Canadian Medical Education Journal

Major Contribution/Research Article

Using Item Analysis to Assess Objectively the Quality of the Calgary-Cambridge OSCE Checklist

Tyrone Donnon, Murray Lee and Sarah Cairncross

University of Calgary, Calgary, Alberta, Canada

Published: 31 March, 2011

CMEJ 2011, 2(1):e16-e22 Available at http://www.cmej.ca

© 2011 Donnon, Lee and Cairncross; licensee Synergies Partners

This is an Open Journal Systems article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: The purpose of this study was to investigate the use of item analysis to assess objectively the quality of items on the Calgary-Cambridge Communications OSCE checklist.

Methods: A total of 150 first year medical students were provided with extensive teaching on the use of the Calgary-Cambridge Guidelines for interviewing patients and participated in a final year end 20 minute communication OSCE station. Grouped into either the upper half (50%) or lower half (50%) communication skills performance groups, discrimination, difficulty and point biserial values were calculated for each checklist item.

Results: The mean score on the 33 item communication checklist was 24.09 (*SD* = 4.46) and the internal reliability coefficient was α = 0.77. Although most of the items were found to have moderate (*k* = 12, 36%) or excellent (*k* = 10, 30%) discrimination values, there were 6 (18%) identified as 'fair' and 3 (9%) as 'poor'. A post-examination review focused on item analysis findings resulted in an increase in checklist reliability (α = 0.80).

Conclusions: Item analysis has been used with MCQ exams extensively. In this study, it was also found to be an objective and practical approach to use in evaluating the quality of a standardized OSCE checklist.

Correspondence: Tyrone Donnon, PhD, Medical Education and Research Unit, Room G13 Heritage Medical Research Building, Faculty of Medicine, University of Calgary, 3330 Hospital Drive, NW, Calgary, Alberta, Canada, T2N 4N1; Tel: (403) 210-9682; Fax: (403) 210-7507; E-mail: tldonnon@ucalgary.ca

Introduction

With the intent of writing quality examinations for medical school, faculty educators will spend a considerable amount of time and effort to design and develop test items for multiple choice question (MCQ) exams. Notwithstanding, an inordinate amount of effort is also spent on creating comprehensive checklists for objective structured clinical examination (OSCE) stations that encompass the clinical skill competencies expectations of our students and residents at various stages in their educational development. As with MCQ items, most faculty often believe that they have written excellent OSCE scenarios and corresponding itemized checklists reflecting the knowledge, skills and attitudinal objectives of recent teaching and learning experiences. The use of item analysis can provide an objective approach to interpret examinees' performance on an OSCE.

In a recent systematic review, Patrico et al.[1] reported that limited information and inconsistency in reporting on the use of OSCEs in medical education research were common. In particular, only 34.6% of the studies reviewed provided evidence of the internal consistency of items within stations (11.5%) or reported data that would allow for the calculation of an aggregated or total reliability coefficient across a number of OSCE stations (23.1%). The use of item analysis to improve reliability of an OSCE was explored in a study by Auewarakul et al.,[2] where the removal of poor performing or 'problem stations' resulted in subsequent increases in generalizability coefficients. While the use of item analysis has become common practice after the administration of an MCQ examination, item quality analysis for OSCE checklists have been limited to reliability analysis as a function of the average of the inter-item correlations or of all the items within a station as a measure of a station's internal consistency. Item analysis can provide a statistical approach to investigating the quality of OCSE checklist items while exploring ways to improve the internal reliability of the test as a whole.

