Threats to validity in the use and interpretation of script concordance test scores

Matthew Lineberry,[1] Clarence D Kreiter[2] & Georges Bordage[1]

CONTEXT Recent reviews have claimed that the script concordance test (SCT) methodology generally produces reliable and valid assessments of clinical reasoning and that the SCT may soon be suitable for high-stakes testing.

OBJECTIVES This study is intended to describe three major threats to the validity of the SCT not yet considered in prior research and to illustrate the severity of these threats.

METHODS We conducted a review of SCT reports available through the Web of Science database. Additionally, we reanalysed scores from a previously published SCT administration to explore issues related to standard SCT scoring practice.

RESULTS Firstly, the predominant method for aggregate and partial credit scoring of SCTs introduces logical inconsistencies in the scoring key. Secondly, our literature review shows that SCT reliability studies have generally ignored inter-panel, inter-panellist and test–retest measurement error. Instead, studies have focused on observed levels of coefficient alpha, which is neither an informative index of internal structure nor a comprehensive index of reliability for SCT scores. As such, claims that SCT scores show acceptable reliability are premature. Finally, SCT criteria for item inclusion, in concert with a statistical artefact of the SCT format, cause anchors at the extremes of the scale to have less expected credit than anchors near or at the midpoint. Consequently, SCT scores are likely to reflect construct-irrelevant differences in examinees' response styles. This makes the test susceptible to bias against candidates who endorse extreme scale anchors more readily; it also makes two construct-irrelevant test taking strategies extremely effective. In our reanalysis, we found that examinees could drastically increase their scores by never endorsing extreme scale points. Furthermore, examinees who simply endorsed the scale midpoint for every item would still have outperformed most examinees who used the scale as it is intended.

CONCLUSIONS Given the severity of these threats, we conclude that aggregate scoring of SCTs cannot be recommended. Recommendations for revisions of SCT methodology are discussed.

Medical Education 2013: 47: 1175–1183
doi: 10.1111/medu.12283

Discuss ideas arising from the article at www.mededuc.com 'discuss'

[1]Department of Medical Education, University of Illinois at Chicago, Chicago, Illinois, USA
[2]Department of Family Medicine, University of Iowa, Iowa City, Iowa, USA

Correspondence: Dr Matthew Lineberry, Department of Medical Education, University of Illinois at Chicago, 808 South Wood Street, M/C 591, CME 973, Chicago, Illinois 60612, USA. Tel: 00 1 312-355-5418; E-mail: MattL@uic.edu

INTRODUCTION

Script concordance tests (SCTs) use written clinical cases featuring elements of uncertainty to assess how well examinees' interpretations of key findings correspond to the interpretations given by a panel of experienced clinicians.[1] For a given case, each SCT item first proposes a hypothesised diagnosis, investigation or management approach, and then provides a finding that might confirm, disconfirm or have no bearing on the hypothesis. Examinees indicate how the new information alters the likelihood or appropriateness of the hypothesis using a 5-point Likert-type scale, ranging from −2 ('strongly refutes') to +2 ('strongly confirms'). To set the scoring key, a panel of experienced clinicians completes each item and the modal panellist response is considered the fully correct response. A unique aspect of SCTs is that non-modal panellist responses are used to award partial credit, a practice referred to as 'aggregate scoring'. For example, suppose that out of seven panellists, four believe that the information on an item refutes the hypothesis somewhat (−1) and three believe it supports the hypothesis somewhat (+1). In this instance, an examinee answer of −1 would receive full credit, an answer of +1 would receive 75% credit (i.e. reflecting the ratio of non-modal to modal panellists for that response), and all other responses would receive no credit.

A recent review[2] deemed the validity evidence for SCTs generally supportive, and another review[3] tentatively suggested their use in high-stakes assessment. Although these reviews are insightful, three important issues which pose serious threats to the valid interpretation of SCT scores remain unaddressed. We consider these issues and offer directions for resolving them.

CONTENT VALIDITY OF SCT SCORING

Typically, content validity evidence focuses on a test's domain coverage as it compares with the relative importance of the domain given the test's intended purpose and intended interpretations. However, the joint *Standards on Educational and Psychological Testing*[4] state that content validity also encompasses 'procedures regarding administration and scoring'. Accordingly, the SCT's unique scoring method must be based on sound logic to support content validity arguments.

