Assessment of Undergraduate Clinical Reasoning in Geriatric Medicine: Application of a Script Concordance Test

Ronaldo D. Piovezan, MD, MSc,[*,†] *Osvladir Custódio, MD, MSc,*[*] *Maysa S. Cendoroglo, MD, PhD,*[*] *Nildo A. Batista, MD, PhD,*[†] *Stuart Lubarsky, MD, MHPE,*[‡] *and Bernard Charlin, MD, PhD*[§]

A challenging aspect of geriatric practice is that it often requires decision-making under conditions of uncertainty. The Script Concordance Test (SCT) is an assessment tool designed to measure clinical data interpretation, an important element of clinical reasoning under uncertainty. The purpose of this study was to develop and analyze the validity of results of an SCT administered to undergraduate students in geriatric medicine. An SCT consisting of 13 cases and 104 items covering a spectrum of common geriatric problems was designed and administered to 41 undergraduate medical students at a medical school in São Paulo, Brazil. A reference panel of 21 practicing geriatricians contributed to the test's score key. The responses were analyzed, and the psychometric properties of the tool were investigated. The test's internal consistency and discriminative capacity to distinguish students from experienced geriatricians supported construct validity. The Cronbach alpha for the test was 0.84, and mean scores for the experts were found to be significantly higher than those of the students (80.0 and 70.7, respectively; $P < .001$). This study demonstrated robust evidence of reliability and validity of an SCT developed for use in geriatric medicine for assessing clinical reasoning skills under conditions of uncertainty in undergraduate medical students. These findings will be of interest to those involved in assessing clinical competence in geriatrics and will have important potential application in medical school examinations. **J Am Geriatr Soc 60:1946–1950, 2012.**

Key words: clinical reasoning; script concordance test; medical education; geriatric medicine; assessment

A particular challenge of geriatric medicine is that many problems that its practitioners encounter are ill defined. The management of chronic diseases and associated comorbidities, the ethical concerns related to life expectancy and palliative care, the multidisciplinary nature of care, and other important geriatric principles and practices are often sources of doubt and uncertainty in the clinical care of older people.[1] Furthermore, speech disorders, neuropsychiatric conditions, and nonspecific symptoms and signs are more common in elderly adults and offer particular challenges in data interpretation and decision-making. These challenges have come into particularly sharp relief in the current demographic climate, in which the population is rapidly aging, and physicians need to be increasingly adept at treating the specific problems associated with caring for elderly adults.

Early in their training, medical students must learn to navigate this daunting world of ambiguity in geriatric medicine, and their preceptors must use sound tools to monitor and evaluate their level of competence in this challenging field.[2] Various assessment methods have been developed to evaluate competencies acquired during clinical geriatric training. The assessment method most commonly used to evaluate learning in medical education is the multiple-choice written test, but a learner's capacity to resolve ambiguous situations is, in general, poorly assessed using this test format.[3]

The Script Concordance Test (SCT), based on script theory from cognitive psychology, is an alternative method for evaluating reasoning under these circumstances. Experienced geriatricians possess networks of knowledge, called illness scripts, that become mobilized to handle the problems they routinely face in their daily practice.[4] Illness scripts develop and continually transform with experience and reflection. They facilitate the efficiency of clinical reasoning processes, and some consider them to be the hallmarks of expertise in a specialty.[3] Scripts begin to appear when students are faced with their first clinical cases and become refined throughout their clinical careers.[5]

From the [*]Division of Geriatrics, and [†]Centre for Higher Education Development in Health, Federal University of São Paulo, São Paulo, Brazil; [‡]Centre for Medical Education and Department of Neurology, McGill University, Montreal, Canada; and [§]Centre for Applied Teaching in Health Sciences, University of Montreal, Montreal, Quebec, Canada.

Address correspondence to Ronaldo D. Piovezan, Division of Geriatrics, Federal University of São Paulo, Rua Prof. Francisco de Castro, 105, Vila Mariana, 04020–050 São Paulo, SP, Brazil. E-mail: rdpiovezan@gmail.com

DOI: 10.1111/j.1532-5415.2012.04152.x

0002-8614/12/$15.00

SCT may be useful for monitoring and evaluating script development and clinical reasoning in situations approximating those frequently encountered in real-life practice—that is, in contexts of uncertainty—in geriatrics. This tool has been developed in diverse educational environments, used in different countries, and translated into several languages.[6–9] It uses written case descriptions as vignettes and requires examinees to make decisions about diagnostic possibilities, investigative options, and treatment alternatives. Answers are provided on a Likert scale intended to capture the variability of responses of experts to clinical problems for which single correct answers may not be evident.[10]

