Advances in Health Sciences Education (2008) 13:59–69
DOI 10.1007/s10459-006-9024-4

Checklist Content on a Standardized Patient Assessment: An Ex Post Facto Review

JOHN R. BOULET[1,*], MARTA VAN ZANTEN[1], ANDRÉ DE CHAMPLAIN[2], RICHARD E. HAWKINS[2] and STEVEN J. PEITZMAN[1]
[1]*Educational Commission for Foreign Medical Graduates (ECFMG®), 3624 Market Street, Philadelphia, PA, 19104-2685, USA;* [2]*National Board of Medical Examiners (NBME®), Philadelphia, PA, USA (*author for correspondence, Phone: 215-823-2227; Fax: 215-386-3309; E-mail: jboulet@ecfmg.org)*

(Received: 26 April 2006; Accepted: 7 July 2006)

Abstract. While checklists are often used to score standardized patient based clinical assessments, little research has focused on issues related to their development or the level of agreement with respect to the importance of specific items. Five physicians independently reviewed checklists from 11 simulation scenarios that were part of the former Educational Commission for Foreign Medical Graduate's Clinical Skills Assessment and classified the clinical appropriateness of each of the checklist items. Approximately 78% of the original checklist items were judged to be needed, or indicated, given the presenting complaint and the purpose of the assessment. Rater agreement was relatively poor with pairwise associations (Kappa coefficient) ranging from 0.09 to 0.29. However, when only consensus indicated items were included, there was little change in examinee scores, including their reliability over encounters. Although most checklist items in this sample were judged to be appropriate, some could potentially be eliminated, thereby minimizing the scoring burden placed on the standardized patients. Periodic review of checklist items, concentrating on their clinical importance, is warranted.

Key words: checklist, inter-rater agreement, reliability, standardized patient, validity

Background

The use of standardized patient (SP)-based assessments to measure the clinical skills of medical students, residents and physicians in practice is widespread (Adamo, 2003; Newble, 2004). These assessments are usually scored using case-specific checklists consisting of both history items that an examinee (e.g., medical student, physician) is expected to ask and physical exam maneuvers that should be performed. For most assessments, the checklist is completed by the performing SP immediately following each encounter. Depending on the nature of the assessment, the length of the

encounter, and the complexity of the simulated patient complaint, the number of items can range from a few to dozens.

While much research has been conducted investigating the scoring and utility of case-specific history taking and physical examination checklists (Boulet et al., 2002;Whelan et al., 2005;), few studies have specifically focused on issues concerning how the cases, and associated checklists, are developed. For a clinical skills exam to be valid, the checklists must consist of items that are truly essential to the task. A review by Gorter et al. (2000) concluded that too few researchers are fully describing their methods of checklist construction. Moreover, there is little, if any, published evidence that specifically links checklist content to the evidence-based medical literature. All this information is necessary for others to be able to effectively evaluate the validity of the checklists, a particularly critical issue when the scores are being used for credentialing or licensure decisions.

While most checklists used for assessment purposes are developed by a group of persons, including medical specialists, there is a lack of research on issues related to the level of agreement amongst individuals on the relative importance, or necessity, of specific items. During the test development process, committee members usually create the checklists together as a team, with either a consensus or majority model for deciding which items to include. Due to the complex nature of clinical medicine, and the sometimes nebulous nature of the simulated patient complaint(s), it is not surprising that there is often disagreement regarding specific content.

Although checklist content is extremely important, there are both clinical and logistical limits on the number of items that can be included for scoring. From a clinical perspective, especially for time-limited patient interviews, the physician examinee can only delve so far into the patient's history. Moreover, for common clinical presentations, the relevant patient history is, by nature, somewhat constrained. From a logistical perspective, it is common for SPs to document which checklist items were attained, but only after the encounter is finished. Keeping in mind content validity concerns, it therefore makes sense to attempt to minimize the amount of information that must be memorized. Since checklists typically have from 15 to 25 items (Boulet et al., 2003), the task of accurately documenting what did and did not occur in the clinical encounter can be onerous, especially when the SP must interact with several physicians in sequence.