The premises for aggregate scoring of SCTs are outlined by Charlin *et al.*[5]: 'Professionals in similar situations do not collect exactly the same data and do not follow the same paths of thought. Professionals also show substantial variation in performance on any particular real or simulated case.' The implicit conclusion drawn from these premises is that professionals' disagreements about data interpretation on SCT items represent a valid divergence of professional opinion.[5]

The premises are true, but the conclusion does not follow from them. Experts may indeed use different means to make decisions, often arriving at the same decision. However, this does not imply that experts correctly use any *one* particular means in opposed ways, as occurs when experts disagree about how a particular piece of information bears on a particular hypothesis on an SCT item. In the argument above, valid professional divergence is cited in the premises but is not carried directly into the conclusion. The conclusion also ignores the second premise. If professionals do not all perform equally well, SCT panels should not include those who, although experienced, may nonetheless hold false factual knowledge or misconceptions in their clinical reasoning, such as the misconceptions identified among cardiologists regarding the ultrastructural basis of myocardial failure.[6]

The scoring methodology of the SCT is readily challenged by a *reductio ad absurdum* argument, which draws attention to the common occurrence in which one group of panellists believes a piece of information supports the hypothesis and another group believes the opposite. Even if it is not yet widely known, there is an objectively correct answer for every SCT item in an actuarial sense. By the laws of probability, a single piece of information will make a hypothesis more likely, less likely, or will have no bearing; it cannot simultaneously make the hypothesis more *and* less likely.[7] When clinicians disagree so fundamentally, the simplest explanation is that we don't know what the right answer is; one camp is wrong and possibly both are. Alternatively, panellists may be making different assumptions about unspecified case particulars or using different incomplete arguments. Anecdotally, we observed the latter in an examination of panellists' responses to a case on the Practicum Script Concordance Test (www.script.edu.es); it is likely that these represent but a few of a variety of reasons for panellists' disagreements across items, none of which justify retaining those disagreements in the scoring key. In one SCT study, asking particularly experienced experts to discard 'widely

deviant responses' led to the exclusion of 6.4% of panellists' responses.[8] We argue that any panellist disagreement about the basic effect of information on the likelihood of a hypothesis that cannot be resolved through discussion renders the item in question unacceptable for use in educational achievement testing.

The incongruity of panellists' diametric opposition on an item is compounded when SCTs award no credit to examinees who respond with a mark of '0' ('neither refutes nor supports') on such items. An examinee with perfect knowledge of experts' contradictory opinions about that particular item could reasonably surmise that splitting the difference is the only way to convey his or her acknowledgement of the divided expert opinion.

Finally, a further scoring incongruity is apparent in that examinees can outperform the majority of panellists on an SCT, which challenges the criteria for whom to include on panels. Charlin *et al.*[9] clarified the relative standing of examinees and panellists by transforming scores onto a common metric, with the panel mean transformed to be equal to 80 with a standard deviation (SD) of 5. An examinee score of 80 is thus 'easily interpretable as "equal to the level of the panel mean"'.[9] This makes it apparent that a number of examinees – as many as 27% of residents in one SCT administration[9] – score *above* the panel mean. By definition, such examinees are more concordant with the panel mean than *most panellists themselves*. As concordance with the panel mean is intended to measure the quality of data interpretation, it could be argued that such examinees should be considered as panellists because they demonstrate superior data interpretation. However, doing so would alter the scoring key, changing everyone's scores and possibly indicating yet another different panel *ad infinitum*. The crux of the issue is that being experienced does not make someone correct on all aspects of data interpretation; some other justification, such as expert consensus or reference to empirical data, is required to establish the correctness of SCT responses.

In sum, SCT scoring methods have fundamental logical inconsistencies, which constitutes a weak foundation for content validity arguments.

