The Psychometric Properties of Five Scoring Methods Applied to the Script Concordance Test

Andrew C. Bland, MD, Clarence D. Kreiter, PhD, and Joel A. Gordon, MD

Abstract

Purpose
The Script Concordance Test (SCT) is designed to measure cognitive ability related to successful clinical decision making. An SCT's usefulness for medical education depends on establishing its construct validity. The SCT's present construct relates examinee's scores to experts' response patterns, which does not require a single-best-answer format. Because medical education assessments do require a single best answer, the authors compared the psychometric properties of two aggregate scoring methods with three single-best-answer scoring methods for an SCT.

Method
A nephrology SCT was developed and administered to 85 examinees. Examinees' scores derived from a key developed using eight experts and a traditional aggregate scoring method on a five-point Likert-based scale were compared with four alternate scoring methods (one method eliminated the multipoint Likert-type scale and three eliminated the Likert-type scale and employed single-best-answer scoring).

Results
Two of the four alternate scoring methods performed as well as the traditional Likert-type aggregate scoring method. Scores from all five methods were highly intercorrelated. In addition, each method produced scores similarly correlated with level of experience, and none exhibited an intermediate effect.

Conclusions
Single-best-answer scoring with three answer choices produced results similar to aggregate scoring on a Likert-type scale. Because SCT items appear to assess an examinee's understanding of the interrelatedness of medical knowledge, single-best-answer scoring on an SCT may be valid as an educational assessment. More research is needed to assess differential validity compared with multiple-choice question exams and the predictive validity related to clinical performance.

Acad Med. 2005; 80:395–399.

Feldovich and Burrows[1] first used the "script" concept to characterize clinical reasoning, and in recent years others have further defined the role of scripts as they function in clinical decision making.[2,3] In medicine, script theory focuses on the network or structure of medical knowledge used by the clinician. According to script theory, experienced clinicians process new clinical encounters by activating networks of prior knowledge, or "scripts," that enable the efficient organization and comparative analysis of information pertinent to the clinical case. As in other areas of scientific inquiry, obtaining accurate measures of the variables defining the phenomena of interest is a prerequisite to advancing the theory. Towards this goal, recent research has focused on assessing the script construct and documenting the development and validation of what has been termed the Script Concordance Test (SCT).[4–6]

To be used as an assessment instrument in medical education, an SCT must assess an examinee's understanding of the interrelatedness of the medical knowledge used to solve clinical problems. Using a Likert-type response scale with aggregate scoring, the developers of the SCT have sought to generate scores that reflect the similarity in response patterns between examinees and a group of experts.[4,7,8] The developers assert that the amount of agreement between an examinee and an expert panel is a measure of script development in the examinee,[8] and that SCT items do not have a single best answer.[7,8]

Psychometric research on the reliability and construct validity of SCT scores suggest they are reliable[8] and linearly related to experience (i.e, do not exhibit an intermediate effect).[4–6] Unfortunately, if a single best answer to an SCT item does not exist, the SCT will be of limited use for in-course assessments. The validity considerations regarding an SCT score reflecting the similarity between experts' and examinees' knowledge structures may restrict SCT testing to research applications.

As a direct consequence of the assumption that SCT items do not have a single best answer, researchers have employed a Likert-type response scale in conjunction with an aggregate scoring method to summarize performance.[3,7,8] With aggregate scoring,[9,10] an answer key is developed by administering the test to a group of experts. Because answers to SCT items are coded using a Likert-type scale, multiple levels of affirmative and negative responses as well as a neutral response are possible. Points for a given response are assigned as a function of the proportion of experts who have made that same response. Hence, an examinee's aggregate-based test score reflects the similarity, or concordance, with that of a group of experts.

Determining an appropriate role for the SCT within medical education depends

Dr. Bland is assistant professor of clinical medicine, University of Illinois College of Medicine Peoria.

Dr. Kreiter is associate professor, Department of Family Medicine and Office of Consultation and Research in Medical Education, University of Iowa College of Medicine, Iowa City.

Dr. Gordon is professor, Department of Internal Medicine Division of Nephrology, University of Iowa Carver College of Medicine, Iowa City.