An item analysis provides insights into the quality of the items or questions that are a reflection of the teaching and learning process that preceded the examination. The quantitative information from an item analysis reveals the student's level of understanding of the content while providing feedback to the instructors on improvements to the writing or revision of items and issues related to the quality of the teaching provided.[3] A well constructed test that meets the content expectations of a course or program should challenge learners appropriately, demonstrate they have a comprehensive understanding of the material or skills taught, and identify whether they are ready to proceed to the next stage of their studies or training. After the administration of an MCQ examination, it is common for most medical schools to have an item analysis prepared and reviewed by faculty members or evaluation consultants. The item analysis can provide important information about the quality of each of the MCQ items used in the exam and allow for faculty to make objective decisions about whether or not the item is functioning at a desired level of expectation. For example, it is common for faculty to believe that an item is satisfactory only to find that the question is intrinsically ambiguous as a result of an undesired response pattern elicited by the item analysis.[4]

Using Item Analysis for a Communications OSCE Checklist

There are two basic indices related to an item analysis: 1) *difficulty* simply refers to proportion of examinees that answered the question correctly, and 2) *discrimination* compares a proportion of more knowledgeable examinees with an equal proportion of less knowledgeable examinees on an item as a function of their overall performance on the examination. More importantly, an item analysis can be used to assess the quality of the items as a function of assessing individual differences, improving the internal reliability of the overall test, and providing feedback to instructors and writers of exams.

Although logical relevance is the principal criterion for test validity (i.e., content validity), an item analysis is essential to determining whether the items on exams are functioning in the desired manner. The purpose of this present study was to investigate the use of item analysis on a well-established, standardized OSCE checklist. In particular, an approach to item analysis was specifically designed to look at multiple option responses on items, the development of difficulty and discrimination indices, and a process for objectively evaluating the quality of checklist items and providing feedback to course instructors.

Methods

Setting and participants

A total of 150 first year medical students familiar with the Calgary-Cambridge Guidelines for interviewing patients participated in a final year end examination which included a 20 minute communication OSCE station. The students' performance on the communication checklist with a standardized patient was assessed by a physician examiner located in an adjacent patient examination room separated by a one-way mirror.

Instrumentation

The Calgary-Cambridge Observation Guides for teaching emphasizes the importance of communication skills as a core competency in becoming an effective doctor.[3] As part of preclinical training for medical students in the development of their patient communication skills, key basic tasks are identified using the patient interviewing guide: 1) initiating the session, 2) gathering information, 3) providing structure to the consultation, 4) building relationship and 4) closing the session.[4] The corresponding checklist designed for the patient interview focuses on both the process taught to the medical students within the context of a specific clinical presentation. At the end of year preclinical OSCE, the communication skills station consists of a 20 minute encounter with a standardized patient trained to present, in this case, with a cough and potential chest infection. During the OSCE, the examiner uses a 33 item checklist to score medical students' performances based on a three point scale: 0 = "No" (not attempted), 0.5 = "Yes, but" (attempted, but not completed), and 1 = "Yes" (completed).

Procedures for OSCE item analysis

Students' performance on the communication OSCE checklist was scanned for data entry, and a local software program was developed to conduct a basic item analysis of the raw scores. The students' total score was calculated as a sum of the total number of items or tasks they had completed correctly on the checklist. Based on a student's total score on the communication checklist, he or she is placed into either the upper half (50%) or lower half (50%) performance groups. To maintain anonymity and confidentiality of the students when results are reviewed, all data are presented in an aggregated format and summarized by checklist item only. To determine a checklist item's discrimination value, the proportion of students in the upper group that completed the task correctly is subtracted by the proportion of students in the lower group that completed the task correctly. To obtain an estimate of the difficulty value, we reported the total proportion of all students that were able to complete the task correctly. A point biserial score was calculated for each item option, which is similar to test item discrimination and can also be used to assess item quality. The point biserial value is a correlation calculation between the item option as a dichotomous variable and the continuous variable represented by the students' total score on the OSCE checklist. This correlation between item score and total score quantifies how well the item contributes to overall reliability of the exam. The overall internal consistency or reliability of the communication OSCE was investigated using Cronbach's alpha (α).

Results

The mean score on the 33 item communication checklist was 24.09 (*SD* = 4.46) and the internal reliability coefficient was α = 0.77. As shown in Table 1, most of the items were found to have moderate (*k* = 12, 36%) or excellent (*k* = 10, 30%) discrimination values. Although there were no negative discriminating items, there were 6 (18%) identified as 'fair' and 3 (9%) as 'poor'.