The results of two previous geriatrics SCTs have been reported. The purpose of one study was to provide evidence of cognitive validity for the test method by investigating subjects' reaction time in response to SCT-style questions presenting typical or atypical features of a case in geriatrics.[11] The study was not designed to discriminate subjects according to level of expertise. The second study showed statistically significant differences between groups according to level of expertise in geriatrics, but sampled a restricted content area (urinary incontinence).[12] No published reports of SCT development specific to the assessment of undergraduate training and encompassing the broad spectrum of geriatric health problems were found. The purpose of the current study was to develop and analyze the validity of results of an SCT administered to undergraduate students in geriatrics.

METHODS

SCT Development

This SCT, developed in Portuguese at a Brazilian academic institution, is the first of its kind. Three geriatricians directly involved in the undergraduate training of students in the Division of Geriatrics at the Federal University of São Paulo were responsible for creating the test. Thirteen cases featuring 115 items were initially created. Common situations encountered in geriatrics were blueprinted for inclusion in the test. (Table 1).

Table 1. Geriatric Themes and Problems Addressed in the Test

Losses in activities of daily living
Care of patients at the end of life
Constipation
Delirium
Dementia
Depression
Dizziness
Dysphagia
Falls
Iatrogenesis
Incontinence
Malnutrition
Polypharmacy
Syncope
Unintended weight loss

The SCT was constructed according to previously published guidelines.[10] It was composed of vignettes based on genuine clinical geriatric cases. As in real-life situations, the case descriptions contained elements of uncertainty or ambiguity and therefore did not present sufficient information to allow uncontestable, single-correct-answer solutions. The scoring system aims to compare responses of students with those of practicing geriatricians; this latter group formed the reference panel used to set the test's scoring grid.[13]

An example of an SCT vignette and related items is shown in Figure 1. Each test item consists of a clinical vignette followed by three columns. The first column provides a hypothesis (e.g., a diagnostic consideration, an investigational strategy, a treatment recommendation, or an ethical dilemma) relevant to the situation described in the vignette. In the second column, a new piece of clinical information (e.g., a symptom, a sign, or a test result) is provided. The third column contains a 5-point Likert-type scale for the examinee to use to indicate the perceived effect of this new information (column 2) on the proposed hypothesis (column 1).

