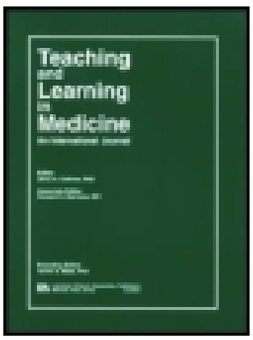


Teaching and Learning in Medicine

ISSN: 1040-1334 (Print) 1532-8015 (Online) Journal homepage: http://www.tandfonline.com/loi/htlm20

Assessment in the Context of Uncertainty Using the Script Concordance Test: More Meaning for Scores

Bernard Charlin , Robert Gagnon , Stuart Lubarsky , Carole Lambert , Sarkis Meterissian , Colin Chalk , Johanne Goudreau & Cees van der Vleuten

To cite this article: Bernard Charlin , Robert Gagnon , Stuart Lubarsky , Carole Lambert , Sarkis Meterissian , Colin Chalk , Johanne Goudreau & Cees van der Vleuten (2010) Assessment in the Context of Uncertainty Using the Script Concordance Test: More Meaning for Scores, Teaching and Learning in Medicine, 22:3, 180-186, DOI: 10.1080/10401334.2010.488197

To link to this article: http://dx.doi.org/10.1080/10401334.2010.488197

Teaching and Learning in Medicine, 22(3), 180–186

ISSN: 1040-1334 print / 1532-8015 online
DOI: 10.1080/10401334.2010.488197

Routledge
Taylor & Francis Group

Assessment in the Context of Uncertainty Using the Script Concordance Test: More Meaning for Scores

Bernard Charlin and Robert Gagnon
Center of Pedagogy Applied to Health Sciences (CPASS), University of Montreal, Montreal, Quebec, Canada

Stuart Lubarsky
Department of Neurology and Nurosurgery, McGill University, Montreal, Canada

Carole Lambert
Center of Pedagogy Applied to Health Sciences (CPASS), University of Montreal, Montreal, Quebec, Canada

Sarkis Meterissian
Department of Surgery, McGill University, Montreal, Canada

Colin Chalk
Department of Neurology and Nurosurgery, McGill University, Montreal, Canada

Johanne Goudreau
Faculty of Nursing, University of Montreal, Montreal, Canada

Cees van der Vleuten
Department of Educational Development and Research, University of Maastricht, Maastricht, The Netherlands

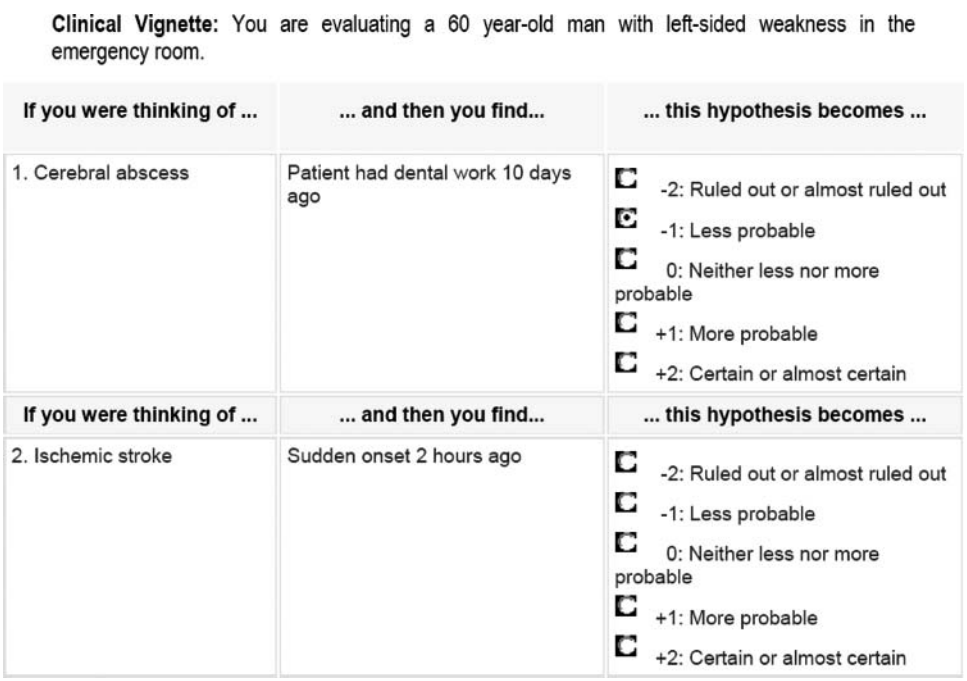
***Background*: The Script Concordance Test (SCT) uses authentic, ill-defined clinical cases to compare medical learners' judgment skills with those of experienced physicians. SCT scores are meant to measure the degree of concordance between the performance of examinees and that of the reference panel. Raw test scores have meaning only if statistics (mean and standard deviation) describing the panel's performance are concurrently provided. *Purpose*: The purpose of this study is to suggest a method for reporting scores that standardizes panel mean and standard deviation, allowing examinees to immediately gauge their performance relative to panel members. *Methods*: Based on a statistical method of standardization, a new method for computing SCT scores is described. According to this method, test raw scores are converted into a scale in which the panel mean is set as the value of reference, and the standard deviation of the panel serves as a yardstick by which examinee performance is measured. *Results*: The effect of this transformation on four data sets obtained from SCTs in radio-oncology, surgery, neurology, and nursing is discussed. *Conclusion*: This transformation method proposes a common metric basis for reporting SCT scores and provides examinees with clear, interpretable insights into their performance relative to that of physicians of the field. We recommend reporting SCT scores with the mean and standard deviation of panel scores set at standard scores of 80 and 5, respectively. Beyond SCT, our transformation method may be generalizable to the scoring of other test formats in which the performance of examinees and those of a panel of reference undertaking the same cognitive tasks are compared.**