Purpose

The initial purpose of this study was to investigate the levels of agreement amongst physicians who independently evaluated the importance of checklist items from standardized patient cases that were previously administered in a high-stakes clinical skills examination. Based on the results of this

investigation, and the analysis of reviewer agreement, the second goal was to determine whether scores based solely on agreed upon indicated history taking questions and physical examination maneuvers were appreciably different from those based on all original items combined.

Methods

Assessment instrument

The Educational Commission for Foreign Medical Graduates (ECFMG®) is responsible for certifying international medical graduates (IMGs) who wish to pursue graduate medical training in the US. There currently are a number of ECFMG certification requirements, including a passing score on a clinical skills examination. The Clinical Skills Assessment (CSA®) fulfilled this requirement from July 1998 to April 2004; it has since been replaced by a similar standardized patient assessment, the United States Medical Licensing Examination (USMLE™) Step 2 Clinical Skills (CS) exam (Federation of State Medical Boards & National Board of Medical Examiners, 2003).

The CSA was an SP-based performance examination that required candidates to demonstrate their clinical skills in a simulated medical environment. Candidates had 15-minutes to interact with each of the SPs, and 10-minutes after encounters to document and interpret their findings. History taking (Hx) and physical examination (PE) skills were assessed via case-specific checklists scored by the SP following the encounter. The combination of Hx and PE is known as data gathering (DG).

The cases used in the CSA were created by a group consisting of 10 physicians and 1 nurse educator who were recruited in 1996 to participate in the ECFMG Test Development Committee (TDC). Prior to each TDC meeting, members were given case assignments based on anticipated test blueprint vacancies. To facilitate the case writing process, they were provided with a case template where they filled in information regarding symptoms, past medical history, family history, findings on physical examination, etc. During the first day of each meeting, committee members would work in subcommittees to enhance and solidify the clinical scenario. On the second day, each of the newly created case scenarios was given to a sub-committee whose members had not yet seen the case. One of the committee members played the part of the examinee (physician), interviewing and assessing a trained standardized patient. Other committee members observed the encounter and recorded every question the doctor asked and all physical exam maneuvers performed. The sub-committees then came back together and developed the checklists for each of the cases based on (a) what actually happened when an "examinee", blinded to the case materials, worked up the patient and (b) their own personal clinical judgment as to what is important.

Assessment data

For this study, we selected 11 out of the over 200 CSA cases developed by the TDC. From a content perspective, these 11 study cases were purposely chosen to model a typical CSA test form. In addition, to help ensure the generalizability of our findings, we only picked cases with sufficiently large numbers of candidate scores ($n > 2000$ per case). Typically, an examinee would need to interview and evaluate 11 standardized patients as part of an assessment. Therefore, even though examinees were differentially exposed to the cases in this study, the 11 modeled encounters could be conceived of as a hypothetical test form.

Panelists

We recruited five staff physicians to provide the initial study data. Three are specialists in emergency medicine, two are trained in internal medicine, one with a subspecialty of nephrology, and one with a subspecialty of infectious diseases. All of the staff physicians had been involved in the development and validation of high-stakes clinical skills assessments, including the ECFMG CSA and the USMLE Step 2 CS.

Instructions

The physician panelists were told to read through the descriptive materials for each of the 11 cases. They were then instructed to indicate the clinical appropriateness of each checklist item. Three levels were provided: ''indicated and essential'' (coded 1), ''indicated but not essential'' (coded 2), and ''neither indicated nor contraindicated'' (coded 3). For the purpose of this study, ''indicated'' denoted items that, given the patient complaint, would be used by a physician/medical student as part of the hypothesis testing paradigm needed to generate a reasonable diagnosis. Essential items were those that were considered indicated, but also judged to be absolutely necessary for patient care.

