Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal of the Chinese Medical Association 74 (2011) 531–538

www.jcma-online.com

Original Article

Assessment of first-year post-graduate residents: Usefulness of multiple tools

Ying-Ying Yang [a], Fa-Yauh Lee [b,*], Hui-Chi Hsu [a], Chin-Chou Huang [c], Jaw-Wen Chen [c], Hao-Min Cheng [c], Wen-Shin Lee [a], Chiao-Lin Chuang [a], Ching-Chih Chang [a], Chia-Chang Huang [c]

[a] *Division of General Medicine, Taipei Veterans General Hospital, Taipei, Taiwan, ROC*
[b] *Department of Medicine, Taipei Veterans General Hospital, Taipei, Taiwan, ROC*
[c] *Department of Medical Research and Education, Taipei, Veterans General Hospital, Taipei, Taiwan, ROC*

Received March 17, 2011; accepted June 29, 2011

Abstract

Background: Objective Structural Clinical Examination (OSCE) usually needs a large number of stations with long test time, which usually exceeds the resources available in a medical center. We aimed to determine the reliability of a combination of Direct Observation of Procedural Skills (DOPS), Internal Medicine in-Training Examination (IM-ITE®) and OSCE, and to verify the correlation between the small-scale OSCE+DOPS+IM-ITE®-composited scores and 360-degree evaluation scores of first year post-graduate (PGY_1) residents.
Methods: Between 2007 January to 2010 January, two hundred and nine internal medicine PGY1 residents completed DOPS, IM-ITE® and small-scale OSCE at our hospital. Faculty members completed 12-item 360-degree evaluation for each of the PGY_1 residents regularly.
Results: The small-scale OSCE scores correlated well with the 360-degree evaluation scores ($r = 0.37$, $p < 0.021$). Interestingly, the addition of DOPS scores to small-scale OSCE scores [small-scale OSCE+DOPS-composited scores] increased it's correlation with 360-degree evaluation scores of PGY_1 residents ($r = 0.72$, $p < 0.036$). Further, combination of IM-ITE® score with small-scale OSCE+DOPS scores [small-scale OSCE+DOPS+IM-ITE®-composited scores] markedly enhanced their correlation with 360-degree evaluation scores ($r = 0.85$, $p < 0.016$).
Conclusion: The strong correlations between 360-degree evaluation and small-scale OSCE+DOPS+IM-ITE®-composited scores suggested that both methods were measuring the same quality. Our results showed that the small-scale OSCE, when associated with both the DOPS and IM-ITE®, could be an important assessment method for PGY_1 residents.

Keywords: assessment; direct observation of procedural skills; first year post-graduate resident; Internal Medicine in-Training Examination; medical school; medical students; Objective Structural Clinical Examination; test

1. Introduction

The outbreak of the severe acute respiratory syndrome (SARS) epidemic that occurred during 2003 exposed serious deficiencies in Taiwan's medical care and public healthcare systems, as well as its medical education system. The Department of Health, Executive Yuan of Taiwan, R.O.C., has had no efforts in promoting its "Project of Reforming Taiwan's Medical Care and Public Healthcare System" since the spread of SARS was controlled. The reform of the medical care system aims to provide holistic medical treatment to people. Its strategies and methods include strengthening the improvement of resident education and quality of medical care. A project titled "Post-graduate General Medical Training Program" was announced by the Department of Health in August 2003. The evaluation of internal medicine first-year post-graduate (PGY_1) residents usually consists of the Objective Structured Clinical Examination (OSCE) because it combines reliability with validity by using multiple testing in a standardized set of appropriate clinical scenarios in a practical and efficient

* Corresponding author. Dr. Fa-Yauh Lee, Department of Medicine, Taipei Veterans General Hospital, 201, Section 2, Shih-Pai Road, Taipei 112, Taiwan, ROC.
E-mail address: fylee@vghtpe.gov.tw (F.-Y. Lee).

1726-4901/$ - see front matter
doi:10.1016/j.jcma.2011.10.002

format.[1] The multiple-choice Internal Medicine in Training Examination (IM-ITE®) is a written test that is believed to be an alternative to performance testing such as the test by OSCE.[2,3] The reliability of the IM-ITE® is known to be good, with less testing time required.[2] The IM-ITE®, covering knowledge in physical examination, laboratory, technical, and communication skills, is relatively cheap and easier to administer compared with an OSCE.[4,5] However, a paper-and-pencil knowledge test will overemphasize the cognitive aspects of clinical skills if the test does not require a resident to actually demonstrate these skills. Direct observation of procedural skills (DOPS) involves direct observation of a resident performing a variety of technical skills.[6] A combination of the OSCE with the IM-ITE® and DOPS could bypass individual undesirable effects of each method and increase the completeness of assessment.[5,7,8]

