Validity of In-Training Evaluation Report (ITER) in Saudi Pediatric Residency Training. Single center experience

Taghreed Shuaib[2]

Corresponding author: Dr Taghreed Shuaib tshuaib@kau.edu.sa
Institution: 2. King Abdulaziz University
Categories: Assessment, Postgraduate (including Speciality Training)

Received: 10/01/2018
Published: 22/01/2018

Abstract

Objective: The study aimed to examine validity evidence of in-training evaluation report (ITER) at King Abdulaziz University Hospital (KAUH)'s pediatric residency program. Predictive validity of ITERs, to assess learner acquisitions of required competencies were studied. We used ITER scores to identify trainees with unprofessional conduct.

Methods: One-year data for 35 residents in pediatric residency training program at KAUH were reviewed. Data were extracted from 360 ITERs completed by faculty members and specialists. ITERs' reliability was evaluated using Cronbach's alpha. ITERs scores were correlated with final Objective Structured Clinical Examinations (OSCE) and written exams scores. Professionalism composite scores from ITERs were compared with scores of trainees reported for unprofessional attitude.

Results: Specialists had high variability in scoring competencies with average standard deviations (SD) of 0.85 compared with 0.15 by faculty. Reliability of ITER scores using Cronbach's alpha was 0.895. Good correlation was observed between competencies ranging between 0.86 to 0.29. There was fair correlation between ITER composite scores and those of final OSCE and written exams. There was no correlation between professionalism scores and residents reported for unprofessional conduct.

Conclusion: ITER scores had good validity and reliability despite variability observed between faculty and specialist when scoring competencies. ITERs failed to detect trainees with unprofessional conduct.

Keywords: In-training evaluation report, validity, pediatric residency
Introduction

The Statute of the Saudi Commission for Health Specialties (SCFHS) was approved in 1992 with the aim of improving the health care system in the Kingdom of Saudi Arabia. Through its supervisory, executive and specialist boards and committees, SCFHS is responsible for setting standards for medical practice, and developing, supervising and evaluating postgraduate training programs which to-date have exceeded 37 programs (Statute of the Saudi Commission for Health Specialties). The pediatric residency training program was one of the first programs to be initiated more than 20 years ago, and has been one that has continued to developed over the past years. In 2014, major changes were implemented to the pediatric residency training program by the SCFHS to meet the evolving needs of the community, developments in medical education, and to follow international medical standards. The SCFHS adopted the Royal College of Physicians and Surgeons of Canada (CanMED) competency framework in designing new curricula including new assessment tools to ensure effectiveness of the new system (Pediatric Saudi Board training booklet 2014). This was considered an important move forward in the postgraduate medical education system in the country.

The Saudi Pediatric Residency Training is a four-year program designed to allow trainees to have broad exposure and experience in a wide range of medical pediatric conditions by rotating through different domains and specialties within pediatrics. Several changes were made to the old training program after adopting CanMED including increasing the number of training blocks from 11 to 12 blocks annually with each lasting 28 days. This aimed at enhancing the trainees’ acquirement of the required competencies as per the CanMED framework. Assessments are guided by the rotation-specific objectives written in the SCFHS’s training manual which is available for both trainee and trainer (Pediatric Saudi Board training booklet 2014). Ongoing assessments used to be done using the In-training evaluation report (ITER) only, then two additional tools, the Mini-CEX and 360ᴼ evaluation or multi-source feedback, were added. The majority of trainees’ assessments are summative with 35% of the residents’ total assessment scores being based on the ITER. Trainees’ performance outcome at the end of the year depend on their results in the final written and clinical examinations which are conducted in the form of multiple-choice questions (MCQs) and Objective Structured Clinical Examination (OSCE) respectively. The training program is internally and externally audited every 2-4 years to ensure that objectives and rotation structure are met.

