REFERENCES

1 Dornan T. *Experienced Based Learning. Learning Clinical Medicine in Workplaces*. PhD Thesis. Maastricht: Maastricht University, 2006.
2 Smith ES, Tallentire VR, Cameron HS, Wood SM. The effect of contributing to patient care on medical students' workplace learning. *Med Educ* 2013;**47**:1184–96.
3 Hattie J, Timperley H. The power of feedback. *Rev Educ Res* 2007;**77** (1):81–112.
4 Ashford SJ, Blatt R, VandeWalle D. Reflections on the looking glass: a review of research on feedback-seeking behaviour in organisations. *J Manag* 2003;**39** (6):773–800.
5 Bok HGJ, Teunissen PW, Spruijt A, Fokkema JPI, van Beukelen P, Jaarsma DADC, van der Vleuten CPM. Clarifying students' feedback-seeking behaviour in clinical clerkships. *Med Educ* 2013;**47**:282–91.
6 Dornan T, Boshuizen H, King N, Scherpbier AJJA. Experience-based learning: a model linking the processes and outcomes of medical students' workplace learning. *Med Educ* 2007;**41**: 84–91.
7 Teunissen PW, Stapel DA, van der Vleuten CPM, Scherpbier AJJA, Boor K, Scheele F. Who wants feedback? An investigation of the variables influencing residents' feedback-seeking behaviour in relation to night shifts. *Acad Med* 2009;**84** (7):910–7.
8 van der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, van Tartwijk J. A model for programmatic assessment fit for purpose. *Med Teach* 2012;**34**:205–14.
9 Slootweg I, Lombarts K, van der Vleuten CPM, Mann K, Jacobs J, Scherpbier AJJA. Clinical teachers' views on how teaching teams deliver and manage residency training. *Med Teach* 2013;**35**:46–52.
10 Cooke M, Irby DM, Sullivan W, Ludmerer KM. American medical education 100 years after the Flexner report. *New Engl J Med* 2006;**355** :1339–44.
11 Driessen EW, van Tartwijk J, Govaerts M, Teunissen PW, van der Vleuten CPM. The use of programmatic assessment in the clinical workplace: a Maastricht case report. *Med Teach* 2012;**34**:226–31.

Scoring the Script Concordance Test: not a black and white issue

Stuart Lubarsky, Robert Gagnon & Bernard Charlin

We wish to thank Lineberry *et al.*[1] for their insightful review of the Script Concordance Test (SCT). Before delving into the important issues raised by the authors in their article,[1] it is instructive to review the fundamental principles upon which the script concordance approach rests. First, script concordance is informed by an established theory of knowledge organisation derived from cognitive psychology, which describes the manner in which knowledge is encoded, structured and mobilised for use during clinical encounters. Second, its item format reflects the way information is processed in authentic clinical problem-solving situations, in accordance with empirical research in the domain of clinical reasoning. Third, it introduces the uncertainty and ambiguity characteristic of daily practice into the realm of assessment (from which uncertainty is customarily excluded). Finally, by contrast with most conventional methods of assessment in current use, the SCT employs a scoring system that accounts for the variability of responses of experienced health professionals to clinical situations, a system referred to as 'aggregate scoring'.

Montreal, QC, Canada

Correspondence: Stuart Lubarsky, Centre for Medical Education, McGill University, Lady Meredith House, 1110 Pine Avenue West, Room 204, Montreal, QC H3A 1A3, Canada. Tel: 00 1 514 934 8060; E-mail: stuart.lubarsky@mcgill.ca

doi: 10.1111/medu.12362

Script concordance is informed by an established theory of knowledge organisation derived from cognitive psychology.

The authors' critique[1] of the script concordance method centres mainly on this latter principle. According to Lineberry and colleagues,[1] accepting divergent responses from members of the reference panel is unjustifiable practice. To illustrate their argument, the authors cite an example

in which the panellists are uniformly divided on the opposing values +1 and −1.[1] We would agree that such diametrically bipolar items, if and when they do crop up, should be carefully reviewed prior to test administration, and discarded if warranted. As with any assessment format, common sense and sound judgement prevail in the development of any particular iteration of a test.

However, one should be careful not to discard the baby with the bathwater. Previous work on the SCT has shown that a certain degree of response variability among the members of an SCT reference panel is, in fact, a key determinant of the test's discriminatory power.[2] Given that statistically significant discrimination (with a moderate to large effect size) between groups of participants is achieved in study after study, it seems defensible to maintain the SCT's aggregate scoring system under the assumption that response differences lie not necessarily in errors or misconceptions, but potentially in richness of thinking about clinical cases.

Response variability among the members of an SCT reference panel is a key determinant of the test's discriminatory power.

Lineberry *et al.*[1] oppose conflicting responses to SCT questions on logical grounds, stating that 'a single piece of information … cannot simultaneously make the hypothesis more *and* less likely'. This assertion is true for any individual respondent, but not necessarily true for a *group* of respondents. The magnitude and direction of impact of a bit of clinical information on a given hypothesis may vary for different individuals depending on the way that their knowledge is organised (i.e. in their illness scripts), which, in turn, depends on their unique prior clinical and learning experiences in health care. Further, although it may be true (depending on one's epistemological bent) that 'there is an objectively correct answer for every SCT item',[1] it is equally arguable that in complex, dynamic, real-world situations, objectively correct answers to clinical problems often reveal themselves only with time, and sometimes never do. Indeed, in most circumstances, clinicians are required to make the best judgements they can with the information currently available to them. The SCT is designed to probe the concordance of examinees' judgements under such conditions of uncertainty with the judgements of those who are considered to represent the reference standard (or "gold standard") of decision making in the field?