High-stakes, large scale-OSCEs are used to assess clinical competence at the performance level of a "show how" method based on Miller's competency pyramid.[9] The format of the OSCE is designed with a circuit of multiple stations in which the candidates accomplish specific tasks within a required time period.[9–11] Replacing some OSCE stations with a written test might save resources and increase overall test reliability.[4] It could offer an adequate compromise between the demands of reliability and feasibility. In post-graduate curriculum, designing a mixed-method assessment is often advised.[12] Additionally, different content, multiple assessors, and a sufficient assessment time seem to be the fundamentals of a reliable assessment in clinical rotations. The 360-degree evaluation (multisource feedback) assesses general aspects of competence, including communication skills, clinical abilities, medical knowledge, technical skills, and teaching abilities of PGY_1 residents.[13] In general, different evaluation tools, including high-stakes, large-scale OSCE, DOPS, IM-ITE®, and 360-degree evaluations have their own particular roles in the assessment of learning outcomes. Thus, the purpose of our study was to determine the reliability of using a small-scale OSCE combined with other tools (DOPS and IM-ITE®) or a 360-degree evaluation to thoroughly evaluate PGY_1 residents.

2. Methods

2.1. Study population

Between 2007 and 2010, 209 PGY_1 residents (trainees) were evaluated by a small-scale OSCE before and after finishing 3 months of PGY_1 internal medicine residency courses of Taipei Veteran General Hospital at Taiwan (Taipei VGH). Taipei VGH is a regional medical center that provides primary and tertiary care to active-duty and retired military members and their dependents. Taipei VGH serves as the primary teaching hospital for its internal medicine residency program. All the raters and senior physicians were recruited from among the clinical faculty and were teachers for the Department of Internal Medicine. The well-trained, non-physician experts for DOPS were independent from the Department of Internal Medicine of Taipei VGH.

2.2. Study setting

The content of the small-scale OSCE, DOPS, IM-ITE®, and 360-degree evaluation were designed by a committee of expert physicians from our system who created the content blueprint and wrote the test questions according to well-established principles of examination construction. The committee members were regularly rotated.

2.3. Small-scale OSCE

The small-scale OSCE consisted of six 15-minute stations. The OSCE consisted of six clinical problems that were made up of six core competencies defined by the Accreditation Council for Graduate Medical Education [ACGME (Appendices 1 and 2)]. The content of each clinical problem consisted of physical examination skills, interpersonal skills, technical skills, problem-solving abilities, decision-making abilities, and patient treatment skills.[14] The examination took place simultaneously at two different sites. At each site, there were two sessions, and the raters at each station changed between the two sessions. Thus, for each station, there were a total of four different raters during the test day. The small-scale OSCE had neither written a component nor a technical skills station, but it was entirely performance-based. At some stations, standardized patients were used to mimic the clinical problems of actual patients. A faculty rater graded each PGY_1 resident according to a given set of 10–12 predetermined items presented in the form of a checklist. The score of checklists included items with a dichotomous scoring, yes/no, and an overall trichotomous scoring of pass/borderline/fail. All faculty raters attended serial training sessions that included extensive instruction on how to use the checklist in practice rating sessions. At each OSCE station, the raters acted as passive evaluators and were instructed not to guide or prompt the PGY_1 residents. The summary score of each station was the sum of all the checklist items. The residents' performance score for each OSCE station was obtained by calculating the percentage of checklist items he or she obtained. The OSCE was performed before the training ($OSCE_{before}$) and at the end of 3 months of training program ($OSCE_{3rd\ month}$). Finally, average OSCE scores were calculated by averaging the three monthly scores for further analysis.

2.4. DOPS

All PGY_1 performed a series of standardized technical skills. For each skill, PGY_1 residents were examined by the direct observation of experts and senior physicians using the technical skill-specific checklist.[15] Four technical skills, including advanced cardiac life support (ACLS), lumbar puncture, central venous catheter insertion, and endotracheal tube insertion, were assessed regularly. Experts and senior physicians were provided with an identical checklist for the four technical skills before the test day and were asked to familiarize themselves with the checklist. In addition, they received a 30-minute orientation session just before

examination. The DOPS checklist included items on communication skills, technical performance, and some theoretical questions, including knowledge of the indications, contraindications, potential complications, and different routes for the procedure that related to the task.[16] All of these items were developed from the 11 domains of the DOPS in presented Appendix 3. Finally, the DOPS scores of each PGY_1 resident were the averages of the ratings from the four experts and senior physicians for the four technical skills.

2.5. Monthly 360-degree evaluation

The 360-degree evaluations were made during the interval between the administration of the small-scale OSCE and DOPS. The 360-degree evaluation assessed general aspects of competence, including communication skills, clinical abilities, medical knowledge, technical skills, and teaching abilities that are shown in Appendices 4 and 5. The Spearman-Brown prophecy formula was used to calculate the number of individuals needed to obtain a reliable rating.[13,16,17] Our preliminary study found that the number of raters to achieve a reliability of 0.7 was 4. Five additional raters were needed to achieve a reliability of 0.8. Accordingly, the results of five raters of 360-degree evaluations were included for final analysis.