A complete list of the 33 checklist items and the item analysis values for the discrimination and difficulty indices are provided in Table 2. As shown by the proportion differences in students from the upper and lower half groups, those items with discrimination values in the excellent category (0.40 or higher) were associated with difficult values that ranged from between 0.57 to 0.83. Items with fair or poor discrimination values, however, were primarily associated with difficulty values that were less than 0.15 or greater than 0.83 (with the exception of item number 24). Item 24 assessed students' ability to structure the consultation by attending to the timing of the information gathering process resulting in a difficulty value of 0.61 and a fair discrimination value of 0.10. Although the majority of students did well on this item, it was decided that this may be related to a teaching issue in that an emphasis on attention to the timing as a function of the quality of the patient encounter needs to be emphasized with students in the future.

Overall, all of the items have elicited the desired response pattern from the students in that the "Yes" (correct) option resulted in positive discrimination values while the other two options ("Yes, but" and "No") were found to discriminate negatively or not at all. There were three items, however, that were flagged for post-examination review based on their poor discrimination values (i.e., 0.01 to 0.09). Items number 12 (difficulty = 0.93, discrimination = 0.04) and 27 (difficulty = 0.98, discrimination = 0.07) were determined to have been item objectives that had been 'mastered' by the students, reflecting the overall student performance achieved on these two items. Item number 15 had a difficulty value of 0.15 (indicating that only 15% of the students received full marks for this item). A greater percent of the students (47%) did receive partial marks (0.5 = "Yes, but") for gathering a portion of the information regarding the patient's "history of pets, travel or asthma." A decision to remove this item was made and a re-run of the analysis improved the overall internal consistency of the communication OSCE checklist to α = 0.80.

Discussion

The main findings of the present study are that 1) item analysis results can be used as an objective and practical approach for medical schools to evaluate the quality of OSCE checklists, 2) discrimination and difficulty indices can be used to identify items that are intrinsically ambiguous and provide instructors with insights on how to improve the quality of the checklist and potentially clarify students' misunderstanding of concepts in the future.

Item analysis procedures have been used primarily for MCQ exams and are only practical when a computerized item analysis program is available. Using a locally designed item analysis software program, we used the data collected from a group of preclinical students' performance on a standardized communication OSCE checklist to derive item statistics on how difficult and well the checklist items discriminate between lower and higher performance examinees. The results from the communication OSCE checklist which is based on the Calgary-Cambridge Guidelines for patient interviewing showed overall positive and a majority of moderate (*k* = 12, 36%) to excellent (*k* = 10, 30%) discrimination values. As a function of the quality of these items, the overall internal reliability of the 33 item checklist was α = 0.77. On a review of items that were poor discriminators in distinguishing between knowledgeable and less knowledgeable examinees, the removal of item 15 resulted in an improved reliability coefficient of α = 0.80.

The use of the communication OSCE checklist item analysis during the post-examination review process provided instructors and student representatives with an opportunity to discuss the examinees' performance in an objective manner. For example, both parties focused more on those checklist items that discriminated poorly (identified as less reliable measures of the examinees' performance overall) than on whether or not students' achieve full marks on each item. Items that were identified for discussion based on their poor discrimination values were shown to fall into three categories: 1) items that reflected overall mastery (93% or more of the students scored the item correctly) and hence little or no discrimination was found between upper and lower performance groups, 2) items that reflected students' misunderstanding of concepts which could be potentially remediated in subsequent teaching sessions, and 3) an item that examinees had not performed well on (15% received full marks for item) and the discussion focused on removal and potential rewriting of the item for future use.

