	**18.3**	0,352	0,636
5.1	0,264	0,561	**12.1**	0,330	0,527	**19.1**	0,306	0,514
5.2	0,269	0,537	**12.2**	0,357	0,508	**19.2**	0,346	0,555
5.3	0,402	0,581	**12.3**	0,353	0,634	**19.3**	0,328	0,456
6.1	0,300	0,582	**13.1**	0,278	0,630	**20.1**	0,339	0,522
6.2	0,230	0,561	**13.2**	0,319	0,584	**20.2**	0,205	0,495
6.3	0,324	0,569	**13.3**	0,344	0,546	**20.3**	0,252	0,669
7.1	0,374	0,654	**14.1**	0,369	0,606			
7.2	0,321	0,620	**14.2**	0,327	0,596			
7.3	0,357	0,612	**14.3**	0,303	0,587			

R : coefficient de corrélation.

valider la compétence de raisonnement clinique des étudiants sages-femmes. Plusieurs explications et possibilités d'amélioration découlent de ce constat.

Propositions d'amélioration

Axe docimologique : amélioration des vignettes en quantité et en qualité

Il est possible d'augmenter la quantité de vignettes pour augmenter la cohérence interne du test. Comme le champ de connaissances à évaluer couvrait les quatre domaines de la spécialité sage-femme, le nombre de questions aurait dû être plus important.

Il est possible d'augmenter la qualité des vignettes. Les cas ayant une corrélation négative entre le score global et la question sont à améliorer. Le test proposé était nouvellement construit en collaboration avec l'équipe enseignante d'une école de sages-femmes. Il est évident, à la lumière des résultats, que certaines questions sont à retravailler. Comme Meterissian le préconise, il aurait été judicieux de prévoir plus de cas (100 questions) au départ, afin de pouvoir sélectionner ceux ayant la meilleure corrélation avec le score total[14]. Un entrainement à la rédaction des personnes qui rédigent les vignettes auraient probablement amélioré la qualité des vignettes. Les résultats d'une étude[15] suggèrent que la formation des auteurs d'examen améliore la qualité des questions à

choix multiple, sans pour autant évaluer de façon précise l'effet que pourrait avoir une telle formation sur les capacités de développement du test. La grille de relecture des vignettes élaborée par Caire[16] et appliquée ici n'a pas permis d'obtenir une qualité optimale du test.

Si nous avions eu plus de questions, des correcteurs automatiques en ligne qui éliminent habituellement 25 % des questions en fonction de leurs qualités métrologiques auraient pu être utilisés comme le conseille Dory[17].

Dory a analysé la composition qualitative du panel. Ses résultats montrent que l'entrainement des membres du panel au TCS n'a pas d'impact sur les qualités psychométriques du test contrairement au mode de pratique professionnelle (public *vs.* privé). Dans notre étude, les professionnels constituaient des panels « homogénéisés » puisque les professionnels exerçaient en établissements de santé publics, n'étaient pas enseignants, n'avaient pas d'entrainement à la pratique du TCS[17]. Cependant, nous n'avons pas analysé l'existence ou non « d'expert déviant ». Cette notion « d'expert déviant » a été développée par Gagnon[18] et pourrait expliquer la faible cohérence interne du test. D'après Gagnon, le nombre de membres du panel doit être supérieur à 15 pour limiter l'impact d'expert déviant en appliquant des méthodes. Nos panels d'experts n'en comprenaient que 12.

Axe lié à la théorie des scripts

Il est possible d'émettre des hypothèses liées à la théorie des scripts pour expliquer la faible cohérence interne du test. L'objectif était d'équilibrer les secteurs d'activité des sages-femmes afin d'améliorer la validité de contenu (Tableau II). En effet, polarisés au départ sur la composition des panels, nous avons eu le souci de couvrir largement les compétences des praticiennes pour construire ce test. Ces compétences relèvent de la maïeutique, champ disciplinaire situé au carrefour de plusieurs spécialités (gynécologie, obstétrique, néonatologie). Par conséquent, les cas étaient forcément moins nombreux par domaine et ceci a pu diminuer la cohérence interne. Il semble nécessaire de développer ce test dans chacun des secteurs d'activité pour recontrôler sa cohérence interne. Cette hypothèse ne semble cependant pas être confirmée par l'analyse du coefficient de corrélation multiple *R* puisque chacune des questions a un lien plus fort avec les questions des autres domaines qu'avec les questions de son domaine.