Analysis

Rater agreement

Given that the CSA was developed to measure skills required at entry to graduate medical education, we felt that the most important delineation for a checklist item was that it was considered ''indicated'' by the physician panelists. For a resident just beginning graduate training, asking history taking questions and performing physical exam maneuvers that are indicated, some of which may not be essential, would be expected. Therefore, the ''indicated and essential'' and ''indicated but not essential'' categories (1 and 2) were

combined. Even though we collapsed these categories, we noted some variability in the raters' use of the individual categories across the total 215 checklist items in the 11 cases. Depending on the individual rater, between 125 and 157 of the 215 items were judged to be indicated and essential. One rater determined that 36 items, across all 11 cases, were neither indicated nor contraindicated; another rater only placed six items in this category.

Various analyses of the physician judgments were completed. First, we tabulated the number of indicated items. A checklist item was considered indicated if all five physician raters judged it to be "indicated and essential" or "indicated and not essential". Next, we divided the data gathering checklist into the history taking and physical examination components and tabulated, within each of these domains, the number (and percent) of indicated items. Finally, we summarized panelist agreement using the Kappa coefficient (Landis and Koch, 1977).

Scoring

Based on the rater agreement analyses, we dropped items where there was no consensus regarding their indication. Various statistical analyses (e.g., estimation of variance components, descriptive statistics – means, standard deviations, correlations) were used to describe the psychometric properties of the new scores (only indicated items) and to summarize the relationships between these scores and those based on the longer, original, checklists. Although no examinees encountered all of the cases in our hypothetical 11-case test form, it was still possible to estimate variance components based on the bivariate relationships between case scores. This analysis was done separately for the original case checklist scores and the scores based only on the indicated items. Since the same SP performed regardless of scoring method, a person x task model was used.

Results

Case descriptions, including the number of indicated items, are provided in Table I. Based on our criterion, only 167 of 215 items (78%) were judged to be indicated. The percentage of indicated items varied by case: for the chest pain and acute abdominal pain cases, over 30% of the checklist items were not indicated; for the fall case, 20 of 21 checklist items (95%) were categorized as indicated. Over all cases, the percentage of indicated items was lower for physical examination (68%) than for history taking (82%).

To summarize rater agreement we calculated a Kappa coefficient for each pair of judgments, over all cases ($n = 215$ items). Although exact agreement, for any given pair of raters, was relatively high, this is influenced by the fact

Table I. Number of items judged to be indicated, by clinical skill and case

Case	Description of the case	Data gathering (Hx & PE items)		History taking (Hx)		Physical exam (PE)	
		n Items	*n* Indicated	*n* Items	*n* Indicated	*n* Items	*n* Indicated
Acute Abdominal pain	A 20-year-old male complains of diffuse abdominal pain.	25	17 (68)	17	13 (76)	8	4 (50)
Fatigue	A 50-year-old male complains of fatigue and constipation.	17	13 (76)	14	12 (86)	3	1 (33)
Lightheadedness	A 60-year-old female complains of feeling faint and lightheaded.	17	12 (71)	12	10 (83)	5	2 (40)
Headache	A 30-year-old female complains of headaches and left arm numbness and weakness.	24	21 (88)	17	14 (82)	7	7 (100)
Upper abdominal pain	A 40-year-old female complains of intermittent upper abdominal pain.	17	14 (82)	14	12 (86)	3	2 (67)
Forgetfulness	A 60-year-old female is brought to the doctor by her daughter who claims her mother (the patient) is becoming forgetful.	20	15 (75)	12	10 (83)	8	5 (63)
Palpitations	A 30-year-old female complains of episodes of rapid heart beating.	19	14 (74)	16	11 (69)	3	3 (100)
Constipation	A 50-year-old male complains of constipation, fatigue and weight loss.	15	12 (80)	10	9 (90)	5	3 (60)
Weakness	A 60-year-old female complains of weakness and trouble speaking.	21	16 (76)	12	10 (83)	9	6 (67)
Fall	An 80-year-old female complains of pain in her hip due to a fall yesterday.	21	20 (95)	12	12 (100)	9	8 (89)
Chest pain	A 30-year-old male complains of chest pain for the past few hours.	19	13 (68)	14	10 (71)	5	3 (60)
	Total	215	167 (78)	150	123 (82)	65	44 (68)

( ) percent.

that most items are indicated. Kappa coefficients, over the 10 possible rater pairs, ranged from a low of 0.09 (rater 1 versus rater 4) to a high of 0.29 (rater 1 versus rater 3). The average Kappa coefficient was 0.21, indicative of relatively poor agreement.

