

# Developing Case-specific Checklists for Standardized-patient–Based Assessments in Internal Medicine: A Review of the Literature

Simone Gorter, MD, Jan-Joost Rethans MD, PhD, Albert Scherpbier, MD, PhD, Désirée van der Heijde, MD, PhD, Harry Houben, MD, PhD, Cees van der Vleuten, PhD, and Sjef van der Linden, MD, PhD

## ABSTRACT

**Purpose.** To review the literature on the methods used in writing case-specific checklists for studies of internal medicine physicians' performances that were assessed by standardized patients.

**Method.** The authors searched Medline, Embase, Psychlit, and ERIC for articles in English published between 1966 and February 1998. The following search string was used: "[standardi\* or simulat\* or programm\*] near (patient\* or client\* or consultati\*)] and internal medicine." The authors then searched the reference lists of papers retrieved from the database searches, as well as those from seven proceedings of the International Ottawa Conference on Medical Education and Assessment.

**Results.** The procedure yielded 29 relevant articles: database searches yielded 14 published reports dealing with case-specific checklists, 11 articles were culled from the reference lists of these papers, and the Ottawa Conference proceedings yielded four articles. Only 12 articles reported specifically on the development of checklists. In general, there were three sources used for developing checklists: panels of experts, the investigators themselves, and

responses from expert physicians to written protocols. No article indicated that researchers had relied exclusively on data from the literature to compose their checklists. Only three articles indicated that literature sources had informed their checklist development. All articles except one relied on explicit criteria for the inclusion of items on the checklists. In 21 of the 29 articles, the checklists had been scored by SPs, but the scoring of specific items on the checklists varied according to the purpose of the SP–physician encounter. Only four of the articles made the checklists available or indicated that the checklists could be obtained from the authors.

**Conclusion.** The development of case-specific checklists for SP examinations of physicians' performance has received little attention. To judge the validity of studies of physicians' performances that use SPs, the development processes for the checklists need to be more fully described to enable readers to evaluate the validity and reliability of the studies.

*Acad. Med.* 2000;75:1130–1137.

*Dr. Gorter is a researcher, Dr. van der Heijde is associate professor of rheumatology, and Dr. van der Linden is professor of rheumatology, all in the Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, Maastricht, The Netherlands. Dr. Rethans is a general practitioner and coordinator, standardized-patient program, Skillslab, and Dr. Scherpbier is professor and scientific director, both at the Institute of Medical Education, Faculty of Medicine, Maastricht University. Dr. Houben is a rheumatologist, Department of Rheumatology, Atrium Medical Center, Heerlen, The Netherlands. Dr. van der Vleuten is professor of education and director, Department of Educational Development and Research, Maastricht, The Netherlands.*

New technologies, aging populations, and the growing demand and increasing costs for health care have all stimulated interest in the assessment of the care being provided to patients. Efforts to as-

*Correspondence and requests for reprints should be addressed to Dr. Rethans, Skillslab, Institute of Medical Education, Faculty of Medicine, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands; telephone: + 31 43-3881790; fax: + 31 43-3884127; e-mail: (j.rethans@sk.unimaas.nl).*

sess patients' care have yielded direct and indirect approaches to measuring physicians' performances.<sup>1</sup> The indirect assessment of physicians' performances involves the evaluation of physicians' referral or prescribing behaviors, the review of morbidity or complication rates, the review of medical records, and the assessment of patients' satisfaction rates. Methods for direct assessment of physicians' performance are the observation of practice, using video or audio equip-

ment, and the use of standardized patients (SPs), introduced to medical education by Barrows.<sup>2,3</sup>

In this literature review, we focus on the direct assessment of physicians through the use of standardized patients. In particular, we sought literature describing the development of case-specific checklists for evaluating physicians' performances in an SP assessment encounter.

## BACKGROUND

Standardized patients portray patients in a standardized and consistent fashion either in a formal examination setting or when sent unannounced into clinical practice to assess physicians' performances.<sup>4,5</sup> They can be asymptomatic, they can have stable, abnormal findings on physical examination, or they may be able to simulate physical findings.<sup>6</sup> The SP method has proven to be a valid and reliable method to assess performance.<sup>6,7</sup>

After a physician encounter, SPs usually report the physician's performance on a checklist, which can be either generic or case-specific. Generic lists are used to assess general skills (for example the Arizona rating scale<sup>8</sup>), whereas case-specific checklists provide detailed information about the physician's skills with medical history taking, physical examination, case management, and/or communication. Case-specific checklists are developed specially for each case and are tailored to the content of the consultation. Generic lists can be used for many different cases.