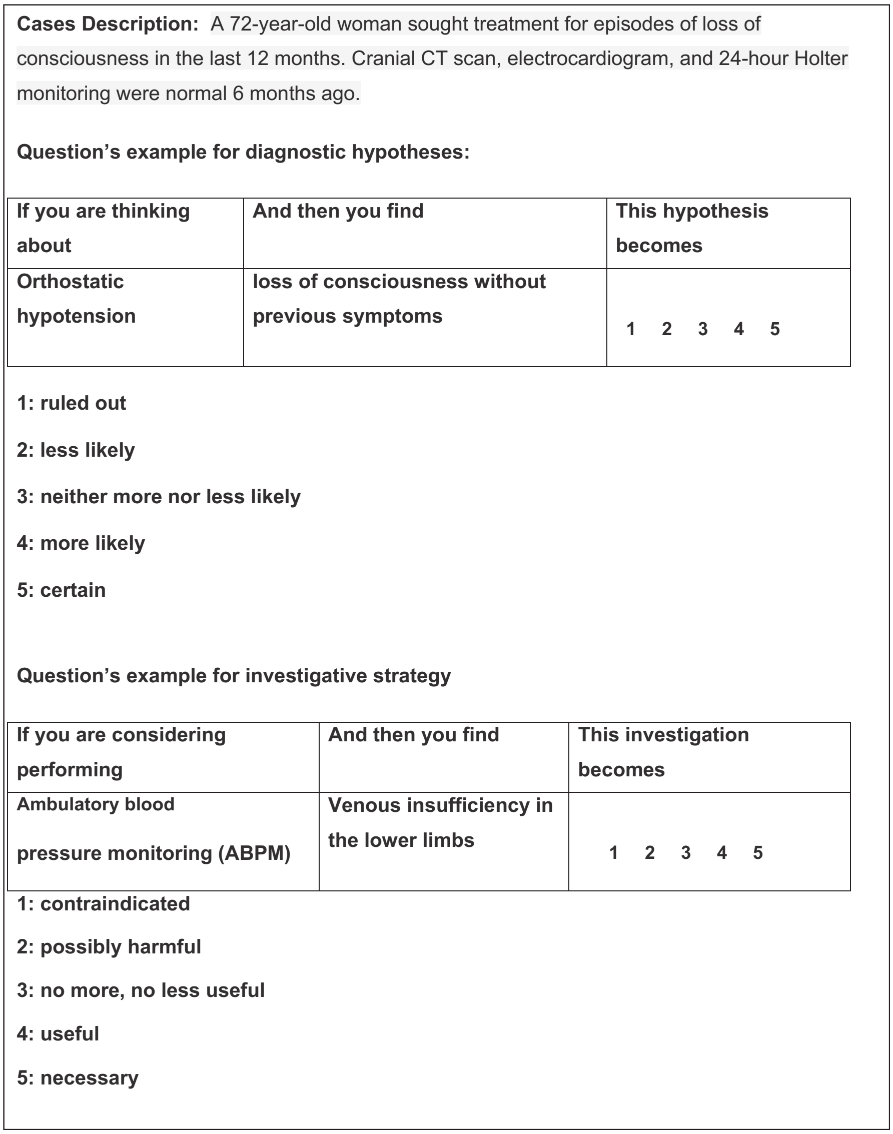
Scoring System

In contrast to many traditional test methods, there are no single-best answers to SCT items; several responses to each item may be considered acceptable. The examinee's response to each item is compared with the responses of the test's panel members. Credit is assigned to each response based on how many of the experts on the panel choose that response. A maximum score of 1 is given for the response chosen by most of the experts (the modal response). Other responses are given partial credit, depending on the fraction of experts choosing them. Responses that experts did not select receive 0. An example of the SCT scoring system is shown in Table 2.

Pilot Test

A preliminary version of the geriatrics SCT, consisting of 115 items nested in 13 clinical cases, was created. This version of the test was piloted by administering it to five experienced geriatricians (who were not otherwise involved in the study). After they completed the test, they were asked to review the cases and items and provide feedback, and those that were unanimously felt to be too vague, misleading, or otherwise unacceptable were discarded.

Frequencies of answers were calculated for each item individually. Concordance analysis of the five participants' answers to each item was conducted following previously described methods.[14] Only concordance values judged to be fair or good (whose values were between 0.70 and 0.85) were considered for use in the final version of the test. Concordance values of less than 0.70 (low concordance) were felt to indicate problems relating to the construction of an item. High concordance levels (>0.85) were judged to represent overly high consensus of the specialists' answers for an item, indicating that insufficient ambiguity or uncertainty was built into the item. These items were thought to be too similar to single-correct-answer multiple-choice questions and were eliminated. The optimized version of the SCT contained 104 items nested in 13 cases,



Cases Description: A 72-year-old woman sought treatment for episodes of loss of consciousness in the last 12 months. Cranial CT scan, electrocardiogram, and 24-hour Holter monitoring were normal 6 months ago.

Question's example for diagnostic hypotheses:

If you are thinking about	And then you find	This hypothesis becomes
Orthostatic hypotension	loss of consciousness without previous symptoms	1 2 3 4 5

1: ruled out

2: less likely

3: neither more nor less likely

4: more likely

5: certain

Question's example for investigative strategy

If you are considering performing	And then you find	This investigation becomes
Ambulatory blood pressure monitoring (ABPM)	Venous insufficiency in the lower limbs	1 2 3 4 5

1: contraindicated

2: possibly harmful

3: no more, no less useful

4: useful

5: necessary

Figure 1. Test format. CT = computed tomography.

which has been shown to be a reasonable case–item distribution for obtaining reliable results.[15]

Subjects and Setting

Reference Panel

Twenty-one experienced geriatricians were recruited to form the reference panel for the test. All panel members held faculty positions at different academic institutions in São Paulo, Brazil, and had been practicing geriatric medicine for at least 10 years. Each panel member independently completed a paper-based version of the SCT.

Students

Medical school is completed in 6 years in Brazil. The curriculum has a spiral design, whereby students undertake an increasingly challenging module in geriatrics in each of the first 5 years. Fifth-year students at the Federal University of São Paulo, Brazil, were recruited to participate in the study. They were invited to answer the test on the last day of the final geriatrics module. All participants were asked to provide informed consent to participate in the study. Anonymity was assured. Students who agreed to participate were given a paper-based SCT, which they completed individually.

Data Analysis

Descriptive statistics, including means, standard deviations (SDs), and minimum and maximum values of participants' scores were obtained. Differences between mean expert and student scores were calculated using *t*-tests, and internal reliability was estimated using Cronbach alpha.

