Supplemental digital content for Donnon T, Al Ansari A, Al Alawi S, Violato C. The Reliability, Validity, and Feasibility of Multisource Feedback Physician Assessment: A Systematic Review. Acad Med. 2014:89(3).

## Supplemental Digital Table 1

Description of the 43 Studies on Physician Multisource Feedback (MSF) Included in a Systematic Analysis of the Literature Published 1975–January 2013

| Study (Origin) | Specialty (No. Participants) | MSF Instrument Personnel type (No. Items)* | Constructs/Factors Assessed* | General Information on Process, Administration, and/or Feasibility* |
|---|---|---|---|---|
| **Physician Assessment Review (PAR)** | | | | |
| Violato et al., 1997[7] (Canada) | Family Physicians (n = 17), Internal Medicine, and Surgery (*n* = 11) (*n* = 28, physicians) | **PAQ**<br>MC (34 items)<br>**SAQ**<br>Self (34 items)<br>**PS**<br>Pt (49 items)<br>**CAQ**<br>CW (18 items)<br>**APCQ**<br>MC (39 items)<br>**ACRPQ**<br>MC (34 items) | <br>Prof, Clin comp, Inter Per<br><br>Prof, Clin comp, Inter Per<br><br>Prof, Mngr<br><br>Prof, Inter Per, Comm<br><br>Prof, Clin comp, Inter Per<br><br>Prof, Clin comp, Inter Per | The results of this study provided evidence of reliable and validity for four of the six (PAQ, SAQ, PS, and CAQ) MSF questionnaires used to triangulate measures of professionalism, interpersonal skills, and clinical competencies between peers or medical colleagues (MC), coworkers (CW), and patients (Pt) with a physician's self (Self) assessment. A precursor to the PAR instruments, the authors concluded that the findings provide evidence that patients, peers, coworkers and medical colleagues can provide reliable and multidimensional theoretically meaningful assessment of physicians in practice. |
| Hall et al., 1999[13] (Canada) | Multiple Specialties (*n* = 308, physicians) | **PAR (Generic)**<br>Self (26 Items)<br>MC (26 Items)<br>CW (17 Items)<br>Pt (44Items)<br>Consultant (23 Items)<br>Referring ( 21 Items) | <br>Prof, Clin comp,Inter Per<br>Prof, Clin comp,Inter Per<br>Prof, Comm, Inter Per<br>Prof, Comm, Mager<br>Prof, Clin comp, Inter Per<br>Prof, Clin comp | In this pilot study of physicians registered with the College of Physicians and Surgeons of Alberta (CPSA) the PAR program was initially introduced. This PAR project was found to be feasible at an estimated cost of $200 per physician, and based on these findings was implemented in the province, where all physicians are required to participate every 5 years. |
| Violato et al., 2003[14] (Canada) | Surgery (*n* = 201, surgeons) | **PAR (Surgery)**<br>Self (34Items)<br>MC (34 Items)<br>CW (19 Items)<br>Pt (39Items) | <br>Prof, Clin comp, Comm, Inter Per<br>Prof, Clin comp, Comm, Inter Per<br>Comm, Inter Per<br>Comm, Inter Per,Mngr | As part of the CPSA PAR process, modified versions of the instruments were developed to be used with surgeons. The authors concluded that an MSF system is feasible, reliable, and valid in assessing key competencies and, moreover, provides feedback to initiate change in surgeons' practice. |
| Lockyer & Violato, 2004[15] (Canada) | Psychiatry (*n* = 101),Pediatrics (*n* = 100)and Internal Medicine (*n* = 103) (*n* = 304, physicians) | **PAR (Specialty Generic)**<br>MC(36 Items) | <br>Prof, Clin comp, Comm | The reliability and generalizability coefficients provide support for the use of the PAR program in Alberta across three different specialties. Although consistency is found in the number of factors measured, percentage of variance accounted for any one factor reflects differences in competencies assessed between the specialties. |
| Lockyer et al., 2006[16] (Canada) | General Practice (*n* = 37, physicians) | **PAR modified (IMG)**<br>Self (21 Items)<br>MC (22 Items) | <br>Prof, Clin Comp<br>Prof, Clin Comp | The findings indicate that the modified PAR tools have acceptable psychometric properties for the assessment of international medical graduates (IMGs), whose knowledge and skills have not been |

| | | | | |
|---|---|---|---|---|
| | | CW (12 Items)<br>Pt (13 Items) | Prof, Comm,<br>Prof, Comm, Mngr | formally assessed through national examination processes. The authors suggest that further research comparing IMGs with a benchmark group of Canadian physicians are needed to achieve a level of authenticity in measuring clinical competency and performance. |
| Lockyer et al., 2006[17] (Canada) | Emergency Medicine (*n* = 187, physicians) | **PAR (Emerg Med)**<br>Self (30 Items)<br>MC (31 Items)<br>CW (20 Items)<br>Pt (16Items) | <br>Prof, Clin comp, mngr<br>Prof, Clin comp, mngr<br>Prof, Clin comp, Inter Per<br>Prof, Comm, Inter Per | As part of the CPSA PAR process, modified versions of the instruments were developed to be used with emergency medicine physicians. The psychometric analysis suggests that the instruments developed were feasible and provided evidence of reliability and validity. |
| Lockyer et al., 2006[18] (Canada) | Anesthesia (*n* = 197, physicians) | **PAR (Anesthesia)**<br>Self (29 Items)<br>MC (29 Items)<br>CW (19 Items)<br>Pt(11Items) | <br>Prof, Clin comp, Comm<br>Prof, Clin comp, Comm<br>Comm, InterPer<br>Prof, Comm | As part of the CPSA PAR process, modified versions of the instruments were developed to be used with anesthesiologists. The authors concluded that it was feasible to develop MSF instruments for anesthesiologists that are psychometrically reliable and valid. |
| Violato et al., 2006[19] (Canada) | Pediatrics (*n* = 100, physicians) | **PAR (Pediatric)**<br>Self (37 Items)<br>MC (38 Items)<br>CW (22 Items)<br>Pt (40 Items) | <br>Prof, Clin comp, Comm<br>Prof, Clin comp, Comm<br>Comm, Inter Per<br>Prof, Comm, Mngr | As part of the CPSA PAR process, modified versions of the instruments were developed to be used with pediatricians. The authors concluded that it was feasible to develop high-quality MSF instruments for pediatricians that are psychometrically reliable and valid. |
| Lockyer et al.,2007[20] (Canada) | Family Medicine (*n* = 250, family physicians) | **PAR (Fam Med)**<br>Self (31 Items) | <br>Prof, Clin comp, Comm, Mngr | Since 1996, the PAR has become mandatory for continued licensure every 5 years for all major clinical disciplines. Physician self-assessment was shown to be stable between Time 1 and Time 2; assessments indicated that the incorporation of feedback over time is limited. |
| Violato et al., 2008[21] (Canada) | Family Medicine (*n* = 250, family physicians) | **PAR (Fam Med)**<br>Med Colleague (31 items)<br>Co-worker (17 items)<br>Patients (40 items.) | <br>Prof, Clin Comp, Inter Per<br>Prof, Comm<br>Prof, Comm, Off Per, DrAcc, PhySp | Since 1996, the PAR has become mandatory for continued licensure every 5 years for all major clinical disciplines in the province of Alberta. The PAR showed evidence for the construct validity and stability of the MC, CW, and Pt instruments over a 5-year period between assessments at Time 1 and Time 2. |
| Violato et al., 2008[22] (Canada) | Psychiatry (*n* = 101, physicians) | **PAR (Psychiatry)**<br>Self (37 Items)<br>MC (38Items)<br>CW (22 Items)<br>Pt(40Items) | <br>Prof, Clin comp, mngr<br>Prof, Clin comp<br>InterPer, Comm<br>Prof, ,Comm, mngr | As part of the CPSA PAR process, modified versions of the instruments were developed to be used with psychiatrists. The authors showed that it was possible to develop a feasible multisource feedback program in psychiatry with evidence of reliability and validity that provides feedback about key clinical competencies. |
| Lockyer et al., 2009[23] (Canada) | Pathology & Laboratory Medicine (*n* = 101, physicians) | **MSF tool**<br>Self (39Items)<br>MC (39 Items)<br>CW (22 Items)<br>Referring (30Items) | <br>Prof, Clin comp, Inter Per<br>Prof, Clinc omp,Inte rPer<br>Prof, Comm<br>Prof, Clin comp, Mngr | Modified from the PAR instruments used with CPSA, an MSF system used with pathologists and laboratory medicine physicians was shown to be reliable, valid, and feasible in providing guided feedback on competencies and behaviors. |
| Overeem et al., 2012[24] (Netherlands) | Multiple Specialties (*n* = 146, physicians) | **PAR (modified for NL)**<br>Self (32 Items)<br><br>MC (33 Items) | <br>Prof, Clin comp, Mngr, Inter Per<br>Prof, Clin comp, Mngr, | Based on the MSF PAR system used with the CPSA in Canada, the Self, MC, CW, and Pt instruments were modified to complement the Dutch health care system. The authors concluded that the use of three MSF instruments produced reliable and valid data for |

