



**HAL**  
open science

## Des tests de concordance de scripts et examens cliniques à objectifs structurés aux serious games

François Carbonnel, Bernard Clary, Jacques Rambaud, David Costa

### ► To cite this version:

François Carbonnel, Bernard Clary, Jacques Rambaud, David Costa. Des tests de concordance de scripts et examens cliniques à objectifs structurés aux serious games. *Exercer*, 2013, 24 (105), pp.32-6. hal-01219646

**HAL Id: hal-01219646**

**<https://hal.science/hal-01219646>**

Submitted on 23 Oct 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



# Des tests de concordance de scripts et examens cliniques à objectifs structurés aux *serious games*

## *From script concordance testing and objective structured clinical examination to serious games*

François Carbonnel, Bernard Clary, Jacques Rambaud, David Costa

exercer 2013;105:32-6.

francois.carbonnel@gmail.com

*Current normative or formative assessments of medical students tend to focus more on skills rather than knowledge. The script concordance test has good psychometric qualities to assess reasoning skills required to resolve complex problems in situations of uncertainty. Objective structured clinical examinations assess clinical competence in realistic situations (clinical situation, standardized and simulated patients, neutral observers, observation check lists) however their organization is complex and costly. The introduction of human factors (panel of experts, simulated and standardized patients and observers) reduces the objectivity and reliability of both these tests. The generation Y (births between 1980 and 1992) has grown up in the digital age with widespread access to technology and as such expect new and innovative teaching aids. Serious games, developed in the health education field, could be used as a student assessment tool. This requires the convergence between medical science and information technology. Moreover, its implementation should meet the needs of the educational spiral (teaching needs, method, objectives and assessment). The move towards digital assessment could open the path towards a new way of evaluating skills in medicine.*

Département de médecine générale –  
UFR Montpellier-1

### Mots-clés

Tests de concordance de scripts
Examens cliniques à objectifs structurés
Serious games
Évaluation
Compétence

### Key words

Script concordance testing
Objective structured clinical examination
Serious games
Evaluation
Skills

Les auteurs déclarent n'avoir aucun conflit d'intérêts concernant les données publiées dans cet article.

## Introduction

Au cours des années 1960, l'influence des sciences de l'éducation et de la sociologie a entraîné une évolution progressive de la pédagogie médicale, qui est passée d'une approche fondée sur les contenus à transmettre à une approche par objectifs. L'idée principale était de clarifier les intentions et de spécifier les connaissances désirées<sup>1</sup>. Cependant, cette approche a parfois abouti à la création de catalogues d'objectifs, au risque de morceler les apprentissages au détriment de la création de représentations plus globales et signifiantes pour les étudiants<sup>2,3</sup>. À partir des années 1980, cette

prise de conscience a fait émerger un nouveau modèle dit de pédagogie par compétence. Est ici entendu par « compétence » un savoir en acte ou des séquences d'action, une « *intégration des habiletés, des connaissances et des capacités nécessaires à l'accomplissement d'une tâche spécifique* »<sup>3,4</sup>. La base de l'enseignement devient ainsi non plus un simple savoir-faire mais plutôt un authentique savoir-agir en situation. Le *curriculum* médical de nombreuses universités, en particulier anglo-saxonnes, a été modifié pour se fonder sur cette approche. Des dispositifs de formation privilégiant la confrontation des étudiants à des tâches en lien avec des situations-problèmes les

plus authentiques possible ont été mis en place. Dans cette perspective, les problèmes dits simples et ceux dits complexes sont différenciés. Deux outils sont particulièrement utilisés pour étudier ce dernier type de situations : les tests de concordance de scripts (TCS) et les examens cliniques à objectifs structurés (ECOS).

Les points forts et les limites de ces méthodes seront successivement présentés, éclairés par une recherche bibliographique. Enfin, un moyen d'évaluation et d'enseignement récent, les *serious games*, sera présenté, et son intérêt dans l'approche par compétences en pédagogie médicale sera discuté.