Correspondence should be addressed to Dr. Bland, RenalCare Associates, 515 NE Glen Oak Ave #108, Peoria, IL 61603; e-mail: ⟨abland1@uic.edu⟩.

on a verifying the assumption that an SCT item lacks a single best answer. Although SCT scores display moderate to high levels of reliability, the validity of using an SCT-based measure as an educational assessment of clinical decision making is questionable if a single best answer cannot be determined. For example, the Likert-type responses are critical in defining the nature of an SCT item. The SCT's developers believed that the SCT items resembled those presented by a performance assessment, so they elected to employ the Likert-type response format. A qualitative evaluation of SCT items, however, might reveal that they share much in common with the proposition statements used in True/False items, which do have a single correct answer. If so, this would not invalidate the SCT, but rather it might lead to a reconsideration of the Likert-type response format and the aggregate scoring method.

Charlin et al.[7] suggest that scoring outcomes provide evidence that SCT items are best viewed as having multiple correct answers. They compared aggregate and consensus SCT scores of medical students and experts in obstetrics–gynecology. On a 45-item exam using a seven-point (range: −3 to +3) Likert-type response scale, they found 59% of independent expert responses disagreed with the expert consensus answer obtained during discussion among the expert panel. Charlin et al. suggest the outcome implies that scoring should not be based on a single best answer. However, the disagreements between aggregate and consensus scores they found may be an artifact of the seven-point response scale, and hence, must partially reflect variability in the experts' use of the Likert-type scale. For example, it is questionable whether +2 and +3 responses, defined as disagreement in the Charlin et al. study, were expressing meaningfully different answers. Although Charlin et al. go on to claim a statistically larger difference between experts' and students' scores for the aggregate method compared with the consensus method, only seven experts were tested and the groups displayed a clear a violation of the homogeneity of variance assumption (expert SD = 3.29, $n = 7$; student SD = 8.02, $n = 150$). Given how critical this assumption is when dealing with small and unequal sample sizes,[11] the t values reported in the study were not valid as a test of the null hypothesis (that the expert and student groups had equal means).

Clinical Information: You are evaluating a 60-year-old female for a serum sodium concentration of 125 mEq/L. She has a history of COPD, hypertension, diabetes, CHF, depression and hypercholesterolemia. She also has a 80 pack-year history of smoking.

The referring doctor ordered a:	And her history reveals:	This test becomes:
Serum Cortisol	Currently on prednisone 5 mg po qd	-2 -1 0 1 2
Serum Uric Acid	She is treated with HCTZ	-2 -1 0 1 2
Blood Glucose	She has 4+ glucose on her urine dipstick	-2 -1 0 1 2
Serum Osmolality	Her last blood draw was lipemic	-2 -1 0 1 2
TSH	She has crackles in the lower 1/3 of her lungs and 3+ sacral edema	-2 -1 0 1 2

-2 much less useful
-1 less useful
0 neither more nor less useful
1 more useful
2 much more useful

Figure 1 Sample script concordance test item used to evaluate the validity and reliability of scoring methods, aggregate, concordance, and variations of single-best-answer scores.

Charlin et al.[7] also found that, when compared with the aggregate key, the consensus answer key produced a higher internal alpha reliability index in the examinee sample (.63 versus .52). They report a correlation of $r = .72$ between consensus and aggregate measures, and they conclude: "The value of the Pearson correlation coefficient (between aggregate and consensus scores) for students (0.72, $p < .001$) indicates the two methods induced differences in the classification of students." Although true, if the error in the two scores is uncorrelated, the reliabilities reported for the two scores (.63 and .52) also indicates that, when the .72 correlation is corrected for the unreliability of either one of the two measures,[12] the correlation between the true scores is very near one (1.0 and .91, respectively). Hence, an important alternative conclusion might be that both scoring methods measure nearly the same ability, but that the aggregate score appears to contain more random error. In summary, the study by Charlin et al. fails to show that rational analysis of consensus agreement on a best-item answer by a group of experts will produce an inferior scoring key. Instead, the results might offer support to a single-best-answer hypothesis related to the SCT item.

More empirical evidence is needed to better understand the relationship between an aggregate and a single-best-answer score. The purpose of our study was to further investigate how single-best-answer scoring of the SCT compares with aggregate scoring. Specifically, we compared two aggregate scoring keys with three other scoring keys that, to varying degrees, reflected a single best answer. We examined the relationship between the total test scores and level of training, the correlation between the total scores, and the relationship between items (internal consistency) for evidence of the validity and reliability of the scoring methods.

