Med.Sci.Educ. (2015) 25:439–446
DOI 10.1007/s40670-015-0178-7

ORIGINAL RESEARCH

Testing for Multilevel Dimensionality: A Higher-Order Factor Analysis of a Script Concordance Test

Adam B. Wilson [1] · **Gary R. Pike** [2] · **Aloysius J. Humbert** [3]

Published online: 29 September 2015

Abstract

Background To date, few script concordance test (SCT) research studies have empirically investigated whether SCTs measure a single test dimension of clinical reasoning. Prior analyses have been inconclusive and unsuccessful at identifying exam dimensions through simple first-order factor analyses. Therefore, the principle objective of this work was to explore the higher-order factor structure of a problem-solving SCT to determine whether the dimensionality of the test conformed to a multilevel construct arrangement.

Methods This retrospective data analysis utilized scores from medical students (n=522) who took a SCT in their fourth year of undergraduate training. Higher-order factor analyses were conducted on six different SCT scoring methods. In addition, Schmid-Leiman solutions were conducted to evaluate the proportion of variance unique to a given level of the model.

Results Five of the six scoring methods yielded a factor structure with three first-order factors and one second-order factor. The total variance in the first-order factors explained by the second-order factor was notable at greater than 40 %. However, Schmid-Leiman solutions unveiled the deceptiveness of the higher-order models in that, across the various scoring methods, very few items contributed unique variance to the second- or first-order factors. Ideally, all items would have contributed unique variance to one or more levels of the model.

Conclusions In this study, SCT scores did not conform to a clear interpretable higher-order factor structure suggesting that SCTs may not measure the meaningful clinical reasoning constructs they are thought to measure. An explanation for these findings is provided, and recommendations for revising the SCT format and scoring procedures are proposed.

Keywords Clinical reasoning · Script concordance test · Psychometrics

Introduction

The French-Canadian scholar Bernard Charlin developed the first rendering of the script concordance test (SCT) in 1998 called the Diagnostic Script Questionnaire [1]. Charlin's SCT design was based on Schmidt's script theory of medical decision making [2]. SCT items begin with a short clinical vignette followed by a proposed diagnosis (or other possible decision). Examinees are then given additional information and asked to rate the impact of the said information on the originally proposed diagnosis. Test takers rate the impact using a five-point Likert scale that ranges from "−2" to "+2", with "−2" indicating that the diagnosis is highly unlikely and "+2" representing a very likely diagnosis. A response of "0" represents no change in the likelihood of the said diagnosis (see Sample 1). Examinees' responses are compared against responses of a reference panel of experienced physicians, and scoring is a function of the degree of concordance with the panel. Test takers receive full credit if they select the most frequent answer chosen by the panel, partial credit if only some panel members made the same selection, and no credit if no panel members selected the response [3].

✉ Adam B. Wilson
Adam_Wilson@rush.edu

[1] Department of Anatomy and Cell Biology, Rush University, 606 S. Paulina St., Suite 505A, Chicago, IL 60612, USA

[2] Indiana University Purdue University Indianapolis, Indianapolis, IN, USA

[3] Department of Emergency Medicine, Indiana University School of Medicine, Indianapolis, IN, USA

To the liking of many, SCT scores simultaneously measured one's level of reasoning competence and clinical experience [1, 4, 5]. Despite an abundance of SCT validity studies that have employed classical test theory, only one has directly explored the factor structure of SCTs with the aim of identifying the latent constructs measured by this instrument [6]. The continued need to conduct factor analyses on SCTs is great and has been recognized by the early forebearers of SCT research [7].

Factor analysis is a useful procedure for evaluating the dimensionality of a set of multiple indicators (e.g., test items) and is thereby a practical objective approach for identifying underlying latent constructs inherent to an instrument's disposition. Construct identification, in turn, influences and gives meaning to instruments' scores. In the absence of clear, stable constructs, the meaning of scores may become misconstrued and highly subjective, rendering the utility of an exam useless [8].