"Competence is developmental" (Leach 2002) and the results of learning eventually become a habit. Assessment of competencies reflects the trainee’s performance and his/her ability to achieve progress and gain new knowledge and skills (Epstein, 2007). Furthermore, assessment guides physicians’ learning, sheds light on unprofessional physician conduct which might potentially endanger patients’ lives, and also aids in identifying physicians for further training programs (Epstein, 2007). There are various assessment tools in medical education with each measuring different aspects of learning. The choice among different tools depends on what is being measured and the validity and reliability of the measuring tool in measuring it. Newble et al. (1994) published “guidelines for assessing clinical competence” in which they described that the cardinal features of any assessment tool as being reliable, valid, efficient, accessible, and able to guide future learning (Newble et al, 1994).

ITER, used for end-of-rotation evaluation, is a tool for assessing several essential competencies which should be achieved by training physicians (Hatala & Norma, 1999). As per the CanMED competency framework, ITER assesses seven physician "roles" (“Medical Expert, Communicator, Collaborator, Manager, Health Advocate, Scholar, and Professional”) (Frank & Danoff, 2007). It is based on observations on the trainee over a period of time by supervising clinicians or faculty members in the clinical settings (Jackson, Kay & Frank, 2015). ITER subjectivity can be overcome by increasing the number of assessments and having multiple raters (Jackson, Kay & Frank, 2015; Park, Riddle & Tekian, 2014). ITER scores and scores’ interpretation need to be supported by validity.
evidence to be interpreted meaningfully (Dowing, 2003; Kane, 2013). Despite the fact that there is insufficient data in the literature concerning the validity and reliability of ITER as an assessment tool, it is still widely-used (Chaudhry, Holmboe & Beasley, 2008). Park et al. (2014) considered ITER a useful tool in particular for measuring a trainee's progress over the years with regards to knowledge more than professionalism with an annual increment in scores of 0.28 and 0.12 points per postgraduate year respectively (Park, Riddle & Tekian, 2014). A retrospective study by Jackson et al. (2015) looking at validity of ITER in 228 internal medicine residents revealed poor construct and predictive validity (Jackson, Kay & Frank, 2015). On the other hand, another study by Kassam et al. (2014) found that ITER used in assessing medical emergency residents revealed strong reliability and evidence of construct validity (Kassam, Donnon & Rigby, 2014) while Ginsburg et al. (2013) found that written comments by raters in ITER showed better correlation with trainee's performance than the scores did (Ginsburg, Eva & Regehr, 2013). Professionalism, which includes a list of attributes and behaviors, is a core competency of physicians and is expected by patients and medical societies (Mueller, 2015). Furthermore, since professionalism is associated with better clinical outcomes, medical trainees should be taught and assessed for professionalism (Mueller, 2015; Pauls, 2012). There are several elements or sub-competencies under professionalism (e.g. communication, ethics, accountability). Assessment of professionalism with its sub-competencies can't be achieved via a single tool and several tools are required to do this (Mueller, 2015). A survey by Pauls (2012) on the assessment of professionalism, reveals that the majority of assessors used ITER as a tool. The top barriers for assessing professionalism identified in the survey were faculty awareness and availability, and the statement of clear objectives for the trainees. Unprofessional attitude among trainees is expected and therefore it is important to identify trainees with such an attitude during their training period for appropriate guidance.

In this study we evaluated the validity evidence of ITER in the pediatric residency training program at the King Abdulaziz University Hospital (KAUH) using Messicks's framework. Moreover, we examined the predictive validity of ITER scores to identify trainees with unprofessional conduct. In the KAUH pediatric residency training program, the professionalism of trainees is monitored by the program's director who receives complaints regarding unprofessional conduct through informal written letters or emails. Identifying trainees with unprofessional conduct was challenging as ITER was the most frequent and sometimes the only used tool. There was also a lack of clear criteria that addresses this competency and a lack of faculty awareness concerning it.