The 12-item, one-page 360-degree evaluation forms were made by the faculty members, including one chief resident, one visiting physician, one chief physician, one nurse, and one head nurse of each of the services that residents rotated through monthly. In other words, every PGY_1 resident received five evaluations by the five raters. The monthly 360-degree score was the average of scores from the five raters. Finally, the average 360-degree evaluation scores was calculated by averaging the three monthly scores (360-degree $\text{evaluation}_{1\text{st month}}$ 360-degree $\text{evaluation}_{2\text{nd month}}$ 360-degree $\text{evaluation}_{3\text{rd month}}$) for further analysis.

2.6. IM-ITE®

The IM-ITE® is designed by the American College of Physicians (ACP) to give residents an opportunity for self-assessment, to give program directors the opportunity to evaluate their programs, and to identify areas in which residents need extra assistance.[2,18] Our multiple-choice IM-ITE® is a modification of the ACP's IM-ITE®. Our IM-ITE® was developed to test required knowledge that PGY_1 residents most frequently encounter during their in-patient rotation. Initially, our IM-ITE® was composed of 80 items. After a first validation of the tool, 50 items were chosen based on experts' and residents' comments and validated again with a group of experts who confirmed the quality of the selected 50 items for our assessment purposes.

2.7. Certification system

At the end of the course, all PGY_1 residents were instructed to complete the DOPS and IM-ITE® as if they were the regular tests, even though the DOPS and IM-ITE® scores had no influence on pass/fail decisions of the OSCE. Additionally, the 12-item, 360-degree evaluation was completed for each PGY_1 for each month. Our research used the averaged 360-degree evaluations, DOPS, IM-ITE®, and averaged small-scale OSCE scores, which had been collected as part of the routine procedure of the Department of Internal Medicine of Taipei VGH.

For the trainees who failed the DOPS and small-scale OSCE, special programs were designed according to their defects by senior physicians. Then, these trainees were re-evaluated until they passed all these tests. For those who failed the IM-ITE® and 360-degree evaluation, special training classes were conducted to re-educate them, program directors monitored their performance in the following 3-year residency (e.g., internal medicine, family medicine, surgery, pediatrics, dermatology, ophthalmology) course.

2.8. Statistic analysis of data

To ensure equal weighting of all evaluations formats, which was needed for further analysis, the scores of separate/averaged 360-degree evaluation, DOPS, IM-ITE®, and separate/averaged small-scale OSCE were transformed onto a similar 100% scale. The borderline group method was used to set the standard of "pass" for 360-degree evaluation, DOPS, IM-ITE®, and small-scale OSCE scores. Each station's "pass" score was the mean of the scores of PGY_1 residents whose scores were rated "borderline."[19,20] To estimate the reliability of the 360-degree evaluation, DOPS, IM-ITE®, and the small-scale OSCE separately, Cronback's alpha (α) coefficient were calculated for each evaluation. Kappa statistics were used to check the inter-rater agreement between expert and senior physician for the four procedure stations of DOPS. An α of < 0.05 was accepted as statistically significant.

The descriptive statistics of the mean scores and standard deviations for each examination tool were analyzed with one sample or two-sample student's t test or analysis of variance when appropriate. Additionally, the correlations between the average OSCE and 360-degree evaluation score, small-scale OSCE + DOPS-composited score, and average 360-degree evaluation score, small-scale OSCE + DOPS + IM-ITE® score and average 360-degree evaluation score were analyzed by Pearson's correlation methods (Version 10.1, SPSS Inc., Chicago, Ill., USA). Comparisons between two correlation coefficients from paired measurements were carried out using the formula created by Kleinbaum and colleagues.[21]

3. Results

Between 2007 January and 2010 January, 245 PGY_1 residents participated and underwent 24 administrations of the OSCE (every 3 months, two OSCE for each PGY_1 resident), 18 administrations of DOPS (every 2 months), 12 administrations of IM-ITE® (every 3 months) and 750 administrations of 360-degree evaluation (every 1 month, three 360-degree evaluations for each PGY_1 resident) in our system. Our study involved 99 specialties and subspecialties in total.

Finally, the complete data of 209 trainees were included for analysis. Thus, the data completeness rate was 85.3%.

