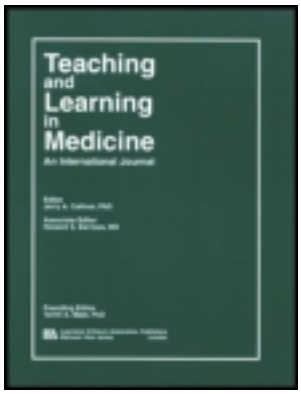

Teaching and Learning in Medicine: An International Journal

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/htlm20

Analyzing Script Concordance Test Scoring Methods and Items by Difficulty and Type

Adam B. Wilson [a] , Gary R. Pike [b] & Aloysius J. Humbert [c]

[a] Department of Surgery , Indiana University , Indianapolis , Indiana , USA

[b] Office of Information Management and Institutional Research , Indiana University-Purdue University Indianapolis , Indianapolis , Indiana , USA

[c] Office of Undergraduate Medical Education , Indiana University , Indianapolis , Indiana , USA

Published online: 04 Apr 2014.

To cite this article: Adam B. Wilson , Gary R. Pike & Aloysius J. Humbert (2014) Analyzing Script Concordance Test Scoring Methods and Items by Difficulty and Type, Teaching and Learning in Medicine: An International Journal, 26:2, 135-145, DOI: 10.1080/10401334.2014.884464

To link to this article: http://dx.doi.org/10.1080/10401334.2014.884464

Teaching and Learning in Medicine, 26(2), 135–145

ISSN: 1040-1334 print / 1532-8015 online
DOI: 10.1080/10401334.2014.884464

Routledge
Taylor & Francis Group

Analyzing Script Concordance Test Scoring Methods and Items by Difficulty and Type

Adam B. Wilson
Department of Surgery, Indiana University, Indianapolis, Indiana, USA

Gary R. Pike
Office of Information Management and Institutional Research, Indiana University–Purdue University Indianapolis, Indianapolis, Indiana, USA

Aloysius J. Humbert
Office of Undergraduate Medical Education, Indiana University, Indianapolis, Indiana, USA

***Background*: A battery of various psychometric assessments has been conducted on script concordance tests (SCTs) that are purported to measure data interpretation, an essential component of clinical reasoning. Although the breadth of published SCT research is broad, best practice controversies and evidentiary gaps remain. *Purposes*: In this study, SCT data were used to test the psychometric properties of 6 scoring methods. In addition, this study explored whether SCT items clustered by difficulty and type were able to discriminate between medical training levels. *Methods*: SCT scores from a problem-solving SCT (SCT-PS; $n = 522$) and emergency medicine SCT (SCT-EM; $n = 1{,}040$) were collected at a large institution of medicine. Item analyses were performed to optimize each dataset. Items were categorized into difficulty levels and organized into types. Correlational analyses, one-way multivariate analysis of variance (MANOVA), repeated measures analysis of variance (ANOVA), and one-way ANOVA were conducted to explore study aims. *Results*: All 6 scoring methods differentiated between training levels. Longitudinal analysis of SCT-PS data reported that MS4s significantly ($p < .001$) outperformed their scores as MS2s in all difficulty categories. Cross-sectional analysis of SCT-EM data reported significant differences ($p < .001$) between experienced EM physicians, EM residents, and MS4s at each level of difficulty. Items categorized by type were also able to detect training level disparities. *Conclusions*: Of the 6 scoring methods, 5-point scoring solutions generated more reliable measures of data interpretation than 3-point scoring methods. Data interpretation abilities were a function of experience at every level of item difficulty. Items categorized by type exhibited discriminatory power providing modest evidence toward the construct validity of SCTs.**

Keywords script concordance tests, clinical reasoning, psychometrics

INTRODUCTION

It is commonly accepted that routine judgments made during clinical reasoning processes can be probed and subsequently measured.[1] One approach to probing these judgments is the script concordance test (SCT) that is theorized to assess one's ability to interpret clinical data. Unlike multiple-choice exams, SCTs measure how the reasoning practices of examinees compare to a panel of experienced physicians in the field under examination; this is collectively referred to as aggregate scoring. The exam answer key is therefore derived according to the perceptions of experienced physicians who traditionally record their responses to items on a 5-point Likert type scale. Typical SCT questions follow a "key features" approach in that they are reflective of those features that physicians find most pertinent for solving commonly encountered clinical scenarios.[2] Despite the longevity and wealth of SCT research, evidentiary gaps remain in the domain of SCT scoring.

Although one benefit of SCTs is their ability to reliably distinguish between medical training levels, this trait is thought to be largely a consequence of the aggregate scoring approach.[3,4] Some contest the aggregate scoring method altogether and advocate for single-best-answer scoring.[5] In an attempt to settle the controversy of aggregate versus consensus scoring, Bland et al. ran statistical analyses on five different scoring keys for marking SCTs. It was found that 5-point and 3-point aggregate scoring keys were similar, as reliability values were nearly identical and correlations of scores against levels of training were statistically significant and moderate in magnitude.[5] Results suggested that 5-point scaling systems added very little

Outcomes of this study were presented at the Association of American Medical Colleges Central Group on Educational Affairs meeting hosted by the University of Cincinnati College of Medicine, March 21-23, 2013. We wish to express our appreciation to Drs. James Brokaw and Mark Seifert of Indiana University School of Medicine for reviewing early drafts of this manuscript and providing thoughtful comments.

Correspondence may be sent to Adam B. Wilson, Department of Surgery, Indiana University School of Medicine, 545 Barnhill Drive, Emerson Hall 543, Indianapolis, IN 46202, USA. E-mail: wilsoadb@iupui.edu

discriminative information and 3-point scales were sufficient. Three-point scoring methods that accounted for differences in distance from either the mean or mode response were reasonably reliable and effective at distinguishing between levels of experience.[5] However, the study by Bland et al. did not explore all scoring method possibilities, and their work was, to some extent, limited by moderate sample sizes. Investigating the concurrent validity of other scoring methods may help to discern if administrators' preferences for or against certain scoring solutions are more or less arbitrary and may shed light on whether it is valuable to have an SCT scoring approach that measures examinees' responses in terms of both direction and degree of impact.