In a previous study, Wilson et al. [6] reported that the three SCT datasets under investigation did not conform to a unidimensional factor structure. Therefore, the ostensible claim that SCTs measure one construct of data interpretation is unlikely and deserves further investigation. Furthermore, it was concluded that the use of a particular scoring method had no apparent effect on the number of constructs the SCTs measured.

As a follow-up to prior work, the present study utilized higher-order factor modeling to assess whether SCTs follow a more complex factor structure than could be identified through simple first-order analyses. Specifically, we asked the following research question, "How well does a SCT conform to a multilevel factor structure, and what are the implications for practice?"

Methods

A comprehensive description of the SCT instrument and scoring methods utilized in this study can be referenced in a previously published article [6]. In summary, data from a problem-solving SCT, taken by fourth-year undergraduate medical students (n=522) and composed of 58 items nested within 16 cases after optimization, were used to calculate a higher-order factor model. This SCT has previously undergone common validity testing [9]. The reliability of the optimized SCT was reported to be 0.802, under conditions of traditional five-point aggregate scoring [6]. Higher-order factor analyses were computed for six different scoring methods that included scoring methods "A" (five-point aggregate scoring), "B" (five-point single-answer scoring), "C" (five-point distance from the mode scoring), "D" (five-point aggregate with distance penalty scoring), "E" (three-point aggregate scoring), and "F" (three-point single-answer scoring) [6]. This study was granted exempt status by the Institutional Review Board of Indiana University-Purdue University-Indianapolis (IUPUI).

Higher-Order Factor Analysis

A second-order factor analysis was computed on the first-order factors extracted from the SCT. All second-order factors analyzed three first-order factors for six different scoring conditions. Three first-order factors were utilized because that was the most common number of factors extracted from this instrument in a prior study [6]. Computations were performed using a principle component solution, and second-order factors were extracted according to Kaiser's criterion (i.e., eigenvalues≥1.00). Promax rotation was used to interpret factor structure. The analysis was conducted in SPSS (version 20) as described by Thompson [10]. To further facilitate factor interpretation, Schmid-Leiman solutions were performed to assess the direct relationships between SCT items and higher-order factors [11, 12].

Results

The higher-order factor analysis specified one second-order factor for scoring methods "A" through "E." The one second-order factor explained 41.51 % or more of the total variance between first-order factors (Table 1). Fig. 1 provides an example of the hierarchical model calculated for scoring method "A" (five-point aggregate scoring). For scoring method "F" (three-point single-answer scoring), two second-order factors were extracted and explained 74.47 % of the total variance between first-order factors (Table 1).

A Schmid-Leiman solution was also computed to more fully understand the relationships between the exam items and the higher-order factors (e.g., Fig. 2). A Schmid-Leiman solution probes how well-observed variables measure second-order factors [10, 12]. Specifically, the variance explained between each item and each factor, regardless of factor level, was explored. In interpreting a Schmid-Leiman solution, items that have a greater second-order loading than first-order loadings are considered to be better measures of second-order factors [12]. Conversely, Schmid-Leiman solutions can also delineate which items are purer measures of first-order factors. The Schmid-Leiman solutions (for all six scoring methods) reported that, of the few SCT items with salient loadings, items often loaded more frequently on the second-order factor than on first-order factors (Table 2). For example, Fig. 2 shows that a Schmid-Leiman solution conducted under conditions of a traditional five-point aggregate scoring system (scoring method A) found seven SCT items to be reflective of the second-order factor, while no items were

