Teaching and Learning in Medicine, 24(3), 187–193
Copyright © 2012, Taylor & Francis Group, LLC
ISSN: 1040-1334 print / 1532-8015 online
DOI: 10.1080/10401334.2012.692239

APPLIED RESEARCH

Comparing a Script Concordance Examination to a Multiple-Choice Examination on a Core Internal Medicine Clerkship

William Kelly, Steven Durning, and Gerald Denton
Department of Medicine, Uniformed Services University of Health Sciences, Bethesda, Maryland, USA

Background*: Script concordance (SC) questions, in which a learner is given a brief clinical scenario then asked if additional information makes one hypothesis more or less likely, with answers compared to a panel of experts, are designed to reflect a learner's clinical reasoning. *Purpose*: The purpose is to compare reliability, validity, and learner satisfaction between a three-option modified SC examination to a multiple-choice question (MCQ) examination among medical students during a 3rd-year internal medicine clerkship, to compare reliability and learner satisfaction of SC between medical students and a convenience sample of house staff, and to compare learner satisfaction with SC between 1st- and 4th-quarter medical students. *Methods*: Using a prospective cohort design, we compared the reliability of 20-item SC and MCQ examinations, sequentially administered on the same day. To measure validity, scores were compared to scores on the National Board of Medical Examiners (NBME) subject examination in medicine and to a clinical performance measure. SC and MCQ were also administered to a convenience sample of internal medicine house staff. Medical student and house staff were anonymously surveyed regarding satisfaction with the examinations. *Results*: There were 163 students who completed the examinations. With students, the initial reliability of the SC was half that of MCQ (KR20 0.19 vs. 0.41), but with house staff ($n = 15$), reliability was the same (KR20 = 0.52 for both examinations). SC performance correlated with student clinical performance, whereas MCQ did not ($r = .22, p = .005$ vs. .11, $p = .159$). Students reported that SC questions were no more difficult and were answered more quickly than MCQ questions. Both exams were considered easier than NBME, and all 3 were considered equally fair. More students preferred MCQ over SC (55.8% vs. 18.0%), whereas house staff preferred SC (46% vs. 23%; $p = .03$). *Conclusions*: This SC examination was feasible and was more valid than the MCQ examination because of better correlation with clinical performance, despite being initially less reliable and less preferred by students. SC was more reliable and preferred when administered to house staff.

BACKGROUND

Although descriptive evaluation (i.e., narrative comments made by clinical teachers regarding student performance) is the predominant method of evaluation used by internal medicine clerkships, written examinations account for up to one third of a student's grade.[1] The National Board of Medical Examiners (NBME) subject examination in Medicine is frequently administered, but in a 2005 national survey,[1] only 20% of internal medicine clerkships used it as the sole examination, compared with 50% in 1999. Over that same period, administration of local, faculty-developed examinations had become more common (36% vs. 27%) although the contribution of these examinations to final grades decreased (14% vs. 21%). The content, reliability, and validity of these examinations have not been fully described. The multiple-choice question (MCQ) format is most common (unpublished data from the 2009 Clerkship Directors in Internal Medicine survey). MCQs can efficiently and reliably assess medical knowledge, but the answer options provided may "cue" the examinee and may seem removed from real clinical situations.[2] MCQs do not precisely mirror clinical practice as physicians do not get to choose a single best answer out of four provided possibilities when evaluating a real patient.

Script theory[3] postulates that in real clinical practice, physicians apply prestored knowledge sets (or "scripts") to understand a patient's clinical presentation and then either accept or reject this hypothesis when presented with additional information.[4, 5] The scripts of experienced clinicians may vary substantially, but essential elements are believed to be similar, and students can be measured by their agreement—or concordance with—this standard.[6] A recent PubMed search returned 48 studies utilizing script concordance (SC) testing in different settings. In addition to various medicine specialties, these studies included assessing intraoperative decision making,[7,8]

