Adv in Health Sci Educ
DOI 10.1007/s10459-008-9120-8

ORIGINAL PAPER

Script concordance testing: more cases or more questions?

Robert Gagnon · Bernard Charlin · Carole Lambert · Benoit Carrière · C. Van der Vleuten

Received: 28 September 2007 / Accepted: 24 April 2008

Abstract Script concordance test (SCT) is a case based assessment format of clinical reasoning in which questions are nested into several cases. Recent results using Q4 format suggest that nested questions contribute more to reliability of measure than cases. The present study aims at documenting variance components associated with SCT cases and nested questions and to determine what are the optimal number and combinations of cases and nested questions. Data from SCT in three different fields are presented. G study and D study methodology are used to estimate variance component and to determine optimal number and combinations of cases and questions. Questions nested into cases contributed a large amount of score variance (more than 70%). D studies with varying samples show that, depending on the reliability of the test, an optimal number of 2–4 questions nested into 15–25 cases represents the best combination. Nested questions contribute to a significant portion of score variance, with the implication that formulation of up to 5 questions per case is an efficient way to optimize the reliability of SCT scores.

Keywords Clinical reasoning · Aggregate scoring · D Study · Test optimisation · Script concordance test · Generalizability

Introduction

Script concordance test (SCT) is a format of case based examination. It probes the multiple judgments that are made in the clinical reasoning process and scoring is a measure of the

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

B. Carrière
University of Montreal, Montreal, QC, Canada

C. Van der Vleuten
Maastricht University, Maastricht, The Netherlands

concordance of these judgments to those of a panel of reference. Each item is made of a case presentation, followed by a series of related questions. Up to now, no study has investigated the effect of questions and cases on the variability of performance on the SCT. Furthermore, for test development purpose, it is important to answer a frequently asked question in the design of SCT energy where should be invested? In designing more cases or more questions into each case?

Using data on Key-Features test—another format of case based examination—Norman et al. have recently questioned the concept of case specificity (Norman et al. 2006). According to the current assumption, error variance due to cases should be high and error variance due to questions nested in cases should be low. They surprisingly found relatively little variance due to differences between cases, and about 80% of the error variance due to variability in performance among questions within cases. The authors concluded that similar results were likely to be found for other examination formats containing caselets, with questions nested in cases. The authors were tempted to conclude that the case specificity phenomenon might not be as universal as expected, and that adding questions to cases may be more efficient in terms of reliability of measure than adding more cases.

The present study reports generalizabity analyses on three SCT data sets to examine the sources of variance in relation to cases and items nested within cases (G analysis). Then using D analysis it looks at the optimal number of cases and questions within cases that maximise test reliability.

Methodology

The study rests on the analysis of three previously collected data sets in three different domains. In Radio-oncology, reasoning skills on patient management was investigated; in nursing, attitude toward patient caring was assessed, and finally, in pediatrics clinical reasoning in emergency situations was studied. Ethical approval was already obtained for the three data sets. All three data set were obtained from paper and pencil tests.

Script concordance test

Concordance tests are designed to assess clinical reasoning in situations that include uncertainty. Cases contain uncertainty as they always offer several solutions. Questions may or may not include uncertainty. Those with uncertainty are those for which there is no absolute right answer. For each item, a clinical case is presented, containing either not enough data to solve the clinical problem (diagnostic, treatment), or data ambiguity, or conflict among values (assessment of ethical reasoning for instance). A series of questions is related to the case. Each contains an option relevant to the clinical problem, followed by the presentation of new data. Examinees' task is to assess the effect the new data have on the status of the option. Subsequent questions within the case explore the effect of other data on other options (Charlin et al. 2000). The scoring system, based on partial credits, has been described in the literature (Charlin et al. 2007). A series of studies has revealed the usefulness of concordance tests to discriminate along levels of experience, their applicability in domains as diverse as surgeons' pre-operative reasoning, choice of treatment protocols in Radio-oncology, reasoning in the emergency room, and reasoning in neurology.

