Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test?

R Gagnon,[1] B Charlin,[1] M Coletti,[2] E Sauvé[1] & C van der Vleuten[3]

PURPOSE The script concordance test (SCT) assesses clinical reasoning in the context of uncertainty. Because there is no single correct answer, scoring is based on a comparison of answers provided by examinees with those provided by members of a panel of reference made up of experienced practitioners. This study aims to determine how many members are needed on the panel to obtain reliable scores to compare against the scores of examinees.

METHODS A group of 80 residents were tested on 73 items (Cronbach's α: 0.76). A total of 38 family doctors made up the pool of experienced practitioners, from which 1000 random panels of reference of increasing sizes (5, 10, 15, 20, 25 and 30) were generated with a resampling procedure. Residents' scores were computed for each panel sample. Units of analysis were means of residents' score, test reliability coefficient and correlation coefficient between scores obtained with a given panel of reference versus the scores obtained with the full panel of 38. Statistics were averaged across the 1000 samples for each panel size for the mean and test reliability computations, and across 100 samples for the correlation computation.

RESULTS For sample variability, there was a 3-fold increase in standard deviation of means between a sample panel size of 5 (SD = 1.57) and a panel size of 30 (SD = 0.50). For reliability, there was a large difference in precision between a panel size of 5 (0.62)

and a panel size of 10 (0.70). When the panel size was over 20, the gain became negligible (0.74 for 20 and 0.76 for 38). For correlation, the mean correlation coefficient values were 0.90 with 5 panel members, 0.95 with 10 members and 0.98 with 20 members.

CONCLUSION Any number over 10 is associated with acceptable reliability and good correlation between the samples versus the full panel of 38. For high stake examinations, using a panel of 20 members is recommended. Recruiting more than 20 panel members shows only a marginal benefit in terms of psychometric properties.

KEYWORDS education, medical, undergraduate/ *standards; clinical competence/*standards; educational measurement/standards; physicians family; physicians role; reproducibility of results; Canada.

[1]Department of Surgery, University of Montreal, Montreal, Canada
[2]Department of Family Medicine, University of Bobigny, Bobigny, France
[3]Department of Educational Development and Research, Maastricht University, Maastricht, The Netherlands

Correspondence: Bernard Charlin, URDESS, Faculté de Médecine-Direction, Université de Montréal, CP 6128, Succursale Centre-Ville, Montréal, Québec H3C 3J7, Canada. Tel: 00 1 514 343 6111 (ext. 14140); Fax: 00 1 514 343 7650; E-mail: bernard.charlin@umontreal.ca

INTRODUCTION

As in other health professions, a significant part of a doctor's competence relies on the capacity to deal with uncertainty.[1] In a clinical encounter, not all the data needed to solve a problem are available. These data must be gathered in order to formulate the problem and then solve it. Furthermore, problems can be confusing, contradictory and ill defined,[2] and are often characterised by imperfect, inconsistent or even inaccurate information. The capacity to reason in contexts of uncertainty and to solve poorly defined problems is a hallmark of professional competence.[3]

Traditional written tools for assessing clinical reasoning, such as rich-context, multiple-choice

Overview

What is already known on this subject

The scoring method of the script concordance test (SCT) is a comparison of examinees' answers with those given by members of a panel of reference.

The method takes into account the variation of answers among panel members. It therefore allows assessment of reasoning in context of uncertainty.

What this study adds

Because of the variations among panel members' answers, the reference panel has to be large enough to obtain reliable panel scores and, thus, reliable students' scores.

It is, therefore, important to investigate how large the panel should be.

The study shows that using a panel of 20 is recommended for high stake examinations. For other examinations any number over 10 is associated with acceptable reliability.

Suggestions for further research

Further analysis using generalisability and decision studies techniques with similar data may produce valuable new information on the optimal number of members to include in panels.

questions, properly and reliably test the ability of students to apply well known solutions to well defined problems. But comprehensive assessment of clinical reasoning should include tools other than those assessing well defined problems, tools which measure the ability to rationally solve ill defined problems. Some evaluation tools, such as oral examinations, can assess this aspect of clinical competence but they have limitations such as difficulty of standardisation, objectivity of scoring, or practicability for large groups of examinees.