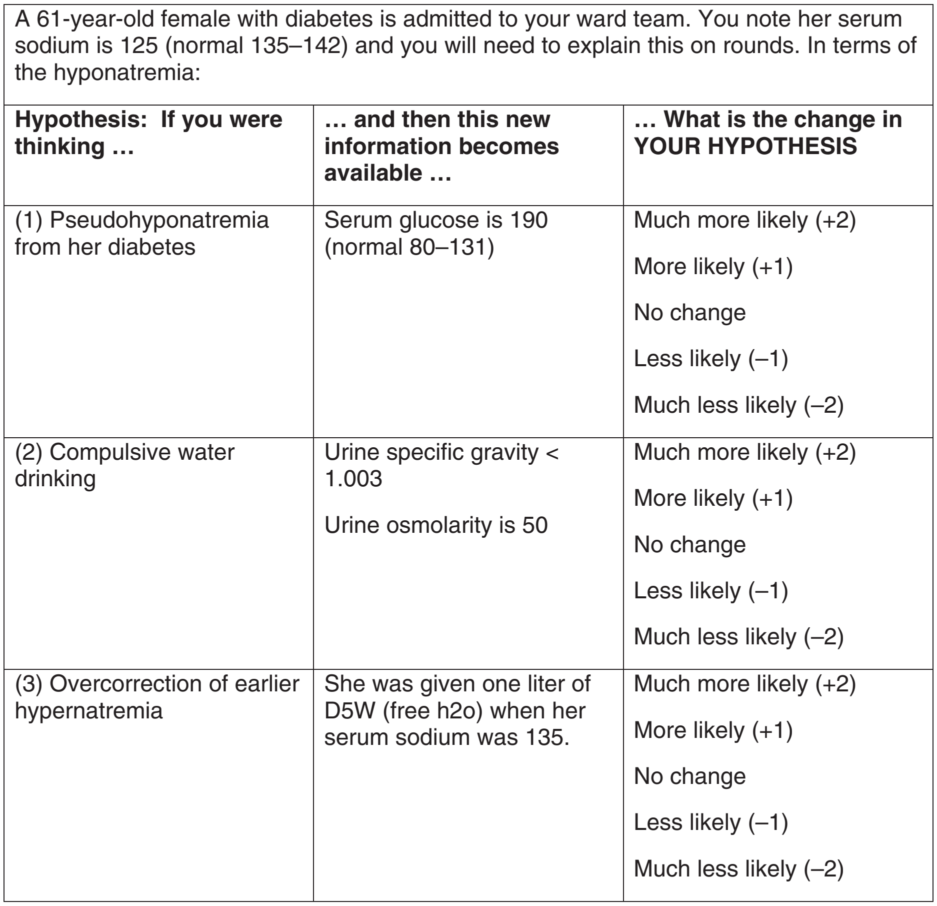
Portions of this data were presented in abstract form at the 2009 CDIM and 2010 AAMC meetings. The opinions in this article are those of the authors and do not represent the official policy of the Uniformed Services University. We thank Mrs. Robin Howard and Dr. Audrey Chang for their assistance with review and statistical analysis.

Correspondence may be sent to William F. Kelly, Medicine-EDP, 4301 Jones Bridge Road, Bethesda, MD 20814-4599, USA. E-mail: william.kelly@usuhs.edu

cultural competence,[9] and as screening for poorly performing physicians.[10] SC testing has been studied in medical students in their preclinical (basic science) years[11] and to assess improvement after curriculum change.[12] Lubarsky et al.[13] recently summarized published validity evidence for SC testing, noting its content and internal structure (reliability) is strongly supported, with response process and educational consequences being less well established.

SC examinations can be constructed using published guidance,[14] starting with brief clinical vignettes. Questions (items) containing supplemental information are then nested in these cases. Rather than choosing one best answer as in MCQs, examinees indicate on a Likert-type scale whether the proposed diagnosis, investigation, or treatment is more or less likely based on new information provided. Thus SC items are hypothesized to test clinical reasoning more directly than MCQs. For a sample SC vignette with nested items, see Figure 1a. The examinee's score for each item is based on how much his or her answers match those of an expert panel. Many test items can be obtained from one vignette, allowing for more items per unit time. Seventy-five SC items can be completed in 1 hour.[14] In pooled analysis of three studies of physicians in residency training and nurse practitioner trainees, Gagnon et al.[15] found that two to four questions nested into each of 15 to 25 cases may be the best combination for content validity and reliability, though this hasn't been studied in medical students. Reports of 20 cases with 60 items (i.e., three SC questions per case) have shown good reliability (Cronbach's $\alpha > .75$).[16,17]