Data sets

Radio-oncology (reasoning skills on patient management) (Lambert 2006)

The instrument was constructed by two radiation oncologists. Representative clinical situations were elaborated in accordance with learning objectives of the residency program. The test was made of 30 patient problems in short vignettes, each of them followed by three related questions about diagnosis, investigation plan, and treatment option, that are relevant to the situation. A total of 90 questions were constructed covering urologic cancer (10 cases), breast cancer (10 cases), and lung cancer (10 cases). Participants were a group of 70 medical students, a group of 38 residents in radiation oncology, and a panel composed of 42 radiation oncologists working in Quebec. Participation in the study was voluntary. Respondents did not receive any remuneration for their participation.

Nursing (attitude toward patient caring) (Deschênes 2006)

A bank of 90 SCT questions for nurse practitioners was developed by researchers from the Faculty of Nursing of the University of Montreal. It covered three main aspects of attitudes toward caring using 90 questions into 29 cases. After elimination of questions with negative correlation with the total score, 16 cases including 48 questions were retained for the present study. Participants were 30 nursing students. The panel was composed of 12 experienced nurse practitioners in active practice, whose presence on a jury is legitimate considering the level of the persons assessed. Panel members were asked to fill out the test exactly as the examinees will do, and their answers were then used to constitute the scoring key. All contacted nurses agreed to participate. Participation in the study was voluntary. Respondents did not receive any remuneration for their participation.

Pediatric emergency (Clinical reasoning on emergency care) (Carrière 2005)

The SCT was made of 38 cases (60 questions) related to pediatric emergency medicine (PEM) and developed according to Royal College learning objectives. Two emergency pediatricians, fully qualified as PEM physicians in Canada, designed a table of specifications and prepared the test. A quality control grid was also followed closely to assure content validity during the test construction. The topics chosen represent important learning objectives where clinical reasoning expertise is considered particularly crucial. The 38 cases comprised a total of 60 questions. Sixteen cases contained a single question, while 22 cases had 2 questions each. In order to keep a balanced design, only these 22 cases (44 questions) were retained for the present study. Participants were residents recruited from various training levels and different residency programs completing a mandatory PEM rotation. Forty-nine of the 51 eligible residents (96.1%) participated. A convenience sample of 12 local attending physicians made the panel of reference. The panel was asked to complete the PEM SCT in a similar condition than residents with the same time period allotted, and without discussion among members.

Statistical analysis

G and D generalizability studies were conducted to identify variance components and reliability for each test, and to optimise the reliability of the given test composition (from 1 to 5 questions per case) (Brennan 2001).

A G study was conducted on each test to estimate variance components and reliability expressed as a relative G coefficient. The model used was R × Q:C where R is for respondents fully crossed with Questions nested in Cases. A random model was used for all facets. Using results from the G study, a D study was conducted to explore the best mix of questions nested into cases. D studies were done using both a domain-referenced perspective. EduG 3.04 (2005) was used for conducting the generalizability study.

Results

G study

Table 1 presents the main characteristics of the three tests. The relative G coefficients are quite satisfactory for tests in nursing and oncology (0.78 and 0.88), while the G coefficient is considerably lower for pediatrics (0.63).

Table 2 presents the results of the estimation of the variance components across the three tests. The main source of variance in scores is associated with the interaction of respondents and nested questions (85–92%). Variance components associated with nested questions are low (0% and 10.2%) still greater than variance components associated with cases (0% and 1.4%). Variance associated with the interaction between respondents and cases is also very low (0% and 1.8%).

D study

Figures 1–3 illustrate the effect of variable number of nested questions and variable number of cases on the relative G coefficient.

