The Script Concordance Test for clinical reasoning in paediatric medicine: Medical student performance and expert panel reliability

A. Morris & D.E. Campbell

Abstract

Background: This study aimed to determine the correlation between student performance in clinical reasoning on the Script Concordance Test (SCT) and a modified essay question (MEQ) exam in a paediatric teaching block and to measure the intra-rater reliability of the expert scoring panel.

Method: A 65-item assessment was developed using the accepted SCT method and scored against the responses of a panel of 10 general and subspecialty paediatricians. Student scores for the summative modified essay question examination at the end of the child and adolescent health block were compared with the score on the SCT. Intra-expert reliability was measured for the 10 paediatricians on the expert panel.

Results: One hundred and two students completed both the SCT and the MEQ examination, with the correlation coefficient indicating moderate correlation (r = 0.46). The weighted Cohen kappa for the paediatricians on the panel ranged from 0.61–0.86, demonstrating good to excellent intra-rater agreement.

Conclusion: We found that the MEQ is not a reliable means of measuring clinical reasoning of medical students, with only moderate correlation with the SCT, and that alternative methods such as SCT should be considered. Our finding of high reliability for paediatricians on the scoring panel is the first published using this methodology. It suggests that for lower stakes examinations, there is no need to re-test examiners. We do, however, propose that this simple method of assessing intra-rater reliability should be considered for high-stakes medical student examinations.

Keywords: Script Concordance Test, student, clinical reasoning, modified essay question.

University of Sydney

Correspondence:
Dr Anne Morris MBBS MPH FRACP
Children's Hospital at Westmead Clinical School
University of Sydney—Discipline of Paediatrics and Child Health
Locked Bag 4001
Westmead, NSW 2145
Email: anne.morris@health.nsw.gov.au

ISSN 1442-1100

Introduction

During medical training, students are required to master skills and knowledge, and to acquire the ability to problem-solve in an ever-expanding curriculum. Assessment methods often focus on content knowledge rather than synthesis of knowledge or clinical reasoning, largely due to the relative simplicity of assessing content. However, the ability to reason and weigh up variables in a clinical setting forms the cornerstone of good medical practice. Therefore, the requirement both to teach and to assess clinical reasoning is an increasingly important component and focus of many medical schools' curricula and assessment.

For many years, the focus of research related to clinical reasoning has been on the process of reasoning rather than on assessment. Clinical reasoning is difficult to define and has been known by several synonyms: problem solving, decision making and judgment are examples. The ability to reason also requires a knowledge base, therefore recall and processing of that knowledge is inherent in the process of reasoning (Norman, 2005). The Script Concordance Test (SCT) was originally described by Charlin, Roy, Brailovsky, Goulet, F. and van der Vleuten (2000) and was developed in order to improve the ability to assess clinical reasoning in clinicians of varying levels of experience, from student to senior practitioners. It is based on script theory, which suggests that clinicians use "scripts" or networks of knowledge when faced with uncertainty in clinical situations in order to develop an approach to diagnosis and management. SCT is the most common proposed and validated means of testing clinical reasoning to date (Lubarsky, Charlin, Cook, Chalk, & van der Vleuten, 2011), however its use as a summative assessment tool for medical students appears to be relatively limited (Duggan & Charlin, 2012; Kelly, Durning, & Denton, 2012).

The SCT presents a series of short clinical scenarios, each with several subsequent items. Each item consists of a hypothesis (e.g., a diagnosis) and a new piece of information from which the examinee must rate how much more or less likely the original hypothesis becomes with that new information. A five-point scale ranging from -2 (as ruled out or almost ruled out) to +2 (as certain or almost certain) is used (Figure 1). Construction of script concordance questions has been well described (Fournier, Demeester, & Charlin, 2008). A total score is generated for examinees' responses according to the degree to which they agree with the aggregate responses of a panel of experts for each item. For

A mother brings in her 8-month-old daughter who has had a fever for 2 days. She is worried that she may have a serious infection.

If you were thinking of:	And then you find:	The hypothesis becomes:				
Otitis media	Discharge from the ear	-2	-1	0	1	2
Meningitis	The child is fully immunised	-2	-1	0	1	2
Meningitis	A rash	-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 SCT scenario and item.

example, the highest score is achieved for choosing the same response as the majority of the experts and the lowest score for selecting the response chosen by the minority, or none, of the experts.