Sensitive clinical reasoning instruments with sound psychometric properties, ostensibly, should be able to detect an increase in data interpretation abilities as experience is gained. However, some literature concerning the relationship between data interpretation and experience is contradictive and cloudy.[6,7] It is not our intent to deliberately debate whether data interpretation skills are a function of experience. Rather, we are taking the position that some instruments or testing methodologies are better equipped to measure data interpretation gains and disparities more so than others. Because this research is focused on investigating the scoring properties of SCTs, our aim was to demonstrate that SCTs are well equipped to differentiate between training levels under conditions of various scoring methods and at the level of item difficulty and item type. Instruments the exhibit dependable discriminatory properties at multiple levels tend to be more valuable and disclose more meaningful information than instruments lacking such qualities.

In its own right, this study is unique in that it compares previously tested scoring methods to methods that are absent or observed less frequently in the literature and delves deeper to understand the discriminatory nature of SCTs at levels other than the test (i.e., composite score) level. Our specific research questions were (a) "How well do non-traditional SCT scoring methods, compared to 5-point aggregate scoring, differentiate between stages of medical training development?" and (b) "To what extent are discriminatory differences heightened or lessened by considering item difficulty and item type?"

METHODS

This large-scale retrospective data analysis included test scores from undergraduate medical students, emergency medicine (EM) residents, and practicing board certified EM physicians who completed either a problem-solving script concordance test (SCT-PS) and/or an emergency medicine script concordance test (SCT-EM) in its entirety.

The SCT-PS was administered to undergraduate medical students ($n = 522$) at Indiana University School of Medicine (IUSM) twice during their enrollment; once during Year 2 (SCT-PS-MS2), while students were dispersed at one of nine IUSM centers, and once during Year 4 (SCT-PS-MS4), at which time all medical students studied at the main centrally located campus. The SCT-PS exam and answer key were created by faculty of IUSM and the State University of New York–Stonybrook as an assessment of problem-solving competence.[8] SCT-PS reference panel participants ($n_{panel} = 13$) were experienced physicians from family medicine and general internal medicine.

In their 4th year, students were required, per clerkship mandates, to complete an SCT-EM in emergency medicine. SCT-EM scores of undergraduate medical students ($n = 988$) comprised the majority of study data. SCT-EM scores from EM residents ($n = 40$) and experienced EM physicians ($n = 12$) were also included in particular analyses. EM residents, postgraduate year 1–3, (EM-PGY1-3) participated on a voluntary basis and were not incentivized for their time. Internal EM physicians comprised the reference panel and participated voluntarily. For the purpose of SCT-EM answer key creation, a panel of board-certified EM physicians, who had at minimum 5 years of clinical experience, were recruited and utilized.[9]

Instrument Blueprints

The SCT-PS, taken by MS2s and MS4s since 2008, had 75 diagnostic-oriented questions nested within 16 cases. From February 2008 to May 2011, the Department of Emergency Medicine administered the SCT-EM to students on EM clerkship composed of 59 items nested within 12 cases. Twenty-three items labeled as "diagnostic" probed how examinees handled new information in light of hypothesized diagnoses, 16 items were of investigational orientation to assess the appropriateness of diagnostic tests, and 20 items focused on therapeutic interventions.

Both instruments followed a traditional SCT format in which examinees responded to items using a 5-point Likert-type scale ranging from –2 to + 2. Negative answer choices were associated, for example, with hypothesis elimination or test or treatment contraindication. Neutrality was indicated by a selection of 0. Useful and absolutely necessary information was designated +1 and +2, respectively. Examinees were allotted 2 hours to complete the SCT-PS and 90 minutes to complete the SCT-EM. SCT scores were initially computed using a 5-point aggregate scoring approach.[10] Incomplete SCTs in which examinees failed to respond to one or more items were excluded from the study. Institutional Review Board approval was obtained prior to data analysis to ensure study compliance with common ethical and procedural standards.

Data Set Optimization

Item analyses are commonly performed to enhance instrument reliability and to reduce the number of items required to measure targeted constructs. Item-total correlations computed the extent to which examinees' responses to individual items were representative of, or consistent with, differences in their total test scores.[11] Item discrimination indices were also calculated to isolate and subsequently discard items that demonstrated

TABLE 1
Description of the six scoring methods

Scoring Method	Description
A. 5-Point Aggregate	This is the traditional aggregate scoring method that awards full credit to examinees who select the mode response from a 5-point response scale. Proportional credit is awarded when examinees' responses align with reference panel members who gave an alternate (nonmode) answer.
B. 5-Point Single Answer	The only response for which examinees receive full credit is the mode response. No partial credit is awarded.
C. 5-Point Distance From Mode	This scoring method renders a weighted penalty to examinees who do not give a mode response. Penalty points are a function of the number of steps examinees are away from the mode response. Equation: $C = 1–(\delta/\Delta)$. Where C = scoring method C; δ = distance between an examinee's response and the mode response; Δ = maximum distance from the mode response to the scale extremes (e.g., 2, 3, or 4).
D. 5-Point Aggregate With Distance Penalty	This scoring method blends methods A and C. In addition to receiving full and partial credit from traditional aggregate scoring, penalties are calculated to account for the distance from the mode response. Instating a penalty prevents examinees who were near to the mode response from receiving the same score of 0 as examinees who were distant from the mode response. Equation: $D = (A+C)/2$. Where A, C, and D = designated scoring methods.
E. 3-Point Aggregate	Responses on a 5-point response scale were recoded to generate a 3-point aggregate score. Responses of +1 and +2 were condensed into a single positive score. Likewise, –1 and –2 were combined to represent a single negative score. Partial credit remained feasible to attain.
F. 3-Point Single Answer	Scoring method E without partial credit.

poor discriminatory power between high- and low-scoring examinees.

Scoring Method Analysis

A description of each scoring method investigated in this study is presented in Table 1, and Table 2 showcases a sample of recoded scores by scoring method. Of the six methods (labeled A–F), methods B, C, and D have not been comprehensively explored. A repeated measures analysis of variance (ANOVA), performed on the SCT-PS data set, assessed whether scoring methods could discriminate between training levels within the same population of examinees. On the SCT-EM data set, a multivariate analysis of variance (MANOVA) was conducted to test whether significant differences between training levels could be detected for each scoring method. In addition, SCT-EM scores were correlated with training levels that were more discretely separated to include MS4s, EM-PGY1s, EM-PGY2s, EM-PGY3s, and EM physicians.