The information gathered from SP checklists can be used for educational purposes, as a needs-assessment tool for developing appropriate continuing medical education programs, or as a tool for summative assessment. In either case, the development of the checklist should be appropriate to the intended use of the data. For example, if checklists are to be used as an assessment tool where decisions about accreditation

may be made, the scoring system needs to be appropriately developed for each checklist. The information on the checklists must be reliable and valid and carefully trained SPs must complete checklists in a consistent way. There is good evidence that the reliability of checklists is sufficient.<sup>9,10</sup> Validity, in general, means that an instrument measures what it is supposed to measure, and for case-specific checklists validity means that the checklist reflects a physician's performance. Checklist construction is, therefore, crucial to the validity and reliability of an SP's assessment of a physician's performance. In order for readers to effectively evaluate the validity of a checklist used in an SP encounter, it is important that they be told who developed the checklist, what procedures were followed, and the order and content of the checklist's items.

This paper presents a review of the literature on the methods used in writing case-specific checklists for studies of internal medicine physicians' performances that were assessed using SPs. We focused primarily on validity aspects, such as who developed the checklist; which development procedure was used; whether or not the development process was based on data from the literature or on data resulting from consensus procedures; what scoring system was used; and, finally, whether or not the checklist was published. Many studies of SPs have taken place in undergraduate settings. Given the focus of our recent project to investigate the performances of rheumatologists in real practice using unannounced SP visits, we reviewed only studies with SPs in internal medicine postgraduate settings.

## METHOD

We searched Medline, Embase, PsycLit, and ERIC for articles in English published between 1966 and February 1998 by using the following string of search terms: "[standardi\* or simulat\*

or programm\*) near (patient\* or client\* or consultati\*)] and internal medicine." We then searched the reference lists of the retrieved papers and all seven proceedings from the International Ottawa Conference on Medical Education and Assessment. For this review we selected only articles that assessed physicians' performances through the use of case-specific checklists.

We structured the review using the following questions:

- Is the process of developing the checklist described?  
If yes,
  - a. Which procedure was used?
  - b. How many and what kind of people (experts, lay people, etc.) were involved in this process?
- Was the development of the checklist based on literature, consensus, or both?
- Was the checklist based on explicit criteria?
- How were results reported and, if applicable, how was the scoring system developed?
- Did standardized patients score the checklist?
- Is the full checklist published or otherwise available?

## RESULTS

The literature search yielded 14 papers dealing with case-specific checklists in studies with standardized patients assessing physicians' performances. In addition, we found 11 articles in the reference lists of these papers. The seven proceedings of the Ottawa Conference on Medical Education and Assessment provided four additional studies. The results for all 29 papers are summarized alphabetically by author in Table 1. The results that follow are organized by research question.

**Descriptions of checklist development.** Of the 29 articles, only 12 reported the procedure of the checklist's development. Checklists were devel-

**Table 1**

Characteristics of Developing Checklists for 29 Articles Retrieved from a Review of the Literature on Standardized Patients' Assessments of Internal Medicine Physicians' Performances							
Reference	Checklist Development Described? Yes/No	Developers/Development	Use of Literature and/or Consensus	Use of Explicit Criteria Yes/No	Results and Scoring System	Scored by SPs? Yes/No	Checklist Published or Available? Yes/No
Boudreau, 1994 <sup>11</sup>	Yes	Expert panel of four senior respiratory specialists	Consensus	Yes	Items assessed on relative importance, 100 points were assigned for each case (SPs scored for satisfaction, expectations, confidence)	Yes	Yes, with article
Bowman, 1992 <sup>22</sup>	No	Not reported	Literature	Yes	% of items obtained	Yes	No
Blackwell, 1985 <sup>37</sup>	No	One faculty member developed checklist	Faculty members	Yes	Different scoring methods: Interviewing skills: poor, satisfactory, excellent (0, 1, 3 points) Data collection: % of information obtained divided by total information available Physical: not done, done incorrectly, done correctly (0, 1, 3); if all important items are identified: 100% Maximum points for single correct diagnosis	Yes	Yes, with article
Calhoun, 1987 <sup>12</sup>	Yes	Panel of three internal medicine staff developed checklist for pertinent medical data	Not reported	Yes	Panel identified items critical for diagnosis and critical psychosocial items; score is total number of items	Yes	No
Carney, 1993 <sup>23</sup>	No	Not reported	Literature and content from a previous continuing medical education program	Yes	% of items obtained	Yes	No
Carney, 1995 <sup>24</sup>	No	Not reported	Literature and content from a previous continuing medical education program	Yes	% of items obtained	Yes	No
Curtis, 1994 <sup>27</sup>	No	Not reported	Not reported	Yes	% of items obtained	Yes	No
Curtis, 1995 <sup>28</sup>	No	Not reported	Not reported	Yes	% of items obtained	Yes	No