3.1. Reliability

Our study included the analysis of the internal reliability of all our evaluation methods. The results showed that the reliability of the different evaluative methods was varied. The before-training OSCE ($OSCE_{before}$) had a reliability of 0.73, the after-training OSCE ($OSCE_{3rd\ month}$) 0.662, DOPS 0.82, IM-ITE® 0.69, 360-degree $evaluation_{1st\ month}$ 0.89, 360-degree $evaluation_{2nd\ month}$ 0.9, and 360-degree $evaluation_{3rd\ month}$ 0.79 (Table 1). Additionally, the inter-rater reliabilities between the expert and senior physicians for DOPS were good (ACLS: Kappa 0.71; lumbar puncture: Kappa 0.69, central venous catheter insertion: Kappa 0.75 and endotracheal tube insertion: Kappa 0.78).

3.2. Consistency of evaluations

Before further correlation studies, the re-evaluation reliability of the small-scale OSCE and 360-degree evaluation were assessed. As seen in Fig. 1, $OSCE_{before}$ and $OSCE_{3rd\ month}$ scores were closely correlated ($r = 0.64$, $p < 0.01$). Meanwhile, the 360-degree $evaluation_{1st\ month}$, 360-degree $evaluation_{2nd\ month}$ and 360-degree $evaluation_{3rd\ month}$ scores were well correlated, ranging from a low correlation of 0.54 between 360-degree $evaluation_{1st\ month}$ and 360-degree $evaluation_{2nd\ month}$ scores and a high correlation of 0.94 between 360-degree $evaluation_{2nd\ month}$ and 360-degree $evaluation_{3rd\ month}$ scores.

3.3. Correlations

Table 2[19] shows that average small-scale OSCE scores was significantly correlated with average 360-degree evaluation scores ($r = 0.37$, $p < 0.05$). Interestingly, the addition of DOPS scores to average small-scale OSCE scores significantly increased its (small scale-OSCE + DOPS-composited score) correlation with the average 360-degree evaluation scores ($r = 0.72$, $p < 0.01$). Furthermore, a combination of the IM-ITE® scores with small-scale OSCE + DOPS scores (small scale-OSCE + DOPS + IM-ITE® scores) markedly enhanced their correlation with 360-degree evaluation scores ($r = 0.83$, $p < 0.01$).

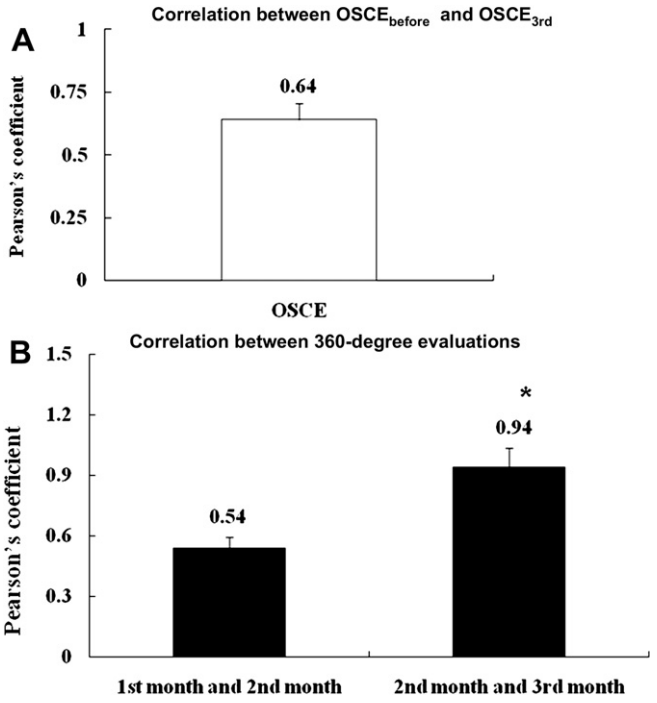

Fig. 1. Correlation between (A) $OSCE_{before}$ and $OSCE_{3rd\ month}$; (B) monthly 360-degree evaluations. * $p < 0.05$ vs. correlation coefficients of 360-degree $evaluation_{1st\ month}$ and 360-degree $evaluation_{2nd\ month}$.

3.4. Difficulty and efficiency of training

Next, we searched for the points that needed to be further improved in the design of the training program. The pass rates and the mean scores were significantly improved after 3 months of internal medicine training course [$OSCE_{before}$: 36% and $OSCE_{3rd\ month}$: 52%, $p < 0.05$ (Fig. 2 and Table 3)]. The pass rate of the DOPS scores was around 70%. Meanwhile, the pass rate of the 360-degree evaluation scores was also progressively improved among three months of internal

Table 1
Various scores of all PGY_1 residents ($n = 209$).

360-degree evaluation scores		Small-scale OSCE scores		DOPS scores	IM-ITE® scores
1st month	83.5 ± 16	Before	74.7 ± 24.1	83.7 ± 27	88.6 ± 31
2nd month	87.3 ± 21*	3rd month	84.6 ± 19.3*		
3rd month	90.2 ± 17*				
Average	86.9 ± 24	Average	79.4 ± 21.1		
	Small-scale OSCE + DOPS-composited scores				
	81.3 ± 21				
	Small-scale OSCE + DOPS + IM-ITE®-composited scores				
	85.9 ± 8				

Data were expressed as mean ± SD.
DOPS = direct observation of procedural skills; IM-ITE = Internal Medicine in Training Examination (IM-ITE®); OSCE = Objective Structural Clinical Examination.
* $p < 0.05$ vs basal level.