Item Difficulty and Item Type Analysis

A secondary aim of this research was to evaluate the combined effects of item difficulty and medical training level on SCT scores. A repeated measures ANOVA and one-way ANOVA compared within and between training level effects of SCT-PS and SCT-EM item scores, grouped by level of item difficulty. Assumptions of the repeated measures model including homogeneity of variance and sphericity were tested to confirm the appropriateness of the statistical approach.

Item difficulty was established according to natural breaks revealed via histogram analysis of student SCT scores. For the SCT-PS, items in which partial or full credit was given to 85.00% or more of examinees were classified as easy items ($n_{easy} = 35$). Items were classified as moderate ($n_{moderate} = 15$) if 60.00% to 84.99% of examinees received partial or full credit. Items in which partial or full credit was awarded to 59.99% of examinees or less were labeled as difficult ($n_{difficult} = 8$). Histogram analysis of the SCT-EM indicated 31 easy items, 16 moderate items, and two difficult items. Items in which partial or full credit was awarded to 80.00% or more of student examinees were easy, between 40.00% and 79.99% were moderate, and 39.99% or less were difficult.

As previously outlined, SCT-EM items were categorized into three types including diagnostic, investigational, and therapeutic items. A repeated measures ANOVA assessed differences in item types within and between training levels.

The level of significance (alpha) was set at 0.05, and eta-squared was utilized to measure the magnitude of the reported

TABLE 2
Scoring samples from 3 of 49 emergency medicine script concordance test items

Items	Response Options −2	−1	0	+1	+2	Student 1 Response	Score
Item 1							
A. 5-Point Aggregate	0.00	0.00	0.71	0.00	1.00	0	0.71
B. 5-Point Single Answer	0.00	0.00	0.00	0.00	1.00	0	0.00
C. 5-Point Distance From Mode	0.00	0.25	0.50	0.75	1.00	0	0.50
D. 5-Point Aggregate With Distance Penalty	0.00	0.13	0.61	0.38	1.00	0	0.61
E. 3-Point Aggregate	0.00		0.71		1.00	0	0.71
F. 3-Point Single Answer	0.00		0.00		1.00	0	0.00
Item 2							
A. 5-Point Aggregate	0.10	1.00	0.10	0.00	0.00	−2	0.10
B. 5-Point Single Answer	0.00	1.00	0.00	0.00	0.00	−2	0.00
C. 5-Point Distance From Mode	0.67	1.00	0.67	0.33	0.00	−2	0.67
D. 5-Point Aggregate With Distance Penalty	0.38	1.00	0.38	0.17	0.00	−2	0.38
E. 3-Point Aggregate	1.00		0.09		0.00	−2	1.00
F. 3-Point Single Answer	1.00		0.00		0.00	−2	1.00
Item 3							
A. 5-Point Aggregate	0.00	0.00	1.00	0.33	0.00	+2	0.00
B. 5-Point Single Answer	0.00	0.00	1.00	0.00	0.00	+2	0.00
C. 5-point Distance From Mode	0.00	0.50	1.00	0.50	0.00	+2	0.00
D. 5-point Aggregate With Distance Penalty	0.00	0.25	1.00	0.42	0.00	+2	0.00
E. 3-Point Aggregate	0.00		1.00		0.33	+2	0.33
F. 3-Point Single Answer	0.00		1.00		0.00	+2	0.00

Composite Score	Item No. 1		2		3		Student 1 Σ Points (Σp)	Score = (Σp/3)(100)
A. 5-Point Aggregate	0.71	+	0.10	+	0.00	=	0.81	27.00%
B. 5-Point Single Answer	0.00	+	0.00	+	0.00	=	0.00	0.00%
C. 5-Point Distance From Mode	0.50	+	0.67	+	0.00	=	1.17	39.00%
D. 5-Point Aggregate With Distance Penalty	0.61	+	0.38	+	0.00	=	0.99	33.00%
E. 3-Point Aggregate	0.71	+	1.00	+	0.33	=	2.04	68.00%
F. 3-Point Single Answer	0.00	+	1.00	+	0.00	=	1.00	33.33%

effects. Eta-squared values of 0.01, 0.06, and 0.14 were used to discern small, medium, and large effects, respectively.[12] Data, as score percentages, are presented as mean (± standard deviation).

RESULTS

Data Set Optimization

Item-total correlations in conjunction with item discrimination indices were used to optimize the reliability of the SCT data sets prior to investigating the study's research questions. Optimization was performed to minimize measurement error and statistical inflation induced by the instruments themselves.

The SCT-PS was optimized by discarding 17 items identified as having negative or modest (i.e., < 0.100) item-total correlations and/or negative discrimination indices. In the same manner, SCT-EM optimization was attained by removing 10 items. Twenty-one diagnostic items, 11 investigational items, and 17 therapeutic items remained on the SCT-EM.

Scoring Method Analysis

Correlations between scores and training level were not conducted on the SCT-PS data set that compared MS2s to MS4s because observations across samples were not independent. As a related aside, the SCT-PS demonstrated moderate predictive validity as correlations between MS2 and MS4 scores were significant ($p < .001$) but modest ($r = 0.381$). A repeated measures analysis reported that all computed scoring methods for the SCT-PS discriminated between training levels. MS4s consistently scored higher than they did as MS2s ($p < .001$, $\eta^2 \geq 0.093$), despite the scoring method employed. Table 3

TABLE 3
SCT-PS descriptive statistics of all scoring methods

	Training Level[a]	Reliability	M Percentage Score (SD)	p Value (MS2s vs. MS4s)	Partial η^2
Scoring Method A	MS2s	0.745	60.2 (10.0)	<.001	0.361
(5-Point Aggregate)	MS4s	0.802	68.8 (10.5)		
Scoring Method B	MS2s	0.809	51.1 (14.3)	<.001	0.102
(5-Point Single Answer)	MS4s	0.778	56.4 (12.3)		
Scoring Method C	MS2s	0.876	78.3 (9.8)	<.001	0.093
(5-Point Distance From Mode)	MS4s	0.745	81.5 (5.2)		
Scoring Method D	MS2s	0.859	70.4 (11.3)	<.001	0.173
(5-Point Aggregate With Distance Penalty)	MS4s	0.798	75.9 (7.7)		
Scoring Method E	MS2s	0.590	82.9 (5.6)	<.001	0.408
(3-Point Aggregate)	MS4s	0.667	88.7 (4.8)		
Scoring Method F	MS2s	0.518	73.8 (6.4)	<.001	0.371
(3-Point Single Answer)	MS4s	0.549	79.7 (5.8)		

Note. SCT-PS = problem-solving script concordance test.
[a]$n = 522$.

summarizes these findings and presents reliability coefficients for each scoring method.