In most circumstances, clinicians are required to make the best judgements they can with the information available to them.

Just who are these "gold-standard" reasoners in the health sciences? For now, they are who we claim they are. In light of the lack of consensus on what actually constitutes expertise in the health professions, 'experts' are simply those individuals who are reputed to have superior knowledge, clinical acumen and experience in a given domain, and whose opinions are sought (and generally accepted) when challenging cases arise within their field. The SCT makes no greater claim toward defining or recognising expertise; for SCT developers, recruiting an appropriate reference panel of 'experts' for a given test remains a tricky, non-evidence-based endeavour.

What is certain, however, is that judicious selection of panellists is critical. For example, in a study cited by Lineberry *et al.*,[1] an SCT in radiation oncology was devised using a reference panel consisting of the entire population of radiation oncologists in Quebec at the time.[3] Given that many of the panellists recruited were subspecialists within their field, it is not surprising that 27% of residents, who harbour undifferentiated knowledge structures during their training years, scored higher than the panellists, whose knowledge structures are tailored for use in their respective subdomains. In another study referenced in the same article, in which panel members (general surgeons) were strategically chosen for their alignment with the specific content domain of the SCT (general surgery), only 1% of residents performed above the panel mean.[3]

In script concordance testing, judicious selection of panellists is critical.

Whereas the 'Who?' question continues to beguile SCT panel selection, the 'How many?' question leads to more satisfying answers. Gagnon and colleagues have shown that the estimated reliability of SCT scores in relation to panellists' responses decreases as the number of panellists diminishes.[4,5] With smaller panels (< 15 members), the less frequently selected responses are more heavily weighted (and therefore may introduce noise), whereas with larger panels (≥ 15 members) modal responses receive proportionately higher weights compared with infrequent responses. These findings call into question the validity of the exercises undertaken by Lineberry *et al.*,[1] whose conclusions regarding the SCT response process and its relationships to other variables were based on a

reanalysis of a single set of scores obtained from an SCT using a sub-optimal eight-member panel. When it comes to panel composition, size matters.

The estimated reliability of SCT scores in relation to panellists' responses decreases as the number of panellists diminishes.

Lineberry *et al.*[1] rebuff the SCT both for straying from the central tenets of classical test theory (in relying on an aggregate scoring system), and for adhering to them too strictly (in relying on classical test theory's most stalwart index of reliability, Cronbach's alpha coefficient). They reveal strong positivist leanings in their call to replace aggregate scoring with methodology that resorts to 'expert consensus or empirical evidence' for determining the 'correctness' of the test's answers.[1] Yet, as indicators of truth, expert consensus and empirical evidence are hardly immune to error. Consensus-building exercises, for example, are subject to the vagaries of group dynamics, coercion and power hierarchies. And one need only be reminded of the recent controversy over hormone replacement therapy to recognise that even what is considered to be empirical medical evidence sometimes rests on a rocky pedestal.

As indicators of truth, expert consensus and empirical evidence are hardly immune to error.

It is undeniably true, however, that the script concordance approach challenges the usual assumptions of classical test theory, and may require a certain degree of thinking outside the psychometric box. Modern measurement models, such as item response theory, may offer promising alternatives. Blais *et al.*, [6] for example, applied the multi-facet Rasch model (a special case of item response theory) to SCT and demonstrated evidence of validity for scoring a test that awards partial credit to non-consensus responses. Further work on Rasch modelling, as well as factor analysis and analysis of expected payoffs for examinees who engage in guesswork, looms large on the script concordance research horizon.

For many learners in the health professions, the hidden curriculum conceals a niggling paradox. In the pre-clinical years, assessments of student knowledge tend to exhibit a bias toward the black and white, extolling single correct answers to well-defined problems. Upon entering the clinical sphere, however, students must quickly learn that there are as many shades of grey as there are patients, and that even experienced clinicians often interpret data, make judgements and respond to complex or uncertain clinical situations in ways that vary (and sometimes conflict). The SCT is an assessment tool designed to acknowledge this important reality in clinical practice; attempts to refine its scoring system would, we hope, remain true to its essence.

REFERENCES

1 Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. *Med Educ* 2013;**47**:1175–83.
2 Charlin B, Gagnon R, Pelletier J, Coletti M, Abi-Rizk G, Nasr C, van der Vleuten C. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Med Educ* 2006;**40**:848–54.
3 Charlin B, Gagnon R, Lubarsky S, Lambert C, Meterissian S, Chalk C, Goudreau J, van der Vleuten C. Assessment in the context of uncertainty using the script concordance test: more meaning for scores. *Teach Learn Med* 2010;**22** (3):180–6.
4 Gagnon R, Charlin B, Coletti M, Sauve E, van der Vleuten C. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? *Med Educ* 2005;**39**:284–91.
5 Gagnon R, Lubarsky S, Lambert C, Charlin B. Optimisation of answer keys for script concordance testing: should we exclude deviant respondents, deviant responses, or neither? *Adv Health Sci Educ* 2011;**16** (5):601–8.
6 Blais JG, Charlin B, Grondin J, Lambert C, Loye N, Gagnon R. Estimation du degré d'accord entre des experts lors du calibrage d'un test de concordance de script avec le modèle à facettes de Rasch. [Estimating inter-expert agreement in a script concordance test calibration using the many-facet Rasch model.] In: Raiche G, Paquette-Cote K, Magis D, eds. *Des Mécanismes pour Assurer la Validité de l'Interprétation de la Mesure en Education*, Vol. **2**. Montreal, QC: University of Montreal Press 2011.