The composition of the panel of experts has been examined in terms of the reliability of the test in varying panel sizes, but it is unclear, however, to what extent a one-off measurement of an expert's responses is reliable, and there are currently no published studies which have examined intra-expert reliability (Gagnon, Charlin, Coletti, Sauve, E., & van der Vleuten, 2005).

The SCT has been reported to be a valid test of clinical reasoning in several specialties, including neurology, urology and paediatric emergency medicine, and it has been demonstrated that clinical reasoning scores increase with years of clinical experience (Carriere, Gagnon, Charlin, Dowling, & Bordage, 2009; Lubarsky, Chalk, Kazitani, Gagnon, & Charlin, 2009; Sibert et al., 2006). However, there are no published data on its use in paediatric medicine and, more specifically, for medical students in paediatric medicine. In our assessment of graduate medical students studying paediatrics in their final year of medicine, we chose to explore the possibility of replacing written modified essay questions (MEQ), which the authors have found may have low inter-rater reliability in marking, with an SCT. We aimed to compare student performance on the MEQ with their performance on a new paediatric SCT. We also aimed to measure the intra-expert (intra-rater) reliability of the expert panel, which consisted of consultant paediatricians.

Method

Development of test questions

The questions were developed according to guidelines for script concordance questions by Fournier et al. (2008). The assessment consisted of a total of four scenarios and 69 items. The case scenarios were based upon the existing MEQs with which the SCT were being compared in order to minimise confounding content knowledge differences (Figure 1). Within each case scenario, the questions were grouped according to subsets of history, examination or investigation results and required the integration of specific clinical information in order to support or refute a diagnostic or management hypothesis. A panel of 10 general and subspecialty paediatricians who were not involved in the SCT question writing undertook the SCT examination on two occasions 3 months apart to generate the answers against which the student answers were compared. Panel members were asked not to discuss test answers with one another in the intervening months. The score was generated using a modified aggregate method based on the modal score (Lubarsky et al., 2011). Cronbach's coefficient alpha for internal reliability of the SCT was calculated using SPSS v19.

Study participants

Medical students undertaking their paediatric rotation were invited to participate in the study and to complete the SCT during the final weeks of their rotation. Those wishing to participate undertook the SCT at the end of a scheduled tutorial after receiving instruction on how to read and answer the SCT questions. The test was taken under examination conditions, and students were given feedback on their performance on the SCT for their own interest and learning.

All students completed the MEQ summative assessment routinely conducted in the final week of the rotation. The MEQ examination consisted of five question booklets, each booklet requiring short answers to questions based on an unfolding clinical scenario. The MEQ was marked by paediatricians involved in teaching the curriculum, using a standardised marking sheet during a single marking workshop. A senior examiner was available for consultation to resolve any uncertainties in marking. The combined score for the five booklets was recorded, and performance on the SCT and the MEQ was compared for those who had completed both tests. Total scores were recorded as a percentage for both examination types and compared using Pearson's correlation coefficient (SPSS v19).

Intra-rater reliability

Experts undertook the same test at two separate intervals 3 months apart. Intra-expert reliability was measured using Cohen's weighted kappa, calculated using http://www.statstodo.com/CohenKappa_Pgm.php and interpreted using the scale of 0.00–0.20, negligible; 0.21–0.40, weak; 0.41–0.60, moderate; 0.61–0.80, good and 0.81–1.00, excellent correlation (Landis & Koch, 1977).

Ethics

Ethics approval was obtained from the University of Sydney Human Research Ethics Committee. No funding was obtained for the study.

Results

Students

Between 2009 and 2010, 366 students completed the Child and Adolescent Health (CAH) curriculum block, with 102 (28%) agreeing to participate in the study and completing both the SCT and the MEQ examinations. The correlation coefficient between the scores (Pearson's r) was 0.46 (Figure 2).

The mean score on the MEQ for students who did not complete the SCT (84.4%) was not significantly different from those who did complete the SCT (86.0%).

Cronbach's alpha for the SCT was 0.6.