RELIABILITY AND INTERNAL STRUCTURE OF SCTS

Reliability refers to the consistency of scores that a measure generates across the various ways scores may be collected and interpreted.[10] Lack of reliability in SCTs can stem from inter-panel or inter-panellist differences, transient particulars about when the measure was administered, lack of coherence among items or cases within a test, and residual error, to name a few. A measure that is unreliable cannot support valid conclusions, which makes it important to logically consider how measurement error may arise, to estimate such errors, and to mitigate errors as efficiently as possible. It is commonly claimed that SCTs with certain methodological features attain satisfactory reliability.[3] However, SCT proponents have not adequately considered inter-panel measurement error and have not considered inter-rater (or 'inter-panellist') errors at all. Additionally, over-reliance on and misinterpretation of coefficient alpha in SCT research reports are likely to have discouraged consideration of potential errors while providing little insight into the internal structure of SCTs.

To evaluate how researchers have approached reliability estimation, we reviewed 77 articles on SCTs, derived from a Web of Science search for available reports published before January 2013 (search terms: topic 'script concordance test' OR title 'script concordance'). Of 41 studies reporting reliability coefficients for SCT administrations, 34 (83%) reported only coefficient alpha or the analogous Kuder–Richardson coefficient of reliability (formula 20 [KR20]). No studies estimated inter-panellist measurement error. Of the seven studies reporting a statistic other than coefficient alpha, two studies used generalisability theory analyses focused only on case or item facets[11,12] and three studies estimated test–retest reliability, with correlations varying from $r = 0.02$ to $r = 0.76$.[13–15] Only one study conducted a generalisability theory analysis modelling both item and occasion facets.[16] For that study's 120-item test, we computed reliability coefficients based on the estimated variance components that were reported; the test–retest reliability computed from the generalisability study output was only $r = 0.45$, and the overall generalisability coefficient for the 120-item test, administered twice, was only 0.40.

For any given SCT, two related but distinct notions of reliability are relevant: (i) the reliability of the panel an estimate of true expert opinion, and (ii) the reliability of examinees' scores as an estimate of their correspondence to true expert opinion. The former is a unique aspect of SCTs unaccounted for by classical test theory (CTT). That is, in typical single-correct-answer testing the scoring key is assumed to be a perfectly reliable and valid indicator of

truth, given that it is based on expert consensus or empirical evidence. Proponents of SCTs have not extended CTT to account for this unique aspect of their tests; consequently, they have used reliability estimation approaches that are not conceptually appropriate to address issues of measurement error in panels.

In CTT, an individual's true score on an item is defined as his or her theoretical expected score – a mean – across infinite hypothetical administrations of the item. A response given by an examinee is a sample from that mean, assumed to partially reflect the examinee's true score as well as some measurement error. Classical test theory and its later theoretical extensions are, at their foundation, theories for making inferences related to that mean, relying on known properties of the sample mean (e.g. that it is an unbiased estimator of the population mean).

For rhetorical purposes, suppose that all experts' responses to SCT items do reflect valid differences of opinion. As such, panels are meant to represent samples not of the mean opinion of experts but of the *distributions* of the opinions expected from the population of relevant experts on each item. The nature and shape of the true distribution are unspecified for any given item: for one item, distribution may be tri-modal; for another it may be uniform, and for yet another it may be normal.

We are unaware of any psychometric theory sophisticated enough to guide estimation of the adequacy or inadequacy of sampling from 'any population frequency distribution that may be observed across a 5-point scale'. Is the sample distribution an unbiased estimator of the population distribution for all possible distributions? If so, in what sense? Is the estimated mode unbiased, or the number of modes? How may true versus error variance be analysed? Without a theory that appropriately describes the sampling distribution of the distribution and its associated inferences, it is impossible to estimate panel error.

Studies have tried to address panel error tangentially by examining the observed reliability of examinees' total test scores obtained with the use of different panels when examinees are held constant. For instance, Gagnon *et al.*[8] reported an analysis in which 20 panels of 15 panellists each were randomly re-sampled from a pool of 45 panelists. They observed that coefficient alpha varied little across panels and concluded that the SCT methodology 'appears to be robust, resistant to deviant answers or members'.[8] However, similarity of coefficient alpha across panels is not informative because two scales with the same coefficient alpha can be measuring very different sets of constructs.[17] Indeed, in Gagnon *et al.*[8]' study, the standardised difference between residents' and panellists' scores – a known-groups validity statistic – fluctuated drastically between panels, from as low as 0.4 SD to as high as 1.8 SD. An even larger range might have been observed if panellists had been sampled without replacement from a larger pool. Thus the validity of their test[8] for distinguishing between experienced and inexperienced respondents varied greatly across panels, calling into question which construct or constructs any particular panel was measuring and making comparisons of reliability across panels inappropriate.