Table 1 Results of the second-order factor analysis conducted on the SCT

	Second-order factor I	Second-order factor II
Scoring method A: five-point aggregate		
First-order factor I	*0.780*	
First-order factor II	*0.741*	
First-order factor III	*0.569*	
Total variance explained	49.39 %	
Scoring method B: five-point single answer		
First-order factor I	*0.661*	
First-order factor II	*0.754*	
First-order factor III	*0.655*	
Total variance explained	47.82 %	
Scoring method C: five-point distance from mode		
First-order factor I	*0.528*	
First-order factor II	*0.761*	
First-order factor III	*0.709*	
Total variance explained	45.39 %	
Scoring method D: five-point aggregate with distance penalty		
First-order factor I	*0.561*	
First-order factor II	*0.777*	
First-order factor III	*0.719*	
Total variance explained	47.87 %	
Scoring method E: three-point aggregate		
First-order factor I	*0.632*	
First-order factor II	*0.509*	
First-order factor III	*0.766*	
Total variance explained	41.51 %	
Scoring method F: three-point single answer		
First-order factor I	*0.828*	
First-order factor II		*0.934*
First-order factor III	*0.709*	*0.407*
Total variance explained by second-order factors I and II	74.47 %	

Loadings <0.2 are suppressed. Salient loadings ≥0.4 are in italics

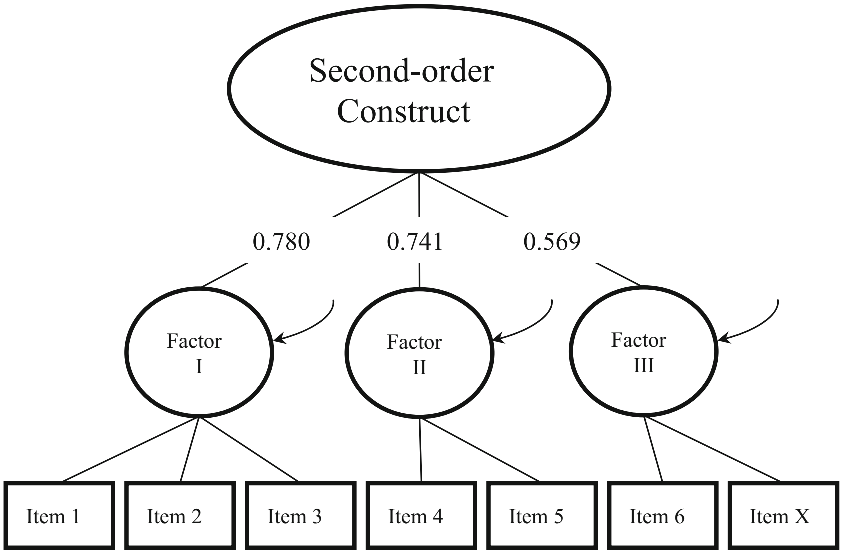

Fig. 1 Sample second-order factor model of the SCT, scoring method "A." *Floating arrows* represent unique extraneous contributions to each factor. First-order factor loadings are not displayed