INTRODUCTION

Problems that doctors encounter in the clinical setting reflect a continuum of specificity.[1] At one end of the spectrum, "well-defined problems" are those for which all of the necessary information is readily available and a clear and correct solution exists (e.g., deciding which dosage of a pediatric medication to administer, based on the patient's weight). At the other end of the spectrum, "ill-defined problems" are those in which the goals may be ambiguous, information may be lacking, and several hypotheses and courses of action may be defensible (e.g., deciding whether to administer drug A or B, based on the patient's symptoms and preferences).[2,3] Under such uncertain

We thank Driss Kazi-Tani for computer engineering to provide online administration of the neurology test.

Correspondence may be sent to Bernard Charlin, Faculty of Medecine, University of Montreal, Pavillion Roger-Gaudry, CP, 6128, Succursale Centre-Ville, Montreal, Quebec, H3C 3J7 Canada. E-mail: bernard.charlin@umontreal.ca



Clinical Vignette: You are evaluating a 60 year-old man with left-sided weakness in the emergency room.

If you were thinking of ...	... and then you find...	... this hypothesis becomes ...
1. Cerebral abscess	Patient had dental work 10 days ago	☐ -2: Ruled out or almost ruled out ◉ -1: Less probable ☐ 0: Neither less nor more probable ☐ +1: More probable ☐ +2: Certain or almost certain
If you were thinking of ...	**... and then you find...**	**... this hypothesis becomes ...**
2. Ischemic stroke	Sudden onset 2 hours ago	☐ -2: Ruled out or almost ruled out ☐ -1: Less probable ☐ 0: Neither less nor more probable ☐ +1: More probable ☐ +2: Certain or almost certain

FIG. 1. Script Concordance Test template (online neurology test).

conditions, doctors must engage in a series of judgments, each estimating the magnitude and direction of the effect of new pieces of information on the status of active hypotheses. These judgments are a crucial part of the clinical reasoning process.

The Script Concordance Test (SCT) attempts to measure the quality of these clinical judgments by comparing the performance of examinees to that of a reference panel of experienced physicians on a series of case-based tasks. These "judgment measures" are used as proxy indicators of clinical reasoning quality, not as measures of clinical reasoning as a whole. Rooted in cognitive psychology, the test is based on a theory of how medical knowledge becomes organized in the mind during a trainee's transition from novice to expert.[4]

In an SCT (see Figure 1), examinees are presented with a brief description of an authentic case, followed by a series of questions asking them to make judgments regarding diagnostic possibilities or management options when new elements of information are provided. Although enough clinical context is provided to allow a meaningful decision to be made, a certain amount of uncertainty, imprecision, or incompleteness is deliberately built into each case in order to simulate real-life clinical settings. SCT scenarios introduce uncertainty at two levels:[5] (a) within the case itself (by design, this level of uncertainty is always present), and (b) within the questions nested in each case, which may vary in the level of uncertainty they encompass.