Methods

Data. Retrospective data from October 2014 through September 2015 of residents in the pediatric residency training program at KAUH were reviewed. Information was extracted from ITERs and the scores of final OSCE and written exams. The ITERs were completed by raters with performance scored on a 5-point scale (1: poor, 2: mediocre, 3: respectable, 4: good, and 5: excellent). According to the CanMED competency framework, the ITER is designed to assess seven roles. These roles are those of "Medical Expert, Communicator, Collaborator, Leader, Health Advocate, Scholar, and Professional". These roles were transformed into ten competencies in the ITER (knowledge, history taking and physical examination, data interpretation, management plan, charts completion, communication skills, attitude, managerial skills, scholar, attendance and punctuality). Three out of ten (30%) of these competencies in the ITER assess professionalism (communication skills, attitude, attendance and punctuality).

Raters completing the ITERs. Raters participated in completing the ITERs were supervising faculty members and specialists (according to our KAUH hierarchy, a specialist is a physician on a level between an attending physician and a resident). None of the raters had received formal training on how to use this assessment form, but the scoring rubric was printed on the back of the assessment form for guidance. The report was issued for each trainee at the end of their training block which lasted 28 days. The evaluation was based on observation and assignment during each
training period. Raters were instructed to discuss the evaluation with trainees and to provide feedback. The unit of analysis in our cohort was training residents. For each trainee, the average of multiple scores for each of the ten competencies per year was calculated in addition to the mean total competencies score, and composite ITERs, and the professionalism composite score. Any unprofessional conduct by trainees was reported in writing to the program director from different sources (nurses, faculty, peers, and patients).

Analysis. Descriptive statistics for faculty members and specialists were used to evaluate the response process and study data characteristics. Cronbach’s alpha was used to assess the internal consistency for reliability of assessment scores. Correlation among competencies was evaluated using the Pearson correlation coefficient. ITER composite scores were compared with OSCE and written exam scores using the \( t \)-test. For residents or trainees reported to the program director for their unprofessional attitude, their professionalism scores were compared to those of other residents using point-biserial correlation. Calculations were done using SPSS version 21.

The study was approved by the institutional review board of King Abdulaziz University Faculty of Medicine, and Department of Pediatrics on the March 26th 2017 (Reference No (134-16).

Results

In the one-year study period between October 2014 and September 2015, 35 pediatric residents were included. The number of trainees in each training level in this cohort is shown in Table 1. A total of 55 raters participated in the assessment, completing 360 evaluations. An average of 9.97 (Standard deviation [SD] 1.46) evaluations or ITERs were completed for each trainee during the study period. The Saudi Commission for Health Specialties policy requires raters to discuss the evaluations with the trainees at the end of their training period and to submit their final assessment not more than three months after the end of the training period.

Table 1 Descriptive statistic: Average score of all competencies by level of training and raters

| Level | Trainees, n | Raters per trainee | Assessments, n | Composite score Mean ± SD | Min | Max |
|-------|-------------|--------------------|----------------|---------------------------|-----|-----|
| **PGY1** | 16 | Consultant 8.62 Specialist 0.94 | 156 25 | 4.45 SD 0.15 4.73 SD 0.84 | 4.16 | 4.64 4.80 |
| **PGY2** | 6 | Consultant 7.67 Specialist 0.33 | 42 13 | 4.49 SD 0.23 4.82 SD 0.89 | 4.24 | 4.85 5.00 |
| **PGY3** | 13 | Consultant 8.85 Specialist 0.38 | 110 15 | 4.57 SD 0.14 4.80 SD 0.80 | 4.33 | 4.78 4.90 |

Response process

Out of 360 assessments (ITERs), 92% were completed by faculty members and 8% by specialists, and their mean assessments per trainee were 8.8 and 1.5 respectively. Internal-consistency reliability across items using Cronbach's Alpha was 0.712. Table 1 shows descriptive statistics of the average composite score for the ten competencies in each postgraduate year (PGY), with a mean score of 4.5 (SD=0.17). Mean composite scores given by specialists
were higher than those given by the faculty in all levels of training. Scoring competencies revealed higher variability among specialists compared with faculty with SD ranging between 0.8-0.89 and 0.14-0.23 respectively. This indicates that faculty members tend to give similar scores when rating competencies compared to specialists who had a wider range of scores in rating. Mean composite scores increased by 0.04 – 0.08 (P= 0.33 and 0.28) across PGY1 to PGY3.

