Threats to Validity in the Use and Interpretation of Script Concordance Test Scores

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

[1] Department of Medical Education, University of Illinois at Chicago, Chicago, IL, USA

[2] Department of Family Medicine, University of Iowa, Iowa City, IA, USA

Address all correspondence to:

Matthew Lineberry, Ph.D.

Department of Medical Education, University of Illinois at Chicago (M/C 591)

808 S. Wood St., CME 973

Chicago, IL 60612

MattL@uic.edu

Phone: 312-355-5418

Fax: 312-413-2048

Key Words: Aggregate scoring; bias; clinical reasoning; coaching; coefficient alpha; reliability; response style; script concordance test; situational judgment test; validity

Word count (main body)**:** 3853

Abstract

Recent reviews have claimed that the Script Concordance Test (SCT) methodology generally produces reliable and valid assessments of clinical reasoning. We describe three major validity threats not yet considered in prior research. First, the predominant method for aggregate and partial credit scoring of SCTs introduces logical inconsistencies in the scoring key. Second, reliability studies of SCTs have generally ignored inter-panel, inter-panelist, 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 artifact of its 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 style. This makes the test susceptible to bias against groups that endorse extreme scale anchors more readily; it also makes the test susceptible to score inflation due to coaching. In a re-analysis of existing SCT data, we found that simulating a strategy whereby examinees never endorse extreme scale points resulted in considerable score inflation (d = 1.51), and examinees that simply endorsed the scale midpoint for every item would still have outperformed most examinees that used the scale as intended. Given the severity of these threats, we conclude that aggregate scoring cannot be recommended. Recommendations for revisions of SCT methodology are discussed.

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 hypothesized 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 on 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 panelist response is considered the fully-correct response. A unique aspect of SCTs is that non-modal panelist responses are used to award partial credit, a practice referred to as *aggregate scoring*. For instance, suppose that out of seven panelists, four believe that the information on an item refutes the hypothesis somewhat (-1) and three believe it supports the hypothesis somewhat (+1). As such, an examinee answer of -1 would receive full credit, +1 would receive ¾ credit (i.e., the ratio of non-modal to modal panelists for that response), and all other responses would receive zero 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. While these reviews are insightful, three important and unaddressed issues remain which pose serious threats to valid interpretation of SCT scores. 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 compared to the relative importance of those domains 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" (p. 11). 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 and colleagues:[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" (p. 849). The implicit conclusion drawn from these premises is that professionals' disagreements about data interpretation on SCT items represent valid divergence of professional opinion.

The premises are true, but the conclusion does not follow from them. Experts may indeed use different means to arrive at 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 the responses of professionals who, while experienced, may nonetheless hold false factual knowledge or misconceptions in

their clinical reasoning, such as cardiologists' misconceptions regarding the ultra-structural basis of myocardial failure.[6]

SCTs' scoring methodology is readily challenged by a *reductio ad absurdum* argument, drawing attention to the common occurrence by which one group of panelists 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 laws of probability, a single piece of information either makes a hypothesis more likely, less likely, or has 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. Alternately, panelists may be making different assumptions about unspecified case particulars or using different incomplete arguments. Anecdotally, we observed the latter in an examination of panelists' responses to a case on the Practicum Script Concordance Test (www.script.edu.es); likely these are but a few of a variety of reasons for panelists' 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 excluding 6.4% of panelists' responses.[8] We argue that any panelist disagreement about the basic effect of information on a hypothesis' likelihood that cannot be resolved through discussion renders the item in question unacceptable for use in educational achievement testing.

The incongruity of panelists' diametric opposition on an item is compounded when SCTs award no credit to examinees who respond "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 their concordance to the divided expert opinion.

Finally, a further scoring incongruity is apparent whereby examinees can outperform the majority of panelists on an SCT, challenging the criteria for whom to include on panels. Charlin and colleagues clarified the relative standing of examinees and panelists by transforming scores onto a common metric, with the panel mean transformed to be equal to 80 with a standard deviation of 5.[9] An examinee score of 80 is thus "easily interpretable as 'equal to the level of the panel mean'" (p. 186). 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 panelists themselves.* Since concordance with the panel mean is intended to measure quality of data interpretation, one could argue that such examinees should be considered panelists, since 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 correctness of SCT responses.

In sum, SCT scoring methods have fundamental logical inconsistencies, constituting 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-panelist 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, making it important to logically consider how measurement error may arise, estimate such errors, and 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-panelist") errors at all. Additionally, over-reliance and misinterpretation of coefficient alpha in SCT research reports has likely discouraged consideration of potential errors while providing little insight into SCTs' internal structure.