## Les tests de concordance de scripts

Le TCS est une méthode d'évaluation développée pour être utilisée en éducation médicale<sup>5</sup>. Les thèses de Donald Schön différencient connaissances « pures », enseignées, et savoir professionnel, utilisé en situation<sup>6</sup>. Pour que les examens puissent discriminer selon la compétence, il faudrait qu'ils contiennent des tâches de résolution de problèmes en situation d'incertitude<sup>7</sup>. La théorie des scripts postule qu'en situation clinique les médecins mobilisent leurs réseaux de connaissance préétablis, dits scripts. Ces scripts sont utilisés pour comprendre la situation et pour agir en visant des objectifs diagnostiques, d'investigation ou de traitement<sup>8</sup>. L'objectif du TCS n'est pas d'évaluer les connaissances mais plutôt leur degré d'organisation, et d'apprécier si elles sont organisées pour agir efficacement dans le contexte clinique<sup>7</sup>.

Un TCS comprend une série de vignettes. Chaque vignette se compose d'un cas clinique restreint, d'une hypothèse initiale, suivie d'une information supplémentaire, d'une question et d'une grille de réponse de type échelle de Likert (figure).

Les vignettes sont de trois types selon qu'elles portent sur une problématique de diagnostic, d'investigation ou de thérapeutique. Le score obtenu à chaque vignette est d'autant plus élevé que le degré de similitude avec les réponses non univoques provenant d'un panel d'experts est important. Le score total correspond à une sommation des scores des vignettes. Le TCS est réalisable sur support papier ou numérique, avec analyse informatisée<sup>9</sup>.

La validité du TCS en tant que test psychométrique, c'est-à-dire le degré avec lequel il mesure ce qu'il est censé mesurer, a été largement évaluée. Dans la littérature, les médecins expérimentés obtiennent généralement des scores supérieurs à ceux des internes, eux-mêmes supérieurs à ceux des externes<sup>7</sup>. Une étude récente a montré que les TCS permettaient de discriminer les capacités au raisonnement clinique d'étudiants n'ayant pas encore effectué de stage clinique, d'étudiants en ayant fait, et différenciaient également

**Un item :** correspond à la façon dont le clinicien mobilise ses connaissances pour traiter et résoudre le problème

**Une vignette :** une situation clinique problématique même pour un expert, mais du niveau d'étudiants à évaluer

**La nouvelle donnée** permet de tester la force qui l'unit à l'option, dans le contexte précis de la vignette

**Une option :** c'est l'option pertinente à la situation clinique présentée (avis d'experts)

**La méthode de correction** permet de mesurer et de comparer chacun des jugements fait par le candidat avec celui d'un panel de cliniciens expérimentés

**Les ancrages de l'échelle de Likert** correspondent à la réalité du raisonnement clinique : une seule donnée permet rarement d'affirmer qu'il ne peut s'agir que de diagnostic, mais mis dans une vignette, une donnée peut rendre pratiquement certain ce diagnostic

**Scénario clinique :**  
Une jeune femme de 22 ans, accompagnée de son mari, consulte pour des métrorragies abondantes faites de sang rouge et de douleurs abdominales importantes. Elle ne prend aucune contraception depuis 6 mois car elle désire un enfant. Elle déclare avoir eu ses règles il y a 6 semaines.