SCTs can be paper based or administered online. Clinicians find SCT appealing because it contains cognitive tasks similar to those they encounter during daily practice. Studies in gynecology, radiology, family medicine, surgery, and neurology have shown high reliability[6–11] and support for some aspects of construct validity with lowest mean scores for medical students, intermediate scores for residents, and higher scores for faculty. The predictive validity of SCT scores in family medicine was reported in a study in which SCT scores at the end of clerkship were correlated with scores on tests of clinical reasoning administered at the end of residency.[12]

In contrast to many conventional forms of testing, such as multiple-choice questionnaires, there are no "correct" answers to the test questions of the SCT; several responses to each question may be considered acceptable. The scoring of the SCT is based on an aggregate method, described by Norman[14] and Norcini,[15] that takes into account the variability of responses of experienced clinicians to particular clinical situations. The examinee's responses to each question are compared with those of a reference panel. Credit is assigned to each response based on how many members of the panel choose that response. A maximum score of 1 is given for the response chosen by most of the panel members (i.e., the mode). Other responses are given fractional scores, depending on the number of panel members choosing them. Responses not selected by panel members receive zero. The final score is meant to reflect how closely the examinee's judgments match, or *concord*, with those of panel members faced with the same set of ill-defined clinical problems.

A disadvantage of the aggregate method is that examinees often have difficulty interpreting their scores in isolation. The scoring schemes of tests with single-right-answer formats, such as multiple-choice questionnaires, provide examinees with an intuitive appreciation of their achievement; for example, a score

of 67 clearly indicates that an examinee has given correct responses to 67% of the test questions. With the aggregate scoring method, SCT scores reflect concordance with those obtained by members of a reference panel. Since scoring is highly dependant of the panel used, for these scores to be meaningful, it is therefore necessary to report the value of the test panel's mean and standard deviation.

The purpose of this article is to propose a standardized method of expressing SCT scores using a method of transformation that is based on the distribution of panel member responses. The effect of this transformation on four data sets obtained from SCTs in radio-oncology, surgery, neurology, and nursing is considered. To our knowledge the transformation method we describe, through straightforward and easy to implement, has not yet been proposed.

METHODOLOGY

Description of the SCT

SCTs[13] comprise a series of short clinical scenarios (cases), each followed by a series of test questions consisting of three parts. The first part ("If you were thinking of") provides a hypothesis in the form of a diagnostic possibility, an investigative option, or a therapeutic alternative that is relevant to the case. The second part ("And then you find") presents new information, such as a physical examination sign, a pre-existing condition, an imaging study, or a laboratory test result, which may (or may not) have an effect on the given hypothesis. The actual question is answered in the third part ("This hypothesis becomes") that contains a 5-point Likert scale, from –2 (*ruled out or almost ruled out*) to +2 (*certain or almost certain*). The examinee indicates in the scale the effect the new information (Part 2) has on the proposed hypothesis (Part 1). Each question related to a particular case is independent of the others. An example of an SCT case and questions is provided in Figure 1.

Scoring

The optimal SCT scoring method is still debated.[16] In this study the usual, aggregate method is used. Physicians with experience in the tested clinical domain are selected as members of the reference panel for a given test. They are asked to complete the test individually, and their answers are used to build the scoring key.[5] Data obtained from the panel are treated anonymously. A panel size of 15 or more members is required to obtain adequate scores reliability.[8] Credit is assigned to each response based on how many of the members on the panel chose that response. To confer the same maximum score for each test question, a slightly modified aggregate method is used.[5] Credit of 1 point is given for the modal answer from the reference panel. Other panel members' choices are attributed a partial credit, proportional to the number of members having provided that answer on the Likert scale divided by the modal value for the item. Answers not chosen by any panel members receive zero. For example (see Table 1), suppose the reference panel comprises 15 members, who respond to a question on the SCT in the following way: none choose the –2 and –1 ratings, 2 choose the 0 rating, 9 choose the +1 rating, and 4 choose the +2 rating. The modal answer in this example is +1. An examinee choosing this rating will receive 1. Selecting the 0 rating will earn 0.22 points (2/9) and the +2 rating 0.44 points (4/9). No points are accorded for selecting the –2 or –1 ratings. To avoid bias, panel members' scores for each question were computed using a scoring key that excluded their own response to that question.