According to script theory,[4–6] clinicians use scripts (i.e. knowledge structures specifically adapted to the task they commonly perform) to actively process information in order to confirm or eliminate hypotheses and management options.[6] Clinicians are, therefore, constantly making qualitative judgements on the significance of the data they collect. Each of these judgements can be measured, providing a method of assessing reasoning on ill defined problems and in contexts of uncertainty.[7] The method is called the 'script concordance approach'.[8]

The approach relies on 3 principles, each of them concerning 1 of the 3 components[9] inherent to any test: the task required of examinees, the way examinees' answers are recorded and the way examinees' performances are transformed into scores. The task is meant to be challenging, even for an expert. It represents an authentic clinical situation and is described in a vignette. It is challenging either because the vignette does not contain all the data needed to provide a solution (for a test on diagnosis or management, for instance) or because several attitudes are defensible (for a test in ethics, for instance). Several options (diagnosis, management or attitude) are relevant. Items refer to the questions experts ask themselves to find a solution. The response format (Fig. 1) is in accordance with what is known on clinical reasoning processes.[10–12] A Likert scale, measuring the judgements that are iteratively made using this process, captures examinees' answers. The scoring method takes into account the variation of answers among members of a panel of reference (Fig. 2). It is an adaptation of the aggregate scoring method.[13,14] Credits on each item are derived from the answers given by the panel of reference.

In a recent study,[15] we compared the effect on scores obtained by students and panel members with the aggregate method versus those obtained with the common method, in which experts are asked to provide a consensus for each item. We found that in the rich context of real clinical life, such as the contexts described in vignettes and items used in the script concordance test (SCT), panel members' answers varied substantially. This is in accordance with the findings of clinical reasoning research, where it has been shown that, in similar situations, professionals do not collect exactly the same data and do not follow the same paths of thought.[10] They also show substantial variation in performance on any particular real or simulated case.[11,12] Therefore, in the aggregate scoring process used in the SCT, the characteristics of the reference panel against which students' performance is compared become a major

If you were thinking of	And the patient reports or you find upon clinical examination	This hypothesis becomes				
Anaphylactic reaction	Respiratory rhythm at 32	−2	−1	0	+1	+2
Asthma	Difficulty swallowing	−2	−1	0	+1	+2
Hyperventilation	A normal pharynx	−2	−1	0	+1	+2
Anaphylactic reaction	Arterial blood pressure = 120/180	−2	−1	0	+1	+2
Asthma	A diffuse arterial II/IV murmur	−2	−1	0	+1	+2
Hyperventilation	Arterial blood pressure = 150/90	−2	−1	0	+1	+2

−2   Ruled out or almost ruled out
−1   Less probable
0   Neither less nor more probable
+1   More probable
+2   Certain, or almost certain

Figure 1 Example of a clinical vignette and format of items used for diagnostic knowledge assessment.

If you were thinking of	And the patient reports or you find upon clinical examination	Number of members for each response				
		−2	−1	0	+1	+2
Anaphylactic reaction	Respiratory rhythm at 32	0	0	1	9	2
Asthma	Difficulty swallowing	5	5	1	0	1
Hyperventilation	A normal pharynx	0	0	7	5	0
Anaphylactic reaction	Arterial blood pressure = 120/180	0	7	5	0	0
Asthma	A diffuse arterial II/VI murmur	0	2	10	0	0
Hyperventilation	Arterial blood pressure = 150/90	0	0	9	3	0

−2   Ruled out or almost ruled out
−1   Less probable
0   Neither less nor more probable
+1   More probable
+2   Certain, or almost certain

Figure 2 Profile of panel members' responses (12 members) on the items presented in Fig. 1.

issue. Because of the variations among panel members' answers, even in homogeneous groups of practitioners,[15] the reference panel has to be large enough to produce reliable panel scores and, therefore, reliable students' scores. No research has yet been conducted about the required number of panel members. This study aimed to answer the question: How many members are needed on the reference panel for the script concordance test to obtain reliable scores for students?

METHOD

Subjects

The SCT technique requires recruiting a panel of reference. This panel is made up of experienced practitioners whose presence on a jury is legitimate considering the level of the persons to be assessed. Panel members are asked to fill out the test exactly as

the examinees will do, and their answers are then used to constitute the scoring key. A total of 38 family doctors in active practice in an urban area and associated with the Faculty of Medicine, University of Bobigny, France (either as teachers or as training supervisors) were asked to participate. All agreed.