**Item 1**

**Si vous pensez à :** une grossesse extra-utérine

**... et que vous trouvez à l'échographie :** une image évoquant un caillot intra-utérin

**L'effet de cette nouvelle donnée sur votre hypothèse diagnostique sera (cocher votre réponse) :**

-2 Elle rend l'hypothèse beaucoup moins probable

-1 Elle rend l'hypothèse moins probable

0 La nouvelle donnée n'aura aucun effet sur l'hypothèse diagnostique

+1 Elle rend l'hypothèse plus probable

+2 Elle rend l'hypothèse beaucoup plus probable

Figure. Exemple de vignette de diagnostic (d'après l'Université de Montréal)

les étudiants des cliniciens experts<sup>10</sup>. La fiabilité du test a été démontrée via le coefficient d'homogénéité alpha de Cronbach. Pour atteindre systématiquement une valeur de 0,80 (caractérisant un test fiable et fidèle), il faut proposer 60 vignettes, soit une heure de test au moins<sup>11</sup>. L'instrument garde ses capacités de discrimination lorsqu'il est utilisé dans une autre culture, mais avec une moins bonne concordance que dans la culture d'origine<sup>11</sup>.

La comparaison entre un examen par QCM et un TCS a montré que les TCS prédisaient bien les résultats aux tests de raisonnement clinique mais moins bien les résultats des tests évaluant les habiletés cliniques, pour lesquels les ECOS étaient supérieurs<sup>12</sup>.

Par ailleurs, cette technique a certaines limites. Le choix et l'organisation du panel d'experts sont un point crucial pour la qualité des TCS. Par exemple, la pertinence de donner des scores à des étudiants à partir de raisonnements d'experts qui sont placés dans un contexte différent de celui dans lequel sont les

étudiants lors de l'examen est discutable<sup>13</sup>. La concordance entre les réponses données par les experts selon qu'elles étaient considérées individuellement et agrégées secondairement ou déterminées par consensus de groupe un an plus tard a été évaluée. La conclusion était que dans 59 % des cas, les membres du panel donnaient des réponses différentes dans les deux contextes. La variabilité des réponses au sein du panel d'experts est importante pour la validité du test<sup>13</sup>. Cependant, des réponses potentiellement « déviantes » pourraient induire une diminution de la fiabilité des scores en introduisant une erreur de mesure. Cette erreur de mesure est cependant négligeable si le panel est suffisamment large (plus de 15)<sup>14</sup>.

La capacité du TCS à détecter un mauvais raisonnement clinique, donc les praticiens faiblement performants, a également été évaluée<sup>15</sup>. Les résultats à un TCS ont été comparés à un examen standardisé évaluant les compétences cliniques durant une journée complète, le *Structured Oral Interview*. Près de la moitié des partici-



pants étaient classés différemment selon les méthodes. Le TCS doit donc être interprété avec précaution, car il ne reflète qu'une partie du processus de raisonnement clinique.

Enfin, l'impact de la méthode des TCS sur l'enseignement et l'apprentissage est insuffisamment étudié, de même que sa place dans des programmes d'enseignement et d'évaluation<sup>16</sup>.

## Les examens cliniques à objectifs structurés

L'ECOS est un outil d'évaluation clinique décrit en 1975<sup>17</sup>. C'est un examen à stations multiples. Chaque station propose un patient réel ou simulé, standardisé et évalue les habiletés, les attitudes ou les aptitudes, ainsi que les aspects cognitifs d'un candidat. Le patient simulé apprend un scénario par cœur, expose son problème de santé et répond aux questions conformément au scénario. Le patient simulé ne doit pas avoir d'antipathie ou de sympathie trop marquée envers les médecins. Il ne doit pas être professionnel de santé, ni acteur professionnel, ni être atteint de la pathologie concernée. Les étudiants sont évalués dans chaque station par des observateurs, qui disposent d'une grille d'évaluation prédéfinie de type check-list ou d'une échelle d'évaluation globale de type Lickert. La pondération n'est pas connue des observateurs. Ce test est de plus en plus utilisé en formation médicale initiale et vise à réduire les biais dans l'évaluation des compétences cliniques. Une revue de la littérature a montré que sa fiabilité, sa validité, son objectivité et sa faisabilité dépendent du nombre de stations, de la construction des stations, de la méthode de notation et du nombre d'étudiants évalués<sup>18</sup>. Son principal inconvénient est sa faisabilité. Sa mise en place et sa réalisation sont chronophages et coûteuses (de 200 à plus de 1 000 \$ selon des études) en raison notamment des besoins en évaluateurs, en patients standardisés et en locaux adaptés.