Table 2
Correlations between evaluative measures.

Evaluation methods	Pearson's coefficient
Average small-scale OSCE score and 360-degree evaluation scores	0.37
Average small-scale OSCE + DOPS-composited score and 360-degree evaluation scores	0.72*
Average small-scale OSCE + DOPS + IM-ITE®–composited score and 360-degree evaluation scores	0.85*

DOPS = direct observation of procedural skills; IM-ITE = Internal Medicine in Training Examination (IM-ITE®); OSCE = Objective Structural Clinical Examination.
* $p < 0.05$ vs Pearson's coefficient of small-scale OSCE score and 360-degree evaluation scores; correlation coefficients were compared using the Kleinbaum formula.[19]

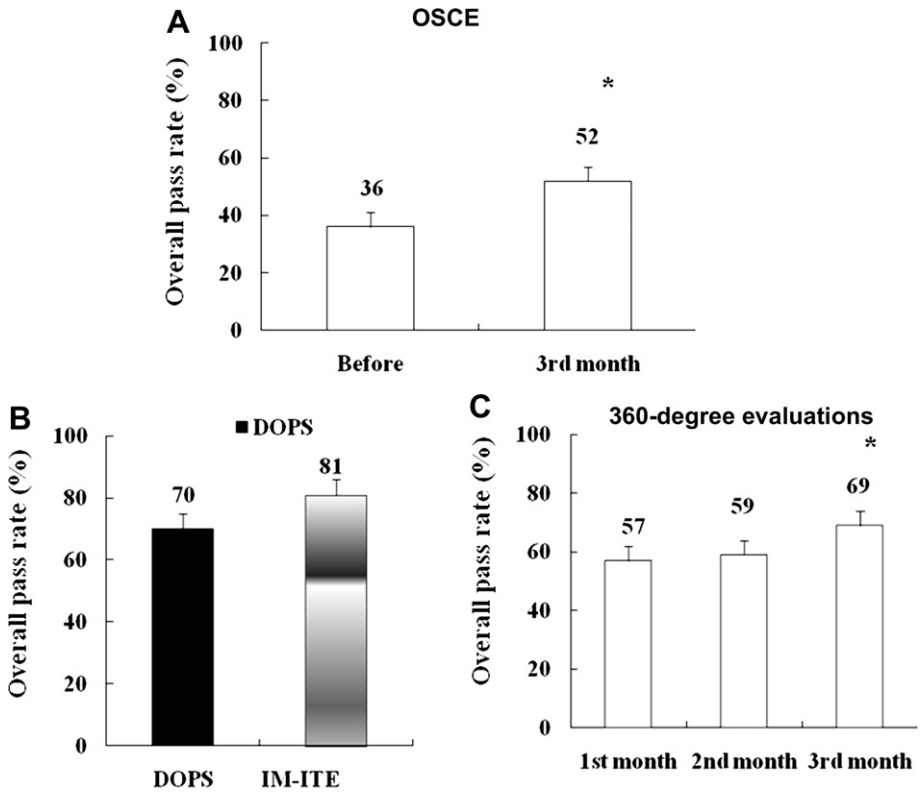

Fig. 2. The overall pass rate (pass students/total students*100%) of (A) OSCE; (B) DOPS and IM-ITE®; (C) 360-degree evaluation of all PGY_1 residents. *$p < 0.05$ vs. $OSCE_{before}$ and 360-degree $evaluation_{1st\ month}$. DOPS = direct observation of procedure skills; IM-ITE = Internal Medicine in Training Examination (IM-ITE®); OSCE = Objective Structural Clinical Examination.

medicine training program (360-degree $evaluation_{1st\ month}$: 57% 360-degree $evaluation_{2nd\ month}$: 59% and 360-degree $evaluation_{3rd\ month}$: 69%, $p < 0.05$). Although the overall pass rates varied between different evaluative methods, the differences did not reach significance level.

4. Discussion

The objective of medical education is to produce excellent medical professionals and performance. To achieve this objective, Taipei VGH introduced and implemented the small-scale OSCE, DOPS, IM-ITE®, and 360-degree evaluations. Previous study suggested that the term "competence" is often used broadly to incorporate the domains of knowledge, skills, and attitudes.[1] No single assessment method can successfully evaluate the clinical competence of PGY_1 residents in internal medicine. It has been reported that the reliability of medical education performance increases with the addition of each different reliable measure.[22] Thus, educators need to be cognizant of the most appropriate application tool. Our study explored whether a combination of assessment tools provides the best opportunity to evaluate and educate PGY_1 resident in Taiwan.