Table 1 (Continued)

Reference	Checklist Development Described? Yes/No	Developers/Development	Use of Literature and/or Consensus	Use of Explicit Criteria Yes/No	Results and Scoring System	Scored by SPs? Yes/No	Checklist Published or Available? Yes/No
Petrusa, 1990 <sup>36</sup>	No	Not reported	Not reported	Yes	One author (faculty) provided scoring protocol. Complete and partially correct answers; scoring maximum = 100% per examination (case). Total score = average of 17 examination scores	Yes	No
Russel, 1991 <sup>31</sup>	No	Not reported	Not reported	Yes	% of physicians performing a certain task	Yes	No
Russel, 1992 <sup>32</sup>	No	Not reported	Not reported	Yes	% of physicians performing a certain task	Yes	No
Stimmel, 1996 <sup>38</sup>	No	Not reported	Not reported	Yes	Not reported	Yes	No
Stillman, 1986 <sup>15</sup>	Yes	Study committee members defined protocol, described key findings and maneuvers, etc.	Consensus among study committee members	Yes	History and physical examination: % of items obtained Weighted item score Critical item score Diagnostic aggregate score, thoroughness score	Yes	No
Stillman, 1991 <sup>16</sup>	Yes	Study committee members defined protocol, described key findings and maneuvers, etc.	Consensus among study committee members	Yes	History and physical examination: % of items obtained Weighted item score Critical item score Diagnostic aggregate score, thoroughness score	Yes	Yes
Tamblyn, 1997 <sup>17</sup>	Yes	Consensus panel of eight academically affiliated physicians	Consensus by panel (based on recommendations just published in literature)	Yes	The panel enumerated and classified drug management decisions into categories: optimal, acceptable, suboptimal, and unsafe; they identified contraindications to NSAID therapy that should be assessed % of actions performed	Yes	No
Wenrich, 1995 <sup>20</sup>	Yes	Investigators developed protocol and checklist, reviewed by larger group of investigators	Not reported	Yes	% of items obtained	Yes	No
Wolf, 1995 <sup>26</sup>	No	Not reported	Not reported	Yes	% of items obtained	Yes	No
Woolliscroft, 1989 <sup>18</sup>	Yes	Expert panel of three faculty internists used a modified Delphi method	Not reported	Yes	Expert panel identified critical items; reported % of items obtained, % of critical items obtained	Yes	No

oped in three ways: by panels of experts, by the study investigators, and by the review of findings after the SP encounter protocol had been sent to a group of expert physicians.

In seven studies<sup>11-18</sup> a panel developed the case-specific checklists and indicated important items. For example, Boudreau et al.,<sup>11</sup> in their assessment of consultative skills in respiratory medicine, arranged four SP encounters for 22 physicians. For all cases, an expert panel consisting of four senior university respiratory specialists had identified specific criteria with respect to the quality of the consultation. The expert panel, working by consensus, also assessed each item on the checklist for its relative importance. Using a similar approach, Gordon<sup>14</sup> published a study in which the performances of interns were assessed with the use of three unannounced SPs. An expert panel of two family physicians, two internists, and two housestaff was used to obtain criteria for these cases. First, the panel had to modify a checklist that was developed by the investigators. Then the members met to review the responses, made individual modifications, and met again to reach consensus on each case.

The second procedure that we encountered entailed investigators' developing the checklists themselves. The investigators then presented the checklists to other physicians or investigators for validation.<sup>19-21</sup>

The third procedure was reported by Norman.<sup>7</sup> A written protocol for each case was sent to a group of expert physicians, who were asked to make a differential diagnosis, propose investigations and management options, and list all essential findings. The checklist was then developed based on the expert physicians' responses to the protocol.

**Use of literature or consensus in checklist development.** None of the studies reported that researchers had relied exclusively on data from the literature to compose their checklists. A total of 16 articles did not describe the

development process, but three of these<sup>22-24</sup> did indicate that literature sources had been used to develop the checklists. Carney<sup>23,24</sup> used the content of a continuing medical education program in addition to information obtained from the literature as a source for the checklists.