**Internal structure (Reliability)**

Pearson correlation coefficient between competencies ranged between 0.86 to 0.29 (P<0.000- 0.08). Competencies reflecting professionalism (attendance, punctuality, and communication skills) had little to fair correlation among them which ranged between 0.47 to 0.19. Reliability of ITER scores using Cronbach's Alpha was 0.895.

**Relation to other variables:**

Mean ITER scores were compared with those of OSCE and end-of-year written exams and using the Pearson correlation, they showed fair correlation of 0.36 and 0.27 (P= 0.035 and 0.11) respectively. However, correlating the mean ITER for each training level revealed fair correlation with OSCE in PGY1, and in PGY3 with r= 0.28 and 0.49 respectively and a negative correlation in PGY2 with r= -0.32. The final written exam showed fair correlation in PGY1 with a score of 0.47 and no correlation at PGY2, and PGY3 with 0.15 and – 0.38 respectively. In figure 1 the mean scores for each of the ten competencies were plotted against the year of training. The “knowledge” competency had the lowest mean value in all PGYs compared to other competencies, and the lowest mean was in PGY1, with an increment of 0.26 (P= 0.12) from PGY1 to PGY2. The “managerial skills” competency increased by an increment of 0.21- 0.28 over one year (P=0.002 and 0.026). The three aspects of professionalism (attitude, communication skills, and attendance and punctuality) had the highest mean value across the three years.

**Consequences:**

Four out of our cohort of 35 residents (11.4%) were reported for unprofessional practices and were given written warning letters by the program director. Aspects of unprofessional conduct reported in these four trainees were related to attitude, attendance, and communication skills. Table 2 shows a comparison between professionalism scores in reported residents versus the other residents. Residents reported for unprofessional conduct had higher professionalism composite scores compared to other residents. Professionalism composite scores had no correlation with residents reported for unprofessional conduct (point-biserial) at 0.098 (P=0.57), and per competency it ranged between 0.14- 0.20. There was no significant difference between residents reported for unprofessional conduct and other residents and no correlation with professionalism composite score.

Table 2: Mean professionalism scores for residents with unprofessionalism conduct (n = 4) compared to the other residents (n = 31)

| Competencies               | Difference (P) | Residents with unprofessional conduct ± SD | All other Residents ± SD |
|----------------------------|----------------|-------------------------------------------|--------------------------|
| Attitude                   | 0.55           | 4.85 ± 0.13                               | 4.90 ± 0.09              |
| Communication Skills       | 0.98           | 4.65 ± 0.20                               | 4.69± 0.18               |
| Attendance and punctuality | 0.13           | 4.92 ± 0.09                               | 4.85± 0.15               |
| Professionalism composite score | 0.34         | 4.81 ± 0.09                               | 4.77 ± 0.13              |
Discussion

This the first study in the region to evaluate ITER validity using Messick’s validity evidence. The validity of ITER or end-of-rotation evaluation was evaluated by several studies, and the majority of these studies found it to be a valid tool under certain conditions (Ginsburg, Eva & Regehr, 2013; Jackson, Kay & Frank, 2015; Kassam, Donnon & Rigby, 2014; Park, Riddle & Tekian, 2014). ITER also gives additional information about trainee progress over the years with regards to knowledge but not professionalism (Park, Riddle & Tekian, 2014) and despite its limitations in this regard, it is still a popular assessment tool (Pauls, 2012).