TABLE 1
Scoring basis (raw scores)

Anchors on Script Concordance Test Question	–2	–1	0	+1	+2
No. of Times Anchor Was Chosen by a Panel Member	0	0	2	9	4
Calculation Based on Modal Answer	0/9	0/9	2/9	9/9	4/9
Points Attributed to Examinee	0	0	0.22	1.0	0.44

From a scoring perspective the unit of measurement is the case. When there is more than one question per case, a case score is calculated by averaging the examinee's scores over the number of questions in the case. Calculating a case score is important because the statistical tests that are used for reliability calculations assume that the items are independent. The case is the unit of measurement (or item in a measurement sense), not the questions. It is also important to average the question scores, rather than to simply add them, because averaging ensures that each case is not weighted by the number of questions it contains. In other words, averaging the question score precludes a case with two questions from having twice the weight as a case with only one associated question. For each case the maximum score is 1; final test scores are calculated by adding case scores. The maximum score for the test represents the total number of cases.

The reference panel is important for score interpretation, a process is needed to optimize the answer key to account for outlying results. In this study we elected to remove outliers from the panel. We considered panel members whose scores were below two standard deviations from the panel mean to be outliers.

Transformation of Scores

Standardization is a method of score transformation that expresses deviation from a mean within a distribution of scores. Common methods of standardization express deviation from the mean of all test-takers. However, for an SCT, the final score is intended to reflect the difference in performance between the individual examinee and that of the reference panel. According to our transformation method, scores for each expert are initially

computed as usual (i.e., by treating each panelist as an examinee, and scoring them against a key set by the remaining panel members). Test raw scores are then converted into a scale based on the mean and standard deviation of the panelists alone, not on the mean and standard deviation of all participants (i.e., panelists and examinees combined). The panel mean thus serves as a reference value, and the panel standard deviation is a yardstick by which examinee performance can be measured.

Transformation is carried out according to principles of score standardization whereby the means and standard deviations are fixed. Several families of standardized scores exist, that is, z scores (0,1), College Entrance Examination Board (500,100), or T scores (50, 10). In a first step, z scores are calculated for examinees, with mean and standard deviation of the panel set at 0 and 1. In a second step, we converted z scores into modified T scores with panel mean and standard deviation set at 80 and 5, respectively. The second scale is just a linear transformation of the z scale, and score correlation between scores expressed in these two scales equals 1. If scores must be expressed at the level of the case, the same transformation can be made on raw case scores.

Data sets

Radio-oncology. A radio-oncology test[10] was constructed with cases taken from the three most prevalent fields in the cancer patient population: pulmonary, urological, and breast cancers (10 cases per field). Each case, presented in a short scenario, was followed by three related questions. Two levels of clinical experience were sampled. The first level consisted of 4th-year medical students ($n = 70$) from the University of Montreal who took the exam immediately after attending a lecture about radio-oncology and related topics. Only 4 of them possessed clinical experience in radio-oncology, acquired through an elective rotation (students taking these elective rotations often consider specializing in radio-oncology). All 70 students agreed to participate. The second level consisted of the population of residents of the three residency programs in radiation oncology in the province of Quebec (Montreal, Laval, and McGill Universities)—a total of 52 residents. All those from University of Montreal (22), half of those from McGill (8/16), and 8 out of 14 from Laval took the test (no reason was provided by those who declined). The 38 participating residents represented 72% of radio-oncology residents of the province. 70% (26) were juniors and 30% (11) were seniors (1 resident did not specify his or her year of residency). The reference panel was a sampling of the whole population of board-certified practitioners in radiation oncology of the province of Quebec ($n = 62$). In total, 47 (76%) agreed to participate.