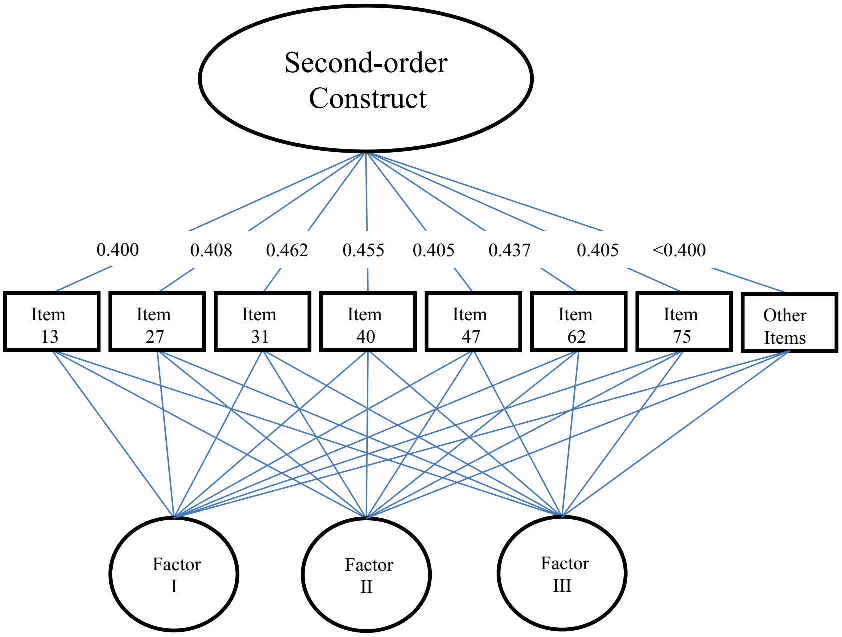

Fig. 2 Schmid-Leiman solution for the SCT, scoring method "A." Contributions of seven items to the second-order factor were ≥0.400. Contributions of all items to the first-order factors were non-significant (<0.400)

found to measure the three first-order factors. Analysis of the other scoring methods (B–F) gave variable results.

Discussion

The inconclusiveness of previous SCT construct validity research warranted the execution of a higher-order factor analysis in an attempt to identity a more complex, multifarious factor structure. The higher-order factor analysis specified one second-order factor for the investigated SCT, with scoring method "F" as the exception. The proportion of variance explained by the second-order factors (at >40 %) was reasonable for a second-order model [13]. However, interpretation of the Schmid-Leiman solution implied that relatively few exam items made a strong unique contribution to the second- or first-order factors. After removing the variance of the second-order factor that was also present in the first-order factors, little detectable variance remained in the first-order factors. Under ideal Schmid-Leiman circumstances, all SCT items would have loaded somewhere, either on the first-order factors, second-order factor, or both. These outcomes demonstrated that SCT items with significant loadings (though few per scoring method) were hierarchical in nature. Moreover, the few items with salient loadings drove the factor structure of the instrument.

An informal qualitative investigation of exam items failed to uncover commonalities among well-performing items. Additionally, when comparing items with significant loadings against items with non-significant loadings, no explicit differences in item characteristics or quality were identified. Weak Schmid-Leiman solutions and a scarcity of shared item characteristics imply that the studied SCT did not follow a meaningful second-order factor structure.

A more thorough investigation of SCT reliability was revealing and, in part, helped to explain the outcomes of this research. Reliability, as calculated using Cronbach's coefficient alpha, is a function of the number of items on an instrument and the degree of covariance among the items. Instruments with items that have small inter-item correlations can demonstrate reasonable reliability in the presence of large numbers of items. For instance, the investigated SCT had small inter-item correlations (mean inter-item correlation=0.067, min=−0.113, max=0.283), yet showed reasonably high reliability (0.802; scoring method A) due to the presence of 58 items. Decreasing the number of exam items from 58 to 10, for example, would drastically reduce the reliability of the instrument because of the lack of strong inter-item correlations. To complicate matters, items with weak inter-item correlations tend not to factor analyze well, meaning they rarely yield a clear comprehensible factor structure. The culmination of this information suggests that large numbers of SCT items are required to produce high-reliability coefficients and also explains why a clear factor structure was not observed in this study. A secondary question that emerged from the findings of this research was "Why do SCT items have small inter-item correlations, and could something inherent in the format or exam structure of SCTs explain this finding?"

A partial explanation for weak inter-item correlations may stem from scoring incongruities. The following quote from an article by Lineberry et al. [14] nicely illustrates this point.

Table 2 Schmid-Leiman solutions for all scoring methods

	Second-order factor						First-order factors							
	Scoring method						I						II	III
Item:case	A	B	C	D	E	F	A	B	C	D	E	F	A–F	A–F
1:1														
2:1														
4:1					0.486									
6:1														
10:2														
13:3	0.400		0.498							0.431				
14:3					0.495									
15:3														
16:4														
17:4				0.445										
19:4														
21:5														
22:5														
23:5		0.433			0.574	0.466								
24:5														
25:5														
26:6					0.458									
27:6	0.408		0.403							0.457				
28:6														
29:6				0.417										
30:6			0.479							0.412				
31:7	0.462	0.418	0.443	0.417	0.557	0.436					0.409	0.556		
32:7														
33:7					0.423									
34:7					0.403							0.460		
35:8														
36:8														
37:8														
38:8					0.413									
39:9										0.473				
40:9	0.455													
41:9			0.432											
42:9														
43:10														
44:10				0.462										
45:10					0.516									
46:10														
47:10	0.405	0.410			0.522							0.410		
50:11														
52:11														
53:12														
54:12										0.408				
55:12					0.429									
56:12														
57:12									0.431					