To evaluate how researchers have approached reliability estimation, we reviewed 77 SCT articles, 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 KR20. No studies estimated inter-panelist measurement error. Of the seven studies reporting a statistic other than coefficient alpha, two studies used generalizability theory analyses focused only on case and/or item facets[11,12] and three studies estimated test-retest reliability, with correlations varying from r = 0.02-0.76.[13-15] Only one study conducted a generalizability theory analysis modeling both item and occasion facets.[16] For that study's 120-item test, the test-retest reliability

computed from their generalizability study output is only $r = 0.45$, and the overall generalizability coefficient for the 120-item test, administered twice, is only 0.40.

For any given SCT, two related but distinct notions of reliability are relevant: (1) the reliability of the panel as an estimate of true expert opinion and (2) 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 its basis on expert consensus and/or empirical evidence. SCT proponents 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 their 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 their true score as well as some measurement error. CTT 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 as samples not of the mean opinion of experts but of the *distributions* of opinions expected from the population of relevant experts on each item. The nature and shape of the true distribution is unspecified for any

given item; for one item, it may be tri-modal, for another it may be uniform, for yet another it may be normal, etc.

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 – that is, is the estimated mode unbiased, and/or the number of modes, etc.? How may true vs. error variance be analyzed? 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 resulting from use of different panels, holding examinees constant. For instance, Gagnon and colleagues reported an analysis in which 20 panels of 15 panelists each were randomly re-sampled from a pool of 45 panelists.[8] 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" (p. 607). However, similarity of coefficient alpha across panels is not informative, since two scales with the same coefficient alpha can be measuring very different sets of constructs.[17] Indeed, in Gagnon and colleagues' study, the standardized difference between residents' and panelists' scores – a known-groups validity statistic – fluctuated drastically between panels, from as low as 0.4 standard deviation to as high as 1.8 standard deviations.[8] An even larger range would likely have been observed if panelists had been sampled without replacement from a larger pool. Thus the validity of their test for distinguishing experienced vs.

inexperienced respondents varied greatly across panels, calling into question what construct or constructs any particular panel was measuring and making comparisons of reliability across panels inappropriate.

Along with panel-level sampling inconsistencies, individual panelists can be inconsistent in their responses to SCTs; for instance, a panelist might give different responses to the same SCT items if re-tested a few weeks or months after an original administration. While some SCT researchers acknowledge that panelists' disagreements might partially reflect measurement error,[3,5] the method effectively ignores this possibility. No attempt is made to analyze the valid vs. invalid components of panelist variance and all panelists' answers are retained in the scoring key. Thus along with panel error, individual panelist 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*, with most variance attributable to general and/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 could 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 panelists and examinees completing an SCT on a single occasion, transient test-retest errors can occur due to random fluctuations in mood, mental sharpness, recent events, etc. Such transient errors cause item errors to be correlated, inflating coefficient alpha.[21] Based on the few SCT reports mentioned earlier that actually administered SCTs on multiple occasions, it is apparent that SCT test-retest errors are far from trivial.

The sum of systematic test-retest error, inter-panel error, inter-panelist error, inconsistencies among items and/or cases, and interactions among these sources of error is likely to be substantial. Coefficient alpha only reflects 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 SCTs is largely unknown. Factor analyses are needed to assess the number of constructs being measured by SCTs, and more thorough generalizability theory studies would be needed to simultaneously estimate the magnitude of various measurement errors.[22] However, there is a conundrum as to how panelist variance should be considered in such generalizability theory studies. If all panelist variance is supposedly valid, one designates panelist error components as fixed (vs. random). This amounts to asserting that the panel for any given SCT includes all the possible panelists of interest (or perfectly represents those panelists), which is not tenable. However, designating panelist error components as random facets would redefine the test as a measure of examinees' deviation from the average panelist 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 alternate scoring method for which this conundrum does not arise.

SCTs' response process and relationships to other variables

Artifacts 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 to SCT scores correlating with a construct they arguably should not relate to, namely, examinees' response style. By extension, this can cause scores to be biased against particular groups and makes the test highly susceptible to score inflation due to coaching.

Three phenomena account for this issue. First, 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 being key criteria for SCT items. Such items are thus less likely to be written in the first place. Second, the fact that the scale is truncated at ±2 causes partial credit to regress to the scale midpoint. Figure 1**Error! Reference source not found.** illustrates this for a hypothetical 5-item test in which each scale point is the fully-correct option for one item, thereby satisfying the admonition of Fournier and colleagues[23] to "spread answers over each anchor of the Likert scale" (p. 20). 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 guessing extremes.