Une méta-analyse récente a démontré que les ECOS sont une méthode bien standardisée et relativement objective, mais qu'ils ne garantissent pas des sco-

res fiables et des décisions adaptées des étudiants en médecine (coefficient alpha de Cronbach régulièrement < 0,60)<sup>19</sup>. Il semblerait que plus l'ECOS a de stations, plus les résultats sont fiables et que deux examinateurs soient préférables.

En termes d'objectivité des ECOS, il a été prouvé que le test pouvait être fortement biaisé par une familiarité entre l'étudiant et l'examineur<sup>20</sup>. Une familiarité « positive » entre eux est corrélée à une amélioration significative des résultats, comparable à la différence entre des étudiants de début et de fin de cursus. Ces résultats devraient être pris en compte dans l'organisation des ECOS. L'objectivité de l'évaluation est également influencée par l'ordre de rotation des étudiants selon les stations. Par exemple, des étudiants ayant participé à une station sur la gynécologie-obstétrique tendent à rechercher des notions en lien avec cette discipline quand ils participent ensuite à une station sur la psychiatrie<sup>21</sup>. Les interactions entre le genre de l'étudiant et celui du patient standardisé induisent également un biais, en particulier pour les stations nécessitant un examen de la poitrine : les femmes ont des résultats significativement supérieurs à ceux des hommes avec des patientes standardisées (mais il n'y a pas de différence avec des patients de sexe masculin)<sup>22</sup>.

Enfin, à titre expérimental, l'utilisation d'un patient virtuel dans une station ECOS, composé d'une simulation sur ordinateur et d'un assistant humain, a été évaluée. Ce patient virtuel est capable de différencier les performances de l'étudiant, mais il induit un facteur confondant lié à l'influence de l'assistant humain<sup>23</sup>.

En résumé, le principal avantage des ECOS, qui est d'évaluer les compétences cliniques d'un étudiant dans un contexte plus réaliste grâce à l'insertion d'un facteur humain, est aussi sa principale faiblesse, car il induit des biais en termes d'objectivité, de fiabilité et aussi de faisabilité.

## Les serious games

Selon Julian Alvares, les *serious games* ou jeux sérieux correspondent à « une application informatique dont l'intention initiale est de combiner, avec cohérence,

à la fois des aspects sérieux tels que [...] l'enseignement, l'apprentissage, la communication, ou encore l'information avec des ressorts ludiques issus du jeu vidéo. Une telle association, qui s'opère par l'implémentation d'un scénario pédagogique, qui sur le plan informatique correspond à implémenter un habillage (sonore et graphique), une histoire et des règles idoines, a donc pour but de s'écarter du simple divertissement. Cet écart semble indexé sur la prégnance du scénario pédagogique, dont les objectifs sur le plan formel se confondent avec ceux du jeu vidéo. Le scénario pédagogique correspond alors à "une fonction" dont l'intention s'apparente soit à l'apprentissage d'une connaissance, soit à celle d'une pratique, ou bien des deux à la fois, et dont la propriété est de susciter l'envie d'apprendre et dont la réalisation dépend d'un jeu vidéo avec lequel elle puisse s'intégrer »<sup>24</sup>. La notion d'apprentissage d'un savoir en action, c'est-à-dire d'une compétence, est ici affirmée. La notion de « jeu sérieux » serait héritée des humanistes de la Renaissance italienne et française, qui liaient l'apprentissage de notions sérieuses à l'amusement. Dans son acception actuelle d'application informatique, ce type d'outil daterait de 2002, avec la sortie du jeu « *America's Army* » développé pour le compte de l'armée américaine en tant que moyen de recrutement<sup>25</sup>. C'est le premier *serious game* à avoir rencontré le succès auprès du grand public. Ce type de technique pédagogique a ensuite été développé dans de nombreuses disciplines, en particulier l'aéronautique pour la formation des pilotes.