Eighty residents in the family medicine residency programme at the University of Bobigny were also recruited to complete the test. Participation in the study was voluntary. Subjects were recruited during the academic year 2002. Respondents did not receive any remuneration for their participation.

Test

A bank of SCT items for family medicine was developed in October 2001 by researchers at the Faculty of Medicine, University of Montreal and the Quebec Board of Physicians.[16] The instrument used in the study (90 items) was created using items from the bank. Some minor revisions were made to adapt the instrument to the context of family medicine in France.

Scoring process

The individual answers of panel members were used to create the scoring keys, following the common methodology used in the SCT.[7] For each item, examinees' answers received a credit mark corresponding to the proportion of panel members who selected it. The maximum score for each item was 1 for the modal answer. Other panel members' choices received a partial credit. Answers not chosen by panel members received 0. To obtain this proportional transformation, the number of members who had provided an answer on the Likert scale was divided by the modal value for the item. If, for example, on a given item, 20 members (out of 38) chose response 1 on the Likert scale, this choice received the maximum score of 1 point (20/20). Then, if 10 members chose response 2, this choice received 0.5 (10/20). Finally, if 8 panel members chose response 0, this choice received 0.4 (8/20). The total score for the test was the sum of credits obtained on each item, which in the end was transformed to obtain a maximum of 100.

Procedure

Panel members were asked to complete the test individually at the beginning of a group meeting. Residents were asked to fill it out at the beginning of one of their courses or during one of their rotations in a private practice office. Residents were given all

the time they needed to complete the task. The panel members completed their test in less than 40 minutes. Respondents were told that data would be used in the context of a comparison of the response patterns of practising doctors versus those of doctors in training (residents). Answers to the test were processed anonymously.

Optimisation of the test

The test written by residents and panel members was made up of 90 items with an α coefficient of 0.64. In order to work with the best possible instrument, an optimisation procedure was used, to ensure that only the best items, those giving the highest α value, were retained for the calculation of a global score. Items with negative item−total correlation were excluded in a stepwise manner, and this procedure was carried out until all items with negative correlations or very low item−total correlations ($r < 0.05$) had been excluded. The final α was 0.76 for the 73 items retained.

Statistical analysis

From the full panel of 38 doctors, a resampling procedure was used to draw 1000 random panels of increasing sizes ($n = 5, 10, 15, 20, 25$ and 30) through successive iterations. The resampling procedure was as follows for each panel size ($n = 5, 10, 15, 20, 35, 30$):

1 the response set of a random sample of n members out of 38 was drawn;
2 based on the response of this panel, individual scores for the 80 residents were calculated, and
3 the mean score and the α coefficient of this distribution were calculated.

This operation constituted 1 iteration and 1000 similar iterations were performed. Thus, this procedure produced a distribution of 1000 means of residents' scores and 1000 test α coefficients based on a panel of 5 members, 1000 means and 1000 α coefficients based on a panel of 10 members, and so on.

The first analysis estimated the mean, the standard deviation, the median, the maximum, the minimum and the range of extreme values for every 1000 score samples. The unit of analysis was the sample mean of all 80 residents. This analysis was planned to estimate the sampling variability of this distribution parameter in relation to the number of members in the panels.

The second analysis estimated the reliability of score distribution with a classical reliability approach using Cronbach's alpha. The unit of analysis was the sample reliability for all 80 residents. Alpha coefficients were averaged across the 1000 samples for each panel size. This analysis was planned to estimate the magnitude, and sampling variability, of the reliability coefficient in relation to the number of members in the panels.

The third analysis was based on a computation of the Pearson correlation coefficient between scores obtained by residents with a given panel versus the score obtained with the full panel of 38. For this analysis the process was not automated and had to be carried out by hand. Correlation coefficients were therefore averaged across 100 samples for each panel size (instead of 1000 as in the other analyses). The analysis was planned to estimate the magnitude, and sampling variability, of the correlation coefficient in relation to the number of members in the panels.

The results of the 3 analyses were plotted against panel size. This kind of graphical representation helps to identify a threshold where results are optimised.