| | | | | |
|---|---|---|---|---|
| | | CW (22 Items)<br>Pt (18 Items) | Inter Per<br>Prof, Clin comp, Comm<br>Prof, Comm, Inter Per | evaluating physicians' professional performance in the Netherlands. |
| Lockyer et al., 2012[25] (Canada) | Surgery (*n* = 216, surgeons) | **PAR (Surgery)**<br>Self: (34Items)<br>MC: (34 Items)<br>CW: (19 Items)<br>Pt: (39Items) | <br>Prof, Comm, Clin Comp, Mngr<br>Prof, Comm, Clin Comp, Mngr<br>Comm<br>Comm, Mngr, Inter Per | The purpose of this study was to compare the performance of practicing surgeons in Alberta who graduated from the University of Calgary (a three-year school) with matched samples from other four-year Canadian medical schools and to determine the reliability and validity of the PAR instrument in assessing surgeons. |
| **Sheffield Peer Review Assessment Tool (SPRAT)** | | | | |
| Archer et al., 2005[26] (UK) | Pediatrics (*n* = 112, residents) | **SPRAT**<br>MC, CW (same 24 items) | <br>Clin Comp, Inter Per | Authors concluded that the use of the SPRAT was a feasible, reliable and valid assessment method in informing the record of in-training assessment for pediatric senior house officers and specialists' registrars. |
| Davies et al.,2008[27] (UK) | Histopathology (*n* = 92, residents) | **PATH-SPRAT**<br>Self, MC, CW (same 21 Items) | <br>Clin comp, Comm | The histopathology specific PATH-SPRAT was developed from the SPRAT and designed to assess the generic competencies in Good Medical Practice (GMP). The authors indicate that specialty-specific MSF was feasible and achieved satisfactory reliability. |
| Archer et al.,2008[28] (UK) | Multiple Specialties *n* = 553, residents) | **mini-PAT(SPRAT)**<br>MC, CW(same 16 Items) | <br>Clin Comp, Inter Per | The mini-PAT (Peer Assessment Tool) was introduced to assess clinical performance of foundation trainees. |
| Crossley et al., 2008[29] (UK) | Multiple Specialties (*n* = 137, residents) | **SPRAT/SHEFFPAT**<br>MC, CW (same 24 items)<br>Pt (13 Items) | <br>Clin Comp, Inter Per<br>Clin Comp, Inter Per | Although the SPRAT/SHEFFPAT MSF system was found to be feasible within a hospital/workplace setting, future trust-based assessment requires further development for administration, confidentiality, patient support, and potentially new instruments for non-clinical specialties. |
| Archer et al.,2010[30] (UK) | Pediatrics (*n* = 577, residents) | **SPRAT**<br>MC, CW (same 24 Items) | <br>Clin Comp, Inter Per | SPRAT was used to measure the generic competencies of Good Medical Practice (GMP) as a national implementation mandate for the assessment within the Pediatric Specialist Registrars (SpRs). |
| Archer & McAvoy, 2011[31] (UK) | Multiple Specialties (*n* = 68, physicians) | **SPRAT/SHEFFPAT**<br>MC, CW (same 24 Items)<br>Pt (13 Items) | <br>Clin Comp, Inter Per<br>Clin Comp, Inter Per | This study was conducted in a conjunction with the National Clinical Assessment Service (NCAS) in the UK and used established MSF and PF instruments to assess doctors in potential difficulty. Although health practitioner colleagues appear to report poor performance using MSF, patients fail to concur. This challenges the validity of the patient survey, as it is designed and used currently. |
| **MSF or 360-degree evaluation** | | | | |
| DiMatteo & DiNicola, 1981[32] (USA) | Multiple Specialties (*n* = 141, residents) | **MSF forms**<br>Self (8 Items)<br>Attending (13 items)<br>MC (9 Items)<br>Pt (3 Items) | <br>Clin comp, Inter Per<br>Clin comp, Inter Per<br>Clin comp, Inter Per<br>Clin comp, Inter Per | The author examined the technical and the interpersonal skills of residents across different specialties by using different forms and four groups of raters, including self. The ratings from four sources were found to be fairly independent, indicating that they provide separate measures of physician performance. The reliabilities of measures from four sources were found to be substation, suggesting the usefulness of these sources for physician evaluation. |
| Risucci et al., 1989[33] | Surgery | **360-degree evaluation** | | The authors concluded that the use of the use of 360-degree |