A comparison of the original scores with those based only on the indicated items is presented in Table II. For both the original and indicated-only scores, candidates, on average, performed better on the history taking component as opposed to physical examination. For most cases, there were small differences (based on the average performance) between the original and indicated-only item scores. This was true for Hx, PE and DG. Overall, the mean Hx score based on indicated only items was slightly higher that for the original score. In contrast, the mean PE score was slightly lower. Overall, based on the DG component, and our hypothetical 11-case form, candidates, on average, would be expected to score only about 2% higher if only the indicated items were counted. For most cases, with the exception of headache, weakness, and chest pain, the mean DG score based on indicated items was higher than that based on the original checklist.

The correlations between the original and indicated-only scores, by case and CSA component, are presented in Table III. For Hx, these correlations

Table II. Case performance with and without non-indicated items

Case	History taking (Hx)		Physical examination (PE)		Data gathering (DG)	
	All items	Indicated only	All items	Indicated only	All items	Indicated only
Acute abdominal pain	75.2 (13.6)	81.3 (13.6)	53.6 (18.2)	59.2 (22.7)	68.9 (12.3)	76.1 (12.8)
Fatigue	59.6 (13.9)	58.9 (15.8)	51.5 (37.7)	60.2 (49.0)	58.6 (13.2)	59.0 (14.6)
Lightheadedness	68.6 (16.0)	73.8 (15.6)	37.1 (26.0)	29.6 (38.5)	60.7 (15.0)	66.5 (15.2)
Headache	64.4 (15.4)	63.0 (17.2)	39.4 (24.7)	39.4 (24.7)	57.9 (14.1)	55.9 (15.1)
Upper abdominal pain	79.6 (11.8)	80.4 (11.8)	78.1 (31.2)	76.3 (31.9)	79.5 (11.8)	80.3 (11.8)
Forgetfulness	65.1 (16.5)	65.0 (16.9)	35.3 (19.5)	35.8 (23.0)	54.2 (14.0)	56.6 (14.3)
Palpitations	59.4 (14.0)	63.6 (14.8)	75.2 (29.8)	75.2 (29.8)	61.2 (13.5)	65.4 (14.0)
Constipation	71.2 (12.7)	74.5 (12.7)	44.1 (21.0)	29.3 (30.2)	63.5 (11.2)	66.3 (12.0)
Weakness	70.4 (15.1)	68.7 (15.9)	72.0 (17.2)	74.8 (21.2)	71.1 (12.4)	70.7 (13.7)
Fall	58.8 (14.4)	58.8 (14.4)	43.6 (16.1)	48.8 (17.8)	52.7 (11.5)	55.1 (12.0)
Chest pain	84.8 (10.6)	84.6 (11.6)	66.2 (22.4)	49.4 (32.9)	80.6 (11.6)	78.7 (11.8)
Mean	68.8	70.2	54.2	52.5	64.4	66.4

( ) standard deviation.

Table III. Correlations between scores with and without non-indicated items

Case	History taking (Hx)	Physical examination (PE)	Data gathering (DG)
Acute abdominal pain	0.92	0.89	0.93
Fatigue	0.94	0.76	0.92
Lightheadedness	0.93	0.76	0.90
Headache	0.97	1.00	0.98
Upper abdominal pain	0.93	0.95	0.93
Forgetfulness	0.95	0.76	0.91
Palpitations	0.87	1.00	0.89
Constipation	0.92	0.87	0.91
Weakness	0.95	0.80	0.92
Fall	1.00	0.98	0.99
Chest pain	0.87	0.82	0.88
Mean	0.93	0.87	0.92

ranged from $r = 0.87$ (chest pain, palpitations) to $r = 1.00$ (fall[1]). For PE, the correlations were, in general, slightly lower, ranging from $r = 0.76$ (e.g., fatigue) to $r = 1.00$ (e.g., headache[2]). For the DG element, which combines the Hx and PE items, the correlations were quite high, ranging from $r = 0.88$ (chest pain) to $r = 0.99$ (fall). Here, the average correlation was $r = 0.92$, indicating that nearly 85% of the variance of the indicated item scores was shared with the original (all item) scores.