RESULTS

The full panel was composed of 38 family doctors. Their mean age was 51.4 years (SD = 4.7, range 41−59 years) and mean number of years in practice was 22.6 years (SD = 6.3, range 9−33 years). They had been associated with the Faculty of Medicine at Bobigny, France for between 5 and 20 years.

Sample variability

There was a 3-fold increase in SD of means when the sample panel size went from 5 (SD = 1.57) to 30 (SD = 0.50). The range of values from a panel size of 5 (range = 10.2) to a panel size of 30 (range = 2.9) was also large. Figure 3 illustrates the reduction of sampling variability and range in relation to increasing panel size.

Reliability

The reliability of the test with the full panel was 0.76. There was a large difference in precision when the panel size went from 5 (0.62) to 10 (0.70). When the panel size exceeded 20, the gain became negligible (from 0.74 for 20 to 0.76 for 38). Figure 4 illustrates the changes in reliability estimates and range in relation to increasing panel size. It shows that the use
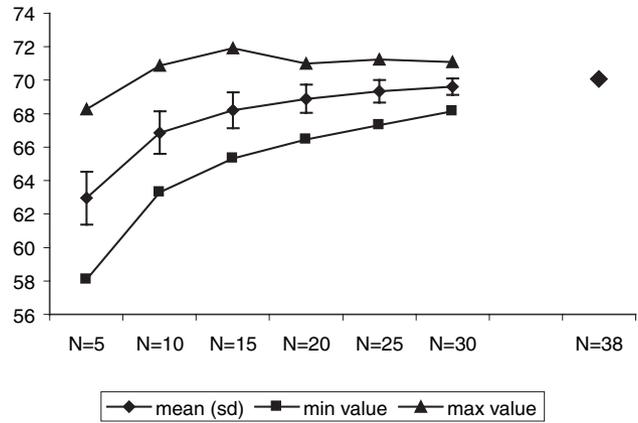

Figure 3 Sampling variability of means taken from different subsets of experts (1000 random samples per panel).
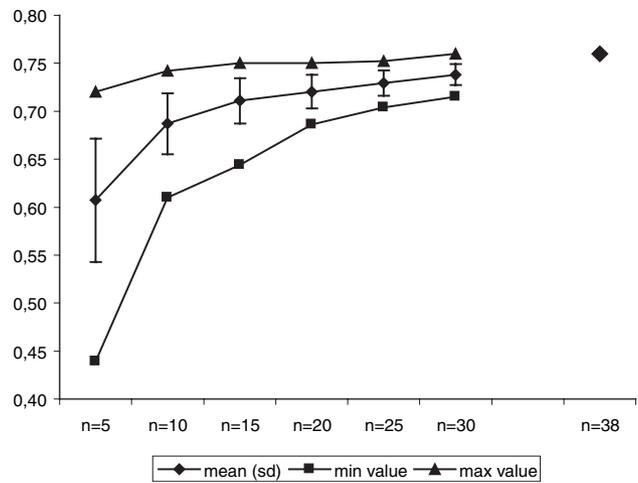

Figure 4 Mean reliability coefficients for each panel size (1000 random samples per panel).

of 20 members in a panel gave a very satisfactory estimate of the reliability observed with the full panel.

Correlation

The average correlation between the individual scores obtained in each sample and the scores obtained with the full number of members was 0.90 with a panel size of 5 (SD = 0.03, range = 0.14), 0.95 with a size of 10 (SD = 0.01, range = 0.07), 0.98 with a size of 20 (SD = 0.004, range = 0.02) and 0.99 with a size of 30 (SD = 0.002, range = 0.008). Figure 5 shows the change in the correlation estimates and range in relation to increasing panel size. Using 15–20 members in the panel resulted in scores that correlated very highly with those produced by the full panel.
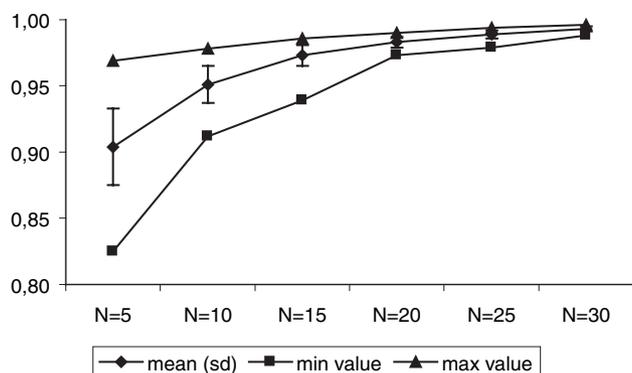
Figure 5 Mean Pearson correlation coefficients for each panel size (100 random samples per panel).