Intra-expert agreement

Cohen's weighted kappa ranged from 0.61–0.86. The kappa with standard error and 95% confidence intervals are presented in Figure 3. All experts demonstrated at least "good" intra-rater agreement and two had "excellent" agreement.

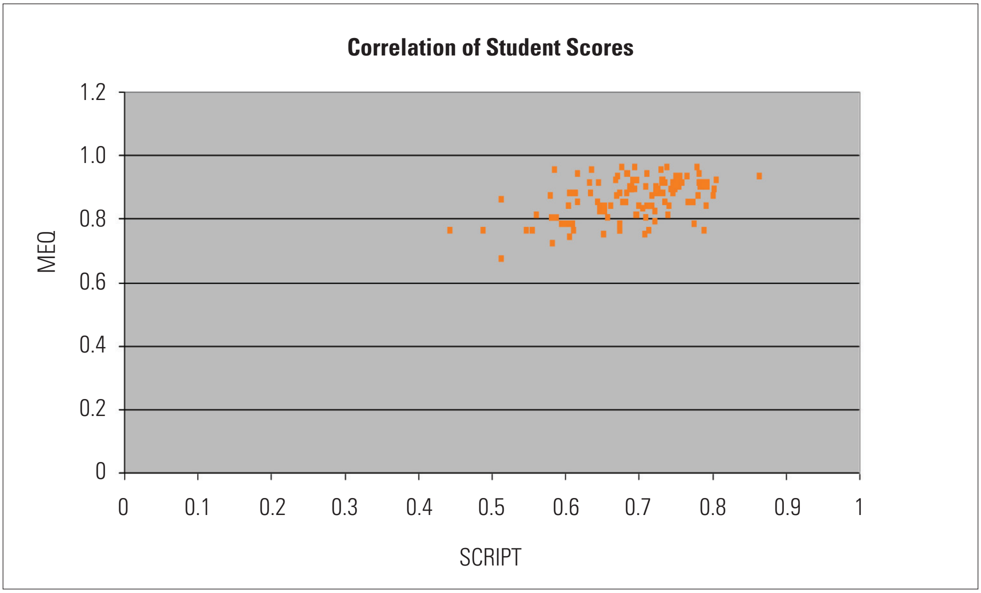

Figure 2. Correlation of student scores for SCT and MEQ (r = 0.46).

Expert #	Cohen's weighted kappa	SE	Lower 95% CI	Error bar value	Upper 95% CI
10	0.861	0.0647	0.7342	0.1268	0.9877
3	0.8198	0.037	0.7472	0.0726	0.8924
1	0.7833	0.0547	0.6761	0.1072	0.8904
6	0.7514	0.0582	0.6373	0.1141	0.8654
8	0.7502	0.0637	0.6252	0.125	0.8751
2	0.7003	0.0518	0.5988	0.1015	0.8018
9	0.6886	0.0679	0.5555	0.1331	0.8216
5	0.6811	0.0574	0.5686	0.1125	0.7936
4	0.6719	0.066	0.5425	0.1294	0.8013
7	0.6124	0.0776	0.4603	0.1521	0.7645

Figure 3. Cohen's weighted kappa for 10 expert panellists.

Discussion

We found that there was, at best, a moderate correlation between students' performance on a modified essay short answer format assessment compared with performance on a script concordance assessment in the same subject domain. This finding suggests that these assessment tasks may, at least to some extent, be measuring different skills in our cohort. This study provides the only published correlation between SCT and MEQ performance in medical students. The degree to which each assessment task is testing clinical reasoning, as opposed to content knowledge or some other attribute, is worthy of further discussion. In a recent review of the evidence supporting the validity of the SCT, Lubarsky et al. (2011) made the assertion that the SCT would be expected to correlate poorly with tests measuring factual recall, such as the MCQ, if the SCT is indeed measuring something other than recall, that is, clinical reasoning. In addition to the few studies comparing SCT with the MCQ in the Lubarsky review, Kelly et al. (2012) found a very poor correlation (r = .22) between the SCT and MCQ for 3rd-year medical students. Duggan and Charlin (2012) have recently reported the introduction of the SCT to replace the MEQ for 5th-year medical students and found that the SCT scores were only weakly to moderately correlated with the scores for MCQ (r = 0.5) and observed structured clinical examination (OSCE) (r = 0.41) respectively, but these authors did not report any correlation with their previously used MEQ assessment (Duggan & Charlin, 2012).

