Adv in Health Sci Educ (2013) 18:779–785
DOI 10.1007/s10459-012-9427-3

Script concordance testing in continuing professional development: local or international reference panels?

E. M. Pleguezuelos · E. Hornos · V. Dory · R. Gagnon · P. Malagrino · C. A. Brailovsky · B. Charlin

Received: 4 July 2012 / Accepted: 15 November 2012 / Published online: 29 November 2012

Abstract *Context* The PRACTICUM Institute has developed large-scale international programs of on-line continuing professional development (CPD) based on self-testing and feedback using the Practicum Script Concordance Test© (PSCT). *Aims* To examine the psychometric consequences of pooling the responses of panelists from different countries (composite panels) and the effect of increasing composite panel size. *Method* 97 cardiologists in Mexico answered a set of 62 PSCT cases/305 questions. A local panel was recruited in Mexico (n = 7). Other panelists were recruited in Argentina (n = 10) and Brazil (n = 11). Together they constituted a composite panel of 28 experts. Random panels of reference of increasing sizes (5, 10, 15, 20, and 25) were generated. Participants' scores were computed for each panel sample. Units of analysis were means of participants' scores per case. Discrimination, ranking and reliability of the scores obtained with each panel were estimated. Descriptive statistics, Pearson correlation coefficient, generalizability analysis, computation of Cronbach's alpha were used in the analyses. *Results* Correlation coefficients between the local and the composite panels ranged from 0.951 to 0.981. Cronbach alpha coefficient values were above 0.85 for all panels. The value of the relative G coefficient from the generalizability analyses varied from 0.91 to 0.93, indicating very high and stable ranking of participants, though absolute value of scores increased with increasing composite panel size. *Conclusions* In CPD environments, and with panel members selected with the highest standards, composite panels can be used.

E. M. Pleguezuelos · E. Hornos · P. Malagrino · C. A. Brailovsky
PRACTICUM Institute of Applied Research in Health Sciences Education, Madrid, Spain

V. Dory
Fonds de la Recherche Scientifique, FNRS and Institute of Health and Society,
Université catholique de Louvain, Brussels, Belgium

R. Gagnon · B. Charlin (✉)
CPASS (Centre de Pédagogie Appliquée aux Sciences de la Santé), Faculty of Medicine,
University of Montreal, CP 6128 Succursale centre-ville, Montreal, QC H3C 3J7, Canada
e-mail: bernard.charlin@umontreal.ca

C. A. Brailovsky
College of Family Physicians of Canada, Montreal, Canada

Panels of all sizes yielded high psychometric qualities. Absolute scores should be interpreted with caution.

Keywords Aggregate scoring · Assessment · Clinical reasoning · Continuing professional development · On-line learning · Practicum script concordance test · Panel of reference · Reliability

Introduction

The Script Concordance Test (SCT) is a test format designed to measure one component of clinical reasoning, i.e. clinical data interpretation (Charlin and van der Vleuten 2004). It was developed according to script theory(Charlin et al. 2007) to capture the reasoning involved in solving ill-defined cases where experiential knowledge is critical (Charlin et al. 1998). The SCT attempts to do so both through its original format and its aggregate scoring method. The SCT's format presents examinees with realistic clinical vignettes, which are purposefully brief. The vignette is ill-defined in that it does not provide sufficient information for a straightforward answer to be given, but rather lends itself to the generation of several hypotheses regarding diagnosis, investigations, or treatment. A few items then follow each vignette, each containing a plausible hypothesis and a new piece of information. Examinees are asked to evaluate the impact of the new datum on the likelihood of the hypothesis. As such, it aims to simulate hypothetico-deductive reasoning in real patient encounters (Charlin et al. 1998). The SCT's other key feature is the use of aggregate scoring. Aggregate scoring refers to methods that compare examinees' responses to those of a reference panel and give credit according to the frequency of the response within the panel (Norman 1985). In other words, examinees that select the same responses as most panel members obtain the highest scores. Aggregate scoring provides a credible way of marking responses in authentic situations, where there is no clear-cut right answer but where different responses can be more or less appropriate. As such, the SCT avoids one of the pitfalls of traditional multiple-choice questions, i.e. the exclusion of controversial areas in medicine (Elstein 1993).