Table 3
Reliability of various methods.

Evaluation methods	Reliability (Cronbach's alpha coefficient)	
Small scale-OSCE	Before	0.73
	3rd month	0.662
DOPS		0.82
IM-ITE®		0.69
360-degree evaluation	1st month	0.89
	2nd month	0.9
	3rd month	0.79

DOPS = direct observation of procedural skills; IM-ITE = Internal Medicine in Training Examination (IM-ITE®); OSCE = Objective Structural Clinical Examination.

It is not clear whether lengthening the written test component (such as IM-ITE®) compensates for the loss of validity due to the use of fewer stations in the OSCE.[4] Nonetheless, the reliability of the OSCE is partly determined by the testing time, and a large-scale OSCE is time- and money- consuming. Accordingly, an expensive large-scale OSCE should still be part of the assessment program.

The 360-degree evaluation have been widely used in several medical and surgical residency training programs, and their usefulness has been very positive.[13,16] Our study observed the increase in rating scores with more months of training (Table 3), which supports the general validity of the 360-degree evaluation in assessing PGY_1 resident competence including knowledge, skills, and attitudes.[12,16,23] For formative purposes, the 360-degree evaluation helps a resident understand how other members of their team view his or her knowledge and attitudes. Thus, the 360-degree evaluation scores also help residents develop an action plan and improve their behavior as part of their training. In our study, we used 360-degree evaluation scores as a standard to assess the efficiency of different

methods, or a combination of them in evaluating the performance of all PGY_1 residents.

Nevertheless, the reliability of 360-degree evaluation in our study was different between the 3 months of the training program. This finding can be explained by the fact that the residents are not working in a stable environment. They change rotation frequently, and there are new raters at the new sites. It is also possible that PGY_1 residents became less homogenous in their abilities during the 3 months of the training program. In fact, the 360-degree evaluation is a method that only provides global rating regarding of the PGY_1 residents' performance; it will not demonstrate the details. In other words, the 360-degree evaluation is a tool for assessing the change of knowledge, skills, and attitude rather than physical examination skills. Actually, a complete evaluation of the PGY_1 performance should include a 360-degree evaluation and an OSCE focusing on physical examination skills.

The reliabilities of the DOPS, IM-ITE®, and 360-degree evaluation were good, indicating a high degree of internal consistency of these assessments. The pass rates of all methods were between 61% and 81% (Fig. 2). In comparison with other tools, the reliability of the small-scale OSCE was not acceptable. Meanwhile, the pass rate was not very high for the OSCE of our study. These results indicate that the structure of the small-scale OSCE used in our study should modify to improve the pass rate in the future. Nevertheless, average small scale-OSCE and 360-degree evaluation scores were still significantly well correlated ($r = 0.37$, $p < 0.05$), suggesting a high reliability of the overall program.[1] Further, we combined the small-scale OSCE with other tools to improve its reliability and reflect the real performance of PGY_1 residents as seen in Table 2. Notably, the correlation between small-scale OSCE + DOPS-composited scores and 360-degree evaluation scores was increased ($r = 0.72$, $p < 0.01$). Finally, a further markedly increase in the correlation between OSCE + DOPS + IM-ITE® and 360-degree evaluation scores was observed ($r = 0.85$, $p < 0.01$). These results can also be explained by the fact that small scale-OSCE, DOPS, and IM-ITE® assess different areas of knowledge and skills. Accordingly, adding all of the three scores showed a high correlation with the 360-degree evaluation because more items were being sampled.

5. Limitations

There are some limitations to our study. First, this was a retrospective study of a single residency program with a relatively small sample size. However, our results are strengthened by the completeness of our data over a 3-year period. The series, small-scale OSCE, DOPS, IM-ITE®, and 360-degree evaluations were 3 years apart in time. This is a long period in a learning environment, and many confounding variables can have an impact on the learning of PGY_1 residents. However, there is always "noise" in educational measurement, and we can postulate that the impact of these confounding variables may be found to be equally distributed among the observed scores of the four evaluations and could explain the results. Despite the noise and 3-year time interval, we still observed a relatively strong correlation among the variables under study.

Second, no long-term follow-up, small-scale OSCE, DOPS, and IM-ITE® measurements during the 3 years of the residents' training were obtained (to evaluate the validity of these tools), and we did not address the durability of the small-scale OSCE and DOPS. Nevertheless, our study showed a strong correlation between the 360-degree evaluation and small-scale OSCE + DOPS + IM-ITE® scores. Accordingly, the following of the core competencies of trainees regularly by IM-ITE® and 360-degree evaluation in our system may be valid on the program level. In OSCE, it was not possible to blind faculty raters to the PGY_1 resident's identity. Our study was included OSCE before and after 3 months of internal medicine training course. In order to avoid the bias coming from the fact that PGY_1 residents with a weaker OSCE performance might have tended to prepare more diligently for their next post-course OSCE, the raters of small-scale OSCE in our study did not give any feedback to PGY_1 residents before they completed the post-course OSCE. Meanwhile, the trainees knew their $OSCE_{before}$ and $OSCE_{3month}$ scores only after finishing the entire testing sequence.