| | | | | |
|---|---|---|---|---|
| (USA) | (n = 32, residents) | Self, MC (same 10 Items) | Prof, Clin comp, Inter Per | evaluation was valid in relation to peer and supervisor ratings of surgical residents. Discrepancies found on the self-assessment with those of the peers and supervisors are suggested to reflect the need for residents to address concerns related to professional, interpersonal, and clinical skill performance. |
| Ramsey et al., 1993[34] (USA) | Internal Medicine (*n* = 314, physicians) | **Peer physician assessment** MC (11 Items) | Clin comp, Inter Per | The findings suggest that it is feasible to use peer-assessment from professional associates to assess practicing physicians in domains such as clinical skills and interpersonal or humanistic qualities that are difficult to measure using other sources. |
| Wenrich et al., 1993[35] (USA) | Internal Medicine (*n* = 232, physicians) | **360-degree evaluation** MC (10 Items) CW (13 Items) | Clin comp, Inter Per | The authors concluded that nurses' ratings appear to provide a feasible and reliable method of evaluating internists' communication skills and humanistic qualities; however, they suggested that this be used in conjunction with ratings provided by peer physicians. |
| Thomas et al. 1999[36] (USA) | Internal Medicine (*n* = 16, residents) | **Peer physician assessment** MC (10 Items) | Clin Comp, Inter Per | The authors concluded that the use of peer review was reliable and feasible when completed by residents, but less so by faculty members. In addition, the authors reported that the residents gave high ratings to the value of the feedback provided by their peers in an end-of-year survey. |
| Lipner et al.,2002[37] (USA) | Internal Medicine (*n* = 356, physicians) | **Peer/patient assessment** MC (11 Items) Pt (10 Items) | Prof, Clin comp Prof, Clin comp, Comm | The patient and peer assessment module was introduced to evaluate the value of MSF in a recertification professional development program for practicing physicians. Participants reported that the module provided feedback that was beneficial for use in improving their practices. |
| Davis, 2002[38] (USA) | Obstetrics/ Gynecology (*n* = 16, residents) | **MSF** Self, MC and CW (same 16 Items) | Clin comp, Inter Per | This evaluation form found support for the use of MSF when used with other medical colleagues (i.e., faculty members and peers), however, showed discrepancies when compared with the ratings given by self and coworker (nurses) assessments. Suggested that residents may benefit from doing the self-assessment to improve their ability to honestly appraise their clinical and interpersonal skills. |
| Joshi et al., 2004[39] (USA) | Obstetrics/ Gynecology (*n* = 8, residents) | **360-degree evaluation** Self, MC, CW, Pt and Medical Students (same 10 Items) | Comm, Inter Per | The authors concluded that the 360-degree evaluation questionnaire appear to be reliable in evaluating residents' competencies in interpersonal and communication skills. Further research on the determining the reliability between evaluator categories and throughout the 4 years of the residency program is suggested. |
| Wood et al.,2004[40] (USA) | Radiology (*n* = 7, residents) | **360-degree evaluation** Self, MC, CW, Pt (same 10 Items) | Prof, Comm | This study shows that the 360-degree evaluation form was a reliable measurement of radiology residents' professionalism and interpersonal/communication skills. Although the time to complete was feasible, there were organizational and analysis challenges. |
| Wood et al., 2006[41] (UK) | Obstetrics/ Gynecology (*n* = 113, residents) | **Team Observation** tool MC (4 items) | Mngr, Inter Per | The Team Observation tool has become mandatory in obstetrics/gynecology training for the past 6 years. The aim was to assist in the facilitation and assessment of the implementation of "Calman's Structured Training" program. |
| Brinkman et al., | Pediatrics | **MSF** | | Adapted from the American Board of Internal Medicine surveys, the |

| | | | | |
|---|---|---|---|---|
| 2007[42] (USA) | (*n* = 36, residents) | | Prof, Comm<br>Prof, Clin Comp, Comm | Parent Satisfaction Questionnaire consists of 10 communication- and humanistic-related questions, and the nurse evaluation consists of 14 items related to professionalism, communication, and clinical competence. These questionnaires were shown to enhance standard feedback on resident performance and improved pediatric resident communication skills and professionalism. |
| Allerup et al., 2007[43] (Denmark) | Internal Medicine (*n* = 42, residents) | **360-degree evaluation**<br>MC and CW (same 15 Items) | Prof, Clin comp, Comm, InterPer | The purpose of this study was to explore the feasibility of 360-degree assessment in an internal medicine residency program in a Danish setting. Although the feasibility and reliability was found to be acceptable, the construct validity of the MSF tool was not determined or verified based on the domains identified in this study. |
| Pollock et al., 2007[44] (USA) | Plastic Surgery (*n* = 6, residents) | **360-degree evaluation**<br>MC, CW (same 60 Items) | Prof, Clin comp, Comm, Mngr, Inter Per | In this study, plastic surgery residents' performance was rated differently by health care professionals. Nevertheless, the residents found the 360-degree evaluation to be beneficial as they received two independent, formative assessments over a number of years of integrated training. |
| Massagli & Carline., 2007[45] (USA) | Physical Medicine & Rehabilitation (*n* =56, residents) | **360-degree evaluation**<br>CW, Rehab Staff, Medical Students (same12 Items) | Prof, Clin comp, Comm, Inter Per | The authors concluded that the use of a Web-based 360-degree evaluation tool is a feasible way to obtain reliable ratings from rehabilitation staff about resident behaviors. This instrument showed adequate reliability and validity in assessing residents in the physical and rehabilitation program. |
| Lelliott et al., 2008[46] (UK) | Psychiatry (*n* = 347, physicians) | **ACP 360**<br>Self, MC (same 57 Items)<br>Pt (17 Items) | Clin comp, Comm, InterPer<br>Clin comp, Comm, InterPer | The 360-degree Assessment of Consultant Psychiatrists (ACP 360) service was implemented by the Royal College of Psychiatrists in the UK in 2005 to provide feedback for individual consultants for performance improvement. The authors reported that the use of the ACP 360 is considered to be a reliable and feasible service in assessing psychiatrists who work in large multi professional teams. |
| Campbell et al., 2008[47] (UK) | Multiple Specialties (*n* = 291, physicians) | **GMC Survey**<br>MC (17 Items)<br><br>Pt (9 Items) | Prof, Clin comp, Comm, InterPer<br>Prof, Clin comp, Comm, InterPer | The authors concluded that the General Medical Council (GMC) patient and colleague questionnaires were reliable and provided a basis for the assessment of professionalism among UK doctors. It is suggested that further research is needed to explore the validity of the questionnaires as reliable indicators of acceptable professional performance, especially for revalidation of physicians' registration. |
| Meng et al.,2009[48] (USA) | Anesthesia (*n* = 15, residents) | **360-degree evaluation**<br>CW (13 Items) | Prof, Comm, Inter Per | This 360-evaluation form may be useful for post anesthetic care unit rotations. It appears to correlate well with traditional global ratings (although coefficients were not provided), was feasible, and provided formative feedback to the residents. |
| Campbell et al., 2010[49] (UK) | Family Physicians (*n* = 179, physicians ) | **CFET/DISQ (CFEP360)**<br>MC (CFET: 18 Items)<br><br>Pt (DISQ: 12 Items) | Prof, Clin comp, Comm, Mngr Inter Per<br>Prof, Clin comp, Comm, Inter Per | The authors concluded that physician performance, as assessed using the Colleague Feedback Evaluation Tool (CFET) and Doctor's Interpersonal Skills Questionnaire (DISQ) or CFEP360 system, should be able to identify physicians who are underperforming, while still being of use to for the majority of physicians for revalidation purposes. |
| Chandler et al., | Pediatrics | **360-degree evaluation** | | Overall, the 360-degree evaluation ratings for the pediatric residents |