**Use of explicit criteria for checklist development.** With one exception,<sup>25</sup> all articles reported using explicit criteria for the inclusion of items on their checklists.

**Results and scoring systems.** All 29 articles reported in detail on the scoring or registration of the encounter between the SP and the physician, but the actual scoring or classification of results varied considerably. Most authors provided only an overview of the physicians' performances. They did not compare these performances with a standard, nor did they describe a scoring system. Fourteen studies<sup>18-20,22-24,26-33</sup> reported the frequencies of items or points obtained. Klass<sup>33</sup> also reported results as percentages of items obtained from the maximum, and mentioned but did not elaborate on the application of a weighting system.

Most studies that assigned different weights to different aspects of the SP-physician encounter did so in one of three ways. First, some studies identified critical items, for example items deemed important to establish a diagnosis, and weighted them more highly.<sup>12-15,18,21</sup> Second, some studies used the classification of items in four categories (essential, important, indicated, and non-contributory) with corresponding numeric weights (3, 2, 1, and 0).<sup>7,11,15-17,34</sup> Third, some studies distributed different scores to partially correct and completely correct answers.<sup>25,35-37</sup> Two studies used their own methods to develop scores. Norman<sup>21</sup> used the performances in a criterion group as a reference standard. Stillman<sup>15</sup> performed a study of 336 internal medicine residents. For its scoring, the percentage of checklist items completed was calculated and an

overall score was created by averaging scores from different cases while correcting for differences between cases.<sup>15</sup>

**Use of standardized patients to score the checklist.** In 21 of the 29 articles, the standardized patients scored the checklists.<sup>12,15-20,22-24,26-38</sup> In the remaining eight articles, either observers scored the checklists or the authors did not provide sufficient information to determine who did the scoring.

**Publication of checklists.** Checklists were published or made available by the authors in only four articles.<sup>11,16,35,37</sup>

## DISCUSSION

In our review of the literature concerning the development of checklists in studies where SPs were used to assess physicians, we found 29 articles that met our selection criteria. Some articles appeared to report different aspects of the same study, but because we could not ascertain whether this was actually the case, we included all articles as separate studies.

Based on our review of the literature, we found that only a limited number of articles, 12 of the 29 (41%), reported the process of checklist development. Reports that did describe checklist development differed depending on the aims of the assessment. Standardized patient encounters can be used as an educational or needs-assessment tool, or as a summative way to assess the performances of interns, residents, fellows, or specialists. When SP encounters are used as a summative instrument with possible implications regarding accreditation for participating physicians, our opinion is that stricter rules should be applied to the development of the checklist than would be required if the instrument were to be used for formative assessment or needs assessment. The validity of case-specific checklists used for both educational and summative goals has been demonstrated in undergraduate medicine. With the use of SPs in the formal assessment of physi-

cian's performances, it is preferable to use evidence-based performance standards. However, since only a few evidence-based performance criteria are currently available, these performance criteria have to be defined in other ways, such as with consensus procedures. In the literature we reviewed, one study used a modified Delphi process to reach consensus<sup>8</sup> and three articles<sup>11,13,14</sup> reported a procedure that comes close to a nominal group procedure.<sup>39,40</sup> Some reports indicated that case-specific checklists were developed by a team of experts, but the exact procedure was not reported. Our study suggests that too few researchers are describing the methods they have used to reach consensus, even when a rather pragmatic approach is used, and as a result readers are unable to draw conclusions about the validity of the checklists.

Further, our review of the literature found that, while some studies had based the contents of their checklists on literature data, in most cases it was not clear what literature source had been used. When checklists are based on data from the literature, it is preferable to use evidence-based data. However, these data are scarce, which, again, suggests that procedures are needed to reach consensus on the cases. These procedures must be explained to enable the reader to estimate the value of the assessment and its results. Unfortunately, few authors described this development process and too many of them did not mention it at all.

Scoring systems were described in the literature we reviewed. When the purpose of the assessment is to evaluate physicians' performances, reporting the "percentage found" may be a sufficient scoring system. In SP encounters with other purposes, however, some aspects of the encounter between physician and SP are more important than others, and the checklist should take into account these differences in significance. We found several different scoring methods that weighted for significance through

the classification of items, the identification of critical items, and the distribution of points for partially and completely correct answers. In general, these weights were assigned by means of the consensus procedure.