Table 1 Reliability results

	Nursing	Pediatrics	Oncology
Number of respondents	30	49	106
Number in panel	12	12	42
Number of cases	16	22	30
Number of questions	48	44	90
Questions per case	3	2	3
Mean, SD	73.9 (10.6)	67.7 (8.9)	61.3 (11.7)
G coefficient	0.78	0.63	0.88

Table 2 Components of variance

	Nursing			Pediatrics			Oncology		
	VC	SE	%	VC	SE	%	VC	SE	%
Respondents	0.008	0.003	7.1	0.005	0.002	3.4	0.012	0.002	7.7
Cases	−0.0001	0.001	0.0	0.002	0.004	1.4	−0.000	0.000	0.0
Questions:cases	0.008	0.001	0.6	0.015	0.005	10.2	−0.000	0.000	0.0
Respondents × cases	0.002	0.003	1.4	−0.003	0.004	0.0	−0.004	0.001	0.0
Respondents × questions:cases	0.113	0.005	90.8	0.129	0.005	85.0	0.146	0.002	92.3

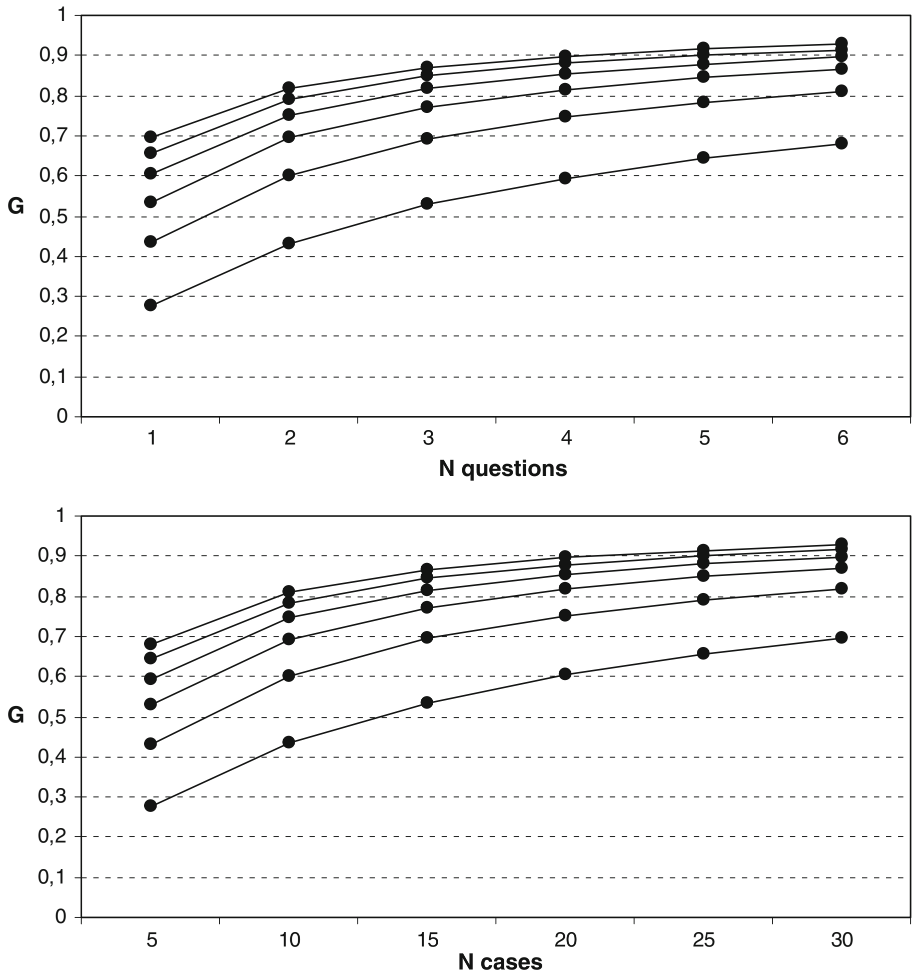

Fig. 1 Nursing. G coefficients in relation to (**a**) number of questions and number of cases (5, 10, 15, 20, 25, 30), and (**b**) number of cases and number of questions per case (1–6)

The impact of adding questions per case appears to be greater than adding cases. For example, in nursing, going from 1 to 3 questions increases the G coefficient by 0.25 point, while adding 5 cases is associated with an increase of about 0.10 point. It can be observed that with over 3 questions by case, the effect of adding more questions is associated with less gain in reliability. This observation is true to the three tests under study. This “saturation” in reliability gain by adding more questions may be compensated by adding more cases, while keeping a maximal number of questions (4 or 5), at the expense of building a much longer test. This is also substantiated by data in Table 3.