Surgery. An SCT in surgery[9] was developed for assessment of intraoperative decision making. Each item (the case and its two or three questions) was designed to foster reflection-in-action. When preparing the cases, an attempt was made to devise authentic clinical scenarios requiring reasoning skills and some degree of experience. After revision for face validity (i.e., whether the question actually addressed a realistic intraoperative dilemma and whether it tested decision-making skills) and content validity (i.e., whether the examination addressed the objectives of training of both the Royal College of Surgeons of Canada and the American Board of Surgery) the test consisted of 100 questions in 35 cases. It was administered to 36 general surgery residents ranging from R1 to R5. The reference panel consisted of 10 board-certified general surgeons who completed the test independently.

Neurology. An SCT of 24 cases,[11] with 3 to 4 questions each, was developed for a total of 94 questions. The cases represented routine, authentic clinical encounters that occur in both inpatient and outpatient settings. Questions were deliberately constructed to explore the ambiguous or uncertain aspects of each case, so that clinical judgment (rather than simply factual knowledge) was tested. To ensure content validity, a test content blueprint was derived from published lists of "symptom complexes" and "specific diseases" deemed essential subjects to be taught during adult neurology clerkship and residency training. Test questions were then developed to sample the broad range of neurological topics, as well as to include a range of judgment issues relating to diagnosis and prognosis, choice of investigations and treatment, and ethical dilemmas. Thirty-four adult PGY1-PGY5 residents and 8 neurology clerkship students from two North American neurology programs (McGill University and Mayo Clinic, Rochester) volunteered to complete the test. The reference panel comprised 16 attending neurologists from McGill teaching hospitals who were at least 3 years postcertification, who regularly attended on consultation or ward services (at least 1 month per year), and who were recognized for their clinical expertise and teaching skills.

Nursing. An SCT assessing attitudes on caring at the Faculty of Nursing of the University of Montreal was developed. It covered three main aspects of caring using 90 questions in 29 cases. Thirty nursing students were tested. The panel was composed of 12 experienced nurse practitioners. All students and nurse practitioners who were contacted agreed to participate.

All four studies received approval from Institutional Review Boards. Participants signed a consent form before taking the tests. All participants volunteered and their responses were treated anonymously. The radio-oncology and nursing tests were administered in French, and the surgery and neurology tests were administered in English.

Statistical Analysis

In the radio-oncology test, 1 student and 1 resident were removed from the study as a result of missing data (more than 4 missing answers in either the pulmonary, urology, or breast sections). For all other participants, there were few missing answers; in these instances unanswered questions were accorded the average score of all the other questions on the test. Reliability was estimated using the Cronbach alpha coefficient.

TABLE 2
Transformed scores from the four tests and distribution of residents and students relative to the panel mean

	N	*M* (*SD*)	Below 2 *SD*	Below 1 *SD*	From 1 *SD* to Mean	Above the Mean
Radio-oncology						
Panel	42	80.0 (5.0)				
Residents	37	73.5 (8.4)	13 (35%)	5 (14%)	9 (24%)	10 (27%)
Students	70	57.8 (7.4)	67 (96%)	2 (3%)	1 (1%)	
Surgery						
Panel	10	80.0 (5.0)				
Residents	34	63.9 (7.2)	25 (74%)	8 (24%)		1 (2%)
Neurology						
Panel	17	80.0 (5.0)				
Residents	53	74.0 (5.4)	14 (26%)	13 (25%)	18 (34%)	8 (15%)
Nursing						
Panel	12	80.0 (5.0)				
Students	30	71.6 (6.4)	13 (43%)	14 (47%)	3 (10%)	

Test optimization was performed by calculating the corrected case-total correlation for each case and then eliminating cases with case-total correlation of less than .10 in a stepwise manner. The process of optimization was stopped when no cases showed case-total correlation of less than .10.

RESULTS

In the radio-oncology test, 3 panel members were considered outliers (final test score below 2 standard deviations from the mean), and 2 had too much missing data. All 5 were removed from analysis. The panel for the radio-oncology SCT therefore consisted of 42 members. In the nursing test, 2 panel members were outliers and were excluded; in the neurology and the surgery SCTs no exclusions were necessary.