Method

We developed a ten-case, 50-question SCT in inpatient consultative nephrology according to methods previously described.[8] Prior to its administration, experts in the field reviewed the test and found it contained a representative sample of problems encountered in inpatient consultative nephrology. Figure 1 shows a sample item from the test. 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 study exempt from IRB review, and we obtained no external funding for the study.

The samples

A group of ten academic nephrologists from a large midwestern medical college and a group of six practicing nephrologists from a large midwestern private nephrology practice served as experts. These experts were randomly divided into two stratified groups, each containing five academic and three practicing experts.

One group generated an answer key by the aggregate scoring method as described by Charlin et al.[8] The other group, the tested experts, took the test as part of the examinee group.

In total, 85 examinees sat for the exam. Their experience ranged from medical student to expert nephrologist: 15 medical students, 17 first-year residents, 16 second-year residents, 20 third-year residents, nine fourth-year residents, and eight expert nephrologists. On average, the eight experts had 21.80 years of experience (range: 8–40 years) in nephrology, and the eight experts who generated the key had 19.13 years of experience (12–36 years). All examinees were from either the University of Illinois Chicago College of Medicine at Peoria or the University of Iowa Roy J. and Lucille A. Carver College of Medicine.

Scoring keys

We developed five scoring keys, each based on experts' responses. Key 1, an aggregate scoring key, used a five-point Likert-type scale and an aggregate scoring calculation technique identical to that recommended and used by Charlin et al.[7] Key 2 employed the same aggregate scoring method as used in Key 1, but all responses on both the expert key and the examinees' response file were recoded to a three-point scale (1, 0, −1). Hence, an affirmative response greater than +1 was recoded as a +1. In a similar fashion, a negative response greater than −1 was assigned a value of –1. Responses indicating 0 retained their value. For Key 3, referred to as the three-point mode method, a correct answer and a score of 1 were awarded for a response that matched the modal value of the experts. If an examinee's response differed from the expert's modal response value, a score of 0 was awarded. A mode-based scoring system defines a single correct or best answer but does not differentiate between responses with varying distances from the best answer (the expert mode). For example, if the expert mode response was +1, both a 0 and a –1 response from an examinee would receive a score of 0. Key 4, referred to as the three-point distance from the mode, provided examinees with credit for being closer to the most frequent response from experts. With this scoring system, a low score (closer to the expert mode) represents a strong performance. Key 5, the three-point absolute distance from the mean method, provides a statistical consensus score and also awards credit for being closer to the best response (experts' mean response). With this method, the single best response is defined as the response closest to the averaged response of the experts on the three-point scale. Scoring uses a single-best-answer method, but it awards more credit the closer the examinee's answer is to the experts' mean. This is achieved by calculating a keyed correct score as the mean of the experts' responses coded on a three-point scale and calculating the points awarded as the absolute difference between an examinee's response and this mean. With this scoring system (and in this report), a low score (closer to the experts' mean) represents a strong performance. If such a scoring technique were put into practice, items should be rescaled between 0 and 1 and the score scale should be reversed (1 –score) for reporting test results to students.

Results

Scores for the five keys are shown in Table 1. Scores generated using Keys 2–5 demonstrated a strong absolute correlation (.88−.93) with the aggregate scoring method (Key 1), and there was also a strong relationship ($r = .90 - .96$) between scores calculated using Keys 2–5. Table 2 presents reliability and validity coefficients for each of the keys. All methods of scoring demonstrated a similarly strong and significant linear correlation ($r = .57-.62$; $p < .0001$) with level of training. Mean scores for each level of training are shown in the last column of Table 2. For each key, scores tended to improve with level of training, and there was no evidence using any key that experts' performances were scored lower than performances in groups with less experience. Hence, the intermediate effect was not observed for any of the scoring methods. Internal reliabilities for the four keys were similar (range = .68−.78). The expert groups were switched and the scoring keys were regenerated with the new group of tested experts. The results were virtually identical (results not shown).

