Adv in Health Sci Educ (2011) 16:601–608
DOI 10.1007/s10459-011-9279-2

Optimization of answer keys for script concordance testing: should we exclude deviant panelists, deviant responses, or neither?

Robert Gagnon · Stuart Lubarsky · Carole Lambert · Bernard Charlin

Received: 20 October 2010 / Accepted: 19 January 2011 / Published online: 1 February 2011

Abstract The Script Concordance Test (SCT) uses a panel-based, aggregate scoring method that aims to capture the variability of responses of experienced practitioners to particular clinical situations. The use of this type of scoring method is a key determinant of the tool's discriminatory power, but deviant answers could potentially diminish the reliability of scores by introducing measurement error. (1) to investigate the effects on SCT psychometrics of excluding from the test's scoring key either deviant panelists or deviant answers; (2) to propose a method for excluding either deviant panelists or deviant answers. Using an SCT in radiation oncology, we examined three methods for reducing panel response variability. One method ('outliers') entailed removing from the panel members with very low total scores. Two other methods ('distance-from-mode' and 'judgment-by-experts') excluded widely deviant responses to individual questions from the test's scoring key. We compared the effects of these methods on score reliability, correlations between original and adjusted scores, and between-group effect sizes (panel-residents; panel-students; and residents-students). With a large panel (n = 45), optimization methods have no effect on reliability of scores, correlation and effect size. With a smaller panel (n = 15) no significant effect of optimization methods were observed on reliability and correlation, but significant variation on effect size was observed across samples. Measurement error resulting from deviant panelist responses on SCTs is negligible, provided the panel size is sufficiently large (>15). However, if removal of deviant answers is judged necessary, the distance-from-mode strategy is recommended.

Keywords Clinical Reasoning · Script concordance test · Optimization · Reliability · Panel

R. Gagnon · C. Lambert · B. Charlin (✉)
CPASS, Faculty of Medicine, University of Montreal, Centre-Ville Station,
C.P. 6128, Montreal, QC H3C 3J7, Canada
e-mail: bernard.charlin@umontreal.ca

S. Lubarsky
Department of Neurology and Neurosurgery and Centre for Medical Education,
McGill University, Montreal, QC, Canada

Introduction

There exists considerable variation in the way that physicians practice medicine. Uncertainty shapes physicians' decisions and colors the lens through which they interpret medical information (Eddy 1984; Allman et al. 1985). One way physicians navigate the world of medical uncertainty is by bringing knowledge from their prior experiences in medicine to bear on present clinical ambiguities. These networks of knowledge, called "illness scripts," are mental structures organized to facilitate rapid mobilization and efficient use (Charlin et al. 2007). Illness scripts develop and become refined with experience (Schmidt et al. 1990). No two physicians possess an identical repertoire of scripts; hence, even experienced practitioners from a common discipline often interpret data, make judgments, and respond to uncertain clinical situations in ways that vary (within an acceptable range of medical practice).

The Script Concordance Test (SCT) is a tool used in medical education for assessing skill in interpreting clinical data under conditions of uncertainty (Charlin et al. 2004; Lubarsky et al. 2011). In contrast to many conventional forms of assessment, there are no single-correct-answers to SCT questions. Instead, several responses to each of the test's questions may be considered acceptable, as determined by a reference panel of experienced practitioners ('experts') from a given domain. The aim of this type of scoring system is to capture the variability of responses of experts to authentic clinical situations (Charlin et al. 2006).

The use of an aggregate scoring method has been shown to be a key determinant of the SCT's discriminatory power (Charlin et al. 2006). However, by awarding examinees partial credit for any answers selected by members of a reference panel, the SCT scoring system also has the potential to introduce unreliability into the measurement. The variability produced by deviant panel responders could conceivably affect the development of SCTs yielding reliable and valid measures (Lubarsky et al. 2011). Furthermore, rewarding atypical or incorrect answers in a scoring key is not a commonly accepted practice in assessment. The present study therefore aims (1) to investigate the effects on SCT psychometrics of excluding from the test's scoring key either deviant panelists or deviant answers; (2) to propose a method for excluding either deviant panelists or deviant answers.

Methods

Script concordance test

SCTs feature a series of short clinical cases, each followed by a set of test questions consisting of three columns (Fournier et al. 2008). For each question, the first column ("If you were thinking of…") provides a diagnostic, investigative, or therapeutic hypothesis relevant to the given case. The second column ("…and then you find…") presents new information, such as a physical examination finding, an imaging study, or a laboratory test result, that may (or may not) have an effect on the initial hypothesis. The question is answered in the third column ("…this hypothesis becomes:"), which contains a five-point Likert-type scale (from -2 to $+2$). Examinees are asked to indicate on this scale the effect they think the new information (part 2) is likely to have on the proposed hypothesis (part 1). SCT scores are presumed to reflect the degree of concordance between examinees' and experts' data interpretation skills under conditions of clinical uncertainty. Table 1 illustrates the test format.