As well as panel-level sampling inconsistencies, individual panellists can be inconsistent in their responses to SCTs; for instance, a panellist might give different responses to the same SCT items if he or she is re-tested a few weeks or months after an original administration. Although some SCT researchers acknowledge that panellists' disagreements might partially reflect measurement error,[3,5] the method effectively ignores this possibility. No attempt is made to analyse the valid versus the invalid components of panellist variance and all panellists' answers are retained in the scoring key. Thus, along with panel error, individual panellist error in SCTs is an unknown quantity.

The vast majority of SCT research evaluates reliability using coefficient alpha. However, a large coefficient alpha only indicates that examinees' responses are *internally consistent* and most variance is attributable to general or group factors rather than particular items. As an index of internal structure, coefficient alpha is largely uninformative for tests with many items such as the SCT. For instance, an 18-item test measuring three uncorrelated factors still yields a large coefficient alpha.[18] Given that SCTs often feature many dozens of items, they may be assessing an even larger number of distinct ability factors and coefficient alpha would not alert us to this. This challenges assertions that the SCT assesses a single common construct.[2]

It is commonly thought that coefficient alpha also estimates test–retest reliability.[19] However, recent scholarship has shown that alpha can considerably underestimate test–retest reliability. The primary issue is that coefficient alpha assumes item-level errors are uncorrelated.[20] For panellists and exami-

nees completing an SCT on a single occasion, transient test–retest errors may occur as a result of random fluctuations in mood, mental sharpness, recent events, and so on. Such transient errors cause item errors to be correlated, inflating coefficient alpha.[21] The few SCT studies reviewed that actually administered SCTs on multiple occasions make it apparent that SCT test–retest errors are far from trivial.[13–16]

The sum of systematic test–retest error, inter-panel error, inter-panellist error, inconsistencies among items and cases, and interactions among these sources of error is likely to be substantial. Coefficient alpha reflects only one of these sources of error and thus gives a very incomplete, upwardly biased assessment of reliability. As such, the internal structure and reliability of the SCT are largely unknown. Factor analyses are needed to assess the number of constructs being measured by SCTs, and more thorough generalisability theory studies will be needed to simultaneously estimate the magnitude of various measurement errors.[22] However, there is a conundrum as to how panellist variance should be considered in such generalisability theory studies. If all panellist variance is supposedly valid, panellist error components are designated as fixed (rather than random). This amounts to asserting that the panel for any given SCT includes all the possible panellists of interest (or perfectly represents those panellists), which is not tenable. However, designating panellist error components as random facets would redefine the test as a measure of the examinee's deviation from the average panellist response for an item. Whether the average of multiple diverging opinions for a given item is a meaningful construct is debatable. Later, we propose an alternative scoring method for which this conundrum does not arise.

THE RESPONSE PROCESS IN SCTS AND RELATIONSHIPS TO OTHER VARIABLES

Artefacts of SCT methodology can readily lead to an unintended consequence: unequal expected credit across the scale anchors (−2 to +2), with anchors at or near the midpoint being associated with greater expected points. This can lead SCT scores to correlate with a construct they arguably should not relate to: namely, examinee response style. By extension, this can cause scores to be biased against particular groups and makes the test highly susceptible to score inflation attributable to coaching.