Discussion and Conclusion

In examining and interpreting the performances of the five keys, it is important to consider Key 1 in relation to Keys 2–5. Key 2 retains the aggregate scoring method and, hence, also retained the no-best-answer assumption. But it also removes all information related to the degree of affirmative and negative response. As Table 2 shows, test performance in relation to reliability and validity (corre-

Table 1

Descriptions and Mean Responses of 85 Examinees on Five Keys for Scoring Script Concordance Tests, The University of Illinois College of Medicine at Peoria and University of Iowa Roy J. and Lucille A. Carver College of Medicine, 2003

Key	Description	Mean (SD)	Range
1 (aggregate five-point)	Traditional weighted expert No correct answer	28.9 (5.4)	12.0–40.1
2 (aggregate three-point)	Recoded weighted expert No correct answer	34.2 (5.2)	19.0–44.8
3 (three-point mode)	Most frequent expert response Correct answer	28.1 (5.7)	14.0–41.0
4 (three-point distance from mode)	Key 3 weighted for distance from response Correct answer	28.1 (7.7)	11.0–49.0
5 (three-point absolute distance from mean)	Average expert response Weighted for distance Correct answer	32.9 (7.6)	18.0–59.75

Table 2

Reliability and Correlation with Level of Training Results for Five Script Concordance Test Scoring Keys, The University of Illinois College of Medicine at Peoria and University of Iowa Roy J. and Lucille A. Carver College of Medicine, 2003

Key	Alpha reliability	Correlation with level of training	Score by level of training	
			Level of training (no. in group)	Score*
1 (aggregate five-point)	.74	$r = .57$ $p < .0001$	Expert (8)	35.9
			Fourth-year resident (9)	32.8
			Third-year resident (20)	30.5
			Second-year residents (16)	26.7
			First-year residents (17)	26.0
			Medical student (15)	26.1
2 (aggregate three-point)	.74	$r = .60$ $p < .0001$	Expert (8)	40.8
			Fourth-year resident (9)	38.2
			Third-year resident (20)	35.9
			Second-year residents (16)	32.4
			First-year residents (17)	31.6
			Medical student (15)	30.8
3 (three-point mode)	.68	$r = .60$ $p < .0001$	Expert (8)	34.5
			Fourth-year resident (9)	32.3
			Third-year resident (20)	30.3
			Second-year residents (16)	26.8
			First-year residents (17)	25.2
			Medical student (15)	23.9
4 (three-point distance from mode)*	.68	$r = -.62$ $p < .0001$	Expert (8)	19.6
			Fourth-year resident (9)	21.7
			Third-year resident (20)	26.4
			Second-year residents (16)	27.5
			First-year residents (17)	32.4
			Medical student (15)	34.7
5 (three-point absolute distance from mean)*	.78	$r = -.58$ $p < .0001$	Expert (8)	24.1
			Fourth-year resident (9)	26.8
			Third-year resident (20)	30.4
			Second-year residents (16)	35.2
			First-year residents (17)	37.2
			Medical student (15)	37.3

* Lower scores reflect less distance from correct answer for key number four and key number five.

lation with experience) using Key 2 was almost identical with that of Key 1. This suggests the five-point scale adds little information. Further anecdotal support for moving away from a five point Likert-type scale is gained from experts' and examinees' comments regarding their sometimes-arbitrary choices between +2 and +1 and –1 and –2. The validity outcome related to the three-point and five-point scale is not surprising since the total scores were almost perfectly correlated ($r = .93$). It appears that using a multipoint scale does little to change the rank ordering of students. Key 3, using single-best-answer scoring, displays a similar validity coefficient (correlation of score with level of training) but is slightly less reliable than either Keys 1 or 2. The lower reliability observed in Key 3 may be related to the loss of information concerning the level of an incorrect response by an examinee. Key 4 also used a single-best-answer score, but it incorporated a statistical technique that scored the two incorrect responses differently (incorrect responses further from the best response received a higher penalty). When this dimension was added, reliability was unchanged but a slightly higher correlation with level of training was observed. Key 5 also scored incorrect responses differently but also awarded credit as a function of experts' agreement on a single best answer. For example, when experts unanimously agreed on one response, the single best answer was awarded the maximum score. However, if there was disagreement among experts regarding the correct answer, the mean expert score could not equal the keyed response. Key 5 generated the highest reliability coefficient, which may be a function of experts' certainty of the correct answer, making items with complete agreement between

experts worth more than those with less agreement.

For an SCT to be used appropriately as a medical education assessment, it needs to measure whether important knowledge associations are present to support clinical reasoning. Aggregate scoring methods present two possible problems to such measurement.

First, the aggregate scoring system does not reward points based on the correctness of the examinees' responses. For example, an item in which all experts answered +2 on a seven-point Likert-type scale (–3 to +3) would award the same score of 0 to examinees answering –3 and +3. Hence, the examinee who agreed with the experts on the direction of the impact but not on the degree of impact will receive the same score one who understood neither the direction nor the degree of impact. Such a scoring scheme does not award points based on the correctness of the response, and no single point on the scale is necessarily considered to be the best or the most correct response.