Third, only four practical consideration stations were included in the DOPS of our study. Previous study had suggested that if the DOPS were to be used for certification, a greater number of skills stations should be included where the consequences of an erroneous pass/fail judgment were serious.[21] Nonetheless, we arranged two raters (both an expert and a senior physician) to increase the reliability by the multisource evaluation. Notably, the inter-rater agreements were quite good for the four technical skills of DOPS in our study. Use of the experts for the DOPS evaluation can also avoid the "halo effect" due to having previous experience with the PGY_1 resident, which could introduce positive or negative bias in scoring.

Finally, previous studies have shown that the reliability of the 360-degree evaluation can be elevated by increasing the number of raters. Our current study only did a rough estimation about the number of raters needed for the reliability of the 360-degree evaluation to reach 0.7–0.8. In fact, a detailed analysis of heterogeneity of raters and PGY_1 residents should also be considered, along with analyses by G-theory, in the future.

In conclusion, the strong correlations between the 360-degree evaluation and the small-scale OSCE + DOPS + IM-ITE® scores suggests that both methods measure the same quality. In the future, a small-scale OSCE associated with DOPS and the IM-ITE® could be an important assessment method in evaluating the performance of PGY_1 residents.

Acknowledgments

We would like to thank the case writers, the clinical faculty, standard patients students and staff of our system for their assistance and participation in implementing the small-scale OSCE, DOPS, IM-ITE®, and the 360-degree evaluation.

Appendix 1

The content of small-scale Objective Structural Clinical Examination stations of PGY_1 residents.

	January 2007–January 2008	February 2008–January 2009	February 2009–January 2010
Utilization of clinical information[a]	○	○	○
Organization and orderliness[a]	○	○	○
Patient safety and ethical issues[a]	●	●	●
Creation of therapeutic relations with patients		●	●
Providing patient-centered care	●		
Counseling and educating patients and family	●		
Decision–making (clinical judgment)		●	
Clinical differential diagnosis			●
Improvement of quality of clinical care		●	
Employing evidence-based practice	○		
Interaction with whole medical system			●

[a] Common stations across three years.

○ = the station was implemented for the year; ● = the station was implemented and standardized patients used for the year; OSCE = Objective Structural Clinical Examination.

Appendix 2

The items that included in the checklist of small-scale Objective Structural Clinical Examination.

1. **Patient care**
   Interviewing; counseling and educating patients and families; physical examination; preventive health service; informed decision–making
2. **Interpersonal and communication skills**
   Creation of therapeutic relations with patients; listening skills
3. **Professionalism**
   Respectful, altruistic; sensitive to cultural, age, gender, and disability issues
4. **Practice-based learning and improvement**
   Analyzing own practice for needed improvement; using evidence from scientific studies (EMB); use of information technology
5. **Systems-based practice**
   Understanding interaction of their practice with the larger system; advocating for patients within the health care system; knowledge of practice and delivery system
6. **Medical knowledge**
   Investigative and analytic thinking; knowledge and application of basic science

Appendix 3

The direct observation of procedural skills domains and items in the checklist.

1. Demonstrates understanding of indications, relevant anatomy, technique of procedure
2. Obtains informed consent
3. Demonstrates appropriate preparation pre-procedure
4. Demonstrates situational awareness
5. Aseptic technique
6. Technical ability
7. Seeks help where appropriate
8. Post-procedure management
9. Communication skills
10. Consideration of patient
11. Overall ability to perform procedure

Appendix 4

Description for each item of 360-degree evaluation.

Item in the checklist	Description
1. Caring behaviors	Demonstrates caring and respectful behavior with patients and families
2. Effective questioning and listening	Elicits information using effective questioning and listening skills
3. Effective counseling	Effectively counsels patients, families, and/ or care gives
4. Demonstrates ethical behavior	Demonstrates ethical behavior
5. Advocates for quality	Advocates for quality patient care, assists patient in dealing with system complexities
6. Sensitive to age, culture, gender, and/or disability	Sensitive to age, culture, gender, and/or disability
7. Communicates well with staff	Communicates well with staff
8. Works effectively as team member/leader	Works effectively as member/leader of teams, understands how own actions affect others
9. Works to improve system of care	Works to improve system of care
10. Participates in therapies and patient education	Participates in rehabilitation therapies, intervention and patient education
11. Committed to self-assessment/ Uses	Committed to self-assessment; uses feedback for self-improvement
12. Teaches effectively	Teaches students and professionals effectively

Appendix 5

Relationship of 12 items on 360-degree evaluation to the Accreditation Council for Graduate Medical Education core competencies.