Three phenomena account for this issue. Firstly, during item writing, items warranting the certainty implied by −2 or +2 ('*strongly* refute' or '*strongly* support') are less likely to be considered sufficiently ambiguous and non-factual, both of which characteristics are key for SCT items. Such items are thus less likely to be written in the first place. Secondly, the fact that the scale is truncated at ±2 causes partial credit to regress to the scale midpoint. Figure 1 illustrates this for a hypothetical 5-item test in which each scale point is the fully correct option for 1 item, thereby satisfying the admonition of Fournier

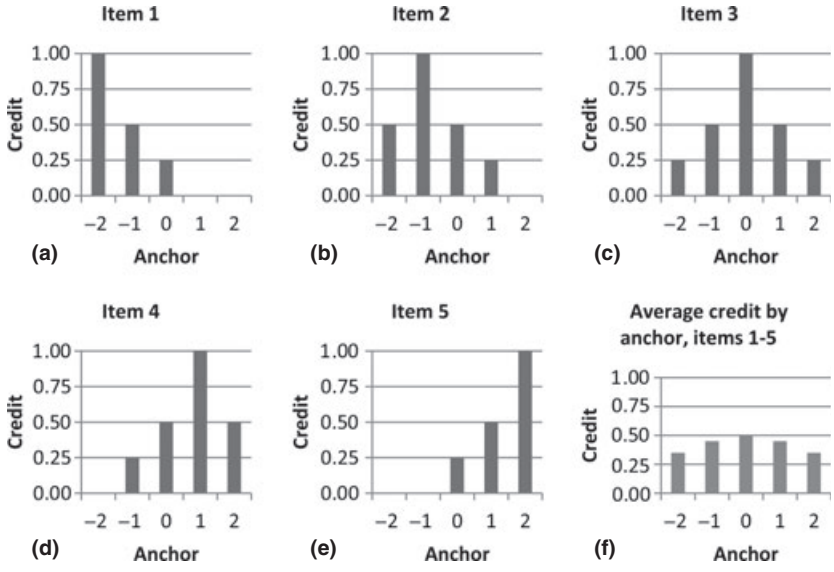

Figure 1 Credit on a hypothetical 5-item test in which each scale point is the fully correct option for 1 item, by (a–e) item and (f) as mean credit by anchor

et al.[23] to 'spread answers over each anchor of the Likert scale'. When an extreme scale point is the modal response, non-modal responses can *only* result in credit being pulled toward the midpoint. Across items, the expected value of guessing the midpoint is thus greater than that of guessing extremes.

Thirdly, standard test refinement processes are likely to favour the elimination of items with correct responses near the extremes. As we have noted, items for which −2 or +2 are correct are more likely to be straightforward, factually based questions with clear answers. Such items will thus be easier and will predominantly reflect examinees' factual knowledge. Consequently, those items will be less discriminating and less internally consistent with the remainder of the test, making them likely candidates for removal. Indeed, SCT items that incur the least variability among panellists tend to be easy and to discriminate poorly; they also more frequently feature correct answers at the extreme scale points.[5] As the removal of items from SCTs is often considerable,[3] biased removal of items at the scale extremes may significantly affect the distribution of expected credit across the scale, and may also result in poor sampling from the tests' intended content domains. At the same time, failure to remove such items would compromise the internal consistency and discrimination of the test.

To investigate whether these issues are manifest in actual SCTs, we reanalysed (with permission) de-individuated data from a previously published SCT report by Bland *et al.*[24] Specifically, for that study's panel and associated scoring key, we computed expected credit across the 5 scale anchors. In the original study, a 50-item SCT was developed and administered to 16 experts and 85 residents.[24] Eight experts served as the panel for score setting and the other eight were analysed as examinees; no items were removed for low item–total correlations.

Using the aggregate scoring method, we plotted the distribution of expected credit for each scale anchor across all 50 items, along with a frequency distribution of the panel's modal responses across all items (Fig. 2a,b). Both show a peaked distribution whereby endorsing the midpoint was associated with greater expected points than selecting the extreme anchors. To investigate how standard test refinement strategies might affect this distribution, we removed 10 items that had item–total correlations (r) of < 0.1, resulting in a 40-item scale. Expected credit was distributed similarly for this reduced test (Fig. 2c,d).