The PRACTICUM Institute of Applied Research in Health Sciences Education is a Spanish non-profit institution (www.practicumfoundation.eu) that organizes on-line continuing professional development (CPD) programs, in partnership with scientific institutions of different disciplines. Its programs aim to develop physicians' reflective abilities on challenging issues within the disciplines (www.script.edu.es). They are built around the Practicum Script Concordance Test© (PSCT) used as a self-testing tool on clinical reasoning (Hornos et al. 2012). After each response, participants receive their score and can access the responses and justifications of panel members, which sometimes include references. As such, the main aim of using assessment with PSCT is formative, and participants are provided with opportunities to retake questions on which they perform poorly.

Being administered on the Internet, through a multilingual educational platform, Practicum Script Concordance Test© for CPD programs enable the provision of distance training for large communities (Hornos et al. 2012). Programs are already in use in several Spanish-speaking countries. For instance, in Mexico, 1,200 pediatricians (2010–2011) and 540 cardiologists (2011–2012) have already taken part in an annual training cycle in which they received a new case per day, 5 days per week, i.e. a total of 240 cases over the course of a year (Hornos et al. 2012). The potential for large-scale international rollouts has

provided the impetus for further research on aggregate scoring, specifically on the impact of panel composition on scores.

The composition of a panel is a key determinant of the validity of scores produced by aggregate scoring. Because clinical practice varies according to local contextual factors but also local epidemiological determinants, experts may very well provide responses that are appropriate in their own setting but would be less appropriate in different settings. SCT developers have therefore logically recommended that panel members be experts who work in the same type of setting as examinees' current or future setting (Fournier et al. 2008; Dory et al. 2012). Some research has examined the impact of panel composition on the results of SCT examinations. Two studies explored the role of type of expertise and found that it had little influence. One of these studies was conducted in midwifery in France and compared scores obtained with a panel of generalists with those obtained with a panel of subspecialists who provided responses only for items pertaining to their particular domain of specialization (Gantelet 2008). Scores did not differ significantly, nor did the psychometric properties of the test and individual items. Another in dermatology found that scores differed more depending on panelists' setting (i.e. hospital versus community) than specialty (i.e. dermatology versus family practice) (Bursztejn et al. 2011). Results from these two studies are consistent with a study on aggregate scoring of traditional multiple-choice and true–false tests in cardiology: level of specialization had little effect on the psychometric properties of the test (Norcini and Shea 1990). Sibert et al. examined the influence of country of practice (Sibert et al. 2002). They analyzed scores of residents in France and Canada generated by two panels, one comprised of experts from France and one of experts from Canada. Although scores were highly correlated, they were significantly higher when generated using the panel of experts from the same country. These studies suggest that the practice setting of panelists is more important than their type of expertise in selecting appropriate panelists in SCT.

Another important issue regarding the panel is panel size. To our knowledge, only one study has specifically explored this question with the SCT. In Gagnon et al.'s study (2005), residents in family medicine sat a SCT. A panel of 38 experts was used to generate scores, which produced a test with an internal consistency of 0.76. Gagnon et al. (2005) used a computer to generate random samples of smaller panels to analyze the impact of panel size on internal consistency findings. They found that smaller samples varied greatly in the resulting internal consistency of scores and that sample sizes of 10 or more (and better 15), provided dependable internal consistency. An unexpected finding was that scores tended to increase with panel size, presumably due to increased opportunities for partial credit for unusual responses.

Up until now Practicum Script Concordance Test programs have used different panels for each country, each composed of 10 experts and sometime fewer (e.g. 7 for the cardiology program in Mexico). While this is a smaller number than is generally recommended, selection criteria are stricter than in most SCT studies to date where board certification is the only formal inclusion criterion. In the Practicum setting, panelists must be board-certified general specialists (not sub-specialized in specific areas), with at least 8 years' experience as a specialist and 10 consecutive years' experience as a physician; they must spend at least 50 % of their working time being involved in patient care (at the time of recruitment and in the previous 5 years); and they must be involved in academic activities (teaching, participation in the organization of CPD activities, publications in the form of journal articles or monographs). These highly selective criteria, that are necessary because experts' answers provide feedback to practicing cardiologists, make recruitment difficult.