In their review of an exit examination composed of MCQ, MEQ and OSCE, Palmer Duggan, Devitt, & Russell (2010) found that the MEQ primarily tested lower-order cognitive skills, such as recall of knowledge, rather than higher-level cognitive skills, such as interpretation and analysis of data and judgement or reasoning. Based on their review of the literature and analysis of the high-stakes exit examination for medical students at the University of Adelaide, a SCT has been introduced to replace the MEQ at that institution. With the current evidence of the validity of the SCT for measuring clinical reasoning, our finding of only moderate correlation with the MEQ would therefore also suggest that the MEQ may be tending towards assessing content and first-order synthesis rather than complex reasoning. The process of the test rather than the validity of the MEQ itself may be another explanation for the only moderate correlation.

The SCT is not currently used in the Sydney medical program, so students are naïve to its structure and format. This inexperience, compared with their existing experience with the MEQ, may have detracted from their SCT performance. Clearly, if the SCT were to be introduced as a summative examination, it would be appropriate for students to have access to practice questions in order to be familiar with the examination style. In addition, there may have been a difference in student attitude to performance on the SCT as a formative assessment compared with the high-stakes MEQ, which formed a major component of their paediatric term assessment. We did, however, find that all students performed very well in the MEQ, with a mean score of 86% for those sitting the SCT and 84.4% for those who did not, which suggests that prior exposure to the cases in the SCT did not give those students an advantage.

Although there has been a considerable amount published on inter-rater reliability, both in SCT and in general medical student assessment literature, it is not clear how reliable the individual expert is when completing a SCT assessment (Dory, Gagnon, Vanpee, & Charlin, 2012). This is important, as the scoring system of SCT relies upon the performance of the "expert" raters and has always been reported as a single-rater assessment, and we have shown that the internal reliability of experts in our study was high. We used the Cohen's weighted kappa, as this statistic takes into account the important fact that the differences between the responses on the Likert scale may not be linear. For example, a difference of 2 points from 0 to -2 does not necessarily have the same clinical significance as that of -1 to 1. In the latter case, the response has shifted from being less probable to more probable and is thus in complete opposition. Experts may have responded differently over the 3 months due to simple errors, such as misreading a question or an error in data entry, and it is possible that there was some recall bias over the 3-month period. We aimed to minimise bias by panel members not being exposed to the SCT during the 3-month interval, asking them to refrain from discussion of the test with one another and not providing them with the combined panel score until after both episodes of testing. It may be found that if the SCT was administered on a regular basis, expert answers would become more congruent over time if questions were repeated, but this was unable to be determined in this study with only one re-test. It is also possible that if the test were to be composed of questions from a bank of questions that are frequently repeated, examiner familiarity may indeed develop, which has implications for the use of such an instrument in a high-stakes assessment.

The range of Cohen's weighted kappa coefficients (0.61 to 0.86) measured in this study is reassuring since, generally, experts have high intra-rater reliability in their answers. The SCT methodology accepts and indeed relies on there being some variability between examiners as a reflection of the inexact nature of clinical reasoning, but reliability within examiner responses has not been reported (Charlin et al., 2006). It has been argued that the internal reliability of the test is not affected if widely deviant answers by panel members are not removed, provided that the panel size itself is large enough (Gagnon, Lubarsky, Lambert, & Charlin, 2011). This condition is not met in our test for intra-rater reliability, and we therefore propose that for a high stakes or credentialing examination, it would be appropriate to consider having the panel convene on two occasions with a threshold set, below which unreliable examiners would be excluded. This approach would ensure that students are not disadvantaged by being measured against clinicians who demonstrate low internal reliability and would partially address the possibility of students being scored against incorrect answers. We believe that even with the requirement for the panel to provide answers on two occasions for such examinations, the overall burden on the clinical school would be less than that required for the marking of the MEQ.