Table 4 presents reliability coefficients, correlation coefficients, mean percentage scores, and the range of scores for all scoring methods computed on the SCT-EM data set. To elicit a balanced design, composite scores derived for each scoring method were weighted so that each of the five training levels equally represented 20% of the studied population. Composite scores were then correlated with training level to test the strength of their associations. All scoring methods demonstrated a significant, positive correlation with level of training ($r = 0.556$–0.784, $p < .001$; Table 4). Correlations among the various scoring methods were moderate to high ($r = 0.675$–0.990, $p < .001$). A one-way MANOVA that included all six scoring methods as dependent variables and training level (MS4s, EM residents, EM physicians) as the independent variable reported significant differences between training levels ($p < .001$, Wilks's $\lambda = 0.864$, $\eta^2 = 0.071$). A power analysis using G*power indicated a 73.9% chance of detecting a medium effect size (as defined by Lomax[12]) at the 0.05 level. A follow-up post hoc test revealed significant pairwise differences between all training level (i.e., MS4s vs. EM Residents, EM Residents vs. EM Physicians, and MS4s vs. EM Physicians) for each scoring method employed ($p \leq .016$). A Box's M test, at $\alpha = 0.001$ per the unbalanced design,[13] was nonsignificant ($p = .004$) indicating that homogeneity of variance–covariance matrices assumption was satisfied.

Item Difficulty and Item Type Analysis

Univariate analyses were conducted on the SCT-PS and SCT-EM data sets to study the effects of item difficulty and medical training level on clinical data interpretation and to explore training level differences by item type. Only data generated with the traditional 5-point aggregate scoring method was used to conduct the item difficulty and item type analyses.

SCT-PS (MS2s vs. MS4s). Scores arranged by level of difficulty were normally distributed with the exception of scores on easy items captured from the second administration of the SCT-PS. Sphericity, assessed because items were grouped into three difficulty levels, was violated warranting the use of the Greenhouse-Geisser correction. The repeated measures analysis reported a significant effect for time ($p < .001$, $\eta^2 = 0.411$), with MS4s outperforming their scores as MS2s, net the effects of item difficulty. Controlling for time, differences in performance between easy, moderate, and difficult items were also found to be statistically significant ($p < .001$, $\eta^2 = 0.509$). A Scheffé procedure revealed statistically significant differences ($p < .001$) for each pairwise comparison (i.e., easy vs. moderate, moderate vs. difficult, and easy vs. difficult) on both administrations of the SCT-PS. Scores on easy items were significantly greater than those on moderate items which were significantly greater than those on difficult items (Table 5). A significant interaction between time and difficulty was also observed ($p < .001$, $\eta^2 = 0.159$). That is, the change in mean performance scores (Δ mean), from MS2s to MS4s, grew in magnitude as item difficulty increased (Table 5). Figure 1 summarizes these findings.

SCT-EM (MS4s vs. EM Residents vs. EM Physicians). Repeated measures between subjects analysis and a least significant difference multiple comparisons procedure[12] reported a statistically significant difference ($p < .001$, $\eta^2 = 0.213$) in overall SCT-EM scores between each training level. Experienced EM physicians significantly outperformed EM residents who significantly outperformed MS4s, net the effects of item difficulty.

Normality and sphericity assumptions were also violated on the SCT-EM data set. As such, a Greenhouse-Geisser correction

TABLE 4
SCT-EM summary of descriptive statistics and correlation coefficients

	SCT-EM				
	Reliability	Correlation With Training Level	Training Level (*n*)	*M* Percentage Score (*SD*)	Range (as Percentage Scores)
Scoring Method A (5-Point Aggregate)	0.556	0.784	EM Physician (12)	82.8 (3.3)	77.7–87.4
			PGY-3 (14)	72.0 (4.4)	66.3–82.5
			PGY-2 (15)	68.8 (5.8)	58.2–74.7
			PGY-1 (11)	63.1 (6.8)	53.5–75.1
			MS4 (988)	60.4 (8.0)	36.2–82.6
Scoring Method B (5-Point Single Answer)	0.464	0.720	EM Physician (12)	68.5 (5.1)	61.2–75.5
			PGY-3 (14)	58.3 (5.8)	49.0–71.4
			PGY-2 (15)	54.8 (6.7)	40.8–65.3
			PGY-1 (11)	48.6 (8.0)	38.8–61.2
			MS4 (988)	47.3 (8.8)	20.4–73.5
Scoring Method C (5-Point Distance From Mode)	0.478	0.721	EM Physician (12)	86.3 (2.8)	81.6–91.2
			PGY-3 (14)	80.6 (3.9)	72.8-88.4
			PGY-2 (15)	80.0 (3.7)	72.8–84.4
			PGY-1 (11)	74.0 (4.9)	69.4–83.7
			MS4 (988)	73.9 (5.1)	57.1–88.4
Scoring Method D (5-Point Aggregate With Distance Penalty)	0.561	0.765	EM Physician (12)	84.6 (2.8)	80.3–88.7
			PGY-3 (14)	76.3 (3.9)	70.0-85.0
			PGY-2 (15)	74.3 (4.6)	65.4–79.4
			PGY-1 (11)	68.5 (6.0)	60.5–79.5
			MS4 (988)	67.3 (6.4)	45.7–84.8
Scoring Method E (3-Point Aggregate)	0.332	0.678	EM Physician (12)	88.7 (4.2)	82.3–96.5
			PGY-3 (14)	84.5 (4.8)	76.8–93.0
			PGY-2 (15)	81.8 (5.3)	72.9–91.2
			PGY-1 (11)	77.7 (3.6)	72.7–84.4
			MS4 (988)	77.0 (5.3)	58.1–90.8
Scoring Method F (3-Point Single Answer)	0.278	0.556	EM Physician (12)	78.6 (6.5)	67.4–89.8
			PGY-3 (14)	75.8 (7.1)	65.3–89.8
			PGY-2 (15)	73.3 (6.8)	61.2–85.7
			PGY-1 (11)	67.5 (4.9)	61.2–75.5
			MS4 (988)	67.8 (6.5)	46.9–85.7

