Advances in Medical Education and Practice

Dovepress

open access to scientific and medical research

Open Access Full Text Article

REVIEW

The construct and criterion validity of the multi-source feedback process to assess physician performance: a meta-analysis

This article was published in the following Dove Press journal:
Advances in Medical Education and Practice
12 February 2014
Number of times this article has been viewed

Ahmed Al Ansari[1]
Tyrone Donnon[2]
Khalid Al Khalifa[1]
Abdulla Darwish[3]
Claudio Violato[4]

[1]Department of General Surgery, Bahrain Defense Force Hospital, Riffa, Kingdom of Bahrain; [2]Medical Education and Research Unit, Department of Community Health Sciences, Faculty of Medicine, University of Calgary, AB, Canada; [3]Department of Pathology, Bahrain Defense Force Hospital, Riffa, Kingdom of Bahrain; [4]Department of Medical Education, Faculty of Medicine, University Ambrosiana, Milan, Italy

Background: The purpose of this study was to conduct a meta-analysis on the construct and criterion validity of multi-source feedback (MSF) to assess physicians and surgeons in practice.

Methods: In this study, we followed the guidelines for the reporting of observational studies included in a meta-analysis. In addition to PubMed and MEDLINE databases, the CINAHL, EMBASE, and PsychINFO databases were searched from January 1975 to November 2012. All articles listed in the references of the MSF studies were reviewed to ensure that all relevant publications were identified. All 35 articles were independently coded by two authors (AA, TD), and any discrepancies (eg, effect size calculations) were reviewed by the other authors (KA, AD, CV).

Results: Physician/surgeon performance measures from 35 studies were identified. A random-effects model of weighted mean effect size differences (*d*) resulted in: construct validity coefficients for the MSF system on physician/surgeon performance across different levels in practice ranged from *d*=0.14 (95% confidence interval [CI] 0.40–0.69) to *d*=1.78 (95% CI 1.20–2.30); construct validity coefficients for the MSF on physician/surgeon performance on two different occasions ranged from *d*=0.23 (95% CI 0.13–0.33) to *d*=0.90 (95% CI 0.74–1.10); concurrent validity coefficients for the MSF based on differences in assessor group ratings ranged from *d*=0.50 (95% CI 0.47–0.52) to *d*=0.57 (95% CI 0.55–0.60); and predictive validity coefficients for the MSF on physician/surgeon performance across different standardized measures ranged from *d*=1.28 (95% CI 1.16–1.41) to *d*=1.43 (95% CI 0.87–2.00).

Conclusion: The construct and criterion validity of the MSF system is supported by small to large effect size differences based on the MSF process and physician/surgeon performance across different clinical and nonclinical domain measures.

Keywords: multi-source feedback system, meta-analysis, clinical performance, construct validity, criterion validity

Introduction

One of the most widely recognized methods used to evaluate physicians and surgeons in practice is multi-source feedback (MSF), also referred to as a 360-degree assessment, where different assessor groups (eg, peers, patients, coworkers) rate doctors' clinical and nonclinical performance.[1] Use of MSF has been shown to be a unique form of evaluation that provides more valuable information than any single feedback source.[1] MSF has gained widespread acceptance for both formative and summative assessment of professionals, and is seen as a trigger for reflecting on where changes in practice are required.[2,3] Certain characteristics of health professionals have been assessed using MSF, including their professionalism, communication, interpersonal relationships, and

Correspondence: Ahmed Al Ansari
Department of General Surgery,
Bahrain Defense Force Hospital,
PO Box 28347, Riffa, Kingdom of Bahrain
Tel +973 1777 6060
Email drahmedalansari@gmail.com

clinical and procedural skills competence.[4] One of the main benefits of MSF is that it provides physicians and surgeons with information about their clinical practice that may help them in improving and monitoring their performance.[5]

The number of published studies on the use of MSF to assess health professionals in clinical practice has increased substantially. In a recent systematic review studying the impact of workplace-based assessment of doctors' education and performance, Miller and Archer[6] reported evidence of support for use of MSF in that it has the potential to lead to improvement in clinical performance. Risucci et al[7] demonstrated concurrent validity for MSF in surgical residents by showing a medium effect size correlation coefficient between MSF scores and American Board of Surgery In-Training Examination (ABSITE) scores. When using MSF with residents at different levels in their program, Archer et al[8] showed modest increases in the performance of year 4 in comparison with year 2 trainees, thereby demonstrating the construct validity of this approach to assessment. Violato et al[9] compared changes in physician performance from time 1 to time 2 (a 5-year interval) using total scores given by medical colleagues and coworkers using the MSF questionnaire and demonstrated a significant improvement in their performance over time. Although MSF has been used in a variety of contexts, the research focus varies on measures across years in programs, differences between assessor groups, or comparisons with other assessment methods, so the validity of MSF needs to be investigated further.

The main purpose of this study was to conduct a meta-analysis by identifying all published empirical data on the use of MSF to assess physicians' clinical and nonclinical performance. We conducted a meta-analysis on the construct and criterion (predictive or concurrent) validity of the MSF system as a function of both summary effect sizes, their 95% confidence intervals (CIs), and interpretation of the magnitude of these coefficients.