The use of item analysis on a standardized communication OSCE checklist based on the Calgary-Cambridge guide for patient interviewing demonstrates that the items designed for the assessment of communication skills of preclinical medical students' have overall positive and good discrimination values and high internal consistency or reliability. In a review of each item's discrimination and difficulty values, we identified potential items of concern and were able to objectively address issues specific to discrepancies we found on examinees' performance based on the item analysis results. Items flagged for discussion were related to poor discrimination values associated with items being too easy (e.g., > 90% complete task correctly) or too difficult (e.g., $\leq$ 15% complete task correctly), and items with moderate difficulty values that poorly discriminated between less and more knowledgeable students. The improvement of overall reliability of the communication OSCE checklist was achieved by the removal of a single item with low difficulty and poor discrimination values.

Although item analysis was found to be an objective and practical approach to evaluating the quality of a standardized checklist, further studies that assess other clinical skills (e.g., physical examination, decision-making, patient management) and involve multiple disciplines or contexts (e.g., emergency medicine, pediatrics, surgery) may lead to a more comprehensive method to using item analysis for assessing the internal reliability of OSCE checklists. In addition, there is the need to investigate content validity or teaching issues related to students' misunderstanding of course material and clinical skill development identified by their poor performance on an OSCE as a function of item analysis results.

Acknowledgements

The authors acknowledge the students, staff and administration at the Office of Undergraduate Medical Education for their continuing support in our research efforts to improve the quality of assessment methods and approaches used at our medical school.

References

1. Patricio M, Juliao M, Fareleira F, Young M, Norman G, Vaz Carneiro A. A comprehensive checklist for reporting the use of OSCEs. *Med Teach.* 2009;31:112-124.

2. Auewarakul C, Downing SM, Praditsuwan R, Jaturatamrong U. Item analysis to improve reliability for an internal medicine undergraduate OSCE. *Adv in Health Sci Educ.* 2005;10:105-113.

3. Ebel RL, Frisbie DA. *Essentials of Educational Measurement*, 5th ed. Englewood Cliffs, NJ: Prentice-Hall 1991.

4. Hopkins KD. *Educational and Psychological Measurement and Evaluation*. Needham Heights, MA: Allyn and Bacon 1998.

5. Kurtz SM, Silverman JD. The Calgary-Cambridge Referenced Observation Guides: an aid to defining the curriculum and organizing the teaching in communication training programmes. *Med Educ.* 1996;30:83-89.

6. Silverman JD, Kurtz SM, Draper J. *Skills for Communicating with Patients*. Oxford, UK: Radcliffe Medical Press 1998.

Table 1. Index of discrimination used for communication OSCE checklist

Index of Discrimination	Item Discrimination Description	Number of Items
0.40 or higher	Excellent	10
0.30 to 0.39	Good	2
0.20 to 0.29	Moderate	12
0.10 to 0.19	Fair	6
0.01 to 0.09	Poor	3
0.00 or negative	Mis-keyed or intrinsically ambiguous	0

Note: Modified from Hopkins[4]

Table 2. OSCE checklist for communication process skills station for preclinical students (*n* = 150)

