Development of a Script Concordance Test using an Electronic Voting System

Paul Duggan†
Discipline of Obstetrics and Gynaecology, University of Adelaide

Abstract

This paper describes our experience using an electronic voting system (EVS) to develop a Script Concordance Test (SCT) to assess clinical reasoning in medical students. One hundred and fifteen questions were reviewed and voted on in a single two hour session by eleven specialist Obstetricians and Gynaecologists ("experts"). Questions were categorised by the level of agreement. The experts completed an evaluation questionnaire. In eighty-four (73%) questions experts were mostly or completely in agreement. In fifteen of these the modal response was that the information provided did not affect the hypothesis or proposed action. Responses were widely divergent in thirty-one (27%) questions. The experts felt that the SCT tested clinical reasoning, the EVS was an appropriate method for answering SCT questions, and that the SCT is a suitable method of assessment of medical students and resident staff. Experts felt that the SCT questions were difficult to write, that questions needed to be written avoiding regression to the mean, that absolute phrasing in the answer options was often inappropriate, and that more experience with writing questions were required. EVS is an efficient method for completing the development of SCT questions but has the disadvantage of requiring synchronous participation of experts.

Introduction

The Script Concordance Test (SCT) is a relatively new method of assessing clinical reasoning that may be suitable for undergraduate and postgraduate medical courses and appears to achieve good reliability with 1-1.5 hours of testing time (Charlin et al., 2000, Brailovsky et al., 2001, Meterissian, 2006). Script theory, on which the SCT is based, hypothesises that the decision making processes of skilled clinicians rely on "scripts" or concepts that are developed from a base of knowledge which is continuously modified by clinical encounters. Acceptable reliability has been demonstrated with a minimum of 10 experts in a SCT reference panel (Gagnon et al., 2005).

An example of a SCT question is shown in Table 1. Whilst there may be superficial similarities to a multi choice question (MCQ), there are important differences. The SCT tests clinical competence in authentic situations and requires higher order cognitive skills. These may also be properties of MCQs but frequently are not. In contrast to the MCQ, there is never a single correct answer for a SCT question and this more accurately reflects real clinical practice. The psychometrics of the SCT give partial credit to candidates who choose an answer that has also been chosen by an expert, in proportion to the number of experts who have selected the same response during question development (Charlin et al., 2000). In contrast to MCQs, SCT questions do not have distractors in the answer options. Writing distractors for MCQ questions is very difficult and it is been unusual for authors to be able to achieve more than one or two plausible distractors in the five option MCQ questions used in final medical examinations in the University of Adelaide.

The SCT can be used for diagnostic, formative and summative assessment. The apparent advantages of the SCT include: satisfactory reliability in 1-2 hours of examination time; automated marking; potential for sharing question development and secure question banks between universities and professional colleges (Sibert et al., 2005; Sibert et al., 2006). The disadvantages of the SCT include: a substantial number of experts are required for standard setting; the development of suitable questions is relatively expensive and time-consuming; writing suitable questions can be difficult; development costs for secure question banking, on-line administration, and automated marking will be significant; expert concordance may be different across national and state borders; more work is required on setting pass marks and performing psychometrics; and most of the published work in support of SCT comes from a single unit in Montreal, Canada (Bland et al., 2005; Charlin et al., 2006; Gagnon et al., 2005; Meterissian et al., 2007).

Table 1: An example of a Script Concordance Test question:

A 32 year old woman who is 8 weeks pregnant presents with a 2 day history of mild cramping lower abdominal pain and light vaginal bleeding.

If you are thinking of…	And you find…	Then this hypothesis becomes….
Threatened miscarriage	An ultrasound scan identified a viable intrauterine pregnancy	1 2 3 4 5

1 = the hypothesis is eliminated
2 = the hypothesis becomes less probable
3 = the information has no effect on the hypothesis
4 = the hypothesis is becoming more probable
5 = it can only be this hypothesis

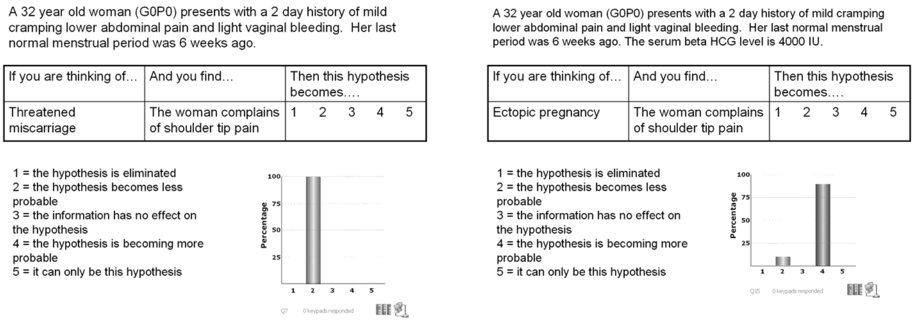
Note: The answer options were configured 1-5 as shown above to conform to the keypad options available with the Electronic Voting System.