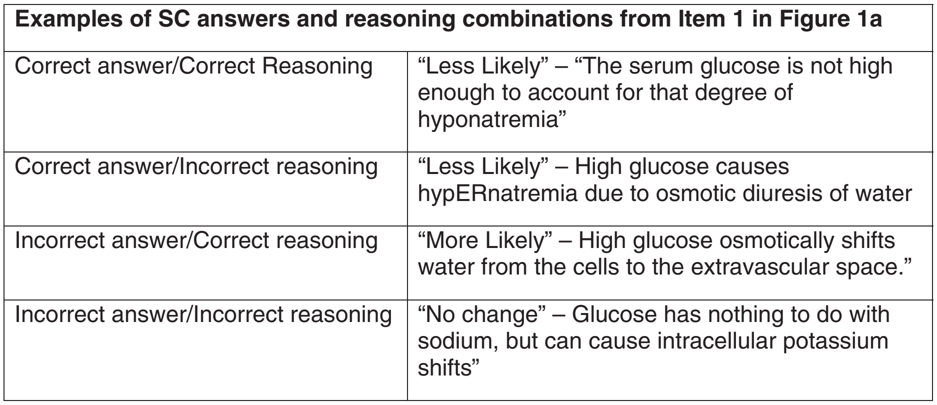
SC questions may use a 3-point or 5-point Likert scale, with 5-point scales ("aggregate scoring, with five-points") in which there is no one, true correct answer being more complex and difficult to answer[10,14] (Figure 1a). Charlin et al.,[18] among their many foundational contributions to SC, have shown that more variability within the expert panel responses is associated with a high effect size for discrimination between groups. Discrimination may come at the expense of reliability (measurement error) if the expert panel includes deviant responses. Gagnon[19] demonstrated that such responses may have a negligible impact if the panel size is large (>15 experts), but if removing them, Gagnon recommended doing so based on their distance from the



A 61-year-old female with diabetes is admitted to your ward team. You note her serum sodium is 125 (normal 135–142) and you will need to explain this on rounds. In terms of the hyponatremia:

Hypothesis: If you were thinking …	**… and then this new information becomes available …**	**… What is the change in YOUR HYPOTHESIS**
(1) Pseudohyponatremia from her diabetes	Serum glucose is 190 (normal 80–131)	Much more likely (+2) More likely (+1) No change Less likely (–1) Much less likely (–2)
(2) Compulsive water drinking	Urine specific gravity < 1.003 Urine osmolarity is 50	Much more likely (+2) More likely (+1) No change Less likely (–1) Much less likely (–2)
(3) Overcorrection of earlier hypernatremia	She was given one liter of D5W (free h2o) when her serum sodium was 135.	Much more likely (+2) More likely (+1) No change Less likely (–1) Much less likely (–2)

FIG. 1a. Example of Script Concordance format. *Note.* Several questions (items) can be nested into one clinical vignette. Scores are based on how close the learner's answers are to a panel of experts that have already taken the test. The common 5-point scale is shown here, though in our study we used 3 points (*more likely*, *no change*, or *less likely*).



Examples of SC answers and reasoning combinations from Item 1 in Figure 1a	
Correct answer/Correct Reasoning	"Less Likely" – "The serum glucose is not high enough to account for that degree of hyponatremia"
Correct answer/Incorrect reasoning	"Less Likely" – High glucose causes hypERnatremia due to osmotic diuresis of water
Incorrect answer/Correct reasoning	"More Likely" – High glucose osmotically shifts water from the cells to the extravascular space."
Incorrect answer/Incorrect reasoning	"No change" – Glucose has nothing to do with sodium, but can cause intracellular potassium shifts"

FIG. 1b. With the addition of free text justification, there are four possible outcomes.

modal score. Bland et al.[20] compared multiple different ways of scoring and showed 3-point mode "single best answer" scoring to have a similar validity coefficient (a score correlating with level of training) with an only slightly lower reliability. As they pointed out, Likert scoring can erroneously judge examinees who fail to recognize the direction of a changing hypothesis (as more or less likely) to be the same as examinees that appreciate the direction but fail to agree with the expert panel on the degree of change. Single-best answer scoring may also be more appropriate when assessing junior learners like medical undergraduates. Three-point scales offer the advantage of ease of standard-setting and test administration and are more appropriate for medical students, whose clinical reasoning is often still in an early developmental stage. More recently, Charlin et al.[21] have recommended transforming raw scores into a scale in which the expert panel's mean is set as the value of reference and the standard deviation is used to measure examinee performance.