Second, valid in-course educational assessments must assess clearly defined behavioral and instructional objectives; it would be poor educational practice to assign students' grades based on scores generated by aggregate scoring. Rather, an educational assessment should seek to reward an examinee's understanding of objectively defined relationships that are important to the solution of a clinical problem. Without a single best answer, behavioral objectives, such as those typically delineated in a table of test specifications, would be difficult or impossible to define. Even the test instructions provided to the examinee would be highly problematic. It is unlikely that asking students to "respond like experts" would provide examinees or test developers with sufficient structure for academic testing.

Because best-answer scoring is better suited for educational assessments, scoring responses on a multipoint Likert-type scale with the aggregate scoring method is not recommended. Our research illustrates that aggregate scoring anomalies that would produce scores different from single-best-answer scoring rarely occur (as evidenced by the high correlation with single-best-answer scoring methods such as Keys 3–5). Given that aggregate scoring and single-best-answer scoring produced very similar total score results and, given the practical difficulties associated with aggregate scoring interpretation, single-best-answer scoring may be the best approach for SCTs.

Clearly, our use of a statistical summary of experts' responses to determine a key for the best answer is somewhat different than using a true consensus method, and it seems likely that a consensus answer, benefiting from discussion and sharing of knowledge between experts, might produce superior results when compared with an aggregate score. However, even when using statistical methods to estimate a consensus on a three-point scale, we found reliability and validity for the best-answer scoring methods were similar to the five-point aggregate method, and for assessment, scores were much easier to interpret. Because our examinee group was more heterogeneous than most found in medical educational test settings, our alpha coefficients were likely to be higher than any obtained within more typically homogenous examinee populations.

Because of its content validity, efficiency, and high reliability, the SCT with single-best-answer scoring is an attractive assessment method for evaluating clinical-reasoning skills. Future research should seek to document predictive validity by discovering whether the associations captured by the SCT are activated in actual clinical encounters. The link between the associations measured by the SCT and the actual utilization of the associations in a true performance assessment has not yet been established. Our statistical methods of arriving at a consensus answer were convenient compared with convening experts for a meeting. Future research should compare these statistical techniques with traditional consensus methods.

The authors thank Lisa Antes, MD for her constructive comments in the construction of the Script Concordance Test for nephrology. They also thank the program directors for internal medicine, Scott Vogelgesang and Lannie Cation, for allowing us to administer the test and the students, residents, and experts who completed the exam.

References

1 Feltovich PJ, Barrows HS. Issues of generality in medical problem solving. In: Schmidt HG, De Volder ML (eds). Tutorials in Problem-Based Learning: A New Direction in Teaching the Health Professions. Assen, The Netherlands: Van Gorcum, 1984.

2 Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise: theory and implications. Acad Med. 1990;65: 611–20.

3 Charlin B, Tardif J, Boshuizen PA. Scripts and medical diagnostic knowledge: theory and application for clinical reasoning instruction and research. Acad Med. 2000;75:182–90.

4 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 exploratory study on the script concordance test. Med Educ. 2001; 35:430–6.

5 Charlin B, Brailovsky CA, Brazeau-Lamontague L, Samson L, Leduc C. Script questionnaires: their use for assessment of diagnostic knowledge in radiology. Med Teach. 1998;20: 567–71.

6 Charlin B, Brailovsky CA, Leduc C, Blouin D. The diagnostic script questionnaire: a new tool to assess a specific dimension of clinical competence. Adv Health Sci Educ. 1998;3: 51–8.

7 Charlin B, Desaulniers M, Gagnon R. Blouin Van der Vleuten C. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. Teach Learn Med. 2002;14:150–6.

8 Charlin B, Brailovsky C, Roy L, Goulet F, van der Vleuten C. The script concordance test: a tool to assess the reflective clinician. Teach Learn Med. 2000;12:189–95.

9 Norman GR. Objective measurement of clinical performance. Med Educ. 1985;19:43–7.

10 Norcini JJ, Shea JA, Day SC. The use of the aggregate scoring for a recertification examination. Eval Health Prof. 1990;13:241–51.

11 Hayes WL. Statistics. 3rd ed. New York: CBS College Publishing, 1981:287.

12 Allen MJ, Yen WM. Introduction to Measurement Theory. Monterey, CA: Brooks/Cole Publishing, 1979:98–9.