Note. SCT-EM = emergency medicine script concordance test; PGY = postgraduate year.

was used to assess differences in item difficulty scores and the interaction between item difficulty and training level. Overall, scores on easy, moderate, and difficult items differed significantly ($p < .001$, $\eta^2 = 0.064$), irrespective of training level. A Scheffé multiple comparisons procedure revealed that each pairwise comparison of item difficulty scores was significant ($p < .001$), net the effects of training level. Bonferroni (Dunn) procedures were also performed independently for MS4s, EM residents, and experienced EM physicians. MS4s performed significantly higher ($p < .001$) on easy items compared to moderate items and significantly higher on moderate items compared to difficult items. Residents performed in a comparable manner ($p \leq .036$). In the case of experienced EM physicians, no differences in performance between easy, moderate, or difficult items were identified ($p = .801$; Table 6).

A one-way ANOVA and a least significant difference post hoc test were conducted to assess differences between training levels for each difficulty category (Figure 2). Although homogeneity of variance was violated for easy ($p = .007$) and difficult items ($p < .001$), under large sample conditions ANOVA is robust with respect to departures.[12] On easy items, experienced EM physicians generated significantly higher scores than EM residents ($p = .001$) who in turn yielded significantly higher scores than MS4s ($p = .043$). Significant differences ($p < .001$) between each medical training level were also reported for moderate and difficult items (Figure 2).

TABLE 5
SCT-PS percentage scores by training level and item difficulty

SCT-PS					
	1st Administration (as MS2s)		2nd Administration (as MS4s)		
Items	M	SD	M	SD	Δ M
Easy[a]	65.26	±10.07	71.64	±10.24	6.38
Moderate[b]	58.51	±13.83	66.41	±14.28	7.90
Difficult[c]	41.97	±20.04	60.80	±18.72	18.83

Note. SCT-PS = problem-solving script concordance test.
[a]$n = 35$. [b]$n = 15$. [c]$n = 8$.

TABLE 6
SCT-EM percentage scores by training level and item difficulty

SCT-EM						
	MS4s[a]		EM Residents[b]		EM Physicians[c]	
Items	M	SD	M	SD	M	SD
Easy[d]	70.43	±8.40	73.15	±6.77	81.99	±4.64
Moderate[e]	46.77	±12.98	61.78	±12.20	84.61	±6.65
Difficult[f]	12.90	±22.64	46.16	±36.33	81.79	±22.71

Note. SCT-EM = emergency medicine script concordance test.
[a]$n = 988$. [b]$n = 40$. [c]$n = 12$. [d]$n = 31$. [e]$n = 16$. [f]$n = 2$.

Lastly, a significant interaction was observed between item difficulty and training level ($p < .001$, $\eta^2 = 0.070$). The combination of the main effects resulted in experienced EM physicians scoring higher than residents and MS4s at any level of difficulty (Table 7). The magnitude of the difference in mean performance scores increased as the distance between training level and item difficulty increased (Table 7).

Item type analysis. A repeated measures analysis on SCT-EM items categorized by type (i.e., diagnostic, investigational, or therapeutic) reported significant differences in scores between item types within each training level ($p < .001$, $\eta^2 = 0.014$) and between training levels ($p < .001$, $\eta^2 = 0.111$). No interaction effect was observed ($p = .066$). Irrespective of

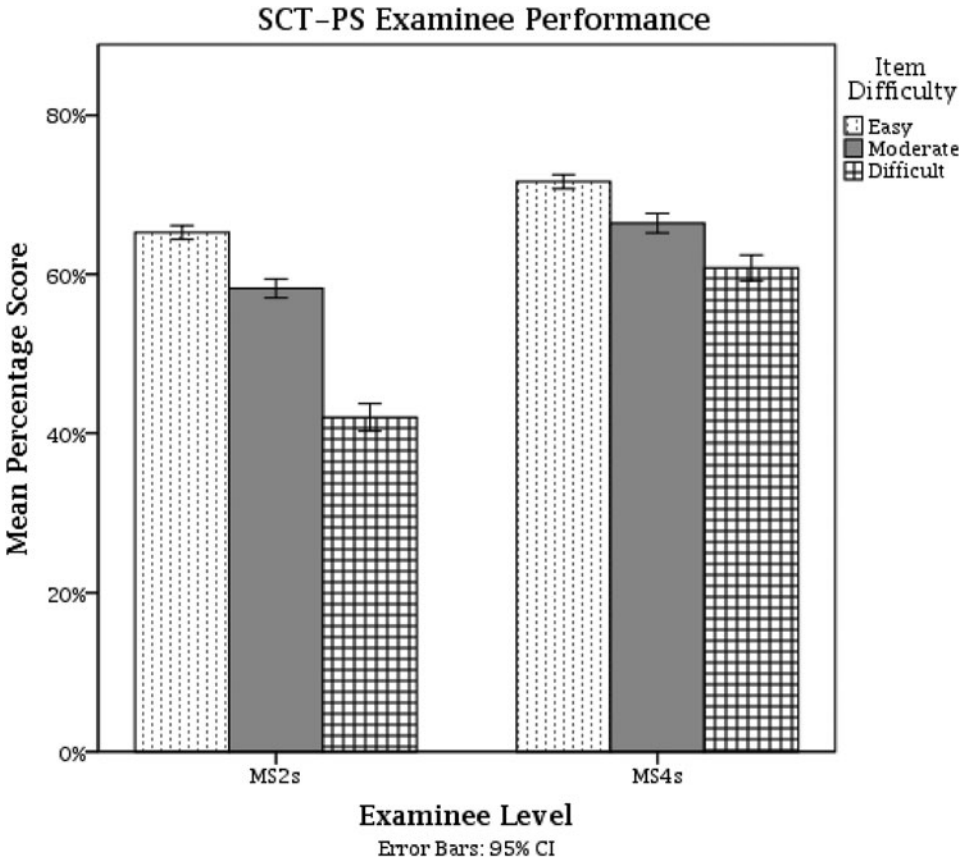

FIG. 1. Bar graph comparing MS2 and MS4 mean percentage scores on the problem-solving script concordance test (SCT-PS). *Note*. Overall and within each difficulty category, MS2s performed significantly lower than MS4s. For both MS2s and MS4s, scores on easy items were significantly higher than scores on moderate items which were significantly higher than scores on difficult items. CI = confidence interval.