We decided to replace our locally developed MCQ examination with an SC examination in the hopes of having better assessment of medical knowledge and clinical reasoning, and a more reliable and valid test instrument, that is, a format perceived as more representative of actual clinical practice by students and teachers. This change was prospectively studied to assess test instrument quality and learner satisfaction.

The purposes of this project were (a) to compare feasibility, reliability, validity, and learner satisfaction of a three-option SC examination with a MCQ examination during a 3rd-year internal medicine clerkship; (b) to compare feasibility and acceptability of SC questions between medical students and a convenience sample of house staff; and (c) to compare learner satisfaction with SC between first- and fourth-quarter medical students.

METHODS

Design

This is a prospective comparison study of faculty-written MCQ and faculty-written SC questions. We replaced our existing, low-stakes 30-item MCQ with two 20-item exams, one MCQ, and one SC, given to all 3rd-year medical students at the end of their internal medicine clerkship during the 2008–2009 academic year. The Institutional Review Board at the Uniformed Services University determined that this project did not require review.

Exam and Survey Development

For the MCQ portion, the 20 best questions from our existing exam that historically performed well psychometrically (in terms of difficulty, discrimination, and interitem reliability) were retained. For the SC portion, questions were created to cover the same content areas as in the MCQs but with 20 questions nested within five clinical vignettes. A published SC exam writing guide[14] was followed including creation by two of the authors (WFK and GDD), adequate content sampling, and item selection and scoring based on a reference standard of 10 experts (our internal medicine clerkship site directors). Questions that the expert panel felt were poorly worded were edited or omitted, with a goal of unanimity in expert responses. Students then answered along a 3-point Likert scale of *less likely*, *no change*, or *more likely*. We used modal scoring in which the most common, or modal, answer given by the experts was considered to be the correct answer.

The authors developed a paper-based learner satisfaction survey regarding the NBME, MCQ, and SC exams, including questions with 7-point Likert-scaled responses concerning test difficulty (*too easy* to *too difficult*), time pressure (*too much time* to *not enough time*), clinical relevance (*not relevant at all* to *completely relevant*), and fairness (*completely fair* to *totally unfair*). The survey included space for participants to write free-text comments and to indicate their format preferences.

House Staff Sampling

Prior to administering the exams and survey to medical students, the MCQ and SC examinations and the survey were administered to a convenience sample of 15 internal medicine residents at a house staff meeting to assess feasibility of the formats, validity of the instrument, and opinions of these learners.

House staff did not take the NBME examination, so they did not answer survey questions regarding that examination.

Exam and Survey Administration

The MCQ and SC tests were administered to medical students independently and sequentially during the final week of their 12-week internal medicine clerkship. The order of administration varied with each quarterly exam administration. Twenty-five minutes were allowed for each of the exams, which were closed-book and proctored. Remaining time could not be used to return to the other exam format. Answers were entered on standard bubble-sheets for computerized scoring. Students were also asked to write a very brief, free-text justification of their clinical reasoning for each SC answer in a separate test booklet. On the same test day, students also completed the NBME examination, which is a 100-question multiple-choice exam that lasts 2.5 hours. At the end of the test day, all students completed the learner satisfaction survey, which was anonymous and paper based.

Analysis

Test questions were machine scored and standard psychometric properties of the two tests were determined. SC scores were calculated dichotomously as correct or incorrect using 3-point Likert modal scoring, and, in a separate analysis, exams were scored using 3-point Likert "distance from mode" scoring in which a student could be 0, 1, or 2 points away from the most common expert answer to each item. For the purposes of this project, free-text answers were reviewed by one of the authors (WFK) and matched with SC questions to create four categories: correct, with or without good clinical reasoning, and incorrect, with or without good reasoning (Figure 1b). Kuder-Richardson 20 (KR20) was used to compare reliability of the MCQ and SC exam scores of students and house staff. KR20 is a measure of internal consistency for instruments with dichotomous choices. It is analogous to Cronbach's alpha, which can be used for continuous measures.[22] Values greater than 0.9 indicate excellent internal consistency.