One reason this is problematic is that across a variety of assessments, respondents show persistent and largely construct-irrelevant differences in the way they respond on Likert-type scales; some favour

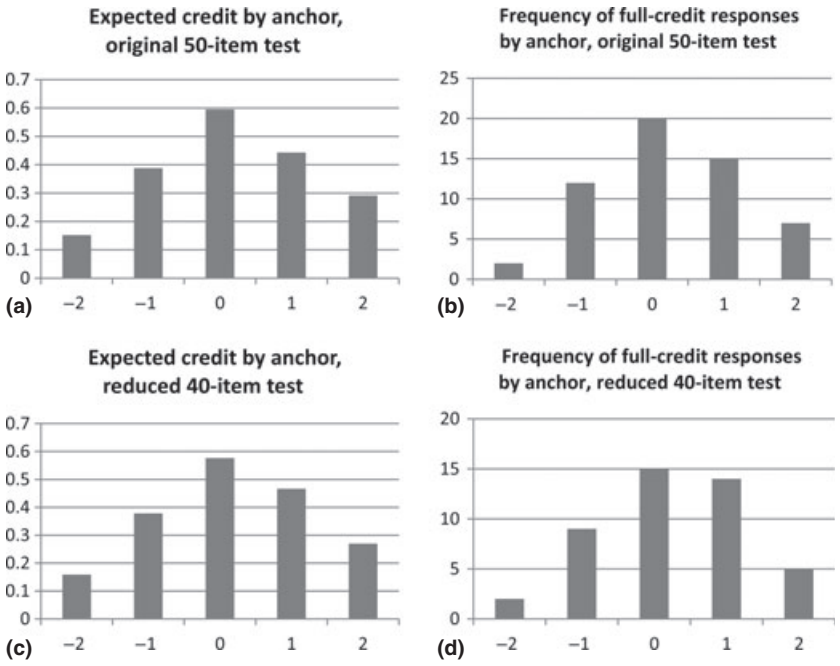

Figure 2 Distributions of expected credit for each scale anchor and frequency distributions of the panel's modal responses across items in the script concordance test reported by Bland *et al.*[24] (a, b) across all items and (c, d) across the 40 items that remained after 10 items with item–total correlations (r) of < 0.1 had been removed

extreme anchors, whereas others rarely endorse them.[25] To the extent that these differences are indeed construct-irrelevant, the validity of the SCT is reduced. Far more concerning is that differences in response styles can be associated with respondents' race and ethnicity. For instance, Asian respondents have been found to avoid extreme scale points, whereas Hispanic and African American respondents tend to favour them.[26–28] Given the methodological artefacts outlined above, members of racial or ethnic groups which tend to favour extreme responses may score lower on SCTs, potentially resulting in adverse impact if scores are used to make high-stakes decisions. Indeed, across two studies of a test of situational judgement similar to the SCT, African American examinees scored as much as 0.56 SD lower than White examinees. Standardising examinees' scores within-persons in those studies (i.e. statistically correcting for examinee variance in response style) largely eliminated those racial differences while simultaneously increasing the test's criterion-related validity.[29] It was not possible for us to run subgroup analyses for different groups in our reanalysis because participant demographics were not recorded. Nonetheless, we believe this issue warrants careful attention in SCT research.

Unequal expected credit also makes the test highly susceptible to response style coaching. Examinees would be wise to avoid extreme responses altogether and to guess values near the midpoint when they are uncertain about an item's correct answer. In situational judgement tests similar to the SCT, this strategy drastically inflates examinee scores by as much as 2.20 SD.[29,30] In order to simulate what might happen if SCT examinees were to use this strategy, we rescored the data for each examinee in Bland *et al.*' study[24] as if the examinee had been coached to completely avoid the extreme anchors; responses of −2 and +2 were recoded as −1 and +1, respectively. For the full 50-item test, the score inflation that results from this simulated coaching is profound ($d = 1.07$) (Table 1a). As described earlier, we also removed 10 items with item–total correlations (r) of < 0.1 and re-ran analyses in order to evaluate how test refinement interacts with the coaching effect. The reduced 40-item test produced a higher coefficient alpha, as expected. However, the effect of coaching for this reduced test was even greater ($d = 1.51$) (Table 1b).

Although the strategy is slightly less effective than that of avoiding the extremes, a hypothetical examinee who simply responds with '0' to every item on the 50-item test would earn credit of 59.5%, more than 10 percentage points higher than the average examinee's score in this sample. For the abbreviated 40-item test, an examinee who gives only responses of '0' would earn 57.6% credit, roughly 8% points higher than the average for that test. Thus examinees who deliberately ignore some or even most SCT response options can outperform examinees who use the scale as it is intended.