Our SCT was robust in that it conformed to the method of development previously reported in terms of question style, scoring method and size of the expert panel (Fournier et al., 2008). Our panel size of 10 members is within the range considered by Gagnon et al. (2005) to provide adequate reliability for a lower stakes examination, with a panel of over 20 experts conveying only marginal additional benefit. There has been some debate in SCT literature about the importance of the panel composition. The principle behind the scoring of the SCT is that the panel should contain members with good overall experience of the field rather than subspecialty experience, but in the review by Dory, Gagnon, Vanpee and Charlin (2012), no clear criteria for how to select panel members were described. Our panel contained both general paediatricians and some subspecialty paediatricians, with varying levels of experience. In addition, all panel members were familiar with the paediatric curriculum and participated regularly in teaching and examination of medical students.

We acknowledge that our test items were presented in a manner slightly different from the original SCT format. Our 45-minute examination consisted of a total of 65 items, arising from four scenarios. Gagnon, Charlin, Lambert, Carriere and van der Vleuten (2009) suggested that the optimal number of items per case for good reliability of the test should be up to five items per case scenario, equating to approximately 15–25 case scenarios for a 1-hour, 75-item examination. We used four medical scenarios as a basis for the SCT and sub-grouped the items for each into history, examination findings and investigation results. This resulted in approximately 15 items in total per scenario in order to better reflect the structure of the MEQ, which had four medical questions and one surgical question. The more conventional SCT would have had more scenarios but fewer constituent items. It is possible that this may have affected the reliability of our SCT, but the construction and scoring of each item was consistent with the conventional SCT.

This study adds to the small body of evidence that suggests that the modified essay question may not be a reliable means of measuring clinical reasoning in the examination of medical students and that alternative methods such as the SCT should be considered, at least as an adjunct to current methods for assessing this crucial clinical skill.

Importantly, we have also demonstrated that expert scorers for SCT have a generally high intra-rater reliability, so that for lower-stakes examinations, there is no need to re-test examiners. However, we do propose that this simple method of assessing intra-rater reliability should be considered for high-stakes medical student examinations.

References

Carriere, B., Gagnon, R., Charlin, B., Dowling, 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., Gagnon, R., Pelletier, J., Coletti, M., Abi-Rizk G., Nasr, C., . . . Van der Vleuten, C. (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., Roy, L., Brailovsky, C., Goulet, F., & van der Vleuten, C. (2000). The Script Concordance Test: A tool to assess the reflective clinician. *Teaching and Learning in Medicine*, *12*(4), 189–195.

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.

Duggan, P., & Charlin, C. (2012). Summative assessment of 5th year medical school students' clinical reasoning by script concordance test: Requirements and challenges. *BMC Medical Education*, *12*, 29. doi: 10.1186/1472-6920-12-29

Fournier, J. P., Demeester, A., & Charlin, B. (2008). Script concordance tests: Guidelines for construction. *BMC Medical Informatics and Decision Making*, *8*, 18. doi: 10.1186/1472-6947-8-18

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.

Gagnon, R., Charlin, B., Lambert, C., Carriere, B., & van der Vleuten, C. (2009). Script concordance testing: More cases or more questions? *Advances in Health Science Education Theory and Practice*, *14*, 367–375.

Gagnon, R., Lubarsky, S., Lambert, C., & Charlin, B. (2011). Optimization of answer keys for script concordance testing: Should we exclude deviant panellists, deviant responses or neither? *Advances in Health Sciences Education Theory and Practice*, *16*(5), 601–608.

Kelly, W., Durning, S., & Denton, G. (2012). Comparing a script concordance examination to a multiple-choice examination on a core internal medicine clerkship. *Teaching and Learning in Medicine: An International Journal*, *24*(3), 187–193.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, *33*, 363–374.

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

Lubarsky, S., Charlin, B., Cook, D. A., Chalk, C., & van der Vleuten, C. P. (2011). Script concordance testing: A review of published validity evidence. *Medical Education*, *45*, 329–338.

Norman, G. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education*, *39*, 418–427.

Palmer, E., Duggan, P., Devitt, P., & Russell, R. (2010). The modified essay question: Its exit from the exit examination? *Medical Teacher*, *32*, e300–e307.

Sibert, L., Darmoni, S. J., Dahamma, B., Hellot, F., Weber, J., & Charlin, B. (2006). On line clinical reasoning assessment with script concordance test in urology: Results of a French pilot study. *BMC Medical Education*, *6*, 45. doi: 10.1186/1472-6920-6-45