Our data concerning the validity of ITER were comparable with what has been published to date about this subject (Ginsburg, Eva & Regehr, 2013; Jackson, Kay & Frank, 2015; Kassam, Donnon & Rigby, 2014; Park, Riddle & Tekian, 2014), although Messick’s validity evidence was not frequently used in most of these published studies. Our study found the ITER score to be valid and reliable, but it had limitations in detecting trainees with unprofessional conduct.

This study can be considered a good local and regional reference for the validity of ITER as an assessment tool for postgraduate training and also confirms the limitation of this tool in assessing professionalism. The study results revealed that the majority of the assessments were completed by faculty members, with internal-consistency reliability across items of 0.712. A higher variability in scoring competencies was observed among specialists compared with faculty members which is contrary to the findings of Park et al. (2014) (Park, Riddle & Tekian, 2014). This variability may be attributed to the fact that specialists work closer with trainees and have longer periods of observation and interaction with them than faculty members do, which gives them an advantage in assessing trainees and giving assessments. However, faculty members do have more experience than specialists, and so having trainee assessments from both of these rater groups would better reflect the actual performance of trainees.

The evidence of internal structure validity for ITER was reflected by the high ITER scores reliability of 0.895, with fair to good correlation among competencies. Plotting mean competencies scores against trainees’ PGY (figure-1) showed that the lowest mean score was given to knowledge at all PGYs, with scores increasing as training level advances. The highest scores were given to the three aspects of professionalism (attitude, communication skills, and attendance and punctuality) with no significant differences when compared at different PGYs, which was comparable with the findings of Park et al. (2014) (Park, Riddle & Tekian, 2014). This proves the validity of ITER as an assessment tool for assessing knowledge but not for professionalism. Comparing mean ITER scores with trainees’ performances at different PGY levels in the final OSCE and written exams showed fair correlation. Looking at each PGY showed that PGY1 had fair correlation with both OSCE and written exams, PGY2 had no correlation with either, and PGY3 had fair correlation only with OSCE. This indicates that ITER scores in PGY1 better predicted the performance of trainees at the end of that year compared to ITER scores in PGY2 and PGY3. This may reflect raters’ greater objectiveness in assessing junior (PGY1) trainees than senior (PGY2, PGY3) trainees, and a tendency to overestimate senior trainee performance.
Figure 1 Mean score by competency and postgraduate year (PGY). The 11 bars per PGY reflect scoring of (from left to right): Knowledge, history taking and physical examination, data interpretation, management plan, charts completion, attitude, communication skills, attendance and punctuality, managerial skills, scholar, and composite score. Bars represent 95% confidence intervals; O = mean score. Y-axis represents mean rating. X-axis represents PGY. The mean score for the three competencies reflecting professionalism are circled in all PGYs.

ITER mean professionalism scores failed to identify trainees with unprofessional conduct who comprised 11.4% of our cohort. The failure of ITER to identify unprofessional conduct can be explained by lack of awareness on how to assess professionalism using this tool. Moreover, professionalism cannot be assessed using a single tool like ITER, but rather multisource feedback is a better reflection about trainee professional conduct.

Trainee assessment using ITER in this study suggests the need for additional tools to give a clear view of trainees’ acquisitions of different competencies, especially professionalism. Raters’ understanding concerning using ITER as an assessment tool was lacking in our cohort and needs more attention when designing continuous professional development (CPD) activities for faculty. The effectiveness of these CPD activities needs to be assessed, monitored, and feedback given to faculty. Having clear objectives for assessing professionalism is important to improve assessment quality and merely stating them in the Saudi Commission for Health Specialties training manual was inadequate. These objectives need to be explained and clarified to both raters and trainees.

This study had certain limitations. It was performed over a short period of time and included a small number of trainees, which could be increased in future studies. Additionally, extending the study period would give a better assessment of trainees’ progress. This study also reflects data from a single center’s experience. However, our assessing of ITER as an assessment tool using Messick’s validity evidence shortly after the implementation in 2014
of other assessment tools (Mini-CEX and 360° evaluation or multi-source feedback) at the KAUH pediatric residency training program will serve as a good reference for future development and utilization of this tool at the national level.