| | | | | |
|---|---|---|---|---|
| 2010[50] (USA) | (*n* = 66, residents) | Self, MC, CW and Pt (same10 Items)<br><br><br> | Comm, Inter Per<br>Comm, Inter Per<br>Comm, Inter Per<br>Comm, Inter Per | were high and provided guidance to them about their interpersonal and communication skills. The authors indicated that the results provide evidence for the use of multiple-evaluator feedback in a residency program that can feasibly be replicated annually. |
| Yang et al., 2011[51] (Taiwan) | Multiple Specialties (*n* = 245, residents) | **360-degree evaluation**<br>MC, CW (same 12 Items) | Prof, Clin comp, Comm | The authors conclude that the use of 360-degree evaluation as a formative method in assessment helped the residents to understand how other members of their team view their knowledge and attitudes. Subsequently, this helped the residents to develop an action plan and improve their behavior. |
| Wall et al., 2012[52] (UK) | Multiple Specialties (*n* = 834, residents) | **TAB**<br>Self: (4 Items)<br>MC, CW (same 4 Items) | <br>Prof, Comm<br>Prof, Comm | The authors concluded that the use of the four-item TAB assessment tool can help some physicians to identify concerns with professional or communication performance. The use of Self-TAB in comparison with the TAB, however, demonstrates physicians' limited ability to self-assess. |
| Qu et al., 2012[53] (China) | Multiple Specialties (*n* = 258, residents) | **EOS Group Tools**<br>Self (21 Items)<br>MC (21 items)<br>Attending (21 items)<br>CW (26 items)<br>Office staff (15 items)<br>Pt (25 items) | <br>Prof, Comm<br>Prof, Comm<br>Prof, Comm<br>Prof, Comm<br>Prof, Comm,<br>Prof, Clin comp, Mngr Inter Per | The authors concluded that the 360-degree evaluation tools developed by the Education Outcomes Service (EOS) group from the Arizona Medical Education Consortium are reliable and valid in assessing resident professionalism and interpersonal communication skills in China. It was suggested that further studies are required to determine how the residents used their data to produce changes in their professional and interpersonal communication skills. |
| Wright et al., 2012[54] (UK) | Multiple Specialties (*n* = 1,065, physicians) | **GMC Survey**<br>MC (18 Items)<br><br>Pt (9 Items) | <br>Prof, Clin comp, Comm, InterPer<br>Prof, Clin comp, Comm, InterPer | The General Medical Council (GMC) has introduced a five-year cycle whereby all licensed doctors must seek "revalidation," in part, through the use of feedback on the Colleague and Patient Questionnaires. Although found to be feasible for formative purposes, concerns about the utility of the Pt and MC feedback as a stand-alone assessment of physician practice are expressed. |

* IMG = International Medical Graduate, PAR = Physician Achievement Review, Prof = Professionalism, Clin Comp = clinical competence, InterPer = Interpersonal Relationship, Comm = Communication, Off Per = Office personnel, Dr.Acc = Access to Doctor, PhySp = Physical Space, MC = Medical colleague, CW = Co-Worker, Pt =Patient, Mngr = manager, SPRAT = Sheffield Peer Review Assessment Tool, SHO = Senior House Officer, SPR =Pediatric Specialists Registrar, PACU = Post Anesthesia Care Unit, PATH-SPRAT = Pathology Sheffield Peer Review assessment Tool, MSF = Multi Source Feedback, OSPE = Objective Structured Practical Examination, F2 = Foundation 2, F1 = Foundation 1, Refphysi = Referring Physician, SHEFFPAT = The Sheffield Patient Assessment Tool, RehStaf = Rehabilitation Staff, TAB = Team Assessment of Behaviors.

Supplemental digital content for Donnon T, Al Ansari A, Al Alawi S, Violato C. The Reliability, Validity, and Feasibility of Multisource Feedback Physician Assessment: A Systematic Review. Acad Med. 2014:89(3).

## Supplemental Digital Table 2

Reliability and Validity Characteristics of the 43 studies on Physician Multisource Feedback (MSF) Included in a Systematic Analysis of the Literature Published 1975–January 2013