International collaboration (facilitated through an Internet platform) and the pooling of experts from several countries may facilitate the composition of large enough panels while maintaining the stringent recruitment requirements. The aims of this paper are to describe the psychometric consequences of (1) using a local panel versus an international panel, and (2) increasing the size of the panel.

Method

An item-writing committee developed a CPD program for cardiologists affiliated to cardiology societies across Latin America. The program was structured into four training sections. Over the course of each section, participants were sent one SCT case (with its 4–5 related questions) per working day. Cases were randomly allocated to avoid participants discussing the cases with colleagues prior to responding. The first section comprised 62 cases with 305 questions (4–5 questions per case). In preparation for administration in different countries, cases and questions were revised for wording and potential national specificities (e.g. medication names). To compute answer keys the material was submitted to experts recruited within three countries, i.e. Mexico, Argentina and Brazil. All experts, test writers as well as panel members, fulfilled the inclusion criteria specified above. Each national society set the rules for attribution of CPD credits. In Mexico for instance, the Mexican Council of Cardiology decided to grant credit if participants obtained a mean score of 60 out of 100 in each section. Participants were given the opportunity to retake cases that they had failed after a period of 5 weeks.

The present study concerns the 2011 PSCT administration in Mexico. Five hundred and fifty two Mexican cardiologists took part in the annual training program. At the time of data collection, 97 participants had responded to the first module. We included participants' initial responses only (i.e. not the responses from cases which they had decided to retake).

A local panel was recruited in Mexico ($n = 7$). Other panelists were recruited in Argentina ($n = 10$) and Brazil ($n = 11$). Together they constituted a composite panel of 28 experts. Random panels of increasing sizes (5, 10, 15, 20, and 25) were generated. Participants' scores were computed for each panel sample. Units of analysis were means of participants' scores per case. The goal of the analyses was to estimate the discrimination, ranking and reliability obtained with the local panel and the composite panels of increasing size. Descriptive statistics, Pearson correlation and Cronbach's alpha coefficients were computed. Generalizability studies were performed using the Edugen program (Cardinet et al. 2012). The facets for the analyses were participants and cases. Generalizability coefficients were performed using a persons × cases (P×C) design.

Results

Correlation coefficients between the local and the composite panels ranged from 0.951 to 0.981. Figure 1 illustrates the high correlation between scores using the Mexican panel and scores using the composite-28 panel. Descriptive statistics, presented in Table 1, show the increase of scores with increasing composite panel size. Cronbach's alpha coefficient value was above 0.85 with panels of all size. The results of generalizability analyses using scores obtained with the local and the composite-28 panel are presented in Tables 2 and 3. Table 4 summarizes data from the generalizabilty analyses performed on scores obtained

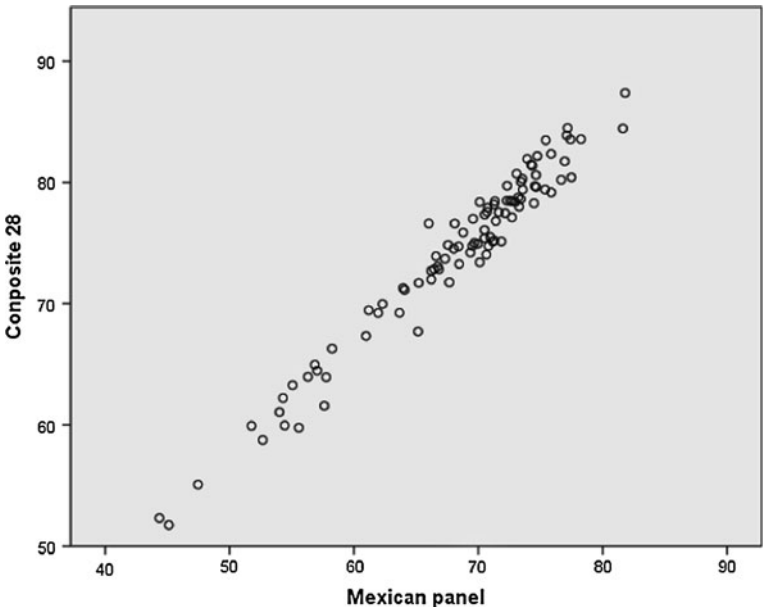

Fig. 1 Scatterplot of scores obtained with the Mexican panel and the composite-28 panel