After completion of the optimization process, the radio-oncology test had 30 cases and 90 questions, the surgery test had 31 cases and 90 questions, the neurology test had 24 cases and 70 questions, and the nursing test had 22 cases and 66 questions. Cronbach's alpha coefficient values were 0.86 for the radio-oncology test, 0.70 for the surgery test, 0.72 for the neurology test, and 0.72 for the nursing test.

Using the proposed SCT score transformation method, mean and standard deviations were set at 80 and 5. The mean, standard deviation, and range for panels and examinees in the four tests are shown in Table 2, with the distribution of scores of residents and students over the mean, under 1 *SD* below the mean and under 2 *SD* below the mean. Scores for the 4 data sets are depicted in Figure 2.

DISCUSSION

As a numerical representation of achievement on a test, a test score should provide examinees with a clear gauge of how well they have performed. However, assessment tools and scales currently used in educational and behavioral science research often report scores that are arbitrary or difficult to interpret. For example, 20-item attitude scales, in which each item provides five responses ranging from *strongly disagree* to *strongly agree*, traditionally give a score of 1 for a response of *strongly disagree* and a score of 5 for a response of *strongly agree*. With 20 items, a respondent might earn a score that ranges from a low of 20 to a high of 100. An artifact of the way the scale was constructed, such a raw score is completely arbitrary, meaningless in isolation without further explanation or transformation.[17]

One way to give meaning to an examinee's raw score is through comparison with the scores earned by other test-takers. This normative approach to scoring can be done using percentile ranks or standard scores. A disadvantage of percentile ranks is that they are ordinal and therefore cannot be manipulated like scores on an interval scale. Standard scores represent an equal-interval scale and can thus be subjected to statistical manipulations, such as computation of averages. Standardization methods are commonly used to compare the performance of individual examinees on a test relative to one another. However, to our knowledge, the use of standardization methods for comparing the scores of individual examinees to those of an aggregate panel assembled for the purpose of setting the scoring grid for a particular test has not been previously explored. Such a method for providing meaning to scores