Checklist Items	Difficulty	Discrim	Lower	Upper
Initiating the Session				
01. Introduces Self and Role	.91	*.17*	.80	.97
02. Identifies and confirms problems list	.73	**.40**	.45	.85
03. Negotiates agenda	.78	.32	.60	.92
Gathering Information (Problem Exploration)				
04. Encourages patient to tell story using two open ended questions	.82	.29	.63	.92
05. Appropriately moves from open to closed questions	.81	**.41**	.53	.93
06. Allows patient to complete statements without interruption	87	.27	.70	.97
07. Facilitates patients' responses verbally/nonverbally	.83	**.43**	.55	.98
08. Clarifies patient's statements	.92	.21	.78	.98
09. Uses closed questions to consider systemic symptoms	.41	.27	.33	.59
10. Identifies two or more details about cough	.87	*.13*	.83	.95
11. Establishes dates (onset of cough, chest infection)	.83	*.16*	.73	.89
12. Smoking history	.93	*.04*	.93	.97
13. Determines marital issues exist, spouse refuses counseling	.30	.28	.18	.46
14. Alcohol history – amount, reasons why, AA consideration	.26	.29	.10	.39
15. No history of pets, travel or asthma	.15	*.08*	.10	.18
16. Sibling history of emphysema, death from same	.92	.20	.80	1.00
Gathering Information (Patient's Perspective)				
17. Determines/acknowledges patient's ideas regarding cause	.49	.26	.30	.56
18. Explores patient's concerns re. problem – how it affects life	.90	*.18*	.78	.95
19. Encourages expression of emotions	.65	**.66**	.23	.89
20. Notes verbal/non-verbal clues (recognizes/attends to cough)	.82	.25	.65	.90
Providing Structure to the Consultation				
21. Summarizes at end of a specific line of inquiry	.65	**.45**	.40	.85
22. Progresses using transitional statements	.85	.24	.73	.97
23. Structures logical sequence	.81	**.42**	.55	.97
24. Attends to timing	.61	*.10*	.58	.67
Building Relationship				
25. Demonstrates appropriate non-verbal behavior	.91	.25	.75	1.00
26. If reads, writes, doesn't interfere with dialogue/rapport	.81	.34	.58	.92
27. Is not judgmental	.98	*.07*	.93	1.00
28. Empathizes with and supports patient	.65	**.64**	.25	.89
29. Appears confident	.87	.27	.70	.97
Closing the Session				
30. Closes interview by summarizing briefly	.57	**.45**	.30	.75
31. Contracts with patient regarding next steps/follow-up	.59	**.43**	.38	.80
32. Avoids jargon during explanation	.87	*.18*	.78	.95
33. Encourages patient to seek clarification/express doubts	.70	**.51**	.38	.89

Note: Checklist based on the Calgary-Cambridge Guide for Patient Interviewing.[6]

Table 3. OSCE checklist for communication process skills station for preclinical students (*n* = 150)

Item No.	Prop. Correct	Disc Index	Point Biserial	Options	Prop Total	Lower (50%)	Upper (50%)	Point Biserial
01	0.91	0.17	0.26	No	0.01	0.03	0.00	-0.08
				Yes, but	0.09	0.18	0.03	-0.25
				Yes	0.91	0.80	0.97	0.26
02	0.73	0.40	0.39	No	0.03	0.05	0.00	-0.07
				Yes, but	0.24	0.50	0.15	-0.37
				Yes	0.73	0.45	0.85	0.39
03	0.78	0.32	0.37	No	0.22	0.40	0.08	-0.37
				Yes, but	0.00	0.00	0.00	0.00
				Yes	0.78	0.60	0.92	0.37
04	0.82	0.29	0.42	No	0.01	0.03	0.00	-0.22
				Yes, but	0.17	0.35	0.08	-0.37
				Yes	0.82	0.63	0.92	0.42
05	0.81	0.41	0.52	No	0.04	0.13	0.00	-0.42
				Yes, but	0.15	0.35	0.07	-0.33
				Yes	0.81	0.53	0.93	0.52
06	0.87	0.27	0.39	No	0.02	0.08	0.00	-0.27
				Yes, but	0.11	0.23	0.03	-0.30
				Yes	0.87	0.70	0.97	0.39
07	0.83	0.43	0.49	No	0.01	0.03	0.00	-0.09
				Yes, but	0.16	0.43	0.02	-0.48
				Yes	0.83	0.55	0.98	0.49
12	0.93	0.04	0.14	No	0.00	0.00	0.00	
				Yes, but	0.07	0.08	0.03	-0.14
				Yes	0.93	0.93	0.97	0.14
15	0.15	0.08	0.17	No	0.39	0.53	0.28	-0.24
				Yes, but	0.47	0.38	0.54	0.11
				Yes	0.15	0.10	0.18	0.17
24	0.61	0.10	0.10	No	0.05	0.08	0.02	-0.06
				Yes, but	0.33	0.35	0.31	-0.07
				Yes	0.61	0.58	0.67	0.10