Table 2 (continued)

	Second-order factor						First-order factors							
	Scoring method						I						II	III
Item:case	A	B	C	D	E	F	A	B	C	D	E	F	A–F	A–F
58:13		0.419						0.500						
59:13		0.421						0.480						
61:13		0.516			0.471			0.482						
62:13	0.437	0.528			0.464			0.534						
65:14		0.497						0.507						
66:14					0.410			0.497						
67:14		0.474						0.509						
68:15								0.527						
70:15		0.401						0.522						
71:15		0.461						0.504						
73:16		0.411						0.671						
74:16		0.512		0.408				0.497						
75:16	0.405	0.494						0.486						
Item loading frequency	A 7	B 14	C 5	D 5	E 14	F 2	A 0	B 13	C 1	D 5	E 1	F 3	A–F 0	A–F 0

The original SCT had a total of 75 items. After optimization, 58 items remained. The 17 discarded items were not analyzed. Loadings <0.400 are suppressed.

A five-point aggregate scoring, *B* five-point single-answer scoring, *C* five-point distance from the mode scoring, *D* five-point aggregate with distance penalty scoring, *E* three-point aggregate scoring, *F* three-point single-answer scoring

> "The incongruity of panelists' diametric opposition on an item is compounded when SCTs award no credit to examinees who respond with a mark of '0' ('neither refutes nor supports') on such items. An examinee with perfect knowledge of experts' contradictory opinions about that particular item could reasonably surmise that splitting the difference is the only way to convey his or her acknowledgment of the divided expert opinion."

A separate, yet related issue that is also likely to influence inter-item correlations is that SCTs are "susceptible to score inflation attributable to coaching" [14]. It has been reported that examinees who guess the midpoint (a response of "0") are subject to receiving more credit than examinees who guess the extreme anchors (such as "−2" or "+2") [14]. Thus, examinees or groups of examinees who favor extreme responses are liable to score lower on SCT exams. The aforementioned study also demonstrated that if an examinee deliberately marks a neutral response of "0" for every item, he/she would score higher than the test average [14]. Unequal distributions in the range of credit awarded across items stemming from score inflation and the idiosyncrasies of aggregate scoring offer another explanation for the findings presented in this study.

A commentary by Clarence Kreiter has caused the authors to critically rethink the exam structure of SCTs. In the commentary, Kreiter [15] argues that examinees must first assess the probability (P1) that the hypothesis (or investigative action, or therapeutic action, etc.) is reasonable in the context of the problem, and then they must calculate the likelihood and usefulness (P2) of the hypothesis given both the scenario and the new information. When examinees respond to an item, they are therefore subjectively rating the magnitude of the difference (P2-P1) on a five-point Likert scale. The interplay, or lack thereof, between P1 and P2 can cause response confusion. If a diagnostic test (P1) is undeniably useful based solely on the information in the case scenario, and the new information adds little to no insight (P2), an examinee is left to determine whether the final response reflects the usefulness of P1 or the difference between P2 and P1, thereby creating response confusion. Here is an example from a SCT constructed for emergency medicine:

Case A 52-year-old Hispanic female with a past medical history of hypercholesterolemia, COPD, and hypertension presents to the emergency department with a chief complaint of chest pain for 3 h. The pain is sharp, substernal, radiates to both arms, and is associated with diaphoresis and nausea.