Student clinical performance, one of our usual methods of student assessment during this clerkship, is a weighted summation of teachers' clinical grade recommendations based on the RIME scheme.[23] The RIME scheme is a validated framework for evaluation widely used in undergraduate medical education. As a measure of validity, we correlated MCQ and SC exam scores to student clinical performance using Pearson's correlation coefficient. For the learner satisfaction survey, we reported means and summarized free-text comments. Mean Likert-scale differences on the learner satisfaction survey between students and house staff and between first and fourth quarter students were compared using parametric and nonparametric tests including Wilcoxon Signed Ranks test. Data analysis was performed using SPSS, with statistical significance level of .05.

RESULTS

All 163 students completed the exams, and all but two participated in the learner satisfaction survey. Student performance on the MCQ and SC exams was similar when scored dichotomously; mean (standard deviation) correct out of 20 items was 13.24 (2.32) versus 13.69 (1.90). Each exam had the same number (15; 75%) of "acceptably difficult" items (ones that no less than 30% of students but no more than 90% answered correctly) with difficulty Index (*M, SD*) being 66.23 (24.77) versus 68.50 (26.91), $p = .67$, respectively. Our MCQ and SC items both had low but similar discriminating power (ability to separate the top and bottom third of students based on overall raw score) with mean (standard deviation) of 8.00 (5.32) versus 6.53 (4.77), $p = .33$, and point-biserial coefficient (correlation between performance on an individual item and performance overall), mean (standard deviation) of 0.28 (0.13) versus 0.24 (0.12), $p = .14$.

Neither exam had good reliability, but the SC exam had half the reliability of the MCQ (KR20 0.19 vs. 0.41). Among internal medicine house staff ($N = 15$), there was no difference in reliability (KR20 = 0.52 for both formats). Test item analysis allowed improvement of reliability of the student's SC examination with elimination of one question (KR20 = 0.32). Reliability of the SC examination did not change when scored as a continuous variable, that is, distance from the mode, in which a student could score 0, 1, or 2 points away from each designated correct answer.

Overall, the SC items were answered correctly 65.0% of the time. The free-text justification for these correct answers demonstrated good clinical reasoning almost all the time (94.8%). Documentation of reasoning was missing or incorrect in the other 5.2%. An example of right-for-the-wrong-reason would be noting that in a renal failure vignette a fractional excretion of sodium (FENA) of 4% would make a prerenal etiology "less likely" but then writing "because the FENA should be higher" or simply writing "I have no idea." When SC items were answered incorrectly, 40.8% still had some correct medical knowledge in the free-text answer. The most common example was in a vignette with a post-op orthopedic surgery patient with chest tightness and shortness of breath. Given that pulmonary embolism was suspected, a normal chest X-ray would make pulmonary embolism "more likely." Many students indicated "no change" or "less likely" while still noting correctly in their free-text comments that "PE doesn't normally show up on x-ray," "unlikely to see Hampton's hump," or "x-ray is insensitive, need a VQ scan or spiral CT." Whereas most students demonstrated sound clinical reasoning, many seemed uncomfortable with real-time diagnostic uncertainty such as being unwilling to let an ECG without ischemic changes decrease their suspicion for myocardial infarction without yet having the cardiac enzymes. Others appreciated lab results strictly as either normal or abnormal rather than as a continuum. After reviewing free-text responses and the psychometric analysis of test performance, we rewrote a few of the questions, and the resulting 30-item SC

TABLE 1
Correlation of exam scores to one another and to clinical performance

	NBME	MCQ	SC	Clinical Performance
NBME		0.352	0.352	0.484
		p <.001	p <.001	p <.001
MCQ			0.360	0.111
			p <.001	p =.159
SC				0.221
				p =.005

Note. NMBE = National Board of Medical Examiners; MCQ = multiple-choice question; SC = script concordance; clinical performance = a clerkship performance measure used in this internal medicine clerkship and a summation of teachers' weighted grade recommendations during the clerkship.

examination had improved reliability (KR20 = 0.54) when prospectively administered to a new group of students.

As seen in Table 1, MCQ and SC exam performance were weakly correlated with each other and with the NBME examination. The NBME examination correlated moderately with student clinical performance (r = .48, p < .01), whereas the SC correlated weakly (r = .22, p < .01), and the MCQ was not correlated with student clinical performance (r = .11, p = .11).