Participant scores were standardized using an established method for expressing deviation from a mean within a distribution of scores.[16] According to this method, scores for each participant are converted to a scale based on the mean and SD of the reference panel, which serves as a reference value. The panel SD therefore serves as a yardstick

Table 2. Example of the Script Concordance Test Scoring System

Answer	1	2	3	4	5
Number of experts who chose this answer	0	0	1	5	4
Number of experts who chose this answer divided by answer provided by the greatest number of panel members (i.e., the modal answer)	0	0	1/5	5/5	4/5
Score for this question	0	0	0.2	1.0	0.8

Suppose a panel of 10 experts was asked to respond to the first question in the example given in Figure 1, and five selected response 4, four selected response 5, and one selected response 3. The scoring for this item would be response 3, 0.2 points (1/5); response 4, 1 point (5/5); response 5, 0.8 points (4/5); responses 1 and 2, both 0 points. An examinee's total score for the test is the sum of the credit obtained for each of the items divided by the total obtainable credit for the test multiplied by 100 to derive a percentage score.

by which students' performance can be measured. This transformation was performed using a two-step process. In the first step, *z*-scores were calculated for students, with mean of the panel of 0 and a SD of 1. In the second step, *z*-scores were converted into modified *T* scores with panel mean and SD set at 80 and 5, respectively.

RESULTS

Forty-one of 60 eligible students agreed to participate in the study. Means, medians, SDs, and ranges for the student and expert groups are shown in Table 3. There was a significant difference between the mean score of experts (80.0) and the mean score of students (70.7) ($P < .001$). The Cronbach alpha for the test was 0.84.

Standardized scores for students were calculated. One student scored above the reference panel mean, 11 students scored between the mean and 1 SD below the panel mean, 11 students scored between 1 and 2 SDs below the panel mean, and 18 students scored more than 2 SDs below the panel mean.

DISCUSSION

The goal was to develop and analyze the validity of the results of an SCT for use in evaluating clinical reasoning of undergraduate students in geriatric medicine. It was anticipated that the SCT, a tool designed to probe examinees' reasoning-under-uncertainty skills in a given domain, would be particularly well suited for use in geriatric medicine, pending further investigation of its psychometric properties in this context. In this study, several sources of validity evidence supported the use of the SCT for assessing this important professional competency in undergraduate geriatric medicine students.

Content validity evidence evaluates the relationship between a test's content and the construct it is intended to measure.[17] To bolster the content validity of the test, a content blueprint was created based on existing standards to ensure that the broad range of topics encountered in geriatric practice was covered during the development of the test's cases and questions.[18–20]

Table 3. Score Comparison Between Groups

Group	N	Mean	Median ± Standard Deviation	Minimum	Maximum	Range
Reference panel	21	80.0	80.9 ± 5.0	70.3	92.1	21.8
Students	41	70.7[a]	72.2 ± 7.3	51.2	84.1	33.0

[a] $P < .001$.

Evidence to support the internal consistency of our assessment tool was also found. The Cronbach alpha value of the test was 0.84, which is generally considered to represent adequate reliability for summative examinations in routine use at the undergraduate level.[21]

A recently described standardization method for comparing the scores of students with those of an aggregate panel was used.[16] According to this method, a high score is meant to indicate that a student reasons through clinical problems similarly to a representative sample of experienced physicians in a field. This method of SCT score transformation provides students with clear, interpretable comparisons between their performance and that of experienced geriatricians.[16] Moreover, this scoring method is well suited for providing formative feedback to educators regarding their learners' academic progress. In the geriatrics curriculum, for example, SCT scores could potentially allow the extent to which the course's educational goals are being met for students at different levels and to which students are beginning to think like experts in the field to be evaluated. SCT performance may also prove useful for developing level- and context-appropriate remediation programs for students in need (e.g., for advanced-level students whose transformed scores are more than 2 SDs below the panel mean).

This study has several limitations. The small number of participants did not allow for determining whether the test is useful for discriminating students based on their year of undergraduate education. Another possible limitation relates to what is known as the "intermediate effect." Traditional "fact-based" assessments such as multiple-choice questions may be prone to this phenomenon, whereby examinees at intermediate levels of training outperform advanced-level examinees.[17] A competency assessment in which examinees with less expertise perform better than those with more expertise is of questionable validity. One theory that accounts for the intermediate effect is that expert knowledge, with time and experience, becomes encapsulated into workable scripts whose details may not be consciously accessible for the purpose of answering examination questions. SCTs have been purported to be less prone than traditional assessments to this phenomenon,[3] but because of the small study population, it was not possible to determine whether the test was successful in overcoming the intermediate effect.