Table 1 Test format for each case and questions. *Clinical case*: A 60 years old patient is referred to you for the treatment of a T1c N0 M0 prostate cancer with a Gleason score of 6/10 and PSA value of 9.0 ng/ml

If you considered recommending…	And that you find out	Your recommendation becomes … (circle)
A. A permanent brachytherapy seed implant	A past history of transurethral resection of the prostate	−2 −1 0 +1 +2
B. A permanent brachytherapy seed implant	A symptomatic active ulcerative colitis	−2 −1 0 +1 +2
C. A permanent brachytherapy seed implant	A prostatic volume of 50 cc	−2 −1 0 +1 +2

−2, contra-indicated; −1, less indicated; 0, it doesn't change your mind; 1, more indicated or 2, a lot more indicated

Scoring

Following common methodology used in SCT scoring, the individual answers of panel members were used to create the scoring key (Gagnon et al. 2005; Lubarsky et al. 2009). For each item, examinees' answers received a score corresponding to the proportion of panel members who selected the same response. The maximum score for each item is 1 for the modal answer. Other panel members' choices receive partial credit. To get this proportional transformation, the number of members who provide an answer on the Likert-type scale is divided by the modal value for the item. Answers not chosen by panel members receive zero. The examinee's total score for the test is determined by the sum of the credit obtained for each of the items.

Data set

This study is based on a data set collected from an SCT designed for use in radiation oncology (Lambert et al. 2009). The instrument was constructed by two radiation oncologists, and comprised 90 items nested in 30 cases covering topics in urologic, breast, and lung cancer. Participants were 70 medical students and 38 residents in radiation oncology. All radiation oncologists were invited to participate to the study (n = 62), and no exclusion criteria were applied. The panel of reference was composed of 45 radiation oncologists with a large array of expertise and clinical experience working in the province of Quebec. Two panel members and 1 resident were excluded from analysis because their tests contained too many missing answers.

Optimization methods

Three methods for optimizing SCT scoring keys were tested in this study. The first method (outliers) prohibited members with aberrantly low total scores from contributing any of their responses to the answer key. The two other methods (distance-from-mode and judgment-by-experts) discarded only widely deviant responses (submitted by any panel member) to individual questions from the test's scoring key.

- The “outlier method” is based on the calculation of panel members’ scores for the whole test. Members that had total scores beyond 2 standard deviations from the mean of the distribution of scores of panel members were excluded.
- The “distance-from-mode” method is based on the identification and exclusion of answers that are (a) more than two anchors (“2 cat. exclusion”) or (b) more than one anchor (“1 cat. exclusion”) away from the modal answer. For instance, if, for a given question, the modal answer is ‘+1’, then (a) ‘−2’ responses are excluded and (b) both ‘−1’ and ‘−2’ responses are excluded.
- For the “judgment-by-experts” method, three radio-oncology experts independently reviewed all answers to determine those they deemed unacceptable. Answers that were classified as unacceptable by at least 2 of the three experts were excluded. The three reviewers were physicians recognized by peers as having particular expertise in the three domains of the test.

Statistical analysis

Two kinds of data sets were analyzed. First, calculations were performed using the whole panel of radio-oncologists (n = 45). Second, since such large panels are rarely assembled for the purpose of setting an SCT answer key, 20 panels consisting of 15 members each were randomly resampled from the whole panel of 45 and submitted to analysis. Fifteen member panels have been shown to be the minimum size to obtain adequate score stability (Gagnon et al. 2005). For both types of data sets, the three optimization methods were applied in calculating test scores for panel members, residents and students.

Metric qualities of the optimization methods were compared on three dimensions: (1) effect on the Cronbach alpha coefficient; (2) correlation between the original and reduced set of panel answers and (3) capability of detecting differences between groups with different level of expertise (measured by effect size).

Effect size was calculated as the ratio of the difference between groups and the mean of the standard deviation of both groups. In all calculations, missing answers were replaced by the mean value of all other valid answers of each respondent with missing data. When there were more than 9 missing values (10% of the 90 test questions), the respondent was excluded from analyses. Sample fluctuation of reliability, correlation with original score and effect size across panels of 15 was studied.

Results

Scores were computed for 45 radiation oncologists, 37 residents and 70 students.

Comparison of methods with a 45-member panel

The Outlier method led to the exclusion of three panel members. By excluding them a total of 270 answers (90 × 3) were eliminated. The distance-from-mode/2 cat. exclusion method led to the exclusion of 93 (2.2%) of the 4,050 answers given by panel members; with the distance-from-mode/1 cat. exclusion 437 (10.7%) answers were excluded. The judgment-by-experts method led to the exclusion of 262 (6.4%) of the 4,050 answers.