Results from the Learner Satisfaction Survey are shown in Table 2. As compared to the MCQ and SC exams, students were more likely to report that the NBME was difficult (mean 7-point Likert scale rating of 5.5 vs. 4.2 and 4.4, respectively; p < .01). Further, students were more likely to report that the NBME was time-pressured compared to the MCQ and SC exams (5.9 vs. 4.4 and 3.7, respectively; p < .01) and that our MCQ was more time pressured than the SC test (4.4 vs. 3.7, p < .01). As compared to the SC, the MCQ exam was more frequently reported as clinically relevant (4.8 vs. 4.2, p < .01) even though the constructs tested were the same. All three exams were rated as equally fair, though the mean Likert-scale ratings were in the neutral category.

Free-text opinions about SC described it as "too subjective" and "too open to interpretation." One comment was especially discouraging, noting, "no single test would change my clinical suspicion, because in real life you order a million tests to get the answer." Several students noted, "It was nice getting to THINK a little bit for once."

TABLE 2
Medical student opinions regarding the National Board of Medical Examiners Subject Examination in Medicine (NBME), a faculty developed multiple-choice examination (MCQ), and a faculty developed script concordance examination (SC)

		Distribution of 7-Point Likert-Scaled Responses, No. of Students (%)			
Question	Exam	1–3	4	5–7	M (SD)
Regarding this exam, Indicate how difficult you thought it was (circle):		Too Easy	Neutral	Too Difficult	
	NBME	2 (1.3%)	21 (13.7%)	130 (85%)	5.5 (0.9)
	MCQ	32 (19.9%)	72 (44.7%)	57 (35.4%)	4.2 (1.0)^{a}
	SC	30 (18.6)%	67 (41.6%)	64 (39.8%)	4.4 (1.0)^{a,b}
Regarding this exam, Indicate how much time was allotted (circle):		Too Much Time	Neutral	Not Enough Time	
	NBME	2 (1.3%)	25 (16.3%)	126 (82.4%)	5.9 (1.1)
	MCQ	22 (20.6%)	69 (43.1%)	58 (36.3%)	4.4 (1.4)^{c}
	SC	61 (37.9%)	67 (41.6%)	33 (20.5%)	3.7 (1.4)^{c}
Regarding this exam, Indicate how relevant the questions were to "actual clinical practice" (circle):		Not Relevant at All	Neutral	Completely Relevant	
	NBME	50 (32.7%)	32 (21%)	71 (46.4%)	4.4 (1.4)
	MCQ	28 (17.5%)	37 (23.1%)	95 (59.4%)	4.8 (1.4)^{d}
	SC	47 (29.2%)	47 (29.2%)	67 (41.6%)	4.2 (1.5)
Regarding this exam, Indicate how fair you thought it was (circle):		Totally Unfair	Neutral	Completely Fair	
	NBME	45 (33.3%)	33 (24.4%)	57 (42.2%)	4.0 (1.3)
	MCQ	55 (38.7%)	38 (26.8%)	49 (34.5%)	3.8 (1.6)
	SC	58 (40.8%)	35 (24.6%)	49 (34.5%)	3.9 (1.6)

Note. All other relationships were non-significant; Wilcoxon signed rank test. ^{a}MCQ and SC reported as significantly less difficult than NBME (p < .001 for both). ^{b}SC not reported as any more difficult than MCQ (p = .117). ^{c}SC reported as significantly less time pressured than MCQ (p < .001); NBME more time pressured than SC or MCQ (p < .001). ^{d}MCQ reported as more clinically relevant than SC (p < .001) but not NBME (p = .10).

Students in the last quarter of their 3rd year of medical school, compared to those in their 1st, reported time pressure less frequently with the SC test (mean 7-point Likert scale rating of 3.2 vs. 4.2, $p < .01$), whereas they were more likely to report time pressure with the faculty-written MCQ (4.7 vs. 4.0, $p = .031$). The MCQ was more likely to be reported as too difficult by fourth-quarter students (4.5 vs. 3.9, $p = .019$). Regarding the NBME, fourth-quarter, 3rd-year students were less likely to report it as time pressured (5.3 vs. 6.4, $p < .01$) and were also less likely to report it as clinically relevant (4.2 vs. 4.9, $p = .046$) than first-quarter students.