In the perspective of a test design, an important question is how many question should be introduced per case to have greater chances to obtain a reliability coefficient greater than 0.80? Data show that one is clearly insufficient, two is better, but three seems to offer the best balance. The present results show that at least 54 questions would be necessary in

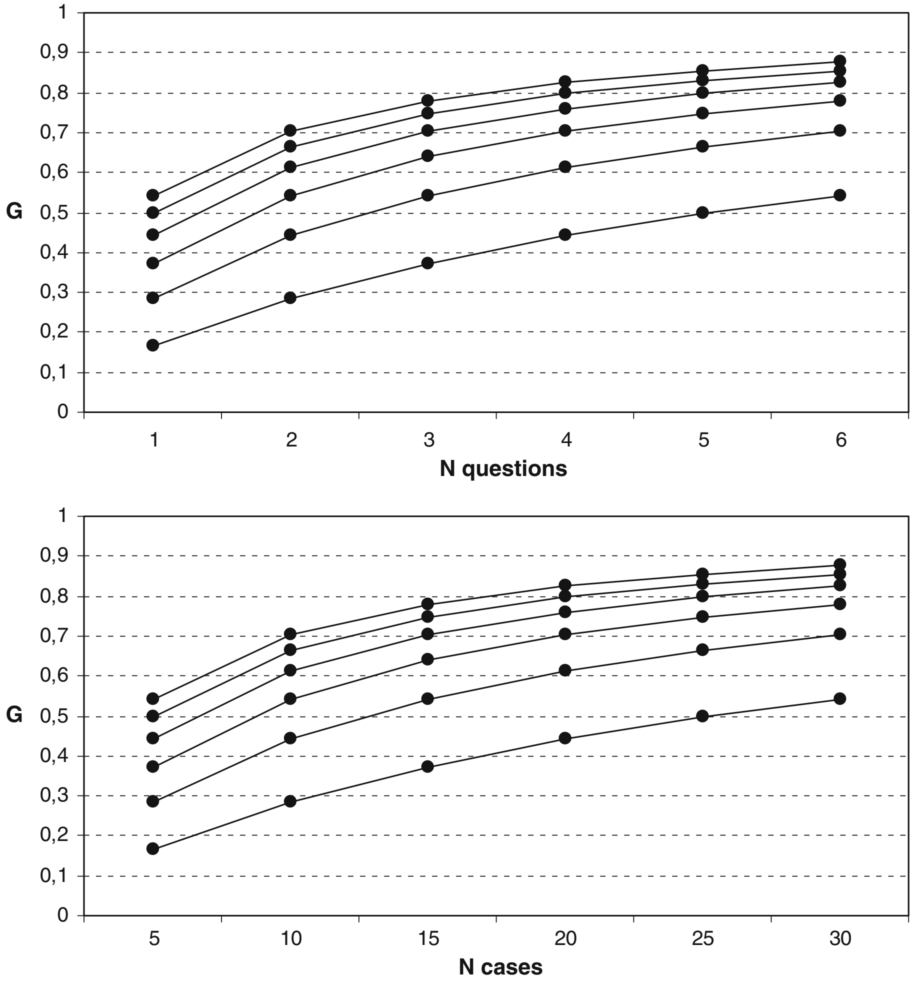

Fig. 2 Pediatrics. G coefficients in relation to (**a**) number of questions and number of cases (5, 10, 15, 20, 25, 30), and (**b**) number of cases and number of questions per case (1–6)

nursing (2 × 27; 3 × 18; 4 × 16), 48 (2 × 24; 3 × 16; 4 × 12) questions are sufficient in oncology, while 102 questions would be necessary in Pediatrics (2 × 51; 3 × 34; 4 × 26).