However, an unexpected result was observed. The mean of residents' scores increased with increasing panel size. For example, the mean of 1000 samples obtained with a panel size of 5 was 62.9, whereas the mean of 1000 samples obtained with a panel size of 30 was 69.6, representing a 7-point difference. In other words, the number of points earned on an SCT is influenced by the number of members used in the panel. No other studies using the SCT have revealed this consequence of the scoring process, making the present observation unexpected, although significant.

It therefore becomes necessary to determine a procedure for allowing a comparison of test results when the number of panel members differs. Equation fitting using regression analysis showed that the difference between the mean score in each subset and the mean score with all panel members can be modelled using the following formula, taking into account the number of panel members:

expected difference
$$= -0.52 + (30.4 \div \text{number of panel members}).$$

There is an inverse relationship between the number of members and the difference between the mean obtained with a reduced panel and the mean obtained with the full panel.

DISCUSSION

The results of the present study contribute towards answering a question frequently asked by developers: How many members are needed on the reference panel to maximise the reliability of the measure and to ensure a good scoring key? The results we have obtained clearly show that recruiting less than 10

members in a reference panel cannot give a reliable assessment of clinical reasoning with the SCT. Any number over 10 is associated with acceptable reliability and good correlation between the samples versus the whole set of panel members. For high stake examinations, using a panel of 20 is recommended. Recruiting more than 20 members shows only a marginal benefit in terms of psychometric properties. This is in line with research that has used generalisability models to estimate the numbers of judges required to obtain a reliable assessment of examinees.[17]

Studies have documented the discriminant validity of the SCT.[18,19] The present study now provides important clues to the optimal number of members to use in constructing the scoring key. On one hand, a minimum of 10 members are clearly needed, while, on the other, recruiting more than 20 members does not show any salient advantage in terms of psychometric properties. The grey area seems to lie in the range of 10–20 members. A smooth progression is observed from 20 to 30 members with all estimates, while 2 significant steps are observed with estimates based on 10 members and estimates based on 15 members.

Considering how difficult it often is to recruit examination jury members, the need to recruit 10–20 members on the panel of reference might be a concern. Actually, the SCT presents an advantage over other test formats: panel members are asked to answer questions that are very similar to those they ask themselves in their own clinical work. Furthermore, in contrast to the many other tests that require knowledge revision for optimal performance, a clinician can fill out the test at any time without any preparation. These 2 reasons explain why, in practice, it is not difficult to recruit members for panels of reference.

Another important issue concerns the composition of the panel. The basic idea behind the SCT is to compare the performances of students or residents with those of persons who are legitimate representatives of the profession (or the specialty) to which the students or residents wish to belong later. Therefore, our view is that panels should be made up of physicians with good clinical experience in the field rather than of experts in narrow parts of the field. Panel composition also depends of the assessment goal. If, for instance, we wish to assess family doctors' clinical knowledge on gynaecology, should the panel be made up exclusively of family doctors with a practice in gynaecology or of gynaecology

specialists? The answer depends on the goals of the test developers.

This study showed an unexpected result. Script concordance test examinees' scores increased with the number of panel members. A reason for this effect might be that more responses in a panel increase the number of categories of responses that may give credits of points to examinees. Thus, mean, median, minimum and maximum are difficult to compare when panel size differs. Therefore, comparing scores on SCTs should also take into account the number of members used. The model we used showed an inverse relationship ($1/n$ members) between the number of members in the panel and the difference between the score in a subset and the score with the full set. Further study will be needed to develop satisfactory adjustment algorithms that can be used for comparing scores based on different numbers of members used in the panels.