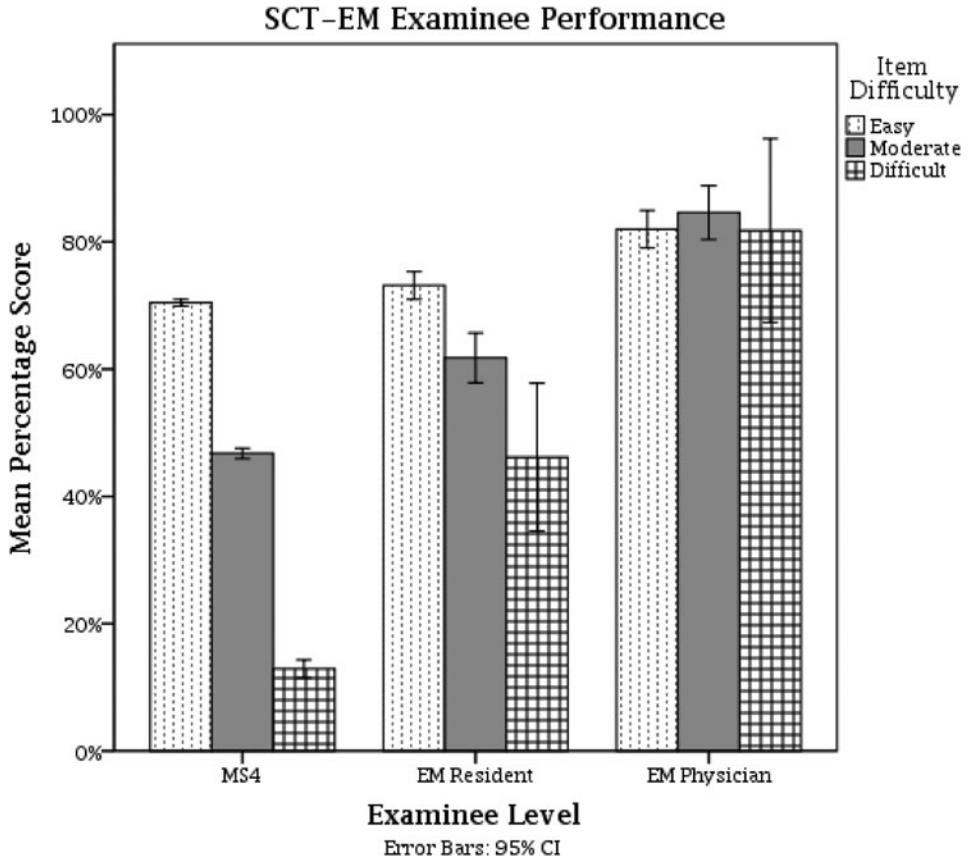

FIG. 2. Bar graph comparing MS4, emergency medicine (EM) resident, and EM physician mean percentage scores on the emergency medicine script concordance test (SCT-EM). *Note*. Overall and within each difficulty category, experienced EM physicians scored significantly higher than residents who scored significantly higher than MS4s. Among MS4s and EM residents scores on easy items were significantly higher than scores on moderate items which were significantly higher than scores on difficult items. No differences in scores categorized by difficulty were observed for experienced EM physicians. CI = confidence interval.

training level, performance on diagnostically oriented items was significantly higher ($p \leq .002$) than investigational or therapeutic items. Overall, no performance differences ($p = .094$) were observed between investigational and therapeutic items. A Scheffé multiple comparisons procedure reported that diagnostically oriented items discriminate between EM physicians, EM residents, and MS4s ($p \leq .003$). On investigational items, MS4s scored as well as EM residents ($p = .090$), whereas EM physicians scored higher ($p < .001$) than MS4s and EM residents. Therapeutic items also exhibited discriminant properties as EM physicians scored significantly higher than EM residents who outperformed MS4s ($p < .001$). For a visual summary of the item type analysis, refer to Figure 3.

TABLE 7
SCT-EM change in percentage scores between training levels organized by item difficulty

	SCT-EM		
Items	MS4s vs. EM Residents |Δ mean|	EM Residents vs. EM Physicians |Δ mean|	MS4s vs. EM Physicians |Δ mean|
Easy[a]	2.72	8.84	11.56
Moderate[b]	15.01	22.83	37.84
Difficult[c]	33.26	35.63	68.89

Note. SCT-EM = emergency medicine script concordance test.
[a]$n = 31$. [b]$n = 16$. [c]$n = 2$.

DISCUSSION

The aim of this project was to compare nontraditional to conventional SCT scoring methods and to evaluate the temperament of SCTs to retain their discriminatory power at the item difficulty and item type levels.

Scoring Method Analysis

A study by Seibert et al.,[14] whose focus was in urology, reported that the SCT was effective at differentiating between examinees at various levels. Novices, residents, and experts demonstrated significantly different levels of reasoning, and SCT scores were positively correlated to training level. Because numerous studies have tested this concept (of concurrent