Des *serious games* se sont ainsi développés dans différents champs de la santé, principalement dans le champ de l'éducation à la santé, incluant parfois des évaluations de la progression des utilisateurs. Ils sont principalement utilisés dans l'éducation à la santé et dans l'éducation thérapeutique, par exemple chez des patients atteints de maladie chronique. Globalement, leurs résultats semblent encourageants en tant que technique pédagogique, mais des études bien menées sur leur validité, leur fiabilité, leur objectivité et leur faisabilité manquent pour apprécier leur intérêt en



# GENERATION



© Thomas Pajot – Fotolia.com

tant qu'outil d'évaluation formative et/ou normative<sup>26-28</sup>. Le *serious game* est donc initialement un outil pédagogique dont la construction dépend de la méthode utilisée (magistrale, interrogative, active ou découverte). Il peut également être envisagé comme un outil d'évaluation formative ou normative, et donc être là aussi construit spécifiquement.

Une étude évaluant l'acceptabilité des jeux vidéo chez des étudiants en médecine nord-américains et les nouvelles technologies en éducation médicale a révélé que 98 % d'entre eux appréciaient l'idée d'utiliser ce type de technologie, et que 96 % estimaient que l'éducation devrait faire meilleur usage des nouvelles technologies<sup>29</sup>. 80 % d'entre eux pensaient également que les jeux vidéo pouvaient avoir une valeur éducative. Cette étude rejoint les conclusions d'un travail sur les évolutions en enseignement médical rendues nécessaires par l'arrivée de la génération Y. Ce terme désigne les jeunes gens nés entre 1980 et 1992 qui ont grandi avec les nouvelles technologies numériques, en particulier Internet<sup>30</sup>. Ceux-ci attendent des innovations pédagogiques et demandent des formations individualisées et stimulantes. Il y a également une forte attente de la systématisation du *feedback*, c'est-à-dire d'un retour d'évaluation sur leurs connaissances et leurs compétences<sup>30</sup>.

Dans cette perspective, un *serious game* permettant d'évaluer des compétences

cliniques, diagnostiques et thérapeutiques pourrait être envisagé. Il permettrait éventuellement de compenser certaines limites des TCS et des ECOS. En termes d'objectivité et de fiabilité, une notation automatisée pourrait limiter les biais induits par les examinateurs humains ou les panelistes « déviants » ou en trop faible nombre. Les conditions de réalisation des tests par les examinateurs et les étudiants pourraient elles aussi être standardisées et ainsi s'affranchir de contingences matérielles ou organisationnelles.

En termes de faisabilité, ce type de test nécessiterait un investissement financier initial important. Sa création obligerait à établir un scénario pédagogique s'appuyant sur un design pertinent et un *gameplay* intéressant pour les étudiants. Le contenu du test serait également important et pourrait initialement s'orienter vers des disciplines médicales plus adaptées à un design ludique telles que la dermatologie ou l'otorhinolaryngologie. Les besoins en programmation devraient faire intervenir une équipe motivée par un travail expérimental organisé en étroite collaboration avec l'équipe pédagogique. Il serait également souhaitable que des membres de ces équipes ayant des compétences avancées dans l'autre discipline fassent le lien entre les deux groupes afin d'optimiser le travail collaboratif.

Enfin, la validité de ce type de test serait à évaluer. Il est probable qu'une diminution de la capacité à déterminer les compétences cliniques des étudiants par ce test apparaisse, ce type d'outil faisant disparaître l'examen clinique « en vrai ». Cependant, des séquences de simulation pourraient être envisagées avec des mannequins pour évaluer certains gestes techniques. Ces séquences pourraient être combinées avec le *serious game* en étant intégrées au scénario pédagogique initial. Il y aurait probablement une perte de liberté d'action pour l'étudiant car la structure du jeu organisera la séquence selon le scénario pédagogique prédéfini. Il est cependant possible d'envisager des situations où le joueur fera différents choix en termes de scripts ou d'action qui orienteront le scénario dans différents sens, afin de recréer une sensation de liberté.