Items on checklist	ACGME core competency					
	Patient care	Medical knowledge	Problem-based learning and improvement	Interpersonal and communication skills	Professionalism	System-based practice
1. Caring behaviors	X			X		
2. Effective questioning and listening			X		X	
3. Effective counseling					X	X
4. Demonstrates ethical behavior		X			X	
5. Sensitive to age, culture, gender, and/or disability	X		X			
6. Communicates well with staff			X	X		
7. Works effectively as team member and leader	X			X		
8. Works to improve system of care	X				X	
9. Participates in therapies and patient education		X				X
10. Advocates for quality					X	X
11. Committed to self-assessment and uses feedback			X			X
12. Teaches effectively			X	X	X	

References

1. Holmboe ES, Hawkins RE. Methods for evaluating the clinical competence of residents in internal medicine: a review. *Ann Intern Med* 1998;**129**:43–8.
2. Garibaldi RA, Subhiyah R, Moore ME, Waxman H. The in-training examination in internal medicine: an analysis of resident performance over time. *Annal Int Med* 2002;**137**:505–10.
3. Vleuten CPM, van Luijk SJ, Beckers HJM. A written test as an alternative to performance testing. *Med Educ* 1989;**23**:97–107.
4. Verhoeven BH, Hamers JGHC, Scherpbier AJJA, Hoogenboom RJI, Vamder Vleuten CPM. The effect on reliability of adding a separate written assessment component to an objective structured clinical examination. *Med Educ* 2000;**34**:525–9.
5. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;**65**:S63–7.
6. Durning SJ, Cation LJ, Jackson JL. Are commonly used resident measurements associated with procedural skills in internal medicine residency training? *J Gen Intern Med* 2007;**22**:357–61.
7. Haber RJ, Avins AL. Do ratings on the American Board of Internal Medicine resident evaluation form detect differences in clinical competence? *J Gen Intern Med* 1994;**9**:140–5.
8. Schwartz RW, Witzke DB, Donnelly MB, Stratton T, Blue AV, Sloan DV. Assessing residents' clinical performance: cumulative results of a four-year study with the Objective Structured Clinical Examination. *Surgery* 1998;**124**:307–12.
9. Chesser AMS, Laing MR, Miedzybrodzka ZH, Brittenden J, Heys SD. Factor analysis can be a useful standard-setting tool in a high-stakes OSCE assessment. *Med Educ* 2004;**38**:825–31.
10. Leach DC. The ACGME competencies: substance or form? *J Am Coll Surg* 2001;**192**:396–8.
11. Swing S. Assessing the ACGME general competencies: general considerations and assessment methods. *Acad Emerg Med* 2002;**9**:1278–88.
12. Watts J, Feldman WB. Assessment of technical skills. In: Nuefeld VR, Norman GK, editors. *Assessing Clinical Competence*. New York: Springer; 1985. p. 259–74.
13. Rodgers KG, Manifold C. 360-degree feedback: possibilities for assessment of the ACGME core competencies for emergency medicine residents. *Acad Emerg Med* 2002;**9**:1300–4.
14. Huang CC, Chan CY, Wu CL, Chen YL, Yang HW, Huang CC, et al. Assessment of clinical competence of medical students using the objective structured clinical examination: first 2 years experience in Taipei Veterans General Hospital. *J Chin Med Assoc* 2010;**73**:589–95.
15. Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;**84**:273–8.
16. Higgins RS, Bridges J, Burke JM, O'Donnell MA, Cohen NM, Wilkes SB. Implementing the ACGME general competencies in a cardiothoracic surgery residency program using 360-degree feedback. *Ann Thorac Surg* 2004;**77**:12–7.
17. Massagli TL, Carline JD. Reliability of a 360-degree evaluation to assess resident competence. *Am J Phys Med Rehab* 2007;**86**:845–52.
18. Carr S. The foundation programme assessment tools: an opportunity to enhance feedback to trainees? *Postgrad Med J* 2006;**82**:576–9.
19. Boursuicot KAM, Roberts TE, Peel G. Using borderline methods to compare passing standards for OSCE at graduation across three medical schools. *Med Educ* 2007;**41**:1024–31.
20. Littlefield J, Pankert J, Schoolfield J. Quantity assurance data for residents' global performance ratings. *Acad Med* 2001;**76**:S102–4.
21. Kleinbaum DG, Kupper LI, Muller KE, Nizam A. *Applied regression analysis and multivariable methods*. 3rd ed. New York, NY: Duxburry Press; 1997.
22. Norcini JJ, Mckinley DW. Assessment methods in medical education. *Teach Teacher Educ* 2007;**23**:239–50.
23. Reznnick K, Regehr G, Mac Rae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg* 1997;**173**:226–30.