| Study (Origin) | Mean no. raters (Response rate)* | Reliability (α), Generalizability ($Ep^2$) and/or Intra-Class Correlation (ICC)* | Validity* |
|---|---|---|---|
| **Physician Assessment Review (PAR)** | | | |
| Violato et al., 1997[7] (Canada) | Self (SAQ): 1 (100%)<br>MC (PAQ): 7.8 (76.8%)<br>Pt (PS): 26.2 (87.4%)<br>CW (CAQ): 8.5 (85.4%)<br>MC (APCQ): 7.4 (73.5%)<br>MC (ACRPQ): 8.6 (85.5%) | Self (SAQ): α = 0.95<br>MC (PAQ): α = 0.95, for 8 raters $Ep^2$ = 0.77<br>Pt (PS): α = 0.95, for 25 raters $Ep^2$= 0.80<br>MC (CAQ): α = 0.95<br>MC (APCQ): α = 0.92<br>MC (ACRPQ): α = 0.89 | ***Construct:*** Principal component factor analysis was conducted for the PAQ (four factor solution), PS (seven factor solution), and CAQ (three factor solution) questionnaires accounting for 73.1%, 70.0%, and 72.8% of the variance, respectively. The mean rating scores were shown to be higher for medical colleagues (MC) or peers ($P < .05$), co-workers, and patients when compared with physicians' self-assessments. |
| Hall et al., 1999[13] (Canada) | Self: 1 (95.8%)<br>MC: Consultant and Referring: 6.4 (79.7%)<br>CW: 5.2 (86.7%)<br>Pt: 22.1 (88.6%) | Self: α = 0.95<br>MC: α = 0.95<br>Consultant: α = 0.93<br>Referring: α = 0.91<br>CW: α = 0.95<br>Pt: α = 0.95 | ***Construct:*** The mean ratings showed that self-assessments were consistently lower than reported by peers (MC, Consultants and Referring), coworkers (CW), and patients (Pt). |
| Violato et al., 2003[14] (Canada) | Self: 1 (96.5%)<br>MC: 7.3 (89.6%)<br>CW: 7.2 (88.2%)<br>Pt: 22.6 (83.2%) | Self: α = 0.97<br>MC: α = 0.98<br>CW: α = 0.95<br>Pt: α = 0.93 | ***Construct:*** A principal component factor analysis showed a five-factor solution for peers (MC) accounting for 69.0% of the variance, three factors for coworker (CW) accounting for 70.9%, five factors for patients (Pt) accounting for 73.5%, and four factors for self accounting for 65.1%.The mean ratings showed that self-assessments were consistently lower than those reported by peers, coworkers, and patients. |
| Lockyer & Violato, 2004[15] (Canada) | MC (Psych): 7.6 (94.6%)<br>MC (Peds): 7.6 (95.5%)<br>MC (IM): 7.6 (94.4%) | MC (Psych): α= 0.98, for 7.6 raters $Ep^2$ = 0.81<br>MC (Peds): α= 0.98, for 7.6 raters $Ep^2$ = 0.88<br>MC (IM): α= 0.99, for 7.6 raters $Ep^2$ = 0.82 | ***Construct:*** Principal component factor analysis was conducted to derive a four-factor solution for MC (psychiatrists) accounting for 70% of the variance, four factors for MC (pediatricians) accounting for 67.6%, and four factors for MC (internal medicine) accountings for 73.4%. |
| Lockyer et al., 2006[16] (Canada) | Self: 1 (91.8%)<br>MC: 5.7 (71.8%)<br>CW: 6.9 (86.1%)<br>Pt: 17.5 (69.9%) | Self: α = 0.83<br>MC: α = 0.98, for 5.7 raters $Ep^2$= 0.67<br>CW: α = 0.91, for 6.9 raters $Ep^2$= 0.59<br>Pt: α = 0.95, for 17.5 raters $Ep^2$= 0.71 | ***Construct:*** Principal component factor analysis showed a two-factor solution for medical colleague (MC) accounting for 71.5 % of the variance, two factors for coworker (CW) accounting for 59.5%, and two factors for patient (Pt) accounting for 74.9% of the variance. Unlike other findings, mean ratings for self-assessment were higher than reported by medical colleague (MC) and near identical to mean ratings that were reported by their patients. |
| Lockyer et al., 2006[17] (Canada) | Self: 1 (100%)<br>MC: 7.7 (95.5%)<br>CW: 7.6 (94.9%)<br>Pt: 21.6 (86.3%) | Self: α = 0.97<br>MC: α = 0.97, for 7.7 raters $Ep^2$= 0.84<br>CW:α = 0.94, for 7.6 raters $Ep^2$= 0.85<br>Pt: α = 0.97, for 21.6 raters $Ep^2$= 0.68 | ***Construct:*** An exploratory factor analysis showed a four-factor solution for the peer (MC), two for the coworker (CW), and two for the patient (PT) instruments that accounted for 71.9%, 62.5%, and 80.0% of the |

| | | | |
|---|---|---|---|
| | | | variance, respectively. The mean ratings showed that self-assessments were consistently lower than reported by peers, coworkers, and patients. |
| Lockyer et al., 2006[18] (Canada) | Self: 1 (100%)<br>MC: 7.8 (94.6%)<br>CW : 7.8 (95.1%)<br>Pt: 17.7 (56.2%) | Self:α = 0.97<br>MC:α = 0.97, for 7.8 raters $Ep^2$= 0.69<br>CW:α = 0.95, for 7.8 raters $Ep^2$= 0.56<br>Pt:α = 0.93, for17.7 raters $Ep^2$= 0.65 | ***Construct:*** An exploratory factor analysis showed a three-factor solution for the peer (MC), two factors for the coworker (CW) and two factors for the patient (PT) instruments that accounted for 74.5%, 67.5%, and 77.6% of the variance, respectively. The mean ratings showed that self-assessments were consistently lower than reported by peers, coworkers, and patients. |
| Violato et al., 2006[19] (Canada) | Self: 1 (100%)<br>MC: 7.6 (95.5%)<br>CW: 7.6 (94.8%)<br>Pt: 23.4 (93.6%) | Self: α = 0.98<br>MC: α = 0.98, for 7.6 raters $Ep^2$= 0.78<br>CW: α = 0.95, for 7.6 raters $Ep^2$= 0.87<br>Pt: α = 0.99, for 23.4 raters $Ep^2$= 0.85 | ***Construct:*** Principal component factor analysis showed a four-factor solution for peers (MC) accounting for 67.6% of the variance, three factors for coworkers (CW) accounting for 63.8%, and four factors for patients (Pt) accounting for 77.6 %. Self instrument is identical to co-worker instrument. The mean ratings showed that self-assessments were consistently lower than reported by peers, coworkers, and patients. |
| Lockyer et al., 2007[20] (Canada) | Self: 1 (100%) | Self, α = 0.96 | ***Construct:*** Principal component factor analysis was conducted to derive a three-factor solution accounting for 71% of the variance.<br>***Predictive:*** The sum of the mean scores calculated for self-ratings between Time 1 and Tune 2 (5-year interval) showed that physicians rated themselves higher in the second iteration, $P < .05$. |
| Violato et al., 2008[21] (Canada) | MC: 7.19 (93%)<br>CW: 7.34 (94%)<br>Pt: 24.09 (97%) | MC: α = 0.96, for 8 raters $Ep^2$= 0.78<br>CW: α = 0.96, for 8 raters $Ep^2$ = 0.83<br>Pt: α = 0.98, for 23 raters $Ep^2$ = 0.80 | ***Construct:*** Confirmatory factor analyses were conducted on the MC (CFI = 0.91), CW (CFI = 0.87), and Pt (CFI = 0.83) instruments.<br>***Predictive:*** From Time 1 to Time 2 (5-year interval) on both the MC and CW total, there was found to be a significant improvement, $P < .001$. From Time 1 to Time 2 (5-year interval) on the Pt total; however, no significant difference was shown. |
| Violato et al., 2008[22] (Canada) | Self: 1 (100%)<br>MC: 7.6 (94.6%)<br>CW: 7.4 (92.1%)<br>Pt: 24.3 (97.3%) | Self: α = 0.96<br>MC: α = 0.98, for 7.6 raters $Ep^2$= 0.81<br>CW: α = 0.96,for 7.4 raters $Ep^2$ = 0.82<br>Pt: α = 0.98, for 24.3 raters $Ep^2$= 0.78 | ***Construct:*** Principal component factor analysis showed a four-factor solution for peers (MC) accounting for 66.8%, three factors for coworker (CW) accounting for 68.8%, and five factors for patients (Pt) accounting for 73.7% of the variance. The mean ratings showed that self-assessments were consistently lower than those reported by peers, coworkers, and patients. |
| Lockyer et al., 2009[23] (Canada) | Self: 1 (100%)<br>MC: 7.6 (91.3%)<br>CW: 7.6 (91.8%)<br>Referring: 7.4 (90.3%) | MC: α = 0.98, for 7.6 raters $Ep^2$ = 0.78<br>CW: α = 0.95, for 7.6 raters $Ep^2$ = 0.80<br>Referring: α = 0.98, for 7.4 raters $Ep^2$ =0.81 | ***Construct:*** Principal component factor analysis showed a five-factor solution for peers (MC) accounting for 68.8% of the variance, three factors for referring physicians (Referring)accounting for 66.9%, and two factors for coworkers (CW) accounting for 59.9%.The mean ratings showed that self-assessments were consistently lower than reported by peers, coworkers, and referring physicians. |
| Overeem et al., 2012[24] (Netherland) | MC: 6.5 (81.3%)<br>CW: 6.7 (83.8)<br>Pt: 15 (51.8%) | MC: α = 0.95<br>CW: α = 0.95<br>Pt: α = 0.94 | ***Construct:*** Principal component factor analysis showed a six-factor solution for peers (MC) accounting for 67 % of the variance, three factors for coworker (CW) accounting for 70%, and a single factor for patient (Pt) accounting for 60%. Physicians with more work experience were rated lower by MC and CW; $P < .05$. MC ratings showed a medium correlation with CW ratings ($r = 0.35$, $P < .01$), a small correlation with Pt |