CONCLUSIONS

The vulnerabilities described here represent serious threats to the valid use of SCT scores. More specifically, almost all of these threats stem from the practice of aggregate scoring. Such scoring precludes the coherent estimation and enhancement of reliability, allows irrational values and clinicians' misconceptions to enter the scoring key, implicitly discourages the seeking of empirical support for the scoring key (as there is supposedly no single correct answer for any item), and risks

Table 1 Effects of simulated coaching on script concordance test scores extracted from Bland et al.[24]

	(a) Original test (50 items)		**(b) Test with item–total r < *0.1* removed (40 items)**	
	Original scores	**Simulated coaching**	**Original scores**	**Simulated coaching**
Coefficient alpha	0.76	0.71	0.79	0.74
Mean credit earned,%	49.2	60.8	49.5	69.2
Standard deviation	12.4	9.2	14.6	11.4
Cohen's d	1.07		1.51	

allowing construct-irrelevant differences in response style to influence scores, possibly resulting in bias or large differences in the scores of coached and uncoached examinees. As such, we conclude that aggregate scoring cannot be recommended.

Replacing the aggregate scoring methodology with a consensus- and evidence-based scoring methodology using a 3-point scale ('refutes', 'neither refutes nor supports' and 'supports') would immediately address most of these issues and facilitate the resolution of the one issue not immediately resolved: namely, over-reliance on coefficient alpha. Such tests would allow the straightforward estimation of reliability in all its complexity, avoid contradictory values in the scoring key, avoid (but not be immune to) embedding clinicians' misconceptions into the scoring key, encourage reference to empirical data when such data are available, and not be subject to construct-irrelevant differences in examinee response style. Reports of SCT-like tests using consensus-based scoring exist[31] and at least one of these also uses a 3-point scale.[32] We did not find any SCT reports that rely on empirical data to justify their scoring keys.

Although we advocate that the SCT should be revised for assessment purposes, it may be insightful to use panellists' responses to SCTs in their current form as a policy-capturing instrument and to explore items for which clinicians' interpretations of data are highly variable. Such instances may illuminate items for which there is a genuine lack of empirical evidence, or for which empirical evidence exists but is not widely known among clinicians.

The SCT methodology has drawn attention to how examinees interpret information in simulated clinical cases. Retaining that focus, while omitting the problematic aspects of the methodology, constitutes a sound way forward for assessment.

Contributors: ML was chiefly responsible for the study conception and design, data acquisition and interpretation, and the drafting of the manuscript. CDK and GB made substantial contributions to the study conception, data acquisition and interpretation, and the drafting of the paper. All authors contributed to the critical revision of the manuscript and approved the final version for publication.
Acknowledgement: None.
Funding: None.
Conflicts of interest: None.
Ethical approval: Not required.

REFERENCES

1 Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance test: a tool to assess the reflective clinician. *Teach Learn Med* 2000;**12** (4):189–95.
2 Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CPM. Script concordance testing: a review of published validity evidence. *Med Educ* 2011;**45** (4):329–38.
3 Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: a systematic review. *Med Educ* 2012;**46**:552–63.
4 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: AERA 1999.
5 Charlin B, Gagnon R, Pelletier J, Coletti M, Abi-Rizk G, Nasr C, Sauve E, van der Vleuten CPM. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Med Educ* 2006;**40** (9):848–54.
6 Coulson RL, Feltovich PJ, Spiro RJ. Foundations of a misunderstanding of the ultrastructural basis of myocardial failure: a reciprocation network of oversimplifications. *J Med Philos* 1989;**14**:109–46.
7 Kreiter CD. Commentary: the response process validity of a script concordance test item. *Adv Health Sci Educ Theory Pract* 2012;**17**:7–9.
8 Gagnon R, Lubarsky S, Lambert C, Charlin B. Optimisation of answer keys for script concordance testing: should we exclude deviant panellists, deviant responses, or neither? *Adv Health Sci Educ Theory Pract* 2011;**16** (5):601–8.
9 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.
10 Gleser GC, Cronbach LJ, Rajaratnam R. Generalisability of scores influenced by multiple sources of variance. *Psychometrika* 1965;**30**:395–418.
11 Brailovsky C, Charlin B, Beausoleil S, Cote 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 experimental study on the script concordance test. *Med Educ* 2001;**35** (5):430–6.
12 Gagnon R, Charlin B, Lambert C, Carriere B, van der Vleuten C. Script concordance testing: more cases or more questions? *Adv Health Sci Educ Theory Pract* 2009;**14** (3):367–75.
13 Holloway R, Nesbit K, Bordley D, Noyes K. Teaching and evaluating first and second year medical students' practice of evidence-based medicine. *Med Educ* 2004;**38** (8):868–78.
14 Giguere A, Labrecque M, Njoya M, Thivierge R, Legare F. Development of PRIDe: a tool to assess