Table 1 Descriptive statistics and values of Cronbach's alpha coefficients for the different panels

Panels	Mean	SD	Min	Max	Cronbach's α
Mexico 7	68.2	7.7	44.4	81.8	0.90
Composite 5	68.4	7.0	45.5	82.9	0.87
Composite 10	72.0	7.1	49.6	84.0	0.89
Composite 15	73.4	7.1	51.7	85.4	0.92
Composite 20	73.1	7.5	50.5	86.4	0.92
Composite 25	74.2	7.3	52.2	87.0	0.92
Composite 28	74.2	7.4	51.7	87.4	0.93

Table 2 ANOVA table: scores calculated with the panel composed of seven mexican experts

Analysis of variance						
Source	SS	*df*	MS	Random	%	SE
P	364,763.03	96	3,799.62	56.27	13.3	8.76
C	353,832.15	61	5,800.53	56.60	13.4	10.66
PC	1,820,117.89	5,856	310.81	310.81	73.4	5.74
Total	2,538,713.07	6,013			100	

Relative G coefficient 0.92

Absolute G coefficient 0.90

with all panels. Variance attributed to participants ranged from 12.5 to 15.3 %. Variance attributed to case difficulty ranged from 10.0 to 17.3 %. Variance attributed to the interaction between cases and participants ranged from 70.2 to 74.7 %. Relative G coefficients

Table 3 ANOVA table: scores calculated with the composite panel of 28 experts

Analysis of variance						
Source	SS	*df*	MS	Random	%	SE
P	326,354.77	96	3,399.53	50.88	15.3	7.83
C	212,088.36	61	3,476.86	33.29	10.0	6.39
PC	1,457,242.73	5,856	248.85	248.89	74.7	4.59
Total	1,995,685.86	6,013			100	

Relative G coefficient 0.93

Absolute G coefficient 0.92

Table 4 Summary of ANOVA results from the generalizability studies

Panels	P (%)	C (%)	PC (%)	Relative G coefficient	Absolute G coefficient
Mexican panel	13.3	13.4	73.4	0.92	0.90
Composite-5	10.4	20.3	69.3	0.90	0.88
Composite-10	12.5	17.3	70.2	0.91	0.89
Composite-15	13.3	13.6	73.1	0.92	0.91
Composite-20	13.8	13.4	72.8	0.92	0.91
Composite-25	14.8	11.9	73.2	0.93	0.92
Composite-28	15.3	10.0	74.7	0.93	0.92

were high (0.91–0.93) as were absolute G coefficients (0.88–0.92). Results using scores obtained from the various panels were remarkably similar.

Discussion

Data depict a test with strong psychometric qualities. The values of reliability coefficients were over 0.90, the test discriminated well between participants (percentage of variance related to participants) and its ranking ability was very stable (relative G coefficient over 0.91) (Cardinet et al. 2012). The length of the test (305 questions nested in 62 cases, the high standard selection criteria for panel inclusion, and the great care taken to construct test items (Hornos et al. 2012) may explain these test qualities.

Whichever panel was used to generate scores, participants' rankings were almost identical, as indicated by high correlations between scores. This indicates that in a continuing professional education, provided all the above-specified conditions are met, it is legitimate to pool panelists from different countries, assuming that epidemiological context and clinical practice do not differ significantly across the involved countries.

The study confirms Gagnon et al. (2005) 's finding that larger panels lead to higher absolute scores on the SCT (and presumably this would hold on other assessments using aggregate scoring). This has practical consequences for the interpretation and decisions made in regards to participants' scores. Setting a passing score was not a preoccupation in this particular context as participants had the opportunity to retake cases for which they failed, but it would be in other continuing professional development contexts. If the objective is to rank examinees according to their scores, panel size has no significant impact. If, on the other hand, the objective is to make decisions (e.g. grant credit) on the basis of the absolute value of scores, then panel size matters.