Third, standard test refinement processes are likely to favor elimination of items with correct responses near the extremes. As discussed above, 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 with the least variability among panelists tend to be easy and to discriminate poorly; they also more frequently feature correct answers at the extreme scale points.[5] Since removal of items in SCTs is often considerable,[3] biased removal of items at the scale extremes could significantly affect the distribution of expected credit across the scale, and might 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 the aforementioned issues are manifest in actual SCTs, we re-analyzed (with permission) de-individuated data from a previously published SCT report by Bland and colleagues.[24] Specifically, we computed expected credit across their scale anchors. In the original study, a 50-item SCT was developed and administered to 16 experts and 85 residents. Eight experts served as the panel for score setting, with the other eight analyzed as examinees. In their study, no items were removed due to 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 (Figure 2, A & B). Both show a peaked distribution whereby endorsing the midpoint is associated with greater expected points than the extreme anchors. To investigate how standard test refinement strategies might affect this distribution, we removed 10 items that had item-total correlations less than $r = 0.1$, resulting in a 40-item scale. Expected credit was distributed similarly for this reduced test (Figure 2, C & 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 favor extreme anchors while others rarely endorse them.[25] To the extent these differences are indeed construct-irrelevant, SCTs' validity 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, while Hispanic and African-American respondents tend to favor them.[26-28] Given the methodological artifacts outlined above, racial or ethnic groups which tend to favor 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 judgment similar to the SCT, African-American examinees scored as much as 0.56 standard deviations lower than White examinees. Standardizing examinees' scores within-persons in those studies (i.e., statistically correcting for examinees' 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 sub-group analyses for different groups in our re-analysis 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 uncertain about an item's correct answer. For situational judgment tests similar to the SCT, this strategy drastically inflates examinee scores, as much as 2.20 *SD*.[29,30] In order to simulate what would happen if SCT examinees used this strategy, we rescored each examinee's data from Bland and colleagues' study as if they 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, score inflation due to this simulated coaching is profound (d = 1.07; Table 1, part A). As described earlier, we also removed 10 items with item-total correlations less than r = 0.1 and re-ran analyses in order to evaluate how test refinement interacts with the coaching effect. The reduced 40-item test has a higher coefficient alpha, as expected. However, the effect of coaching for this reduced test is even greater (d = 1.51; Table 1, part B).

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

Conclusions and Recommendations

The threats described here are serious vulnerabilities for the valid use of SCT scores. More specifically, almost all of these threats stem from the practice of aggregate scoring. Such scoring precludes coherent estimation and enhancement of reliability, allows irrational values and clinicians' misconceptions to enter the scoring key, implicitly discourages seeking empirical support for the scoring key (since there is supposedly no single correct answer for any item), and risks allowing construct-irrelevant differences in response style to influence scores, possibly resulting in bias and/or large differences

between coached vs. uncoached examinees' scores. 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", or "supports") would immediately address most of these issues and facilitate resolution of the one issue not immediately resolved – namely, over-reliance on coefficient alpha. Such tests would allow 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 it is available, and not be subject to construct-irrelevant differences in examinee response style. Reports of SCT-like tests using consensus-based scoring exist,[31] at least one of which also uses a 3-point scale.[32] We did not find any SCT reports that rely on empirical data to justify their scoring keys.

While we advocate revising the SCT for assessment purposes, it may be insightful to use panelists' responses to SCTs in their current form as a policy-capturing instrument, to explore items for which clinicians' data interpretation is 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.

Ethical approval

Archived, non-identifiable data collected by Bland et al. (2005) were re-analyzed for this study. Both the University of Illinois College of Medicine at Peoria and the University of Iowa Roy J. and Lucille A. Carver College of Medicine declared the original study under which data were collected exempt from IRB review.

Acknowledgements

We are most grateful to Dr. Andrew Bland and colleagues for giving us permission to use their data set and to Drs. Rachel Yudkowsky, Yoon-Soo Park, and Jack Boulet for helpful comments on an earlier version of this manuscript.