Dès la première étude en maïeutique, Demeester a introduit dans le test la dimension pronostique alors que seules les dimensions diagnostiques, d'investigation ou de thérapeutique du raisonnement médical avaient été évaluées jusqu'alors[4]. Dans notre test, nous avons remarqué que la vignette n° 20, qui explore la dimension pronostique en post-natal, n'est pas satisfaisante tant avec le panel A qu'avec le panel B. La question suivante pourrait faire l'objet d'une étude : le raisonnement mené pour élaborer un pronostic a-t-il une particularité qui serait moins bien évaluée par le TCS ? Si cela était le cas, cela pourrait en partie expliquer le manque de cohérence interne du test. Plusieurs questions méritent ainsi d'être examinées : le raisonnement utilisé pour établir un pronostic est-il différent du raisonnement de type hypothético-déductif souvent utilisé pour établir un diagnostic ou une conduite à tenir ? Le mode de raisonnement utilisé a t-il une influence sur la portée du test ? Il serait intéressant de regrouper les cas par dimension, pour voir si on améliore la cohérence interne dans chaque dimension. De même, il serait intéressant de développer un TCS dans un seul domaine et une seule dimension, pour voir si ceci augmente la cohérence interne du test en maïeutique.

D'autre part, les recherches de Demeester sur le raisonnement clinique spécifique des sages-femmes seraient intéressantes à poursuivre pour améliorer notre enseignement et notre capacité à l'évaluer[19].

Conclusion

L'étude n'a pas permis de montrer que la spécialisation du panel améliorait la cohérence interne du TCS mais elle a montré que cette spécialisation ne modifie pas les qualités psychométriques du test.

La variabilité du panel n'a pas affecté les clés de correction et les scores des étudiants. Indépendamment de la variabilité du panel, la cohérence interne

de ce test n'est pas encore satisfaisante. Évaluer le raisonnement clinique des sages-femmes dans tous les domaines (gynécologie, obstétrique, néonatologie) de la profession et dans toutes les dimensions du raisonnement clinique (diagnostic, investigation, thérapeutique, pronostic) semble diminuer la cohérence interne du test.

Les résultats de cette étude montrent qu'il est nécessaire de développer ce test dans chacun des domaines de la filière sage-femme pour en améliorer la cohérence interne.

Cette étude ouvre des perspectives de recherche dans les axes suivants : le raisonnement spécifique des sages-femmes, d'une part et la particularité du raisonnement permettant d'élaborer un pronostic, d'autre part. Cette recherche permettra d'améliorer le test et l'enseignement du raisonnement clinique dans la filière maïeutique.

Contributions

Madeleine Gantelet et Anne Demeester ont participé à la conception du protocole de recherche, au recueil des données, à l'interprétation des résultats, à l'analyse statistique et à l'écriture du manuscrit. Vanessa Pauly a participé à l'interprétation des résultats et à l'analyse statistique. Robert Gagnon a participé à la conception du protocole de recherche, à l'interprétation des résultats et à l'analyse statistique. Bernard Charlin a participé à la conception du protocole de recherche, à l'interprétation des résultats, à l'analyse statistique et à l'écriture du manuscrit.

Déclaration d'intérêt

Aucun auteur ne déclare de conflit d'intérêt en lien avec le contenu de cet article.

Approbation éthique

Le protocole de recherche n'a pas fait l'objet d'une demande d'approbation préalable par un comité d'éthique. Tous les participants ont été informés des objectifs du travail et l'anonymat a été respecté lors de l'exploitation des données.