## Conclusion

Les TCS ont montré leur intérêt pour évaluer le raisonnement clinique en situation complexe et la gestion de l'incertitude. Ils n'étudient cependant pas tous les aspects du raisonnement médical et ont des résultats inférieurs aux ECOS pour les habiletés cliniques. Ces derniers nécessitent un investissement en moyens humains, matériels et financiers conséquents qui peut être un frein pour l'université. Ces deux tests ont des limites, en particulier en termes d'objectivité et de fiabilité, du fait de l'introduction de facteurs humains.

Les *serious games* médicaux commencent à émerger en tant qu'outils éducatifs pour les patients. Ils pourraient être utilisés pour l'évaluation des compétences des professionnels de santé en formation.

À ce jour et à notre connaissance, il n'existe pas de *serious game* utilisé comme outil pédagogique et d'évaluation des compétences des professionnels de santé. Un tel projet nécessiterait la coopération d'équipes de programmation et d'équipes médicales avancées dans le domaine de la pédagogie pour écrire un scénario adapté aux besoins des étudiants, tant en évaluation formative que normative.

Il nécessiterait aussi un investissement financier plus conséquent que les traditionnels moyens d'évaluation précédemment décrits, mais permettrait de disposer d'un outil ayant un contenu bien plus riche, adaptable aux niveaux des enseignements et s'interfaçant avec des outils de simulation.

L'acceptabilité de ce type d'évaluation par la génération Y qui a grandi avec les technologies des nouveaux médias et d'Internet a déjà été démontrée. Sa mise en place serait d'un intérêt majeur pour les universités en standardisant les différentes contraintes d'une évaluation et en ouvrant de nouvelles voies de recherche en pédagogie. Basculer sur un modèle d'évaluation numérique de ce type modifierait la façon d'enseigner la médecine et d'évaluer les compétences, et répondrait aux attentes nouvelles des étudiants.



## Résumé

*L'évaluation normative et formative actuelle des étudiants en médecine tend à se centrer sur les compétences plutôt que sur les connaissances. Le test de concordance de scripts a de bonnes qualités psychométriques pour évaluer le raisonnement devant des problèmes complexes en situation d'incertitude. L'examen clinique à objectifs structurés (ECOS) évalue les habiletés cliniques en situation réaliste (situation clinique, patient simulé et standardisé, observateurs neutres, grille d'observation). Son organisation est complexe et coûteuse. L'introduction de facteurs humains (panel d'experts, patients simulés et observateurs) diminue l'objectivité et la fiabilité de ces deux tests. La génération Y (personnes nées entre 1980 et 1992), qui a grandi avec les technologies numériques, est en attente de nouveaux outils pédagogiques. Les serious games développés dans le champ de l'éducation pour la santé pourraient être utilisés dans le champ de l'évaluation. Ils nécessitent une convergence entre les sciences informatiques et médicales et ont une exigence pédagogique avec des besoins de formation, de méthodes, d'objectifs, et d'évaluation. La bascule vers l'évaluation numérique pourrait ouvrir la voie vers une nouvelle façon d'évaluer les compétences en médecine.*