physicians' preference of role in clinical decision making. *Patient Educ Couns* 2012;**88** (2):277–83.
15 Park AJ, Barber MD, Bent AE, Dooley YT, Dancz C, Sutkin G, Jelovsek E. Assessment of intraoperative judgement during gynaecologic surgery using the Script Concordance Test. *Am J Obstet Gynecol* 2010;**203** (3):240. e1–6.
16 Ramaekers S, Kremer W, Pilot A, van Beukelen P, van Keulen H. Assessment of competence in clinical reasoning and decision making under uncertainty: the script concordance test method. *Assess Eval Higher Educ* 2010;**35** (6):661–73.
17 Schmitt N. Uses and abuses of coefficient alpha. *Psychol Assess* 1996;**8**:350–3.
18 Cortina JM. What is coefficient alpha? An examination of theory and applications. *J Appl Psychol* 1993;**78**:98–104.
19 Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ* 2004;**38**:1006–12.
20 Green SB, Yang Y. Commentary on coefficient alpha: a cautionary tale. *Psychometrika* 2009;**74**:121–35.
21 Becker G. How important is transient error in estimating reliability? Going beyond simulation studies. *Psychol Methods* 2000;**5**:370–9.
22 Brennan RL. *Generalizability Theory*. New York, NY: Springer 2001.
23 Fournier J, Demeester A, Charlin B. Script concordance tests: guidelines for construction. *BMC Med Inform Decis Mak* 2008;**8** (1):18–24.
24 Bland AC, Kreiter CD, Gordon JA. The psychometric properties of five scoring methods applied to the script concordance test. *Acad Med* 2005;**80** (4):395–9.
25 Eid M, Rauber M. Detecting measurement invariance in organisational surveys. *Eur J Psychol Assess* 2000;**16**:20–30.
26 Bachmann JG, O'Malley PM. Yea-saying, nay-saying, and going to extremes: Black–White differences in response styles. *Public Opin Q* 1984;**48**:491–509.
27 Lee C, Green RT. Cross-cultural examination of the Fishbein behavioural intentions model. *J Int Bus Stud* 1991;**22**:289–305.
28 Marín G, Gamba RJ, Marín BV. Extreme response style and acquiescence among Hispanics: the role of acculturation and education. *J Cross Cultur Psychol* 1992;**23**:498–509.
29 McDaniel MA, Psotka J, Legree PJ, Yost AP, Weekley JA. Toward an understanding of situational judgement item validity and group differences. *J Appl Psychol* 2011;**96** (2):327–36.
30 Cullen MJ, Sackett PR, Lievens F. Threats to the operational use of situational judgement tests in the college admission process. *Int J Self Assess* 2006;**14** (2):142–55.
31 Williams RG, Klamen DL, White CB, Petrusa E, Fincher RE, Whitfield CF, Shatzer JH, McCarty T, Miller BM. Tracking development of clinical reasoning ability across five medical schools using a progress test. *Acad Med* 2011;**9**:1148–54.
32 Kelly W, Durning S, Denton G. Comparing a script concordance examination to a multiple-choice examination on a core internal medicine clerkship. *Teach Learn Med* 2012;**24**:187–93.

Received 27 March 2013; accepted for publication 6 June 2013