Discussion

The present study shows an optimal strategy in terms of enhancing reliability is, instead of multiplying the number of cases, to increase the number of questions nested into cases. While the sampling of an adequate number of cases is still a primordial concern for content validity matters, this study shows that in terms of reliability it is efficient to use 3–5 questions within each case.

Our results suggest that reliable Script concordance tests (reliability coefficients greater than 0.80) may be composed of few questions per case using 20–30 cases. Less reliable

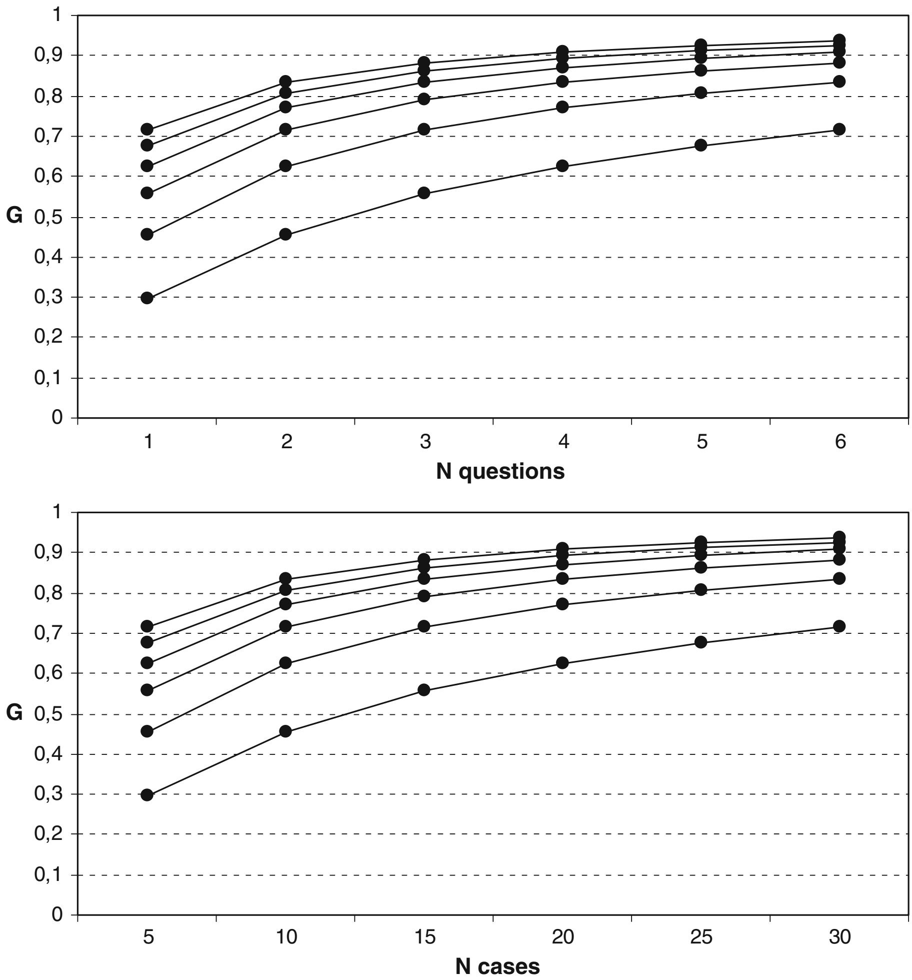

Fig. 3 Oncology. G coefficients in relation to (**a**) number of questions and number of cases (5, 10, 15, 20, 25, 30), and (**b**) number of cases and number of questions per case (1–6)

instruments may need more questions per case (4 or 5) in the perspective of the construction of a reasonable number of cases (25–30 cases). One meaningful aspect of the present results is the fact that the conclusions seem to hold true even for three different domains of testing with variable test quality.