Two limits to the present methodology must be declared. The data used here were collected in France with French panel members and French residents. In a previous study,[19] in which examinees and panel members came from different cultures (Canada and France), we showed that students' ranking was not statistically changed when the panel of reference was made up of people from the other culture. That study therefore showed that results obtained with the SCT are robust across cultures, but we cannot be sure that a study held in another cultural environment would have yielded the same numbers of panel members to obtain reliable students' scores. In addition, for convenience considerations we recruited a group of 38 panel members. All sampling of estimates was based on this population. We believe that 38 is a sufficiently large number of panel members to provide a robust basis of comparison to our approach. Given that we considered a panel of 38 members to represent an 'optimal' panel, we wanted to find out how estimates of parameters would vary with smaller subsets of this optimal group. We do not know and do not have the means to ascertain to what extent our conclusions would have been different had we used 50 or even 100 panel members.

For the present study, an empirical approach based on resampling from a finite population of observations was used. Further analysis using generalisability and decision studies techniques with the same kind of data may bring to light valuable new information on the optimal number of members to include in panels.

CONCLUSION

The results of this study show that it is possible to recruit a limited number of members in a panel for the construction of an SCT scoring key and still obtain reliable data. Using a panel of 20 members is recommended for high stakes examinations, but for other examinations, a panel of any number over 10 members is associated with acceptable reliability.

REFERENCES

1 Fox R. Medical uncertainty revisited. In: Albrecht G, Fitzpatrick R, Scrimshaw S, eds. *Handbook of Social Studies in Health and Medicine.* London: Sage Publications 2000.

2 Schön D. *The Reflective Practitioner: How Professionals Think in Action.* New York: Basic Books 1983.

3 Johnson E. Expertise and decision under uncertainty: performance and process. In: Chi M, Glaser R, Farr M, eds. *The Nature of Expertise.* Hillsdale, New Jersey: Lawrence Erlbaum Associates 1988;209–28.

4 Feltovich P, Barrows H. Issues of generality in medical problem solving. In: Schmidt H, Volder MD, eds. *Tutorials in Problem-based Learning: a New Direction in Teaching the Health Professions.* Assen, the Netherlands: Van Gorcum 1984.

5 Schmidt H, Norman G, Boshuizen HPA. A cognitive perspective on medical expertise: theory and implications. *Acad Med* 1990;**65**:611–21.

6 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.

7 Charlin B, Roy L, Braïlovsky C, Goulet F, van der Vleuten C. The Script Concordance Test: a tool to assess the reflective clinician. *Teach Learn Med* 2000;**12**:189–95.

8 Charlin B, van der Vleuten C. Standardised assessment of reasoning in context of uncertainty: the script

concordance approach. *Eval Health Prof* 2004;**27**: 304–19.

9   Norman G, Swanson D, Case S. Conceptual and methodological issues in studies comparing assessment formats. *Teach Learn Med* 1996;**8**:208–16.

10   Grant J, Marsden P. Primary knowledge, medical education and consultant expertise. *Med Educ* 1988;**22**: 173–9.

11   Elstein A, Shulman L, Sprafka S. *Medical Problem Solving: an Analysis of Clinical Reasoning*. Cambridge, Massachusetts: Harvard University Press 1978.

12   Barrows HS, Feightner JW, Neufeld VR, Norman GR. *Analysis of the Clinical Methods of Medical Students and Physicians*. Hamilton, Ontario: McMaster University 1978.

13   Norman G. Objective measurement of clinical performance. *Med Educ* 1985;**19**:43–7.

14   Norcini J, Shea J, Day S. The use of the aggregate scoring for a recertification examination. *Eval Health Prof* 1990;**13**:241–51.

15   Charlin B, Desaulniers M, Gagnon R, Blouin D, van der Vleuten C. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. *Teach Learn Med* 2002;**14**:148–54.

16   Abi-Risk G, Nasr C, Jacques A, Goulet F, Charlin B. *Test de Concordance de Script en Médecine Familiale*. Montreal: University of Montreal and Quebec Board of Physicians 2001.

17   Verhoeven BH, Steeg AFW, Scherpbier AJJA, Muijtjens AMM, Verwijnen GM, van der Vleuten CPM. Reliability and credibility of an Angoff standard setting procedure in progress testing using graduates as judges. *Med Educ* 1999;**33**:832–7.

18   Braïlovsky 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 exploratory study on the Script Concordance Test. *Med Educ* 2001;**35**:430–6.

19   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.