Variance components, based on a $P \times T$ (person by task) design, were calculated for both the original and indicated-only checklist scores. Based on the original DG case scores, with an 11-case form, the generalizability coefficient was 0.62. Here, the case variance component was quite large (32% of the total variance), indicating that the cases vary in average difficulty. When the non-indicated items were dropped, the generalizability coefficient was 0.61. The choice of case accounted for 33% of the total variance.

Discussion

The development of checklist items for simulated clinical encounters can take many forms, but generally involves some judgment by content experts. Since Delphi-based processes are commonly used to determine, based on the

[1] All of the original Hx items were judged to be indicated.
[2] All of the original PE items were judged to be indicated.

presenting complaint, what questions should be asked of the patient and which physical examination maneuvers should be performed, it is not surprising that there can be some disagreement as to what data gathering activities are necessary for proper care. For checklist developers, factors such as medical specialty, familiarity with the purpose of the assessment, morbidity associated with the presenting complaint, and clinical experience could all play some role in what is judged to be necessary. Furthermore, given the ever changing, and complex nature, of medical care, checklist content should not be considered to be a static entity. What a physician should ask or do today may be quite different in 5 or 10 years.

The results of our study show that most checklist items were at least indicated. This suggests that, in general, the case development processes were appropriate. Nevertheless, based on a conservative criterion for categorization, there were still some cases where some of the items could be questioned. For example, on the acute abdominal case there was disagreement amongst the panelists regarding the essentialness of the PE item ''checks for signs of anemia''. This lack of consensus was likely the result of panelists recognizing that blood tests would be expected as part of the diagnostic workup, and inspection is not an entirely reliable means of evaluation. Certainly, a more thorough review is required for items where all raters agreed that the item was neither essential nor indicated.

Based on agreement statistics, there were numerous items where not all physicians concurred. From a quality assurance standpoint, these items, especially those where the majority of physicians suggested that the history taking question or physical examination maneuver was not indicated, demand additional review. Here, changes over time in the way medicine is conducted may yield items that, although appropriate when the case was originally constructed, are no longer valid. At the very least, this suggests that a periodic review of the scoring criteria for all cases be performed.

Scoring examinees based only on the consensus indicated items yielded about the same average scores as using the original checklist with all items. Interestingly, for most cases, the DG scores based only on the indicated items were higher than those based on the original checklist. This is most likely due to the elimination of superfluous content; history taking questions and physical examination maneuvers that, given the presenting complaint, would be less likely to be asked or performed. Unfortunately, for this study, different sets of examines encountered each of the cases, making it difficult to determine how any scoring changes would impact the individual candidate, especially in terms of pass/fail status. Nevertheless, the fact that the average case DG scores were only slightly greater and the correlations between the two scores were very high, suggests that, from a psychometric perspective, little is lost by using a somewhat shorter measurement instrument for each

case. To investigate this more fully, it would be necessary to undertake a similar study where all cases seen by a cohort of examinees were reviewed.

Given the documented findings on SP scoring accuracy (De Champlain et al., 1999), eliminating a number of checklist items may be advantageous, both from a measurement perspective and logistically. In terms of score validity, as long as there is evidence to suggest that removed items are not fundamentally necessary for determining candidate abilities, a shorter, more relevant, set of items would seem apropos. While this will often yield a less reliable individual case checklist score, the reproducibility of candidate scores is most important, and more so a function of the number of cases in the assessment (Norcini and Boulet, 2003; van der Vleuten et al., 1991). Based on the results of the variance components analyses, there was no appreciable difference in the generalizability of the original scores and those based on only the indicated items.