## Références

1. Pugh EW, Lloyd GJ, McIntyre N. Relevance of educational objectives for medical education. *BMJ* 1975;3:688-91.
2. Nguyen DQ, Blais J-G. Approche par objectifs ou approche par compétences ? Repères conceptuels et implications pour les activités d'enseignement, d'apprentissage et d'évaluation au cours de la formation clinique. *Pédagogie médicale* 2007;8:232-51.
3. Vorhees RA. Competency based learning models: a necessary future. *New Directions for Institutional Research* 2001;110:5-13.
4. Tardif J. L'évaluation des compétences. Documenter le parcours de développement. Montréal : Chenelière Éducation, 2006.
5. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CPM. Script concordance testing: a review of published validity evidence. *Med Educ* 2011;45:329-38.
6. Schön DA. *The reflective practitioner: how professionals think in action*, New York: Basic books, 1983.
7. 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.
8. Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Acad Med* 2000;75:182-90.
9. Kania RE, Verillaud B, Tran H, et al. Online script concordance test for clinical reasoning assessment in otorhinolaryngology: the association between performance and clinical experience. *Arch. Otolaryngol. Head Neck Surg* 2011;137:751-5.
10. Humbert AJ, Johnson MT, Miech E, Friedberg F, Grackin JA, Seidman PA. Assessment of clinical reasoning: a script concordance test designed for pre-clinical medical students. *Med Teach* 2011;33:472-7.
11. 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 from different countries. *Eur Urol* 2002;41:227-33.
12. 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 experimental study on the script concordance test. *Med Educ* 2001;35:430-6.
13. Charlin B, Gagnon R, Pelletier J, et al. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Med Educ* 2006;40:848-54.
14. Gagnon R, Lubarsky S, Lambert C, Charlin B. Optimization of answer keys for script concordance testing: should we exclude deviant panelists, deviant responses, or neither? *Adv Health Sci Educ Theory Pract* 2011;16:601-8.
15. Goulet F, Jacques A, Gagnon R, Charlin B, Shabah A. Poorly performing physicians: does the script concordance test detect bad clinical reasoning? *J Contin Educ Health Prof* 2010;30:161-6.
16. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. *Med Educ* 2012;46:552-63.
17. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *Br Med J* 1975;1:447-51.
18. Barman A. Critiques on the objective structured clinical examination. *Ann Acad Med Singapore* 2005;34:478-82.
19. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ* 2011;45:1181-9.
20. Stroud L, Herold J, Tomlinson G, Cavalcanti RB. Who you know or what you know? Effect of examiner familiarity with residents on OSCE scores. *Acad Med* 2011;86(Suppl):S8-11.
21. Blaskiewicz RJ, Park RS, Chibnall JT, Powell JK. The influence of testing context and clinical rotation order on students' OSCE performance. *Acad Med* 2004;79:597-601.
22. Carson JA, Peets A, Grant V, McLaughlin K. The effect of gender interactions on students' physical examination ratings in objective structured clinical examination stations. *Acad Med* 2010;85:1772-6.
23. Courteille O, Bergin R, Stockeld D, Ponzer S, Fors U. The use of a virtual patient case in an OSCE-based exam--a pilot study. *Med Teach* 2008;30:e66-76.
24. Alvarez J. Du jeu vidéo au serious game : approches culturelle, pragmatique et formelle. Thèse en science de la communication et de l'information. Toulouse 2 et 3, 2007.
25. America's army. Disponible sur [www.americasarmy.com](http://www.americasarmy.com).
26. Lewis MW. Analysis of the roles of « serious games » in helping teach health-related knowledge and skills and in changing behavior. *J Diabetes Sci Technol* 2007;1:918-20.
27. Thompson D, Baranowski T, Buday R, et al. In pursuit of change: youth response to intensive goal setting embedded in a serious video game. *J Diabetes Sci Technol* 2007;1:907-17.
28. Thompson D, Baranowski T, Buday R. Conceptual model for the design of a serious video game promoting self-management among youth with type 1 diabetes. *J Diabetes Sci Technol* 2010;4:744-9.
29. Kron FW, Gjerde CL, Sen A, Fetzters MD. Medical student attitudes toward video games and related new media technologies in medical education. *BMC Med Educ* 2010;10:50.
30. Boulé F. Hautement différente : la génération Y, un défi de taille pour l'enseignement médical. *Pédagogie médicale* 2012;13:9-25.