**Take Home Messages**

The In-Training Evaluation Report (ITER) is a useful tool for program directors to monitor trainees’ acquisition of required competencies and to improving training. ITER, however, cannot be used to assess trainees’ professionalism where more sensitive tools like multisource feedback are a better choice. Importantly as well, for trainee assessments to be effective, assessors need to be trained in applying rating methods reliably according to clear and specific criteria.

**Notes On Contributors**

Dr. Taghreed Shuaib is an Assistant Professor and Deputy Program Director of Pediatric Postgraduate Training Program at King Abdulaziz University Hospital.

**Acknowledgements**

I'm thankful to Dr. Maha Bamehrez, MD program director of Saudi Pediatrics residency training program at King Abdulaziz University hospital for the help and support she provided.

**Bibliography/References**

Chaudhry SI, Holmboe E, Beasley BW. 2008. The state of evaluation in internal medicine residency. J Gen Intern Med. Jul;23:1010-15

Downing SM. 2003. Validity: on meaningful interpretation of assessment data. Med Educ Sep;37:830–837

Epstein RM.2007. Assessment in medical education. N Engl J Med. Jan 25;356:387-96.

[https://doi.org/10.1056/NEJMra054784](https://doi.org/10.1056/NEJMra054784)

FranK JR, Danoff D. 2007. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Med Teah. Sep;29:642-7.

[https://doi.org/10.1080/01421590701746983](https://doi.org/10.1080/01421590701746983)

Ginsburg S, Eva K, Regehr G. 2013. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITERscores and narrative comments. Acad Med. Oct;88:1539-44.

[https://doi.org/10.1097/ACM.0b013e3182a36c3d](https://doi.org/10.1097/ACM.0b013e3182a36c3d)
Hatala R, Norman, GR. 1999. In-training evaluation during an internal medicine clerkship. Acad Med. Oct;74:S118–S120

Jackson JL, Kay C, Frank M. 2015. The validity and reliability of attending evaluations of medicine residents. SAGE Open Med. [internet]. [cited 2015 June 12];3. Available from : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4679281/ doi: 10.1177/2050312115589648.

Kane MT. 2013. Validating the Interpretations and Uses of Test Scores J Educ Meas. Mar; 50: 1–73

Kassam A, Donnon T, Rigby I. 2014. Validity and reliability of an in-training evaluation report to measure the CanMEDS roles in emergency medicine residents. CJEM. Mar;16:144-50.

Leach DC. 2002. Competence is a habit. JAMA. Jan 9;287:243-4.

Mueller PS. 2015. Teaching and Assessing Professionalism in Medical Learners and Practicing Physicians. Rambam Maimonides Med J. [internet]. [cited 2015 Apr 29]; 6:e0011.

Newble D, Dawson B, Dauphinee D, Page G, Macdonald M, Swanson D, Mulhollan H, Thomson A, van der Vleuten C. 1994. Guidelines for assessing clinical competence. Teach Learn Med; 6:213-220. Available from : https://www.researchgate.net/publication/254315744.

Park YS, Riddle J, Tekian A. 2014. Validity evidence of resident competency ratings and the identification of problem residents. Med Educ. Jun;48:614-22.

Pauls MA. 2012. Teaching and evaluation of ethics and professionalism In Canadian family medicine residency programs. Can Fam Physician. Dec;58:e751-6

Statute of the Saudi Commission for Health Specialties (SCFHS). Available from: http://www.scfhs.org.sa/en/about/Pages/default.aspx
Appendices

Declarations

The author has declared that there are no conflicts of interest.

This has been published under Creative Commons "CC BY 4.0" (https://creativecommons.org/licenses/by-sa/4.0/)

AMEE MedEdPublish: rapid, post-publication, peer-reviewed papers on healthcare professions’ education. For more information please visit www.mededpublish.org or contact mededpublish@dundee.ac.uk.