Materials and methods

Selection of studies

In this present study, we followed the guidelines for reporting of observational studies included in a meta-analysis.[10] In addition to PubMed and MEDLINE, the CINAHL, EMBASE, and PsychINFO databases were searched from January 1975 to November 2012. We also manually searched the reference lists for further relevant studies. The following terms were used in the search: "multi-source feedback", "360-degree evaluation", and "assessment of medical professionalism". Studies were included if: they used at least one MSF instrument (eg, self, colleague, coworker, and/or patient) to assess physician/surgeon performance in practice; they described the MSF instrument or its design; they described factors measured by the MSF instrument; they provided evidence of construct-related and/or criterion-related validity (predictive/concurrent); and they were published in an English-language, peer-reviewed journal. The main reason for restricting the search to refereed journals was to ensure that only studies of high quality were included in the meta-analysis. On the other hand, we excluded studies if they used nonmedical health professionals, did not provide a description or breakdown of what the MSF instrument was measuring, did not provide empirical data on MSF results, reported data on feasibility and/or reliability only, and/or focused on performance changes after receiving MSF feedback.

Data extraction

The initial search yielded 1,137 papers, as shown in Figure 1. Of these, 623 papers were excluded based on the title, 292 were excluded based on a review of the abstract, 97 were removed as they were duplicates, and a further 90 were eliminated after a review of the full-text versions. Finally, we agreed on a total of 35 papers to be included for meta-analysis. A coding protocol was developed that included each study's title, author(s) name(s), year of publication, source of publication, study design (ie, construct or criterion validity study), physician/surgeon specialty (eg, general practice, pediatrics), and types of raters (ie, self, medical colleague, consultants, patients, and coworkers). All 35 articles were independently coded by two authors (AA and TD) and any discrepancies (eg, effect size calculations) were reviewed by a third author (KA, AD, or CV). Based on iterative reviews and discussions between the five coders, we were able to achieve 100% agreement on all coded data.

Statistical analysis

The statistical analysis of all effect size calculations was done using the Comprehensive Meta-Analysis software program (version 1.0.23, Biostat Inc, Englewood, NJ, USA). Most of the studies reported mean differences (Cohen's *d*) between MSF scores as effect size measures. However, there were some studies that reported the Pearson's product-moment correlation coefficient *(r)*. For these studies, and in order to preserve consistency in the data that were reported, *r* was converted to Cohen's *d* using the following formula: $d = 2r/\sqrt{(1 - r^2)}$.[11]

We selected MSF domains or subscale measures as the variables of interest and either contrasted these scores