| | | | |
|---|---|---|---|
| | | | ratings ($r$ = 0.21, $P$ < .01), and CW ratings showed a small correlation with Pt rating ($r$ = 0.22, $P$ < .01). |
| Lockyer et al., 2012[25] (Canada) | Self: 1<br>MC: 7.67<br>CW: 7.60<br>Pt: 24 | Self: α = 0.97<br>MC: α = 0.98, for 7.27 raters $Ep^2$ = 0.61<br>CW: α = 0.95, for 7.20 raters $Ep^2$ = 0.70<br>Pt: α = 0.98, raters 22.63 raters $Ep^2$ = 0.81 | ***Construct validity***: Principal component factor analysis showed a four-factor solution for medical colleague (MC) accounting for 75% of the variance, two factors for coworker (CW) accounting for 72%, and four factors for patient (Pt) accounting for 77% of the variance. The mean ratings showed that self-assessments were consistently lower than reported by peers, coworkers, and patients. |
| **Sheffield Peer Review Assessment Tool (SPRAT)** | | | |
| Archer et al., (2005)[26] (UK) | Combined MC and CW: 8.2 (82.0%) | SEM for 4 raters ± 0.50 (95% CI) | ***Construct:*** The mean ratings for specialist registrars were significantly higher than for senior house officers, $P$ < .001. In a hierarchical regression, the rating of the residents by the peers (MC) accounted for 7.6% of the variation in the mean ratings. |
| Davies et al., 2008[27] (UK) | Self: 1 (100%)<br>Combined MC and CW: 9.2 (92%) | SEM for 8 raters ± 0.37 (95% CI) | ***Construct:*** Principal component factor analysis was conducted to derive a two-factor solution accounting for 78% of the variance. Pearson's correlation for self versus assessor ratings was shown to be negative ($r$ = –0.13, $P$ > .05). Consultants marked trainees lower than other occupational groups, $P$ < .001.<br>***Predictive:*** A medium correlation was found between the trainees' PATH-SPRAT aggregated and Objective Structure Practical Examination scores; $r$ = 0.48, $P$ < .001. |
| Archer et al., 2008[28] (UK) | Combined MC and CW: 6.7 (67%) | Combined MC and CW: α = 0.98<br>SEM for 8 raters ± 0.45 (95% CI) | ***Construct:*** Principal component factor analysis was conducted to derive a two-factor solution accounting for 81% of the variance. Consultants scored trainees significantly lower than other assessors; $P$ < .001.The mean scores showed that year-one (F1) trainees were rated significantly lower than year-two (F2) trainees; $P$ < .001. |
| Crossley et al., 2008[29] (UK) | Combined MC and CW: 14 (100%)<br>Pt: 9.7 (27.4%) | Combined MC and CW: SEM for 9 raters ± 0.37 (95% CI)<br>Pt: SEM for 15 raters ± 0.29 (95% CI) | ***Construct:*** Patients (Pt) rated female physicians significantly higher than male physicians for their relational skills, $P$ < .05. The least stringent professional group (foundation doctors/pre-registration house officers) rated the residents higher on average than the most stringent professional group (allied health professionals), $P$ < .05. |
| Archer et al., 2010[30] (UK) | Combined MC and CW: 8.26 (83%) | SEM for 8 raters ± 0.40 (95% CI) | ***Construct:*** Principal component factor analysis was conducted to derive a two-factor solution accounting for 76.5% of the variance. Consultants marked trainees significantly lower than all groups of raters ($P$ < .05), whereas senior house officers and foundation doctors scored trainees significantly higher than consultants ($P$ < .05).<br>***Predictive:*** The mean scores for Year 4 were significantly higher than for Year 2, $P$ < .01. |
| Archer & McAvoy, 2011[31] (UK) | Combined MC and CW: 12.0<br>Pt: 22.8 | NR | ***Construct:*** The mean ratings showed that the assessors identified by the physicians were rated significantly higher than those that were identified by the referring body; $P$ < .001. Nevertheless, patients scored the physicians higher than all assessors; $P$ < .001. The mean ratings showed that these physicians in difficulty when compared to a normative reference group scored significantly lower; $P$ < .001. |