If you were considering treating with…	and you find the patient has a…	…this treatment becomes…
Aspirin	history of GERD	−2 −1 0 +1 +2

−2, contraindicated totally or almost totally; *−1*, not useful, possibly detrimental; *0*, neither more nor less useful; *+1*, useful; *+2*, necessary or absolutely necessary

In this example, aspirin would appear to be a useful therapeutic approach for treating a possible myocardial infarction, as surmised from the case. Secondly, a history of gastroesophageal reflux disease (GERD) is likely to have little to no impact on one's decision to treat with aspirin. As such, should an examinee's response reflect the usefulness of administering aspirin in the context of the scenario or should it reflect the effect the new information had on one's decision to give an already useful medication? Such perplexities raise the question, "What is being measured?" Is the therapeutic action (i.e., giving aspirin) in the context of the scenario alone being measured or is the therapeutic action in the context of both the scenario and the new information being measured? Interestingly, for this particular item, the mode response of the reference panel was "+2", likely a reflection of aspirin's usefulness in a possible heart attack as opposed to the impact the history of heartburn had on using the medication. If SCTs are, in fact, intended to measure P2-P1, then a response of "0" would have been a more appropriate response. However, in this instance, examinees who selected "0" were only awarded partial credit per the answer key derived from the reference panel. Perhaps this type of discrepancy is another contributing factor that explains why irregularities in SCT factor structure were observed in this and previous research. This example demonstrates that response confusion can also surface within the reference panel and that incongruities are capable of making their way to the exam answer key. Irrespective of the frequency with which response confusion occurs, the fact remains that a single SCT item has two components (P1 and P2); perhaps each ought to be measured independently.

Restructuring the format of SCT items may be valuable to consider in future iterations of research as it may strengthen inter-item correlations and enhance the overall factor structure of the instrument. For instance, requiring an extra yet separate response that independently measures P1 may bring clarity to the response process. Also, having examinees entertain multiple hypotheses simultaneously may add to the realism of the exam. Lastly, we postulate that reformatted SCTs may benefit from a scoring method that awards credit based on response patterns across a collection of items versus awarding credit per item. Alternate scoring systems and variations of exam formatting may also benefit from the application of item response theory that adheres to a different set of assumptions than classical test theory. We welcome and invite item response theory experts to join in this discussion and concerted research effort.

Limitations The generalizability of this study is limited because this work analyzed only one locally administered SCT. Secondly, the fixed number of first-order factors (i.e., three) used to compute the higher-order models was based on the minimum number and most common number of extracted factors calculated in a prior study [6]. Though unlikely, due to weak inter-item correlations, it is possible to attain a different higher-order factor structure upon factor analyzing a greater number of first-order factors, such as the four or five factors that were also reported in the previous study. Lastly, a more sophisticated thematic analysis of SCT items may be more revealing in how item commonalities influence factor loadings.

Conclusion

A higher-order factor analysis on multiple scoring methods revealed that the investigated script concordance test did not conform to a meaningful higher-order factor structure. No test dimensions (i.e., constructs) were identified because the multilevel model was largely uninterpretable. Based on these findings and the outcomes of other research, the number and types of dimensions measured by SCTs remain elusive. The inability to empirically decipher the dimensionality of SCTs is problematic. It creates a cascade of ambiguity that undermines the meaning and interpretation of SCT scores. Without a clear and stable factor structure, there is concern that SCT scores will be used irresponsibly to make misleading and inaccurate judgments about examinees. Until this issue is resolved, intentions to utilize SCTs in high-stake assessments are threatened. Moreover, these outcomes serve as a reminder that respectable reliability coefficients alone are not sufficient evidence of validity. Our findings coupled with a critical review of the SCT literature compel us to think that indeterminate dimensions are a by-product of scoring incongruities, weak commonalities between items, and response confusion. While it may be possible to overcome these shortcomings by restructuring the format and scoring practices of SCTs, continued research in this area is needed.

Ethical Approval Indiana University-Purdue University-Indianapolis (IUPUI) Institutional Review Board approved this study.