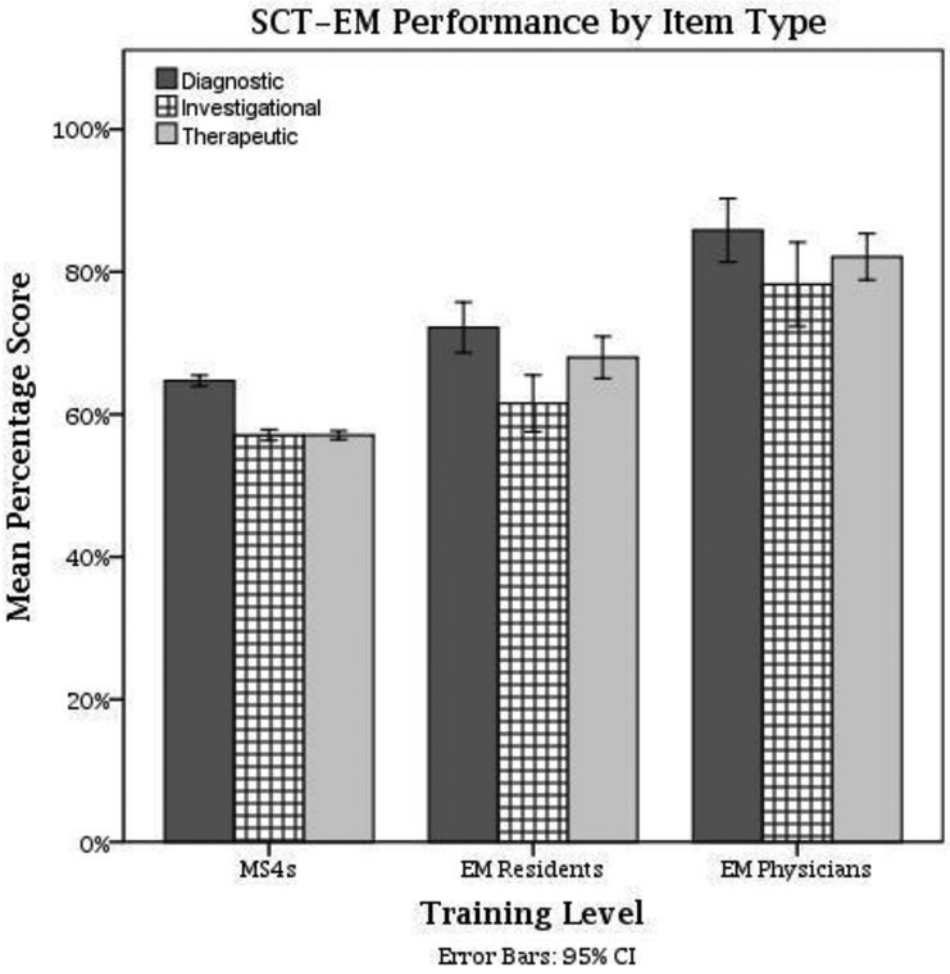

FIG. 3. Examinee scores on the emergency medicine script concordance test (SCT-EM) grouped by item type. *Note*. Diagnostic and therapeutic items were successful at discriminating between training levels ($p \leq .003$), whereas intermediate effects were observed on investigational items between MS4s and residents ($p = .090$). CI = confidence interval.

validity) at the composite score level and reported similar findings,[15–19] we were not surprised to observe that all six scoring methods differentiated between training levels and that scores correlated with level of development.

The reliability of the 3-point scoring methods was consistently lower than that of all 5-point scoring methods. This finding contradicted reports by Bland et al., whose study contained a comparatively smaller sample of 85 examinees. Of the more reliable 5-point scoring methods, methods A and D regularly reported moderate to large measures of effect size ($\eta^2 \geq 0.104$) and demonstrated the highest correlation coefficients. This suggested the efficacy of methods A and D to discriminate between training levels was marginally superior to other methods.

One disadvantage of using either 5- or 7-point Likert scales in traditional aggregate scoring is that test administrators cannot readily distinguish examinee responses that were near the mode response versus those that were distant from it.[5] For instance, if the mode response of the reference panel was +2, examinees who answer –1 receive the same score of 0 as those who answered +1 (presuming no other panel member answered –1 or +1). It is therefore possible for examinees who agree with panel members on the response direction but not the impact to receive the same score as someone who fails to identify both the direction and the impact.[5] This contingency was the impetus for testing the efficacy of scoring method D (5-point aggregate with distance penalty). The properties of scoring method D were similar to traditional aggregate scoring (method A) with the benefit of simultaneously measuring both response direction and impact.

Employing a 3-point scaling system would all together eliminate "degree of correctness" concerns. However, our findings demonstrated that 3-point scoring methods were less reliable, and with 3-point scoring approaches the value of differing expert opinions is minimized.[4] Qualitative data of student perceptions imply that 5- or 7-point scaling systems should be avoided, as students reported at times arbitrarily choosing between +1 and

+2 and –1 and –2.[5] In addition to concerns regarding degree of correctness, Bland contended that "if a single best answer to an SCT does not exist, the SCT will be of limited use for in-course assessment" (p. 395).[5] The rationale is that novices are expected to perform like experts to attain the best possible score. Customarily, course assessment instruments are designed to assess specific course objectives or behaviors. Without a single best answer, it becomes difficult to define attainable objectives.[5] The complexities of aggregate scoring are enough for some practitioners to forego the use of this method entirely. Although we do acknowledge the aforementioned legitimate concerns, based on the findings of this research, we recommend using either a 5-point aggregate (method A) or 5-point aggregate with distance penalty (method D) approach when scoring SCTs because they exhibited stronger reliability and validity coefficients than the other tested methods; keeping in mind that scoring method D accounts for "degree of correctness" unlike scoring method A. Furthermore, methods A and D align better than single-correct-answer scoring schemes (e.g., methods B and F) with the overarching philosophy of the script concordance approach (i.e., to probe reasoning skills under the conditions of uncertainty that typically characterize daily practice).

Item Difficulty and Item Type Analysis

Our findings demonstrated that MS4s, who have greater clinical knowledge and exposure to patients through clerkship experiences, performed significantly higher at all difficulty levels than they did as MS2s. Likewise, on the SCT-EM, experienced physicians outperformed residents who outperformed MS4s in all difficulty categories. Because residents and practicing physicians have increased exposure to rare and atypical presentations, they are theoretically able to build, refine, and link illness scripts in a more organized, purposeful manner than undergraduate medical students.[20]

This retrospective study was performed at a large, multicenter institution. Each of the nine IUSM centers autonomously delivers instruction to medical students during Years 1 and 2 of undergraduate training. As such, we believe that aspects of our study are representative of a large-scale multi-institutional study. In support of this claim, our results echoed a cross-sectional multi-institutional study that investigated differences in clinical reasoning skills of undergraduate medical students. In this study, Williams et al.[21] reported clinical data interpretation gains at each level of undergraduate medical training, though gains in the 3rd year were not as substantial. They also reported that medical school elements (e.g., curriculum, instructional delivery systems, faculty, etc.) account for only a small percentage of variation in data interpretation scores. The study by Williams et al., however, did not explore how the nuances of item difficulty or item type might reshape the interpretation of clinical reasoning performance.

Items categorized by type were also found to distinguish between training levels, with the exception of investigational items not being able to differentiate between MS4s and EM residents. In our view, the outcomes of the item type analysis suggested that (a) residents within the EM program at IUSM do not perform as well on investigational items as might be expected, (b) MS4s are not as well trained on investigational and therapeutic items as they are on diagnostic items, and (c) categorizing items into subconstructs may prove useful for evaluating specific cognitive skill sets and holds promise as an additional marker for program evaluation.