Table 2 presents the comparison of the optimization methods on the panel of 45. Reliability is unaffected and correlations between corrected scores and original scores are

Table 2 Comparison of optimization methods with panel of 45

	Alpha value	Correlation with original data	Effect size		
			Students-residents	Students-panel	Residents-panel
No exclusion	0.86		2.08	2.57	0.48
Removal of members	0.86	0.99	2.05	2.44	0.45
Distance-from-mode					
2 Anchors	0.86	0.99	2.06	2.53	0.47
1 Anchor	0.86	0.99	2.08	2.52	0.46
Judgment-by-experts	0.86	0.99	1.99	2.47	0.43

Table 3 Comparison of optimization methods on panels of 15 (20 samples of 15 out of 45)

	Mean alpha value (SD)	Mean correlation with original data (SD)	Mean effect size (SD)		
			Students-residents	Students-panel	Residents-panel
No exclusion	0.86 (0.01)	0.99 (0.01)	1.97(0.12)	3.44 (0.70)	1.18 (0.33)
Removal of members	0.86 (0.01)	0.98 (0.01)	2.08 (0.11)	3.07 (0.74)	0.88 (0.32)
Distance-from-mode					
2 Anchors	0.86 (0.01)	0.98 (0.01)	1.93 (0.11)	3.16 (0.75)	1.05 (0.37)
1 Anchor	0.85 (0.01)	0.98 (0.01)	2.01(0.12)	2.32 (0.17)	0.37 (0.07)
Judgment-by-experts	0.85 (0.01)	0.98 (0.01)	2.03 (0.12)	3.73 (0.53)	1.24 (0.26)

near perfect ($r > 0.99$). Effect size is minimally affected by optimization. It appears that power to detect difference between groups is slightly lowered by optimization.

Comparison of methods with 15-member panels

Table 3 presents the comparison of the optimization methods using mean value of reliability, correlation and effect size calculated on 20 random samples of 15 panel members. Reliability is almost unaffected by any kind of optimization. All correlations between corrected scores and original scores are high ($r = 0.98$). All methods slightly lower the effect size in all group comparisons. Figure 1 illustrates fluctuations of effect size estimates from one random sample to another on differences between residents and panel; note that the distance-from-mode/one anchor method demonstrates the most conspicuous fluctuation.

Discussion

Script concordance testing has consistently shown good reliability and the capacity to discriminate among levels of experience in domains such as neurology, general surgery, radiation oncology, and emergency medicine (Lubarsky et al. 2009; Meterissian et al. 2007; Lambert et al. 2009; Carriere et al. 2009). However, its current scoring system, based

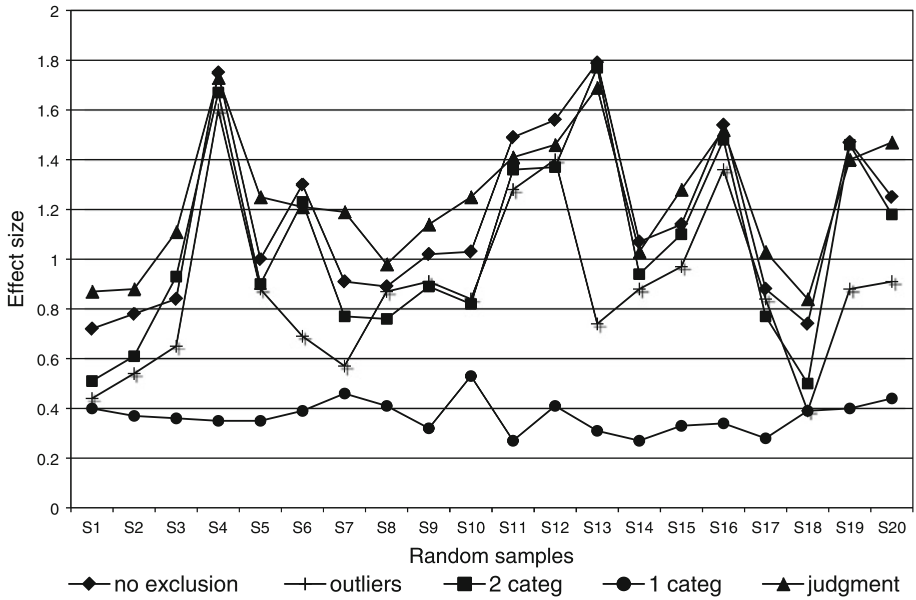

Fig. 1 Sampling variability of effect size values (panel and resident means). *Each line* corresponds to a method of panel optimization. The graphic points out wide variability of effect size value from one sample to another. With most methods effect size value are quite similar for any given sample, but the 1-category method has a very detrimental effect on effect size

on the principle that all answers selected by panel members have inherent value and should be retained in the answer key, remains controversial (Bland et al. 2005). Challengers of the current system contend that obviously 'incorrect' answers should not be accepted, whereas proponents argue that if a panel member, based on his or her expertise, interprets a question in a particular way, it is legitimate to credit examinees for holding a similar—presumably expert—interpretation (Charlin et al. 2004).