Our interest in the SCT stems from problems with the reliability of our existing clinical examinations. Currently, Year 5 MBBS students in the University of Adelaide are assessed in Obstetrics and Gynaecology in two end of term clinical oral examinations. These assessments rely on the availability of senior clinicians who are not employed by the University and who volunteer their time and effort. At present we run these examinations at full capacity based on the availability of examiners and space yet we can not extend the examination time sufficiently to reliably sample candidates' abilities. This existing problem will be seriously exacerbated with the projected swell of student numbers in 2011. The SCT appears to be an attractive alternative to replace one of our current oral assessments, as it appears likely that we can develop sufficient SCT questions to run an examination with high reliability.

This paper describes our initial experience with the development of SCT questions in the Discipline of Obstetrics and Gynaecology and involving senior academics from Adelaide's two medical schools and senior consultant obstetricians and gynaecologists practicing in the public and private sector in Adelaide. We have had substantial experience with an electronic voting system (EVS) in small group tutorials and large group lectures and the use of EVS in the development of SCT questions is a novel approach (Duggan et al., 2007).

Methods

Eleven specialist Obstetricians and Gynaecologists involved in medical student teaching and assessment at the University of Adelaide and the Flinders University participated in a workshop in August 2007. All participants were provided with a copy of a paper describing the SCT (Meterissian, 2006) approximately 8 weeks before the workshop and invited to submit questions for the workshop. A total of 190 SCT questions written by 6 of the participants were submitted. The questions were formatted into a PowerPoint presentation with a single question per slide (Figures 1 and 2). The questions were presented at the workshop in the order of submission. No formal peer review process was undertaken for the questions prior to submission. The participants used an electronic voting system (EVS) to anonymously register their answer to the first 120 questions over a 2 hour period. The EVS responses were displayed immediately for feedback. The first 5 questions were used to test the system and for discussion. Discussion proved to be quite vigorous and time-consuming and it was decided that the remaining 115 questions would be voted on without discussion beforehand. The responses were collated in an Excel spreadsheet and analysed using descriptive statistics.



A 32 year old woman (G0P0) presents with a 2 day history of mild cramping lower abdominal pain and light vaginal bleeding. Her last normal menstrual period was 6 weeks ago.

If you are thinking of…	And you find…	Then this hypothesis becomes….
Threatened miscarriage	The woman complains of shoulder tip pain	1 2 3 4 5

1 = the hypothesis is eliminated
2 = the hypothesis becomes less probable
3 = the information has no effect on the hypothesis
4 = the hypothesis is becoming more probable
5 = it can only be this hypothesis

A 32 year old woman (G0P0) presents with a 2 day history of mild cramping lower abdominal pain and light vaginal bleeding. Her last normal menstrual period was 6 weeks ago. The serum beta HCG level is 4000 IU.

If you are thinking of…	And you find…	Then this hypothesis becomes….
Ectopic pregnancy	The woman complains of shoulder tip pain	1 2 3 4 5

1 = the hypothesis is eliminated
2 = the hypothesis becomes less probable
3 = the information has no effect on the hypothesis
4 = the hypothesis is becoming more probable
5 = it can only be this hypothesis

Figure 1. An example of an outlier (a single selection of option "2" on the right slide) that is most likely due to error in option selection using the EVS keypad. This conclusion is drawn because the significance of shoulder tip pain in the clinical situations described is widely appreciated, as evidenced by 100% agreement regarding its significance in the question on the left. The question on the left would not be used because there is unanimous agreement.

Questions were categorised into 4 groups:

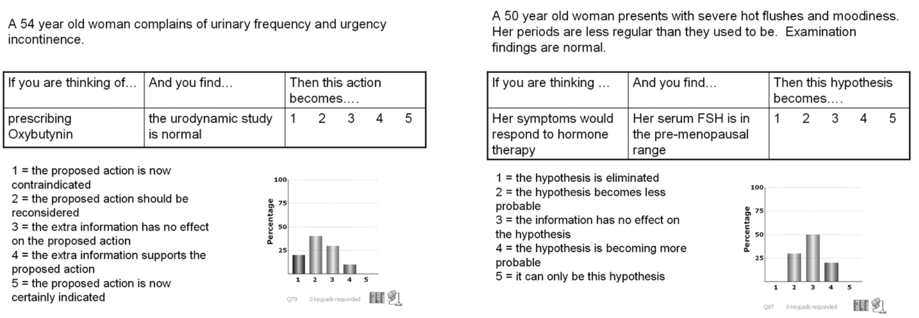
1. total agreement: where all voters voted for a single option (this would then be unsuitable as a SCT question)
2. clearly in agreement: where all voters voted 1-2, 2-3, 3-4 or 4-5 for the question.
3. mostly in agreement: where 2. was satisfied apart from a single outlier.
4. mostly in disagreement: where voting was spread across most or all of the range.