Limitations

Although this large-scale study included data from two SCTs for the majority of analyses, it was not without limitations. Correlation coefficients could not be cross-validated with the SCT-PS data set that consisted of only two training levels. The number of difficult items identified on each exam was restrictive. Also, performance of experienced physicians on the SCT-EM was not ideal, as they performed equally well on easy, moderate, and difficult items. This may suggest that either greater disparity between difficulty categories could have been attained or a natural clinical reasoning plateau was reached by experienced physicians. This finding may have been a result of using only student SCT scores to identify natural breaks between levels of item difficulty. Finally, the presented outcomes may not translate to all SCT instruments.

It is thought that this study was largely resistant to the effects of case specificity due to the presence of multiple cases (i.e., 16 cases for the SCT-PS and 12 cases for the SCT-EM) within each instrument. Case specificity occurs when problem-solving ability is dependent on the attributes of a specific case.[22] According to Norman et al.,[23] the overall effects of case variance are smaller than the effects of item variance. Tests comprising 15 to 20 cases, with two to five nested questions each, are thought to represent the best combination for obtaining sufficiently high reliability estimates.[24] The SCT-PS instrument aligned with these recommendations. However, the number of SCT-EM cases (i.e., 12) fell just short of the 15 to 20 cases per test ratio; perhaps an explanation for the SCT-EM's lower reliability.

CONCLUSIONS

Scoring approaches on a 5-point scale exhibited greater reliability and stronger correlation coefficients than 3-point scoring methods. Upon item level analysis, experienced clinicians outperformed residents who outperformed medical students on easy, moderate, and difficult clinical data interpretation problems, as measured by an SCT-EM. Likewise, MS4s outperformed their own MS2 scores on an SCT-PS at every level of item difficulty. Finally, items meaningfully categorized by type discriminated between training levels and provided more detailed information about the data interpretation abilities of medical students and residents.

REFERENCES

1. Charlin B, van der Vleuten C. Standardized assessment of reasoning in contexts of uncertainty. *Evaluation of the Health Professions* 2004;27:304–19.

2. Brailovsky C, Charlin B, Beausoleil S, Cote S, Van der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. *Medical Education* 2001;35:430–6.
3. Charlin B, Desaulniers M, Gagnon R, Blouin D, van der Vleuten C. Comparison of an aggregate scoring method with a consensus scoring method in a measure of clinical reasoning capacity. *Teaching and Learning in Medicine* 2002;14:150–6.
4. Charlin B, Gagnon R, Pelletier J, Coletti M, Abi-Rizk G, Nasr C, et al. Assessment of clinical reasoning in the context of uncertainty: The effect of variability within the reference panel. *Medical Education* 2006;40: 848–54.
5. Bland AC, Kreiter CD, Gordon JA. The psychometric properties of five scoring methods applied to the script concordance test. *Academic Medicine* 2005;80:395.
6. Groves M, O'Rourke P, Alexander H. Clinical reasoning: The relative contribution of identification, interpretation and hypothesis errors to misdiagnosis. *Medical Teacher* 2003;25:621–5.
7. Chimowitz MI, Logigian EL, Caplan LR. The accuracy of bedside neurological diagnoses. *Annals of Neurology* 1990;28:78–85.
8. 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.
9. Humbert A. Assessing the clinical reasoning skills of emergency medicine clerkship students using a script concordance test. *Acadademic Emergency Medicine* 2008;15:S230–1.
10. Fournier J, Demeester A, Charlin B. Script concordance tests: Guidelines for construction. *BMC Medical Informatics Decision Making* 2008;8:18.
11. Furr RM, Bacharach VR. *Psychometrics: An introduction*. Thousand Oaks, CA: Sage, 2008.
12. Lomax RG. *Statistical concepts: A second course*. Mahwah, NJ: Erlbaum, 2007.
13. Tabachnick BG, Fidell LS, Osterlind SJ. *Using multivariate statistics*. Needham Heights, MA: Allyn and Bacon, 2001.
14. Sibert L, Charlin B, Corcos J, Gagnon R, Lechevallier J, Grise P. Assessment of clinical reasoning competence in urology with the script concordance test: An exploratory study across two sites from different countries. *European Urology* 2002;41:227–33.
15. Khonputsa P, Besinque K, Fisher D, Gong WC. Use of script concordance test to assess pharmaceutical diabetic care: A pilot study in Thailand. *Medical Teacher* 2006;28:570–3.
16. Lubarsky S, Chalk C, Kazitani D, Gagnon R, Charlin B. The Script Concordance Test: A new tool assessing clinical judgement in neurology. *The Canadian Journal of Neurological Sciences*. 2009;36:326–31.
17. Carrière B, Gagnon R, Charlin B, Downing S, Bordage G. Assessing clinical reasoning in pediatric emergency medicine: Validity evidence for a Script Concordance Test. *Annals of Emergency Medicine* 2008;53:647–52.
18. 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. *Radiation Oncology* 2009;4:7.
19. Humbert AJ, Besinger B, Miech EJ. Assessing clinical reasoning skills in scenarios of uncertainty: Convergent validity for a script concordance test in an emergency medicine clerkship and residency. *Academic Emergency Medicine* 2011;18:627–34.
20. Charlin B, Tardif J, Boshuizen H. Scripts and medical diagnostic knowledge: Theory and applications for clinical reasoning instruction and research. *Academic Medicine* 2000;75:182.
21. Williams RG, Klamen DL, White CB, Petrusa E, Fincher RM, Whitfield CF, et al. Tracking development of clinical reasoning ability across five medical schools using a progress test. *Academic Medicine* 2011;86:1148.
22. Kreiter CD, Bergus GR. Case specificity: Empirical phenomenon or measurement artifact? *Teaching and Learning in Medicine: An International Journal* 2007;4:378–81.
23. Norman G, Bordage G, Page G, Keane D. How specific is case specificity? *Medical Education* 2006;40:618–23.
24. Gagnon R, Charlin B, Lambert C, Carriere B, Van der Vleuten C. Script concordance testing: More cases or more questions? *Advances in Health Sciences Education* 2009;14:367–75.