Sample 1: A Sample Vignette and Questions from a typical Emergency Medicine SCT

A 22-year-old female presents to the Emergency Department complaining of lower abdominal pain for the last 12 h. She describes the pain as sharp in the right lower quadrant. She has some nausea with one episode of vomiting. Her last menstrual period was 6 weeks prior, but she is irregular. She has only one sexual partner, who is male.

Given the above case scenario, answer the following questions:

Diagnostic questions

	If you were thinking of the following diagnosis…	…and you find the following evidence….	…the hypothesis becomes…
1	Appendicitis	Normal WBC count	−2 −1 0 +1 +2
2	Tubo-ovarian abscess	Unilateral right adenexal tenderness with palpable mass on pelvic exam	−2 −1 0 +1 +2
3	Urinary tract infection	History of dysuria and frequency	−2 −1 0 +1 +2

−2 ,highly unlikely; *−1*, less likely than before; *0*, neither more nor less likely; *+1*, more likely than before; *+2*, very likely

Investigational questions

	If you were considering asking for…	…and you find the following evidence….	…this investigation becomes…
4	Pelvic ultrasound	Serum hCG=650 mIU/ml	−2 −1 0 +1 +2

−2, not useful at all; *−1*, less useful; *0*, neither more nor less useful; *+1*, useful; *+2*, absolutely necessary

Therapeutic questions

	If you were considering asking for…	…and you find the following evidence….	…this treatment becomes…
5	IV morphine	Positive urine pregnancy test	−2 −1 0 +1 +2
6	Ceftriaxone and azithromycin	Bilateral adnexal and cervical motion tenderness	−2 −1 0 +1 +2

−2, contraindicated totally or almost totally; *−1*, not useful, possibly detrimental; *0*, neither more nor less useful; *+1*, useful; *+2*, absolutely necessary

References

1. Charlin B, Brailovsky C, Leduc C, Blouin D. The diagnosis script questionnaire: a new tool to assess a specific dimension of clinical competence. Adv Health Sci Educ. 1998;3(1):51–8.
2. Schmidt HG, Norman GR, Boshuizen HPA. A cognitive perspective on medical expertise—theory and implications. Acad Med. 1990;65(10):611–21.
3. Fournier J, Demeester A, Charlin B. Script concordance tests: guidelines for construction. BMC Med Informatics Decis Making. 2008;8(1):18.
4. Charlin B, Brailovsky C, Brazeau-Lamontagne L, Samson L, Leduc C, Van der Vleuten C. Script questionnaires: their use for assessment of diagnostic knowledge in radiology. Med Teach. 1998;20(6):567–71.
5. Brailovsky C, Charlin B, Émond C, Maltais P. Script questionnaire as a method of assessing clinical reasoning after educational programs 1999.
6. Wilson A, Pike G, Humbert A. Preliminary factor analyses raise concerns about script concordance test utility. Med Sci Educator. 2014;24(1):51–8.
7. Lubarsky S, Gagnon R, Charlin B. Scoring the script concordance test: not a black and white issue. Med Educ. 2013;47:1152–61.
8. Messick S. Validity. In: Linn RL, editor. Educational measurement. 3rd ed. New York: Macmillan; 1989. p. 13–103.
9. 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. Med Teach. 2011;33(6):472–7.
10. Thompson B. Exploratory and confirmatory factor analysis. Washington: American Psychological Association; 2004.
11. Schmid J, Leiman JM. The development of hierarchical factor solutions. Psychometrika. 1957;22(1):53–61.
12. Wolff H-G, Preising K. Exploring item and higher order factor structure with the Schmid-Leiman solution: syntax codes for SPSS and SAS. Behav Res Methods. 2005;37(1):48–58.
13. Gorsuch RL. Factor analysis. 2nd ed. Hillsdale: Lawrence Erlbaum Associates; 1983.
14. Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. Med Educ. 2013;47:1175–83.
15. Kreiter CD. Commentary: the response process validity of a script concordance test item. Adv Health Sci Educ. 2012;17(1):7–9.

This manuscript has not been published elsewhere nor has it been submitted simultaneously for publication elsewhere.