The use of more content-relevant checklist items also makes sense from a sampling perspective. The checklist (history taking items, physical examination maneuvers) can be thought of a sample of items from the population of all possible questions and maneuvers. Keeping in mind content relevance and associated validity concerns, the issue becomes "how many items do we need to sample to achieve a reproducible examinee score?" As part of the CSA checklist development process, the committee watched a physician interact with a SP and noted the questions that were asked and physical examination maneuvers that were performed. Although this was only a single interaction, only a subset of these items (a non-random sample) ended up on the checklist. However, if checklists are being used to generate history taking and physical examination scores, there is likely to be a limit on the number of items that can be parsed without negatively affecting the psychometric properties of the overall assessment scores. If only highly content specific items remain after review, one would expect that task sampling variability (individual examinee performance differences from case to case) would increase, thereby reducing the reproducibility of the skill-based score. Here, at least based on a reliability criterion, Monte Carlo studies, sampling from existing data sources, could be used to establish the lower and upper thresholds for checklist length. For example, if we had a number of case checklists, and data from a large number of test takers, we could randomly select various subsets of checklist items, of differing lengths, to score. By contrasting the reliability of these various case and assessment scores, it should be possible, at least from a psychometric perspective, to establish some reasonable bounds for checklist length.

From a logistical perspective, shorter checklists hold numerous advantages. First, they are easier to construct and score. Second, reducing checklist length is consistent with short-term memory capabilities of most human beings. From the SP's perspective, it takes less effort to remember what was

asked and what was done, likely improving the accuracy of the scores. This is especially important for multi-station assessments where the SPs encounter a series of candidates, all asking related sets of questions and performing similar types of physical examination maneuvers. Finally, there can be significant costs associated with training SPs to respond properly to queries and to recognize the appropriateness and correctness of physical examination maneuvers. Limiting checklist content could certainly reduce expenses associated with examination delivery.

The results of this study suggest that the selection of checklist content for a standardized patient examination can be somewhat subjective. However, if a sampling framework is embraced, it may not be necessary, at least from a psychometric perspective, to incorporate all expert selected history taking questions or physical examination maneuvers.

References

Adamo, G. (2003). Simulated and standardized patients in OSCEs: achievements and challenges 1992–2003. *Medical Teacher* **25**: 262–270.

Boulet, J.R., McKinley, D.W., Norcini, J.J. & Whelan, G.P. (2002). Assessing the comparability of standardized patient and physician evaluations of clinical skills. *Advances in Health Sciences Education: Theory and Practice* **7**: 85–97.

Boulet, J.R., McKinley, D.W., Whelan, G.P. & Hambleton, R.K. (2003). Quality assurance methods for performance-based assessments. *Advances in Health Sciences Education: Theory and Practice* **8**: 27–47.

De Champlain, A.F., Macmillan, M.K., King, A.M., Klass, D.J. & Margolis, M.J. (1999). Assessing the impacts of intra-site and inter-site checklist recording discrepancies on the reliability of scores obtained in a nationally administered standardized patient examination. *Academic Medicine* **74**: S52–S54.

Federation of State Medical Boards & National Board of Medical Examiners (2003). *2004 USMLE Step 2 CS Content Description and General Information Booklet* Philadelphia: FSMB and NBME.

Gorter, S., Rethans, J.J., Scherpbier, A., Heijde, D., Houben, H. & Vleuten, C., et al. (2000). Developing case-specific checklists for standardized-patient-based assessments in internal medicine: A review of the literature. *Academic Medicine* **75**: 1130–1137.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**: 159–173.

Newble, D. (2004). Techniques for measuring clinical competence: objective structured clinical examinations. *Medical Education* **38**: 199–203.

Norcini, J. & Boulet, J. (2003). Methodological issues in the use of standardized patients for assessment. *Teaching and Learning in Medicine* **15**: 293–297.

van der Vleuten, C.P., Norman, G.R. & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of reliability. *Medical Education* **25**: 110–118.

Whelan, G.P., Boulet, J.R., McKinley, D.W., Norcini, J.J., van Zanten, M. & Hambleton, R.K., et al. (2005). Scoring standardized patient examinations: lessons learned from the development and administration of the ECFMG Clinical Skills Assessment (CSA). *Medical Teacher* **27**: 200–206.