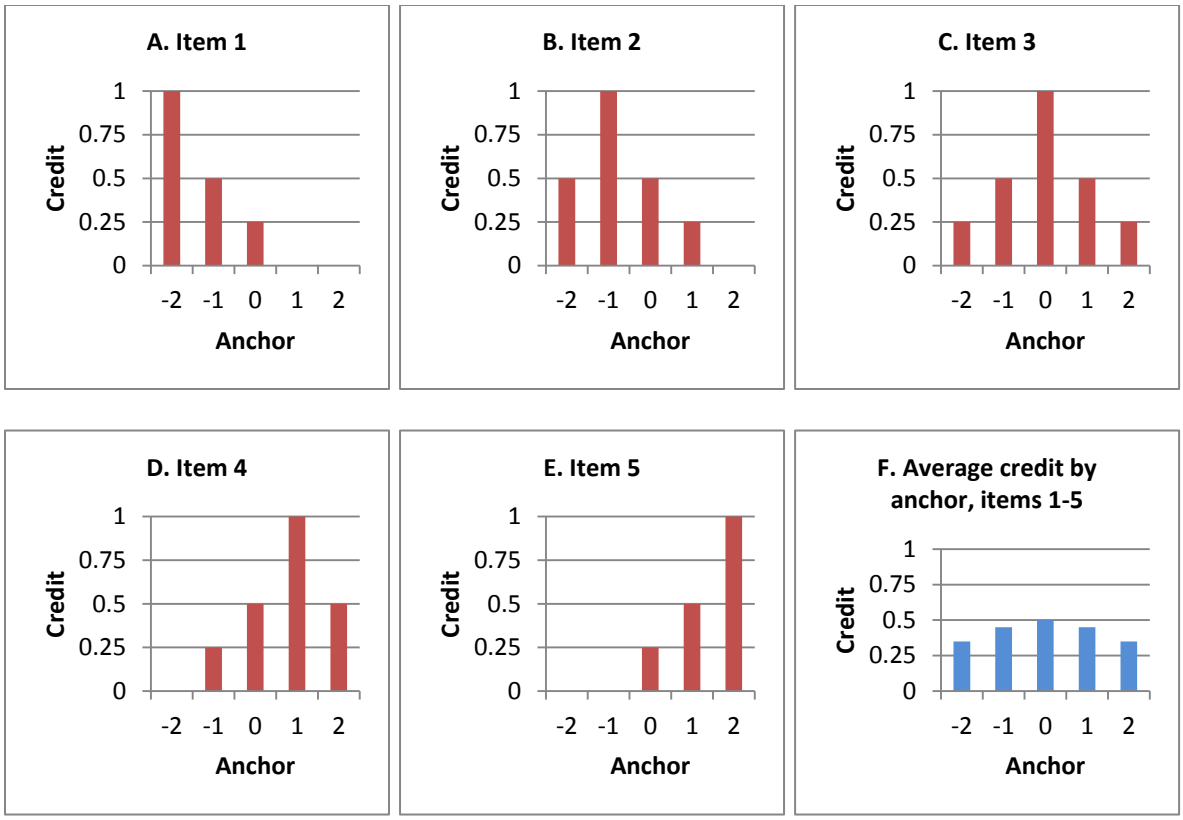

Figure 1

Item scoring keys (graphs A-E) and overall expected credit (graph F) for a hypothetical 5-item Script Concordance Test.