When asked, a majority of students preferred MCQ (55%, 78 students), 18% (26 students) preferred SC, and 26% (37 students) were indifferent. Students cited familiarity and perceived objectivity with the MCQ format. Students' opinions were significantly different from the small group of internal medicine house staff, in which 46% (6 house staff) preferred SC, 23% (3 house staff) MCQ, and 31% (4 house staff) had no preference ($p = .03$). In the opinion survey, the house staff also appreciated the SC as more clinically relevant (7-point Likert scale mean 6.2 vs. 4.2, $p < .001$) than the medical students.

DISCUSSION

In transitioning formats of our faculty-written internal medicine clerkship examination from MCQ to SC, we found that, after taking both formats, students still preferred the familiar MCQ but perceived SC as no more difficult and much more quickly answered—even with the added requirement of free-text justification of each answer. This likely reflects the fact that several items can be asked under each clinical vignette.

In terms of feasibility for faculty, SC exam construction was readily aided by published guidelines[14] and Web-based tools,[17] and scoring could be done by machine as with traditional MCQ. We added the requirement of brief, free-text justifications for each SC answer. This discouraged guessing and offered insight into students' clinical reasoning and, at times, highlighted their inexperience.

Reliability for the initial version of our SC test was poor relative to MCQ. Through standard computerized identification of items with unacceptable difficulty, discrimination, or reliability indices, and aided by review of those free-text SC answer justifications, the questions were revised and reliability of the examination improved. Because more SC than MCQ can be answered in the same period, reliability of SC can be improved by adding more SC questions to the examination without having to increase the duration of the examination. Reliability of the SC examination could be improved to approximate the NBME, if the number of questions and duration of the exam were increased sufficiently. This would be necessary if ever used for high-stakes purposes.

Because the NBME correlated the best with actual clinical clerkship performance and is also a highly reliable test instrument, locally developed faculty written exams may seem superfluous, regardless of format. The NBME is a long exam (currently 2.5 hr), with many questions (currently 100); these are two reasons for its good reliability. Internal medicine clerkship directors continue to develop and administer their own exams according to as-yet unpublished 2009 national survey results, for many reasons. Perhaps locally developed examinations are shorter, correlate better to clinical performance, or predict future performance. Reasons to continue locally developed examinations include having more local control of exam content and more detailed feedback on student performance than provided by the NBME alone.

House staff preferred SC, whereas medical students preferred MCQ. Although we can only speculate about the reasons for this, perhaps house staff are more comfortable with making interpretation and management decisions and are further out from the ubiquitous MCQ formats of medical school questions, whereas medical students find the MCQ format comfortable and familiar. Or perhaps house staff saw the SC test as much closer to actual clinical practice, where one has to decide how much value new information adds to the evaluation of an existing patient's disease process.

Our study has several limitations. First, sample size of 163 was reasonable, but the use of any locally developed test instrument from a single institution limits generalizability. Second, the MCQ and SC exams were short (20 items each) due to test day time constraints. This limits adequate sampling of domains and internal reliability calculations and other comparisons that are directly related to the number of test items. Furthermore, the MCQ was selected based on prior psychometric performance, whereas the SC was new. Third, the opinion survey domains of "difficulty" and "time pressure" are likely corelated. We asked about test time pressure but did not actually record time required to complete the tests. All survey domains including "exam fairness" are not well defined and instead were left to the perception of the individual examinee. The actual significance of the statistically significant Likert-scaled opinion differences may be debated. However, more than twice as many students reported having "too much time" to answer the SC compared with the MCQ.

Our study also has several strengths. First, it is novel in that it directly compares SC and MCQ formats that are covering the same subject matter and within the same population. Another study[24] used both SC and true/false questions but did not obtain learner feedback. Second, to our knowledge ours is the first to require free-text justification of SC answers. Although this generated some interesting insights similar to short-answer and structured essay questions and identified potential problems with test items, further study is needed including establishment of interrater reliability to assess for the value added, if any, to current SC scoring methods. Third, learner satisfaction and opinions were anonymously obtained and made relative to the national standard, the medicine NBME subject examination. A control group of medicine house staff was also used. Finally, exams were administered under authentic test conditions in which the students were invested in their performance as scores contributed to their clerkship grade.