Our results contribute a psychometric perspective to this ongoing debate. We found that, irrespective of panel size (45 vs. 15 members), none of three optimization methods used to reduce panel response variability affected test score reliability. Based on this observation we conclude that, from a purely psychometric viewpoint, it is not necessary to exclude either deviant panel members or deviant answers from the scoring of an SCT (provided that a panel of at least 15 members is assembled).

However, it may trouble test-makers to reward answers that are highly deviant or clearly incorrect, even if credit awarded for these answers is small and has no effect on test reliability. Educators may wish to remove these answers. For this endeavor the three described optimization methods appear equivalent in terms of psychometric consequences. Since exclusion of panel members may be a disincentive for future panel member recruitment, methods based on exclusion of deviant answers may be preferable. Among these methods, the judgment-by-experts method has better face validity but demands a considerable time commitment on the part of experts assigned to the task. The distance-from-mode method (either beyond 1 or 2 categories) is easy to implement and demonstrates equivalent psychometric qualities. It is therefore the method we recommend.

The panel used in this study is highly representative of the entire population of radio-oncologists (45 out of 62 of these specialists in the province of Quebec). This is an ideal but rather unusual assessment situation. In general panels are smaller in size, with a minimum size of 15 for higher stake examinations (Gagnon et al. 2005). In this context fluctuations of scores appear, but these are small with respect to test reliability or correlation with a larger panel of 45.

When comparison of means between groups is needed, effect sizes fluctuate depending on the method used for panel optimization. The SCT used in this study was designed to test educational objectives of radio-oncology residency, therefore the most interesting comparison concerns residents versus panel members. In this respect the detrimental effect of one of the distance-from-mode methods has to be pointed out. With the one anchor method the random fluctuation of effect size across samples (see Fig. 1) disappears, but group standardized differences are considerably decreased.

Conclusion

This study examines a single SCT administered in a specific context, therefore results need to be confirmed in other settings. Nonetheless, our results indicate that, with respect to reliability, the current scoring methodology appears to be robust, resistant to deviant answers or members as long as the panel size is large enough (15 or more members). There is therefore no psychometric requirement to remove deviant panel members or deviant answers. However if panel answer optimization is desired, removing deviant answers only is as efficient as removing deviant experts, and less prejudicial for future panel recruitment. Among methods of answer removal, the method matters: the judgment-by-experts method is superior, but distance-from-mode/2 categories is nearly as efficient and far more practical.

References

Allman, R. M., Steinberg, E. P., Kerulv, J. C., & Dans, P. E. (1985). Physician tolerance for uncertainty: Use of liver-spleen scans to detect metastases. *JAMA, 254*, 246–248.

Bland, A., Kreiter, C., & Gordon, J. (2005). The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine, 80*, 395–399.

Carriere, B., Gagnon, R., Charlin, B., Downing, S., & Bordage, G. (2009). Assessing clinical reasoning in pediatric emergency medicine: Validity evidence for a script concordance test. *Annals of Emergency Medicine, 53*, 647–652.

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

Charlin, B., Gagnon, R., Pelletier, J., et al. (2006). Assessment of clinical reasoning in the context of uncertainty: The effect of variability within the reference panel. *Medical Education, 40*, 848–854.

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

Eddy, D. M. (1984). Variations in physician practice: The role of uncertainty. *Health Affairs (Millwood), 3*, 74–89.

Fournier, J. P., Demeester, A., & Charlin, B. (2008). Script concordance tests: Guidelines for construction. *BMC Medical Informatics and Decision Making, 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.

Lambert, C., Gagnon, R., Nguyen, D., Charlin, B. (2009). The script concordance test in radiation oncology: Validation study of a new tool to assess clinical reasoning. *Radiation Oncology, 9*, 4–7. http://www.ro-journal.com/content/4/1/7.

Lubarsky, S., Chalk, C., Kazitani, D., Gagnon, R., & Charlin, B. (2009). The script concordance test: A new tool assessing clinical judgment in neurology. *Canadian Journal of Neurological Sciences, 36*, 326–331.

Lubarsky, S., Charlin, B., Cook, D.A., Chalk, C., & van der Vleuten, C. (2011). Script concordance method: A review of published validity evidence. *Medical Education* (in press).

Meterissian, S. (2007). Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *The American Journal of Surgery, 193*, 248–251.

Schmidt, H. G., Norman, G. R., & Boshuizen, H. P. A. (1990). A cognitive perspective on medical expertise: Theory and implications. *Academic Medicine, 65*, 611–621.