The questions were reviewed to identify the characteristics that resulted in agreement or disagreement. The participants completed an evaluation questionnaire with 9 Likert type questions (7 point range, where 7=strongly agree, 1=strongly disagree, 4=undecided) and a free response question.

Results

Inter-rater agreement

Of the 115 questions reviewed, in ten (8.7%) there was total agreement, in thirty-three (28.7%) there was clear agreement, in forty-one (35.6%) examiners were mostly in agreement and in thirty-one (27%) examiners were mostly in disagreement. Thus, reasonable agreement was reached in eighty-four (73%) questions. In fifteen of these eighty-four questions (18%) the modal response was "3", i.e. that the information provided did not affect the proposed hypothesis or action. Review of questions in which there was a single outlier indicated that the most likely explanation for the outlier was accidental selection of the wrong button on the hand held responder (Figure 1). Where responses were widely divergent in some cases the question was inappropriate or poorly written, in other cases the question appeared unambiguous and relevant, yet there were genuine differences of opinion. (Figure 2).



A 54 year old woman complains of urinary frequency and urgency incontinence.

If you are thinking of…	And you find…	Then this action becomes….
prescribing Oxybutynin	the urodynamic study is normal	1 2 3 4 5

1 = the proposed action is now contraindicated
2 = the proposed action should be reconsidered
3 = the extra information has no effect on the proposed action
4 = the extra information supports the proposed action
5 = the proposed action is now certainly indicated

A 50 year old woman presents with severe hot flushes and moodiness. Her periods are less regular than they used to be. Examination findings are normal.

If you are thinking …	And you find…	Then this hypothesis becomes….
Her symptoms would respond to hormone therapy	Her serum FSH is in the pre-menopausal range	1 2 3 4 5

1 = the hypothesis is eliminated
2 = the hypothesis becomes less probable
3 = the information has no effect on the hypothesis
4 = the hypothesis is becoming more probable
5 = it can only be this hypothesis

Figure 2. The responses to these questions show a divergence of opinion. However, sub-specialists would insist that there is a correct answer. For the question on the left, urodynamic studies have a significant false negative rate and it is reasonable to prescribe Oxybutynin in this situation. For the question on the right, serum FSH is not recommended as a test in this context as it fluctuates widely from day to day. However, clearly other experts who are not sub-specialists disagree and would likely teach their students accordingly. These are examples of where a Script Concordance Test question that awards partial credit in proportion to expert responses may be fairer than a single correct answer multi-choice question.

Questionnaire Likert responses

The examiners were in agreement that the SCT tests clinical reasoning, the electronic voting system is an appropriate method for answering questions, that they would like to be involved in more meetings like this, the SCT questions were easy to understand, the SCT questions are suitable for assessment of medical students and resident staff, and disagreed with the statement "the SCT questions were easy to write".

Questionnaire free responses

The free responses were as follows:

- More experience required with writing unambiguous questions
- Needs more development
- Consider remote examiner review on web site - will solve timing considerations
- Rather exhausting after 100 questions
- Need further sessions to write SCT - with direct group discussion
- Need to avoid regression to mean i.e. 3 - more 1 and 5 questions
- Avoid phrasing in 1 and 5 that is absolute - better the diagnosis is very likely or mostly excluded

- A very interesting new concept (to me). I really think there is something of value in this.
- It requires staff and student teaching/familiarisation.
- The stems should be brief.

Discussion

A novel feature of this work was the use of an electronic voting system (EVS) to display questions and record expert responses. This generally worked well although there was evidence of occasional loss of attention with some clearly aberrant responses from 1 or 2 experts that were best explained by accidental pressing of the wrong button. The literature does not indicate how such obvious errors should be dealt with but it is clearly not appropriate to award credit to a response that would be dangerous. It would be best to have such questions reviewed and standardised again before their use. An important positive feature of the EVS was that experts could see the divergence of opinion in real time. This made the session more interesting than it otherwise might have been and this was an efficient way of getting through a relatively large number of questions in a relatively short time. Getting through 115 questions in 2 hours and having automatic collation of results would not have been achievable without the EVS or an equivalent, computer-based technology. The most significant important negative feature of the use of the EVS was the need to have experts assembled together at one time in the one facility. Although 11 may seem a small number, it exceeded the then full time clinical academic staff complement in the Discipline at both South Australian medical schools. There are only approximately 90 specialist obstetricians and gynaecologists throughout South Australia. Thus, arranging a quorum of 10 is a challenging thing to achieve for such a small group. This problem could potentially be alleviated by developing on-line satellite linkages to various sites each equipped with EVS technology but that would not remove the need for continuing synchronous involvement of the experts. This is a particularly important issue given the need for significant on-going development of our SCT questions and we have decided that the next iteration will be asynchronous, on-line development of questions. This will require development of a secure web based database. The next iteration will also include other disciplines within the medical school that are interested in SCT and will share the web based resource with our discipline. We expect in time that the SCT will become a significant part of the assessment in the MBBS program as a whole.