In conclusion, this study provides further evidence of the reliability and validity of an SCT for evaluating undergraduate clinical reasoning skills in geriatrics.[22] To the knowledge of the authors, this is the first SCT developed in Portuguese in a Brazilian medical education institution and covering a broad spectrum of geriatric issues. This study therefore demonstrates stability of scores in a different educational environment and linguistic setting than has been demonstrated in previous studies. It was also found to be a feasible tool to develop; SCT construction is simpler than traditional written tools, such as multiple-choice questions. Easy to administer and correct, the SCT is also an efficient and practical tool for assessing large numbers of students at a time. The scoring system is unique and reflects the range of possible decisions among experts. Further studies could potentially contribute to the investigation of this promising approach to evaluation of undergraduate clinical reasoning in geriatric medicine.

ACKNOWLEDGMENTS

Conflict of Interest: The editor in chief has reviewed the conflict of interest checklist provided by the authors and has determined that the authors have no financial or any other kind of personal conflicts with this paper.

Author Contributions: Ronaldo D. Piovezan, Maisa S. Cendoroglo, and Nildo A. Batista contributed to the study concept and design, acquisition of subjects and data. All authors contributed by analyzing and interpreting data, drafting the manuscript, and approving the revised final version of the manuscript.

Sponsor's Role: None.

REFERENCES

1. Gill TM. Geriatric medicine: It's more than caring for old people. Am J Med 2002;113:85–90.
2. Lally F, Crome P. Undergraduate training in geriatric medicine: Getting it right. Age Ageing 2007;36:366–368.
3. Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: The script concordance approach. Eval Health Prof 2004;27:304–319.
4. Charlin B, Boshuizen H, Custers E et al. Scripts and clinical reasoning. Med Educ 2007;41:1179–1185.
5. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: Theory an implication. Acad Med 1990;65:611–621.
6. Meterrisian SH. A novel method of assessing clinical reasoning in surgical residents. Surg Innov 2006;13:115–119.
7. Brailovsky C, Charlin B, Beausoleil S et al. 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;35:430–436.
8. Carrière B, Gagnon R, Charlin B et al. Assessing clinical reasoning in pediatric emergency medicine: Validity evidence for a script concordance test. Ann Emerg Med 2009;53:647–652.
9. Lubarsky S, Chalk C, Kazitani D et al. The Script Concordance Test: A new tool assessing clinical judgment in neurology. Can J Neurol Sci 2009; 36:326–331.
10. Fournier JP, Demeester A, Charlin B. Script concordance tests: Guidelines for construction. BMC Med Inform Decis Mak 2008;8:18.
11. Gagnon R, Charlin B, Roy L et al. The cognitive validity of the Script Concordance Test: A processing time study. Teach Learn Med 2006;18:22–27.
12. Ruiz JG, Tunuguntla R, Charlin B et al. The Script Concordance Test as a measure of clinical reasoning skills in geriatric urinary incontinence. J Am Geriatr Soc 2010;58:2178–2184.
13. Norman GR. Objective measurement of clinical performance. Med Educ 1985;19:43–47.
14. Cicchetti DY, Showalter D, Rosenheck R. A new method for assessing interexaminer agreement when multiple ratings are made on a single subject: Applications to the assessment of neuropsychiatric symptomatology. Psychiatry Res 1997;72:51–63.
15. Gagnon R, Charlin B, Lambert C et al. Script concordance testing: More cases or more questions? Adv Health Sci Educ Theory Pract 2009;14: 367–375.
16. Charlin B, Gagnon R, Lubarsky S 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.
17. Schmidt HG, Boshuizen HP. On the origin of intermediate effects in clinical case recall. Mem Cognit 1993;21:338–351.
18. Halter JB, Ouslander JG, Tinetti ME et al. Hazzard's Geriatric Medicine and Gerontology, 6th Ed. New York: McGraw Hill Medical, 2009.
19. Inouye SK, Studenski S, Tinetti ME et al. Geriatric syndromes: Clinical, research, and policy implications of a core geriatric concept. J Am Geriatr Soc 2007;55:780–791.
20. Sleeper RB. Geriatric primer—common geriatric syndromes and special problems. Consult Pharm 2009;24:447–462.
21. Downing SM. Reliability: On the reproducibility of assessment data. Med Educ 2004;38:1006–1012.
22. Downing SM. Validity: On the meaningful interpretation of assessment data. Med Educ 2003;37:830–837.