| | | | |
|---|---|---|---|
| **MSF or 360-degree evaluation** | | | |
| DiMatteo et al., 1981[32] (USA) | Self: 1<br>Attending : 15<br>MC: 15<br>Pt: 15 | Self: α = 0.56 (Clin comp) and 0.78 (Inter Per)<br>Attending: α = 0.90 (Clin comp and Inter Per)<br>MC: α = 0.67 (Clin comp) and 0.92 (Inter Per)<br>Pt: α = 0.79 (Inter Per) | ***Construct:*** Principal component factor analysis for Internal Medicine (IM) I group showed a two-factor solution for Attending accounting for 68.7 % of the variance, two factors for peers (MC) accounting for 87.5%, and two factors for self (Self) accounting for 57.2% of the variance. Results from similar forms used with the IM II, surgery, and family medicine residents found similar factor solution results.<br>***Concurrent:*** Correlations on the two factors between self (Self) with Attending ($r$ = 0.08 to 0.31), peers (MC) ($r$ = 0.06 to 0.38), and patients (Pt) ($r$ = -0.07 to 0.44) are negative to moderate. |
| Risucci et al., 1989[33] (USA) | Self: 1 (84.4%)<br>MC (peers): 27<br>MC (supervisors): 4 | NR | ***Construct:*** Principal component factor analysis showed a three-factor solution for self (Self) accounting for 68.7 % of the variance, two factors for supervisors (MC) accounting for 80.3%, and a single factor for peers (MC) accounting for 85.3 % of the variance. The mean ratings showed that self-assessments were consistently higher than reported by peers and supervisors, and supervisors' mean ratings were higher than peers'.<br>***Concurrent:*** Supervisor and peer ratings strongly correlated ($r$ = 0.92, $P$ < .001).<br>***Predictive***: The peer and supervisor (MC) 360-degree evaluation showed large correlations with the American Board of Surgery In-Training Examination, $r$ = 0.52 and $r$ =0.55 ($P$ < .01), respectively. |
| Ramsey et al.,1993[34] (USA) | MC: 8.7 (51.6%) | MC: For 11 raters $Ep^2$ = 0.70 | ***Construct:*** Principal component factor analysis showed a two-factor solution accounting for 88.7 % of the variance. |
| Wenrich et al., 1993[35] (USA) | CW: 8.01 (68.2%) | CW: Based on a range of 6.6 to 13.9 raters (depending on item) $Ep^2$= 0.70 | ***Construct:*** Principal component factor analysis showed a two-factor solution for the combined nurse (CW) and peer (MC) evaluation forms based on the 10 common items. The mean ratings showed that nurses scored the physicians lower on humanistic qualities ($P$ < .01) but higher on medical knowledge ($P$ < .001) than the peer (MC) raters. |
| Thomas et al. 1999[36] (USA) | MC: 11.1 (49.2%) | MC: α = 0.94 | ***Construct validity***: Principal component factor analysis showed a two-factor solution for medical colleague (MC) accounting for between 84.4% (senior residents) to 88.2% (junior residents) of the variance. The mean ratings showed that faculty members scored the junior residents consistently lower than senior residents or peers. |
| Lipner et al., 2002[37] (USA) | MC: 10 (100%)<br>Pt: 25 (100%) | MC: For 10 raters$Ep^2$ = 0.61 (95% CI $\pm$ 0.41)<br>Pt: For 25 raters $Ep^2$=0.67 (95% CI $\pm$ 0.14) | ***Construct:*** The mean rating of patients (Pt) was found to be higher than the ratings received from peer (MC) assessments. |
| Davis, 2002[38] (USA) | Self: 1 (93.7%)<br>MC (Peers): 16 (100%)<br>MC (Faculty): 16 (92.9%)<br>CW (Nurses): 16 (83.3%) | MC (Faculty): ICC = 0.66 to 0.84<br>MC (Peers): ICC = 0.78 to 0.90<br>CW (Nurses): ICC = 0.23 to 0.45 | ***Concurrent:*** Pearson correlation coefficients between the MC faculty members and MC peers showed moderate to large correlations on both factors ($r$ = 0.72 and 0.80, $P$ < .01) and on the overall clinical assessment item ($r$ = 0.86, $P$ < .001). In comparison with MC faculty members ratings, however, the correlations with the Self and CW (Nurses) were non-significant and ranged between $r$ = -0.12 to 0.36 and $r$ = 0.04 to 0.24, respectively. |
| Joshi et al., 2004[39] (USA) | Self: 1 (100%)<br>MC: 16 (100%) | MC: For 16 raters ICC = 0.72 | The authors recognize that validity of the question was achieved by "expert opinion" only. |

| | | | |
|---|---|---|---|
| | CW: 25 (100%)<br>Pt: 10 (100%)<br>Medical Students: 12 (100%) | CW: For 25 raters ICC = 0.86<br>Pt: For 10 raters ICC = 0.54<br>Medical Students: For 12 raters ICC = 0.82 | ***Concurrent:*** Faculty (MC) ratings showed a large correlation with nurse coworkers (CW) ratings ($r$ = 0.55, $P$ = .16), a small correlation with Pt ratings ($r$ = 0.21, $P$ = .61), and CW ratings showed a medium correlation with Pt rating ($r$ = 0.43, $P$ = .29). |
| Wood et al., 2004[40] (USA) | Combined MC, CW and Pt: 8.14 (57%) | MC: α = 0.85<br>CW: α = 0.87<br>Pt: α = 0.86 | ***Construct:*** In an analysis of variance, it was found that the Pt mean score ratings of the trainees were significantly higher when compared with MC and CW, $P$ < .001.<br>***Concurrent:*** The correlation coefficients were calculated between a 5-item global ratings form (used as a gold standard) and the (1) Pt 360 degree evaluation ($r$ = 0.70, $P$ = .08), (2) MC ($r$ = 0.46, $p$ = 0.30), and (3) CW ($r$ = 0.62, $P$ = .14) were medium-to-large, however, not significant. |
| Wood et al., 2006[41] (UK) | MC: 12.52 | MC: For 8 raters ICC = 0.80 | ***Construct:*** Principal component factor analysis was conducted on the Team Observation tool to derive a one-factor solution accounting for 76% of the variance.<br>***Predictive:*** Spearman's correlation coefficients were calculated between Time 1 to Time 2 (6-7 month interval); $r$ = 0.77, $P$ < .001. |
| Brinkman et al., 2007[42] (USA) | Parents: 19.3<br>CW: 15.8 | Parents: α= 0.95<br>CW: α = 0.96 | ***Construct:*** Although statistical results between groups at Time 1 and Time 2 were not reported at both Time 1 and Time 2, the multisource feedback group achieved higher ratings from parents and nurses on average than the control group at Time 2. |
| Allerup et al., 2007[43] (Denmark) | Self: 1 (97.6%)<br>MC: 4.7 (94.0%)<br>CW: 2.8 (55.0%) | Combined MC and CW, α = 0.46 to 0.89 | ***Construct:*** The mean correlation ratings between self and coworkers (CW) indicated that nurses on average rate the residents (Self) higher. The mean correlation ratings between self and peers (MC), however, show that other physicians (MC) on average rate the residents (Self) lower. Note that the construct validity of the measures used was not provided and, therefore, the domains identified were not confirmed. |
| Pollock et al., 2007[44] (USA) | MC: 12<br>CW: 28 | NR | ***Construct:*** The mean ratings by peers (MC) was significantly lower than the nurse coworkers (CW) across all competencies areas identified. |
| Massagli & Carline., 2007[45] (USA) | CW: 3.7<br>Rehab Staff: 9.9<br>Medical Students: 3.0 | Combined CW, Rehab Staff and Medical Students: α = 0.89<br>CW: For 5 raters $Ep^2$= 0.80<br>Rehab Staff: For 4 raters $Ep^2$= 0.80<br>Medical Students: For 23 raters $Ep^2$= 0.80 | ***Construct:*** Principal component factor analysis showed a single factor-solution accounting for 84.0% of the variance. The mean scores for fourth-year residents (Self) were shown to be higher than for second- and third-year residents. |
| Lelliott et al., 2008[46] (UK) | Self: 1 (100%)<br>MC: 12.7 (85.0%)<br><br>Pt: 19.2 (63.9%) | Self: α = 0.98<br>MC: α = 0.98, for 13 raters $Ep^2 \geq$ 0.75, ICC = 0.75<br>Pt: α = 0.97, for 25 raters $Ep^2 \geq$ 0.75, ICC = 0.70 | ***Construct:*** Principal component factor analysis showed a seven-factor solution for peers (MC) accounting for 70.2 % of the variance and a single-factor solution for the patient (Pt) tool accounting for 66.8 % of the variance. The mean ratings showed that self-assessments were consistently lower than reported by peers and patients; $P$ < .001. |
| Campbell et al., 2008[47] (UK) | MC:13.8 (69.1%)<br>Pt: 36.2 (92.1%) | MC: α = 0.95, for 12 raters $Ep^2$ = 0.76<br>Pt: α = 0.96, for 36 raters $Ep^2$ = 0.75 | ***Construct:*** Principal component factor analysis showed a three-factor solution for peers (MC) for the 17 performance-based items accounting for 61.0 % of the variance, and two factors for patients (Pt) for the 9 performance-based items accounting for 76.8 % of the variance. On mean ratings patients (Pt) scored the physicians higher than peers (MC), and |