Our review indicates that, although many people have struggled with the development of these lists, the use and development of case-specific checklists for SP encounters have so far received little attention in the published literature. We found only 29 papers on SPs in postgraduate internal medicine settings; only 40% of them described the development of the checklists in any way, and the details varied widely. To judge the validity of studies with SPs, we believe that it is essential that the development process of these lists be described in the literature for both postgraduate and undergraduate settings. The publication of checklists should be encouraged. We hope that this review will stimulate careful consideration of the validity aspects of checklists as well as encourage researchers to describe the applied procedures in their papers on this topic. Such descriptions will help us further in developing reliable and valid instruments to measure physicians' clinical performances.

This project was supported by the Dutch Arthritis Association.

## REFERENCES

1. Rethans JJ, Westin S, Hays RH. Methods for quality assessment in general practice. *Fam Pract.* 1996;13:468-76.
2. Barrows HS. Simulated patients in medical teaching. *Can Med Assoc J.* 1968;98:674-6.
3. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Acad Med.* 1993;68:443-51.
4. Hoppe RB, Farquhar LJ, Henry R, Stoffelmayr B. Residents' attitudes towards and skills in counseling: using undetected standardized patients. *J Gen Intern Med.* 1990;5:415-20.
5. Rethans JJ, Saebu L. Do general practitioners act consistently in real practice when they meet the same patient twice? Examination of intradoctor variation using standardised (simulated) patients. *BMJ.* 1997;314:1170-3.
6. Van der Vleuten C, Swanson D. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med.* 1990;2:58-76.
7. Norman GR, Tugwell P, Feightner JW. A comparison of resident performance on real and simulated patients. *J Med Educ.* 1982;57:708-15.
8. Stillman PL, Brown DR, Redfield DL, Sabers DL. Construct validation of the Arizona Clinical Interviewing Rating Scale. *Educ Psychol Meas.* 1977;37:1031-8.
9. Tamblyn RM, Klass DJ, Schnabl GK, Kopelew ML. The accuracy of standardized patient presentation. *Med Educ.* 1991;25:100-9.
10. Vu NV, Marcy MM, Colliver JA, Verhulst SJ, Travis TA, Barrows HS. Standardized (simulated) patients' accuracy in recording clinical performance check-list items. *Med Educ.* 1992;26:99-104.
11. Boudreau D, Tamblyn R, Dufresne L. Evaluation of consultative skills in respiratory medicine using a structured medical consultation. *Am J Respir Crit Care Med.* 1994;150:1298-304.
12. Calhoun JG, Woolliscroft JO, Ten Haken JD. Internal medicine house officers' performance as assessed by experts and standardized patients. *J Med Educ.* 1987;62:754-60.
13. Gordon JJ, Saunders NA, Sanson-Fisher RW. Evaluating interns' performance using simulated patients in a casualty department. *Med J Aust.* 1989;151:18-21.
14. Gordon JJ, Saunders NA, Hennrikus D, Sanson-Fisher RW. Interns' performances with simulated patients at the beginning and the end of the intern year. *J Gen Intern Med.* 1992;7:57-62.
15. Stillman PL, Swanson DB, Smee S, et al. Assessing clinical skills of residents with standardized patients. *Ann Intern Med.* 1986;105:762-71.
16. Stillman P, Swanson D, Regan MB, et al. Assessment of clinical skills of residents utilizing standardized patients. A follow-up study and recommendations for application. *Ann Intern Med.* 1991;114:393-401.
17. Tamblyn R, Berkson L, Dauphinee WD, et al. Unnecessary prescribing of NSAIDs and the management of NSAID-related gastropathy in medical practice. *Ann Intern Med.* 1997;127:429-38.
18. Woolliscroft JO, Calhoun JG, Billiu GA, Stross JK, MacDonald M, Templeton B. House officer interviewing techniques: impact on data elicitation and patient perceptions. *J Gen Intern Med.* 1989;4:108-14.
19. Day RP, Hewson MG, Kindy P Jr, Van Kirk J. Evaluation of resident performance in an outpatient internal medicine clinic using