A pragmatic approach could also be used to interpret the present results. It may be argued that it is usually practicable to construct between 15 and 30 cases for a given evaluation. Experience shows that it takes about 1 h to complete 75 SCT questions. So for a test designer what then is the optimal balance for one hour of testing time? Data show that it is clearly not efficient to use 1 question per case. Another argument against the use of 1 question per case is that experience tells us that is this cognitively very demanding for students to read a new case to answer each new question. So the issue then becomes: is it better to built 2 (37–38 cases), 3 (25 cases) or 4 (18–19 cases) questions per case. More than 30 cases may represent a serious workload both for test constructors and for

Table 3 D study of reliability by number of questions and cases

Number of questions per case	Number of cases	Total number of questions	Approximate testing time in minutes	G coefficient		
				Oncology	Pediatrics	Nursing
1	30	30	25	0.72	0.54	0.70
2	25	50	40	0.81	0.66	0.79
2	30	60	50	0.83	0.70	0.82
3	20	60	50	0.83	0.70	0.82
3	25	75	60	0.86	0.75	0.85
4	15	60	50	0.83	0.70	0.81
4	20	80	65	0.87	0.76	0.85
5	12	60	50	0.83	0.69	0.81
5	15	75	60	0.86	0.74	0.84

examinees taking the test; the present results show limited gain in reliability with more than 25 cases. Table 3 presents many scenarios where combination of number of cases and nested questions are presented. Again, it should appear to the test constructor that the utilisation of 2 or 3 questions per case is associated with good generalizability with a moderate workload of 20–25 cases to develop.

The present results show that it is important to gather a reasonable amount of information per case (3 questions ± 1 question) to obtain a reliable assessment of SCT scores. Also, acceptable G coefficients may be obtained with the utilisation of as few as 15 cases.

The profile of variance components also suggests that to get a generalizable assessment of SCT performance, one needs to build a test on a collection of enough cases (20–25 cases) using 3 or 4 questions per case. The developer could expect to obtain reliability and generalizability coefficient higher than 0.70 which may be viewed as acceptable. Better reliability (higher than 0.80) may be achieved to the cost of at least 4 questions and 25 cases.

In this study, in accordance with recent observations by Norman et al. questions nested into cases generated more true score than cases by themselves (Norman et al. 2006). In that sense, questions contribute more to reliability than cases at least up to a certain point. Beyond that point adding cases is wiser. A trade-off is in operation in case based testing. Case sampling within the case needs to be increased until a certain optimum. Finding this optimum for each format seems an important criterion for studying reliability of case based examinations. A similar trade-off issue has been found in OSCEs as well (Boegels et al. 1995) What needs to be verified further is to what extent this specific information contained into the questions is also associated to better validity of the measure. More studies on this topic should be devised to answer this question.

References

Boegels, S. M., van Mourik, T., & Van Der Vleuten, C. (1995). Authentic assessment of interviewing and counseling skills: Effect of testing time per station on generalizability and validity. *Teaching and Learning in Medicine, 7*(3), 155–162.

Brennan, R. (2001). *Generalizability theory*. New York: Springer.

Carrière, B. (2005). *Development of script concordance test in pediatric emergency medicine.* Master degree dissertation MHPE. University of Illinois, Chicago.

Charlin, B., Tardif, J., & Boshuizen, H. P. (2000). Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Academic Medicine, 75*(2), 182–190. doi: 10.1097/00001888-200002000-00020.

Charlin, B., Gagnon, R., Sauvé, E., & Coletti, M. (2007). Composition of the panel of reference for concordance tests: Do teaching functions have an impact on examinees' ranks and absolute scores? *Medical Teacher, 29*, 43–53. doi:10.1080/01421590601032427.

Deschênes, M. F. (2006). *Développement d'un TCS en sciences infirmières*. Dissertation, Faculty of Nursing, University of Montréal.

Lambert, C. (2006). *Développement d'un TCS en radio-oncologie*. Dissertation, Faculty of medicine, University of Montréal.

Norman, G., Bordage, G., Page, G., & Keane, D. (2006). How specific is case specificity? *Medical education, 40*(7), 618–623. doi:10.1111/j.1365-2929.2006.02511.x.