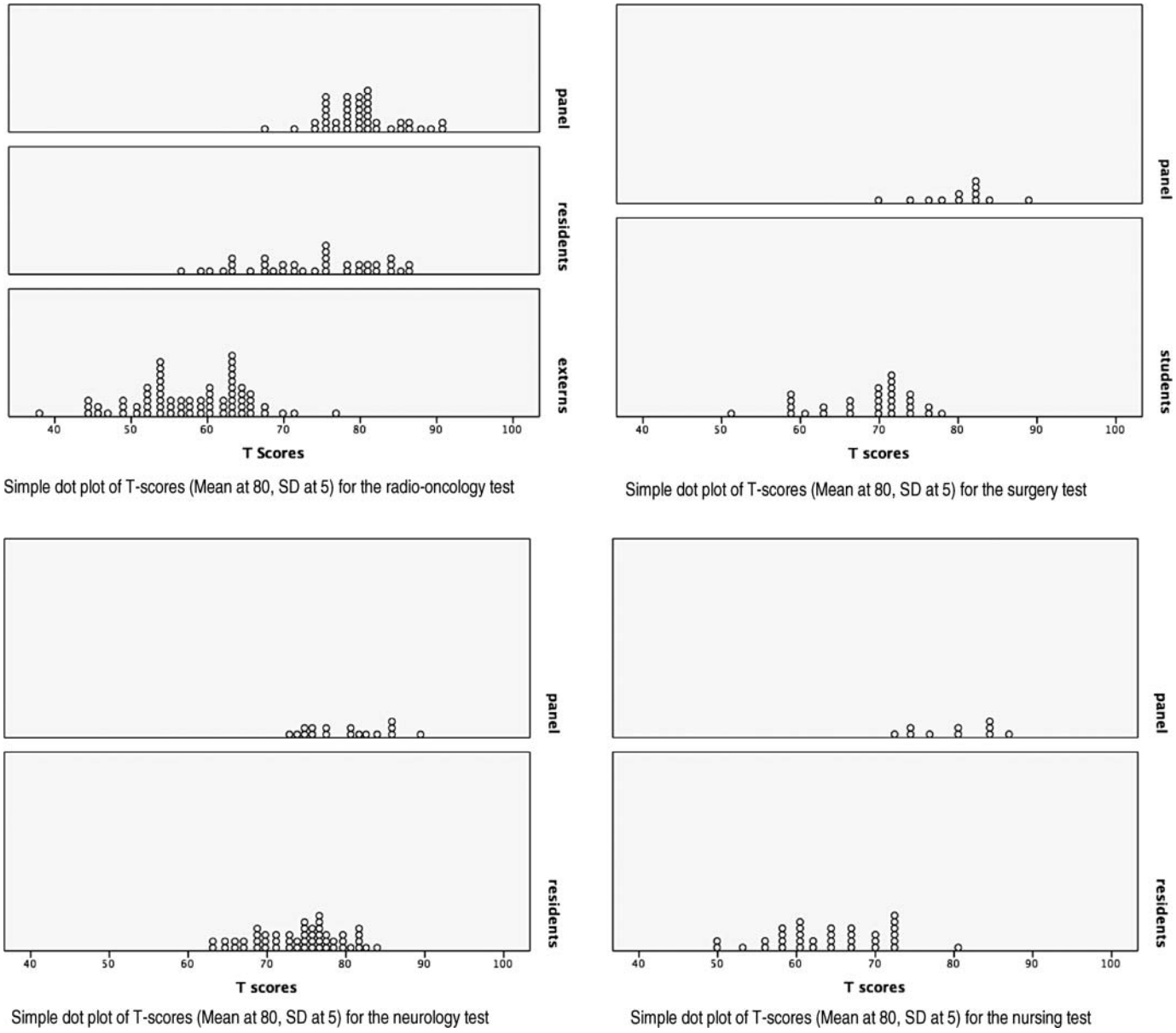


FIG. 2. Distribution of scores within the four tests.

would seem well suited to tests using the script concordance approach.

For each SCT, the scoring key is set by a panel of domain experts. Because aspects of uncertainty are deliberately embedded within each item, SCT questions are not considered to have a single "correct" or "consensus" answer. Instead, the SCT scoring scheme assumes that, for each question, the answer provided by the greatest number of panel members (i.e., the modal answer) reflects optimal reasoning under the given circumstances, while other panel members' answers reflect a difference of interpretation that is still clinically valuable and merits fractional credit. Examinee scores, then, reflect the degree of concordance with the expert panel: the more examinees select modal answers, the higher their final scores will be. A high score indicates that an examinee interprets information pertaining to ill-defined clinical problems similarly to a majority of experienced physicians in the field. (Note that variability in some answers provided by the panel may reflect measurement errors; procedures for rejecting these answers and optimizing the scoring key are currently under investigation.)

The method of SCT scoring in current use issues scores that may be difficult to interpret in isolation. In this study, we present a method of SCT score transformation that uses the mean of panel members' scores as the value of reference, and the standard deviation of the panel serves as a yardstick by which examinee performance is measured. We believe that this

transformation renders SCT scores more meaningful as indicators of examinee performance. Compared with raw scores, the standardized scores provide a familiar normative interpretation, although here the norms refer to the expert group and not the examinee group. This information is useful for gauging how far examinee performance is from the mean of the reference population of practicing professionals they aspire to join. Based on our transformation method, 10 (27%) residents in radio-oncology, 1 (2%) resident in surgery, and 8 (15%) residents in neurology achieved scores above 80, reflecting a mastery of knowledge in these domains that suggests a reasoning capacity compatible with autonomous professional practice.

Two features of the SCT scoring process should be pointed out. First, it is important to acknowledge that SCT scores are not directly comparable across tests. Each SCT compares examinees' performance on specific tasks, in specific contexts, with that of a specific panel. Differences across tests may be the consequence of factors such as variations in task difficulty, composition of resident groups, or panel characteristics. The second feature relates to the sample size of the reference group. Standard scores are unstable with small sample sizes and may yield misleading results. Previous work has shown that a minimum panel size of 15 members is needed to produce an adequate score reliability,[8] and is therefore necessary for confident interpretation of examinee performance using our standard scoring method. Two of the panels we studied included fewer than 15 members (10 for the surgery test, 12 for the nursing test), which may limit the interpretation of transformed scores on these tests.