A potential limitation of the study is the rather small size of the local panel made up of only seven members. It was a concern at the beginning of the study as the literature recommends higher size for a panel, but the data here did not confirm this limitation as the local panel yielded excellent test data, as well as the composite-5 panel. This is probably related to the test quality factors mentioned above and other studies are needed to confirm the validity of the use of panel size smaller that generally recommended (Gagnon et al. 2005). The particular psychometric qualities of the test may limit the generalization of our results. It is in fact unusual to have such a test length and such quality control in test construction and panel selection.

Beside expanding knowledge about aggregate scoring (Norman 1985), this study results have implications for CPD practice. CPD activities are generally devised by local organizations who use experts to develop content and when relevant participant assessment. This tradition is partly based on the assumption that clinical practice differs significantly according to regional factors. Rather than having local reference panels, the alternative of using a single composite panel for activities in different countries, could contribute to the optimization of on-line CPD-SCT activities.

Acknowledgments The authors would like to thank Leandro Harillo for his help in collecting the data and preparing this manuscript.

References

Bursztejn, A. C., Cuny, J. F., Adam, J. L., Sido, L., Schmutz, J. L., de Korwin, J. D., et al. (2011). Usefulness of the script concordance test in dermatology. *Journal of the European Academy of Dermatology and Venereology, 25*, 1471–1475.

Cardinet, J., Johnson, S., & Pini, G. (2012). *Applying generalisability theory using EduG*. New York, London: Routledge, Taylor and Francis Group.

Charlin, B., & van der Vleuten, C. (2004). Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. *Evaluation in the Health Professions, 27*, 304–319.

Charlin, B., Brailovsky, C., Leduc, C., & Blouin, D. (1998). The Diagnosis Script Questionnaire: A new tool to assess a specific dimension of clinical competence. *Advances in Health Sciences Education, 3*, 51–58.

Charlin, B., Boshuizen, H. P., Custers, E. J., & Feltovich, P. J. (2007). Scripts and clinical reasoning. *Medical Education, 41*, 1178–1184.

Dory, V., Gagnon, R., Vanpee, D., & Charlin, B. (2012). How to construct and implement script concordance tests: insights from a systematic review. *Medical Education, 46*, 552–563.

Elstein, A. S. (1993). Beyond multiple-choice questions and essays: The need for a new way to assess clinical competence. *Academic Medicine, 68*, 244–249.

Fournier, J. P., Demeester, A., & Charlin, B. (2008). Script concordance tests: Guidelines for construction. *BMC Medical Informatic Decision Making*, *8*, 18, http://www.biomedcentral.com/1472-6947/8/18.

Gagnon, R., Charlin, B., Coletti, M., Sauve, E., & van der Vleuten, C. (2005). Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test?Medical. *Education, 39*, 284–291.

Gantelet, M. (2008). *Impact du panel de référence sur les résultats d'un Test de Concordance de Script (TCS) développé en formation initiale des sages-femmes*. Dissertation for the completion of a diploma of the Ecole des cadres sages-femmes. Dijon: Ecole de Cadres Sages-Femmes.

Hornos, E., Pleguezuelos, E. M., Brailovsky, C. A., Harillo, L. D., Dory, V., & Charlin, B. (2012). The Practicum Script Concordance Test: an online continuing professional development format to foster reflection on clinical practice. *Journal of Continuing Education in the Health Professions*, in press.

Norcini, J. J., & Shea, J. A. (1990). The effect of level of expertise on answer key development. *Academic Medecine, 65*(Suppl), S15–S16.

Norman, G. R. (1985). Objective measurement of clinical performance. *Medical Education, 19*, 43–47.

Sibert, L., Charlin, B., Corcos, J., Gagnon, R., Grise, P., & van der Vleuten, C. (2002). Stability of clinical reasoning assessment results with the Script Concordance test across two different linguistic, cultural and learning environments. *Medical Teacher, 24*, 522–527.