Remerciements

Nous remercions, pour leur participation active à ce travail : les équipes pédagogiques des écoles de sages-femmes de Besançon, Brest, Clermont-Ferrand, Marseille, Paris Saint-Antoine et Toulouse ; toutes les sages-femmes et étudiants sages-femmes sollicités pour passer le test.

Références

1. Collectif des Associations et de Syndicats de Sages-Femmes (avec la participation du Conseil National de l'Ordre des Sages femmes). Référentiel métier et compétences des sages-femmes. 2010 [On-line] Disponible sur http://www.ordre-sages-femmes.fr/NET/img/upload/1/666_REFERENTIELSAGES-FEMMES2010.pdf
2. 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.
3. Charlin B, Brailovsky CA, Leduc C, Blouin D. The Diagnostic Script Questionnaire: A new tool to assess a specific dimension of clinical competence. Adv Health Sci Educ Theory Pract 1998;20:51-8.
4. Demeester A. Evaluation du raisonnement clinique des étudiants sages-femmes par le test de concordance de script. Mémoire pour la maîtrise universitaire de pédagogie des sciences de la santé, Paris 13, 2004.
5. Charlin B, Boshuizen HP, Custers EJ, Feltovich PJ. Scripts and clinical reasoning. Med Educ 2007;41: 1178-84.
6. Nendaz M. Le raisonnement clinique : données issues de la recherche et implications pour l'enseignement. Pédagogie Médicale 2005;6:235-54.
7. Feltovich PJ, Barrow HS. Issues of generality in medical problem solving. In: Schmidt HG, De Volder ML. Tutorials in problem-based learning: A new direction in teaching the health professions. Assen: Van Gorcum, 1984,128-142.
8. Raynal F., Rieunier A. Schème. In : Raynal F & Rieunier A. Pédagogie : dictionnaire des concepts clé (5ème éd.). Paris : ESF édition, 2005, 332-3.
9. Fournier J-P, Demeester A, Charlin B. Script Concordance Tests : Guidelines for construction. BMC Med Inform Decis Mak 2008;8:18.

10. Gagnon R, Charlin B, Coletti M, Sauvé E, van der Vleuten C. Assessment in context of incertainty: How many members are needed on the panel of reference of a script concordance test. Med Educ 2005;39:284-91.

11. Charlin B. Le test de concordance de script, un instrument d'évaluation du raisonnement clinique. Pédagogie Médicale 2002;3:135-44.

12. Sibert L, Charlin B, Corcos J, Gagnon R, Grise P, van der Vleuten C. Stability of clinical reasoning assessment results with the script concordance test across two different linguistic, cultural and learning environments. Med Teacher 2002;24:522-7.

13. Demeester A, Pauly V, Gagnayre R. Le Test de Concordance de Script en formation initiale sage-femme : intérêt, acceptabilité du test et perspectives de développement. Communication orale. 1er Congrès de la Société internationale francophone d'éducation médicale. Beyrouth, 2006.

14. Materissian S, Zabolotny B, Gagnon R. Is the Script Concordance Test a valid instrument for assessment of intraoperative decision-making skills? Am J Surg 2007;193:248-51.

15. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations.Acad Med 2002;77: 156-61.

16. Caire F, Sol JC, Charlin B, Isidori P, Moreau JJ. Le test de concordance de script (TCS) comme outil d'évaluation formative des internes en neurochirurgie : implantation du test sur Internet à l'échelle nationale. Pédagogie Médicale 2004;5:87-94.

17. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: Insights form a systematic review. Med Educ 2012;46:552–63.

18. Gagnon R, Lubarsky S, Lambert C, Charlin C. Optimization of answer key for script concordance testing: should we exclude deviant panelists, deviant responses, or neither. Adv Health Sci Educ Theory Pract 2011;16:601-8.

19. Demeester A. Mode raisonnement des sages-femmes en situation clinique. Mémoire pour le master de recherche en sciences de l'éducation : Université d'Aix-Marseille, 2005.

Correspondance et offprints : Madeleine Gantelet, École de sages-femmes de Besançon CHRU de Besançon, 2 Place St Jacques, 25000 Besançon, France.
Mailto : mgantelet@chu-besancon.fr