There were a large number of divergent responses. In some of these cases the question was inappropriate or poorly written and in the future we will include an option for experts to indicate that they were not happy with the wording or relevance of the question and thus would not vote. However, in other cases the questions were not ambiguous or poorly written and were quite relevant and the difference of opinion reflected genuine differences in opinion or practice. This raised an interesting philosophical challenge for some of the faculty, particularly the sub-specialists in topics where there were divergent answers. Sub-specialists are clearly of the view that their opinion is the correct one and that other specialists who disagree are wrong. Perhaps this is an important, hidden advantage of the SCT compared with a single correct answer MCQ as partial credit would be awarded to these “wrong” answers in proportion to the number of experts who selected that response. This would seem to be fairer to our MBBS candidates most of whom are not taught by sub-specialists. However, it seems prudent in the further development of SCT questions to allow “experts” not to vote on a particular question and instead to select an option such as “not my area of expertise”.

The original SCT work describes full credit for the modal response and partial credit for other “correct” responses with the partial credit proportional to the number of experts who chose the other (non-modal) options. (Brailovsky et al., 2001) An alternative psychometric approach is to apply 100% of the credit weighting for the question to the modal response. (Bland et al., 2005) More work will be required on the psychometrics once we have better developed our questions. Our results also suggest that intra-rater variability should be assessed in future studies.

The problem of regression to the mean (i.e. option "3" being the commonest modal response) was mentioned in the written feedback and also verbally during the session. Whilst this occurred in only 18% of the questions in which there was acceptable agreement it is challenging to write questions where the "correct" response is to the left or to the right of option 3. Improving the questions by changing the extreme descriptors (1 = the hypothesis is eliminated and 5 = it can only be this hypothesis) to (1 = the hypothesis is most unlikely or eliminated and 5 = the hypothesis is most likely or it can only be this hypothesis) would increase the number of option 1 and option 5 responses, as most experts can think of a reason why an hypothesis can not be absolutely correct or absolutely incorrect.

Conclusions

The Script Concordance Test (SCT) shows promise as a test of clinical reasoning but requires substantial effort and care in its development. The Electronic Voting System (EVS) is an excellent method of efficiently completing SCT question development but has the disadvantage of requiring synchronous participation of experts. Poor inter-rater agreement often reflects poorly written questions but in some cases it is due to expert disagreement, which is a normal part of clinical practice and is managed in the SCT by awarding partial credit in proportion to the number of experts who select a particular response. This may be a fairer method of assessing medical students, who are taught by a range of experts, than using single correct answer MCQ questions.

Acknowledgements

Dr David Morris provided the EVS software, reformatted the questions and ran the EVS session.

References

Bland Andrew C., Kreiter Clarence D., Gordon Joel A. (2005). The Psychometric Properties of Five Scoring Methods Applied to the Script Concordance Test. *Acad Med.*; 80:395–399.

Brailovsky C, Charlin B, Beausoleil S, et al., (2001). Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. *Medical Education*;35:430-436

Charlin, Bernard, Brailovsky Carlos, Roy Louise et al., (2000) The Script Concordance Test: A Tool to Assess the Reflective Clinician. *Teach Learn Med*; 12(4), 189-195

Charlin Bernard, Gagnon Robert, Pelletier Jean, 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 doi:10.1111/j.1365-2929.2006.02541.x

Duggan Paul, Palmer Edward, Devitt Peter (2007). Electronic voting to encourage interactive lectures: a randomised trial. *BMC Medical Education*; 7:25 doi:10.1186/1472-6920-7-25

Gagnon R, CharlinB, Coletti M, et al., (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 doi:10.1111/j.1365-2929.2005.02092.x

Meterissian, Sarkis H (2006). A Novel Method of Assessing Clinical Reasoning in Surgical Residents *Surg Innov*; 13; 115 DOI: 10.1177/1553350606291042

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

Sibert Louis, Darmoni Stefan J, Dahamna Badisse, et al., (2005). Online clinical reasoning assessment with the Script Concordance test: a feasibility study. *BMC Medical Informatics and Decision Making*, 5:18 doi:10.1186/1472-6947-5-18

Sibert Louis, Darmoni Stefan J, Dahamna Badisse, et al., (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.

Duggan, P. (2008). Development of a script concordance test using an electronic voting system. *ergo*, 1 (1), pp. 35-41.

† Corresponding author: paul.duggan@adelaide.edu.au