Our proposed method of SCT score transformation describes a new form of score standardization that could be termed "panel-centered modified *T* scores." The arbitrary selection of 80 as a set value for the panel mean is in accordance with previous work on SCTs in multiple domains, in which reference panel means (expressed as percentages) generally hover around this score. Setting the panel standard deviation at 5 may also be considered reasonable, since a performance level greater than 4 *SD*s above the panel mean would have to be achieved to obtain a score of 100 or more, a level that has never been attained by either examinees or panel members in prior studies using SCT.

This transformation method proposes a common metric basis for reporting SCT scores and provides examinees with clear, interpretable insights into their performance relative to that of physicians of the field. For example, a score of 80 is easily interpretable as "equal to the level of the panel mean," whereas a score of 65, at "3 *SD*s below the panel mean," more clearly conveys that an examinee is far from performing at a panel members' level. Beyond SCT, our transformation method may be generalizable to the scoring of other test formats in which the performance of examinees and a reference panel undertaking the same cognitive tasks is compared.

REFERENCES

1. Fredericksen N. Implications of cognitive theory for instruction in problem-solving. *Review of Educational Research* 1984;54:363–407.
2. Johnson E. Expertise and decision under uncertainty: Performance and process. In M. Chi, R. Glaser, M. Farr (Eds.), *The nature of expertise* (pp. 209–28). Hillsdale, NJ: Erlbaum, 1988.
3. Fox R. Medical uncertainty revisited. In G. Albrecht, R. Fitzpatrick, S. Scrimshaw (Eds.), *Handbook of social studies in health and medicine* (pp. 409–25). London: Sage, 2000.
4. Charlin B, Boshuizen HPA, Custers EJFM, Feltovich PJ. Scripts and clinical reasoning. *Medical Education* 2007,41:1178–84.
5. Fournier JP, Demeester A, Charlin B. Script Concordance Tests: Guidelines for construction. *BioMedCentral, Medical Informatics, and Decision Making* 2008;8:18.
6. Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty: The script concordance approach. *Evaluation & the Health Professions* 2004;27:304–19.
7. Charlin B, Gagnon R, Pelletier J, et al. Assessment in context of uncertainty: The effect of variability within the panel of reference. *Medical Education* 2006;18:22–7.
8. Gagnon R, Charlin B, Coletti M, Sauvé E, Van der Vleuten C. Assessment in the context of uncertainty: How many members are needed on the panel of reference of a script concordance test. *Medical Education* 2005;39: 284–91.
9. Meterissian S, Zabolotny B, Gagnon R, Charlin B. Is the script concordance test a valid instrument for assessment of intraoperative decision-making skills? *American Journal of Surgery* 2007;193:248–51.
10. Lambert C, Gagnon R, Nguyen D, Charlin B. The script concordance test in radiation oncology: validation study of a new tool to assess clinical reasoning. *Radiotherapy and Oncology* 2009;4:7.
11. Lubarsky S, Chalk C, Kazitani D, Gagnon R, Charlin B. The Script Concordance Test: A new tool assessing clinical judgment in neurology. *Canadian Journal of Neurological Sciences* 2009;36:326–31.
12. Braïlovsky C, Charlin B, Beausoleil S, Coté 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. *Medical Education* 2001;35:430–6.
13. Gagnon R, Charlin B, Lambert C, Carrière B, Deschênes MF, van der Vleuten C. Script concordance testing: More cases or more questions? *Advance in Heath Sciences Education* 2008;14:367–75. Accessed at: http://dx.doi.org/10.1007/s10459-008-9120-8
14. Norman GR. Objective measurement of clinical performance. *Medical Education* 1985;19:43–7.
15. Norcini J, Shea J, Day S. The use of the aggregate scoring for a recertification examination. *Evaluation and the Health Professions* 1990;13:241–51.
16. Bland AC, Kreiter CD, Gordon JA: The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine* 2005;80:395–9.
17. Jaeger RM. *Statistics, a spectator sport* (2nd ed.). Thousad Oaks, CA: Sage, 1990.