- standardized patients. *J Gen Intern Med.* 1993;8:193-8.
20. Wenrich MD, Paaup DS, Carline JD, Curtis JR, Ramsey PG. Do primary care physicians screen patients about alcohol intake using the CAGE questions? *J Gen Intern Med.* 1995; 10:631-4.
  21. Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL. Knowledge and clinical problem-solving. *Med Educ.* 1985;19:344-56.
  22. Bowman MA, Russell NK, Boekeloo BO, Rafi IZ, Rabin DL. The effect of educational preparation on physician performance with a sexually transmitted disease-simulated patient. *Arch Intern Med.* 1992;152:1823-8.
  23. Carney PA, Dietrich AJ, Freeman DH Jr, Mott LA. The periodic health examination provided to asymptomatic older women: an assessment using standardized patients. *Ann Intern Med.* 1993;119:129-35.
  24. Carney PA, Dietrich AJ, Freeman DH Jr, Mott LA. A standardized-patient assessment of a continuing medical education program to improve physicians' cancer-control clinical skills. *Acad Med.* 1995;70:52-8.
  25. Lloyd JS, Sener SF, Duncan ED, Bundra SJ, Niedenthal AE. Examining the clinical competence of residents in two specialties. In: Hart IR, Harden RM (eds). *Further Developments in Assessing Clinical Competence.* Quebec, Canada: Can-Heal Publications, 1987:609-21.
  26. Wolf FM, Sisson JC, Zweifler AJ. A standardized patient program to evaluate summarization skills in patient interviews. *Acad Med.* 1995;70:443.
  27. Curtis JR, Paaup DS, Wenrich MD, Carline JD, Ramsey PG. Internal medicine residents' skills at identification of HIV-risk behavior and HIV-related disease. *Acad Med.* 1994; 69(10 suppl):S45-S47.
  28. Curtis JR, Paaup DS, Wenrich MD, Carline JD, Ramsey PG. Ability of primary care physicians to diagnose and manage *Pneumocystis carinii* pneumonia. *J Gen Intern Med.* 1995; 10:395-9.
  29. Curtis JR, Paaup DS, Wenrich MD, Carline JD, Ramsey PG. Physicians' ability to provide initial primary care to an HIV-infected patient. *Arch Intern Med.* 1995;155:1613-8.
  30. Gonzalez Willis A, Rafi I, Boekeloo B, et al. Using simulated patients to train physicians in sexual risk assessment and risk reduction. *Acad Med.* 1990;65(9 suppl):S7-S8.
  31. Russell NK, Boekeloo BO, Rafi IZ, Rabin DL. Using unannounced simulated patients to evaluate sexual risk assessment and risk reduction skills of practicing physicians. *Acad Med.* 1991;66(9 suppl):S37-S39.
  32. Russell NK, Boekeloo BO, Rafi IZ, Rabin DL. Unannounced simulated patients' observations of physician STD/HIV prevention practices. *Am J Prev Med.* 1992;8:235-40.
  33. Klass DCC, Hassard T, Kopelow M, Schnabl G. Influence of level of training on performance in a standardized test of clinical abilities. In: Bender W, Hiemstra RJ, Scherpbier AJJA, Zwierstra RP (eds). *Teaching and Assessing Clinical Competence.* Groningen, The Netherlands: Boekwerk Publications, 1990:327-32.
  34. McLeod PJ, Tamblyn RM, Gayton D, et al. Use of standardized patients to assess between-physician variations in resource utilization. *JAMA.* 1997;278:1164-8.
  35. Mangione S, Peitzman SJ, Gracely E, Nieman LZ. Creation and assessment of a structured review course in physical diagnosis for medical residents. *J Gen Intern Med.* 1994;9: 213-8.
  36. Petrusa ER, Blackwell TA, Ainsworth MA. Reliability and validity of an objective structured clinical examination for assessing the clinical performance of residents. *Arch Intern Med.* 1990;150:573-7.
  37. Blackwell TA, Petrusa ER, Saydjari CL. Performance in general internal medicine: comparative study using an objectively scored evaluation system. In: Hart IR, Harden RM, Walton HJ (eds). *Newer Developments in Assessing Clinical Competence.* Quebec, Canada: Can-Heal Publications, 1986:229-37.
  38. Stimmel B, Colliver JA, Cohen DS, Smith L, Swartz L, Swartz MH. Using standardized patients for teaching and assessing clinical skills for encounters with patients with problems of substance abuse, pain control, and AIDS. In: Scherpbier AJJA, van der Vleuten CPM, Reithans JJ, van der Steeg AFM (eds). *Advances in Medical Education.* Dordrecht, The Netherlands: Kluwer Academic Publishers, 1998: 722-4.
  39. Jones J, Hunter D. Consensus methods for medical and health services research. *BMJ.* 1995;311:376-80.
  40. Fink A, Kosecoff J, Chassin M, Brook RH. Consensus methods: characteristics and guidelines for use. *Am J Public Health.* 1984; 74:979-83.