| | | | |
|---|---|---|---|
| | | | younger physicians were rated higher than older physicians by both their peers and patients; $P < .05$. |
| Meng et al., 2009[48] (USA) | CW: 28.6 (88%) | CW: (Nurses) ICC = 0.87<br>CW: (Secretaries) ICC = 0.79<br>CW: (Nurse Aids) ICC = 0.83<br>CW: (Technicians) ICC = 0.86 | ***Construct:*** The average mean ratings across all items from post anesthetic care unit nurses were higher than secretarial staff.<br>***Concurrent:*** Although the authors indicated that residents who were ranked highly by global ratings were also ranked highly by the 4 categories of 360-degree evaluation ratings, no correlations were provided. |
| Campbell et al., 2010[49] (UK) | MC: 13.9<br>Pt: 47.3 | MC: α = 0.84, for 14 raters $Ep^2$= 0.82<br>Pt: α = 0.95, for 25 raters $Ep^2$= 0.81 | ***Construct:*** Principal component factor analysis showed a two-factor solution for medical colleague (MC) CFET form accounting for 66.0% of the variance, and a single factor for the patient (Pt) DISQ form accounting for 94.0% of the variance. The mean ratings for patients were slightly higher on average than reported by peers (MC). |
| Chandler et al., 2010[50] (USA) | Self: 1 (100%)<br>MC: 2.6<br>CW: 7.4<br>Pt: 1.2 | NR | ***Construct:*** The mean ratings showed that self-assessments were consistently lower than reported by peers (MC) and nurse coworkers (CW); $P < .001$. Self mean ratings were, however, not significantly different from the patients (Pt). |
| Yang et al., 2011[51] (Taiwan) | Combined MC and CW: 4.3 (85.3%) | Combined MC and CW: α = 0.86 | ***Predictive***: The 360 degree evaluation show a medium correlation with the small scale OSCE ($r = 0.37$, $P < .05$). Moreover, adding the DOPS score to small-scale OSCE scores increased it to a large correlation at $r = 0.72$ ($P < .05$), and adding the IM in-training examination increased it to $r = 0.85$, $P < .05$. |
| Wall et al., 2012[52] (UK) | Self: 1 (100%)<br>Combined MC and CW: 11.6 | NR | ***Concurrent:*** The self ratings compared with combined peer (MC) and coworker (CW) ratings showed a small correlation on minor concerns ($r = 0.20$, $P < .001$) and major concerns ($r = 0.26$, $P < .001$). |
| Qu et al., 2012[53] (China) | Self: 1 (100%)<br>MC: 2 (100%)<br>Attending 1(100%)<br>CW: 3 (100%)<br>Office staff: 2 (100%)<br>Pt: 7 (100%) | Self: α = 0.92<br>MC: α = 0.93<br>Attending: α = 0.91<br>CW: α = 0.92<br>Office staff: α = 0.90<br>Pt: α = 0.93 | ***Construct:*** Principal component factor analysis showed a two-factor solution for self (Self) accounting for 71.0% of the variance, two factors for the attending (Attending) accounting for 70.9 % of the variance, two factors for peers (MC) accounting for 70.7%, two factors for nurses (CW) accounting for 75.5%, two factors for Office staff accounting for 74.6%, and four factors for patients (Pt) accounting for 72.7% of the variance. The mean ratings showed that self-assessments were consistently lower than reported by MC and Pt, but were higher when compared with the CW (nurses). |
| Wright et al., 2012[54] (UK) | MC:13.8 (69.1%)<br><br>Pt: 36.2 (92.1%) | MC: α = 0.94, ICC = 0.85, for $\geq$ 15 raters $Ep^2 \geq 0.70$<br>Pt: α = 0.87, ICC = 0.83, for $\geq$ 34 raters $Ep^2 \geq 0.70$ | ***Construct:*** Principal component factor analysis showed a three-factor solution for peers (MC) for the 18 performance-based items accounting for 58% of the variance, and two factors for patients (Pt) for the 9 performance-based items accounting for 79% of the variance. Convergent validity was shown with correlations between the Pt and Doctor's Interpersonal Skills Questionnaire (DISQ), rho = 0.63, $P < .001$; and between the MC and Colleague Feedback Evaluation Tool (CFET), rho = 0.81, $P < .01$. |

* SAQ = Self-Assessment Questionnaire, PAQ = Peer Assessment Questionnaire, PS = Patient Survey, CAQ = Coworker Assessment Survey, AQCQ = Assessment of Physician by Consultant Questionnaire, ACRPQ = Assessment of Consultant by Referring Physician Questionnaire, MC = Medical colleague, CW = Co-Worker, Pt =Patient, SEM =

Supplemental digital content for Donnon T, Al Ansari A, Al Alawi S, Violato C. The Reliability, Validity, and Feasibility of Multisource Feedback Physician Assessment: A Systematic Review. Acad Med. 2014:89(3).

Standard Error of Measurement, NR = Not Reported, CFI = Confirmatory Fit Index, ICC = Intraclass correlation coefficient, $Ep^2$ = Generalizability Coefficient, CFET = Colleague Feedback Evaluation Tool, DISQ = Doctor's Interpersonal Skills Questionnaire