CONCLUSIONS

SC was preferred over MCQ by only a minority of students but was no more difficult, could be answered with less time pressure, and correlated better with actual clinical performance. Available resources and machine-scored 3-point modal scoring made SC construction highly feasible. SC reliability was prospectively improved using item analysis. Free-text answers to SC questions provided interesting insights and aided in test item quality control.

REFERENCES

1. Hemmer PA, Papp KK, Mechaber AJ, Durning SJ. Evaluation, grading, and use of the RIME vocabulary on internal medicine clerkships: Results of a national survey and comparison to other clinical clerkships. *Teaching and Learning in Medicine* 2008;20:118–26. PubMed PMID: 18444197.
2. Epstein RM. Assessment in medical education [Review]. *New England Journal of Medicine* 2007;25;356:387–96. PubMed PMID: 17251535.
3. Feltovich PJ, Barrows HS. Issues of generality in medical problem solving. In HG Schmidt, ML De Volder (Eds.), *Tutorials in problem-based learning: A new direction in teaching the health professions* (pp. 128–142). Assen, The Netherlands: Van Gorcum, 1984.
4. Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Academic Medicine* 2000;75:182–90. PubMed PMID: 10693854.
5. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance test: A tool to assess the reflective clinician. *Teaching and Learning in Medicine* 2000;12:189–95. PubMed PMID: 11273368.
6. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: A ten-year retrospective. *Evaluation & the Health Professions* 1990;13:5–36.
7. Park AJ, Barber MD, Bent AE, et al. Assessment of intraoperative judgment during gynecologic surgery using the Script Concordance Test. *American Journal of Obstetrics and Gynecology* 2010;203:240 e1–6.
8. 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.
9. Ross PT, Uijtdehaage S, Lypson ML. Reflections on culture: Views on script concordance testing. *Medical Education* 2010;44:505–6.
10. Goulet F, Jacques A, Gagnon R, Charlin B, Shabah A. Poorly performing physicians: Does the Script Concordance Test detect bad clinical reasoning? *Journal of Continuing Education in the Health Professions* 2010;30: 161–6.
11. Humbert AJ, Johnson MT, Miech E, Friedberg F, Grackin JA, Seidman PA. Assessment of clinical reasoning: A Script Concordance test designed for pre-clinical medical students. *Medical Teacher* 2011;33:472–7.
12. Jacobson K, Fisher DL, Hoffman K, Tsoulas KD. Integrated Cases Section: A course designed to promote clinical reasoning in year 2 medical students. *Teaching and Learning in Medicine* 2010;22:312–6.
13. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CP. Script concordance testing: A review of published validity evidence. *Medical Education* 2011;45:329–38.
14. Fournier JP, Demeester A, Charlin B. Script concordance tests: Guidelines for construction. *BMC Medical Informatics and Decision Making* 2008;8:18.
15. Gagnon R, Charlin B, Lambert C, Carrière B, Van der Vleuten C. Script concordance testing: More cases or more questions? *Advances in Health Sciences Education: Theory and Practice* 2009;14:367–75.
16. Marie I, Sibert L, Roussel F, Hellot MF, Lechevallier J, Weber J. The Script Concordance Test: A new evaluation method of both clinical reasoning and skills in internal medicine. *Revue de Médecine Interne* 2005;26:501–7.
17. Script Concordance Tests. Montreal, Quebec, Canada: University of Montreal.Available at: http://www.cpass.umontreal.ca/sct.html. Accessed December 31, 2010.
18. Charlin B, Gagnon R, Pelletier J, et al. Assessment of clinical reasoning in the context of uncertainty: The effect of variability within the reference panel. *Medical Education* 2006;40:848–54.
19. 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? *Advances in Health Sciences Education: Theory and Practice* 2011;16:601–8.
20. 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.
21. Charlin B, Gagnon R, Lubarsky S, et al. Assessment in the context of uncertainty using the script concordance test: More meaning for scores. *Teaching and Learning in Medicine* 2010;22:180–6.
22. Cortina, JM. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychiatry* 1993;78:98–104.
23. Pangaro L. A new vocabulary and other innovations for improving descriptive in-training evaluations. *Academic Medicine* 1999;74: 1203–7.
24. Collard A, Gelaes S, Vanbelle S, et al. Reasoning versus knowledge retention and ascertainment throughout a problem-based learning curriculum. *Medical Education* 2009;43:854–65.