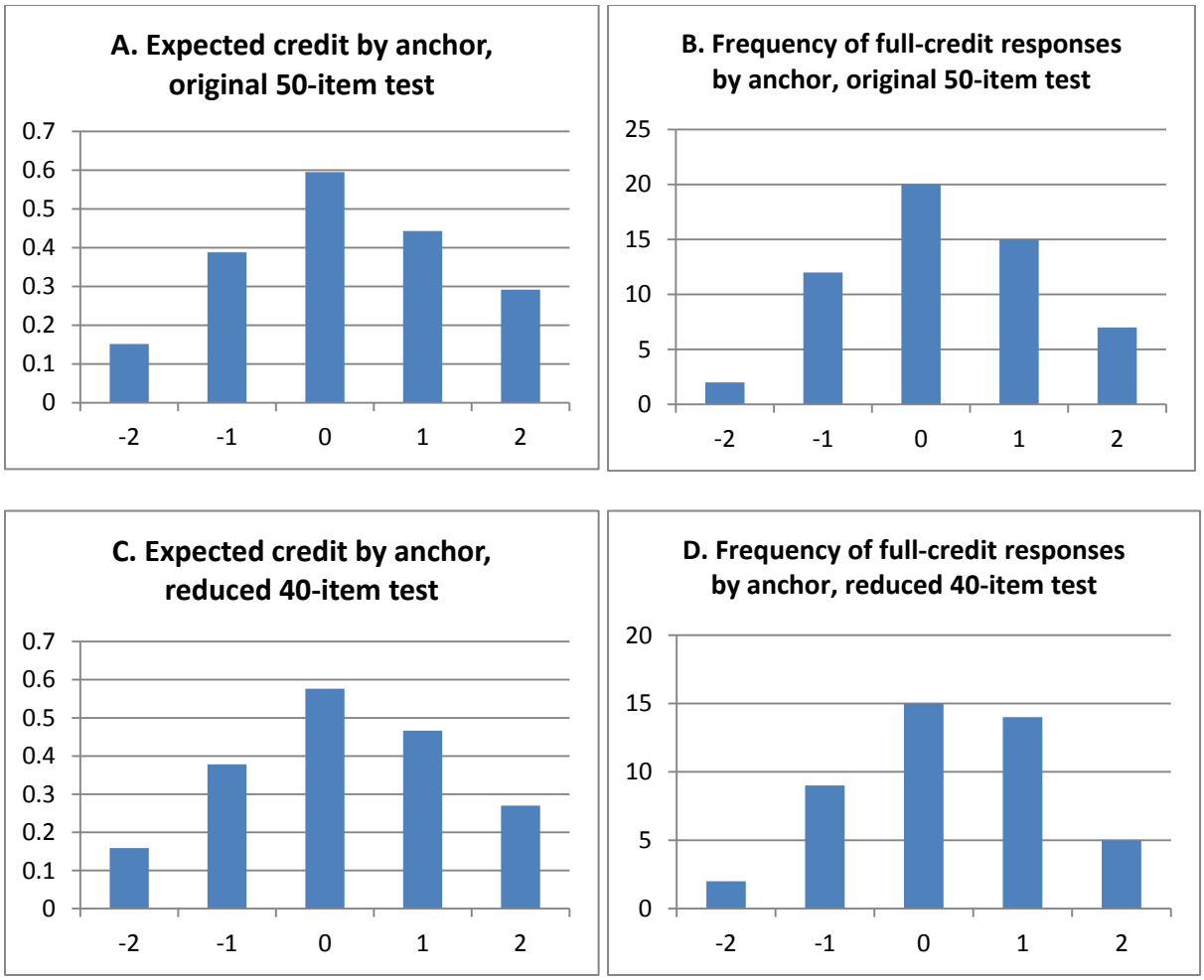

Figure 2

Distributions of expected credit and fully-correct responses in Bland et al. (2005)'s Script Concordance Test

Table 1. Effects of simulated coaching on Bland et al. (2005) Script Concordance Test scores.

	A. Original test (50 items)		B. Items 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 percent credit earned	49.2%	60.8%	49.5%	69.2%
SD	12.4	9.2	14.6	11.4
Cohen's *d*	1.07		1.51	

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):189-195.

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):329–338.

3 Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: a systematic review. *Med Educ* 2012;**46**:552-563.

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, et al. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Med Educ* 2006:**40**(9);848-854.

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**;109-146.

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. Optimization of answer keys for script concordance testing: should we exclude deviant panelists, deviant responses, or neither? *Adv Health Sci Educ Theory Pract* 2011:**16**(5);601-608.

9 Charlin B, Gagnon R, Lubarsky S, Lambert C, Meterissian S, Chalk C, et al. Assessment in the context of uncertainty using the script concordance test: more meaning for scores. *Teach Learn Med* 2010;**22**(3):180-186.

10 Gleser GC, Cronbach LJ, Rajaratnam R. Generalizability 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):430-436.

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):367-375.

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):868-878.

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):277-283.

15 Park AJ, Barber MD, Bent AE, Dooley YT, Dancz C, Sutkin G, et al. Assessment of intraoperative judgment during gynecologic surgery using the Script Concordance Test. *Am J Obstet Gynecol* 2010;**203**(3):240.e1-e6.

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);661-673.

17 Schmitt N. Uses and abuses of coefficient alpha. *Psychol Assess* 1996;**8**:350-353.

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**:1006-1012.

20 Green SB, Yang Y. Commentary on coefficient alpha: a cautionary tale. *Psychometrika* 2009;**74**:121-135.

21 Becker G. How important is transient error in estimating reliability? Going beyond simulation studies. *Psychol Methods* 2000;**5**:370-379.

22 Brennan RL. Generalizability theory. New York: 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);395-399.

25 Eid M, Rauber M. Detecting measurement invariance in organizational 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 behavioral 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 judgment item validity and group differences. *J Appl Psychol* 2011;**96**(2):327-336.

30 Cullen MJ, Sackett PR, Lievens F. Threats to the operational use of situational judgment tests in the college admission process. *Int J Sel Assess* 2006;**14**(2):142-155.

31 Williams RG, Klamen DL, White CB, Petrusa E, Fincher RE, Whitfield CF, et al. Tracking development of clinical reasoning ability across five medical schools using a progress test. *Acad Med* 2011;**9**:1148-1154.

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**:187-193.