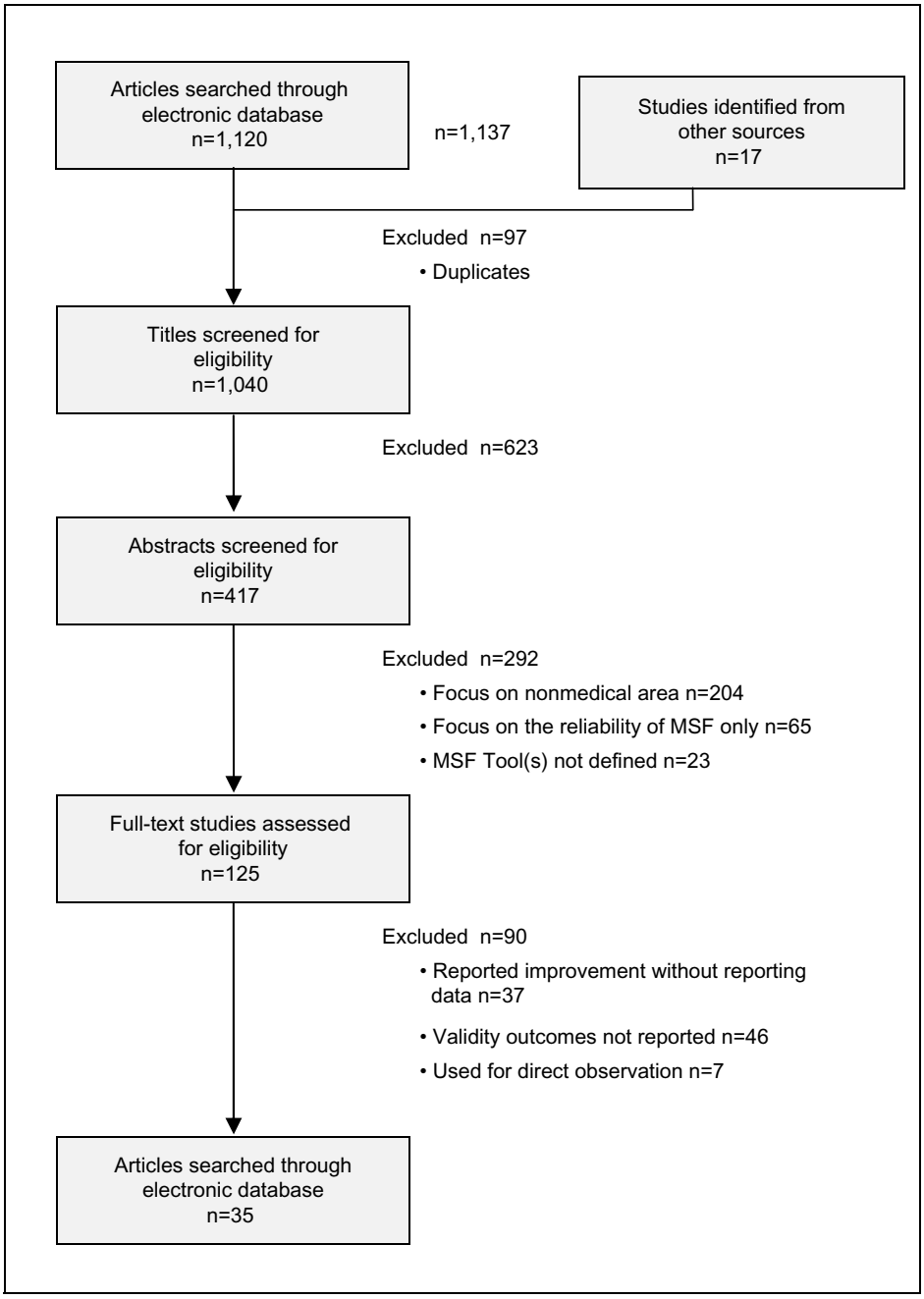

Figure 1 Selection of studies for the meta-analysis.

between assessor groups (eg, different personnel ratings, in-training year, or postgraduate year of practice) or with other measures of clinical performance competencies (eg, ABSITE or Objective Structured Clinical Examination [OSCE]).

On combination of results from studies that used different research designs (eg, different physician year in practice) or different personnel ratings (eg, medical colleagues, coworkers, patients) and methods of analysis between assessor groups (eg, MSF in comparison with ABSITE, as well as an objective structured practical examination [OSPE]), we used a random-effects model in combining the unweighted and weighted effect sizes. The fixed-effects model assumes that the summary effect size differences are the same from study to study (eg, use of MSF with different questionnaires). In contrast, the random-effects model calculation reflects a more conservative estimate of the between-study variance of the participants' performance measures.[12]

In this meta-analysis, residents in different years of rotation and the attending physicians/surgeons were treated

equally in that they represent treating physicians at different stages of their year of practice. Therefore, we are evaluating the performance of these 'physicians/surgeons' that had a more or less similar trajectory in achieving clinical competency as a function of their performance by using the multi-source feedback system.

To assess for the heterogeneity of effect sizes, a forest plot with Cochran *Q* tests was conducted. Absence of a significant *P* value for *Q* indicates low power within studies rather than the actual consistency or homogeneity across studies included in the meta-analysis. In addition, the distribution of the studies in the forest plots was an important visual indicator to measure the consistency between studies. Interpretation of the magnitude of the effect size for both mean differences and correlations are based on Cohen's[13] suggestions, ie, d=0.20 – 0.49 is "small", d=0.50 – 0.79 is "medium", and $d \geq 0.80$ is considered to be a "large" effect size difference.

Results

The characteristics of the 35 studies included in the meta-analysis were based on four groups (Table 1) that reported contrasts between different physician years in practice (group A), differences between physician performance levels on two occasions (group B), rating differences between self, medical colleague, coworker, and patients (group C), and comparisons between MSF and other measures of performance (group D). The reported MSF domain measure (ie, items 1 through 5) and the corresponding unweighted effect sizes based on either the contrast or comparison variables are presented in Table 1. Different approaches to testing the validity of MSF were demonstrated by studies included in this meta-analysis. In groups A and B, we investigated the construct validity of the domains' measures of MSF by showing that physicians at different levels of experience or on two separate occasions tend to obtain higher clinical performance scores. In groups C and D, the criterion validity of MSF is compared with other similar assessments of clinical performance or different raters as either a concurrent or predictive validity measure.

The sample size of the studies range from six plastic surgery residents[14] to 577 pediatric residents[15] who had been assessed using MSF with as few as 1.2 patients and 2.6 medical colleagues[16] and as many as 47.3 patients completing forms per individual.[17] Questionnaire items used as part of MSF ranged from as few as four items[18] to as many as 60 items[14] per questionnaire. Information on specific demographic characteristics, such as students' sex or age was not reported, but level of training and years of practice as a physician were typically identified. In each study, the unweighted mean effect size difference (Cohen's d) was provided or calculated based on the MSF domain measures as a contrasting variable (eg, years spent as a physician in practice) or with a comparison measure (eg, OSPE).

Construct validity of MSF system

Of the 35 studies that reported data on physician/surgeon performance, 31 (88%) demonstrated results in support of the construct validity of the MSF system. As shown in Table 2, we combined five of the studies (group A) to show that for each of the five MSF domains the effect size differences in performance between a year of practice (eg, change in performance as a function of post-graduate year 1 to year 2, Senior House Officer to Specialist Registrar)[8,15,19–21] ranged from d=0.14 (95% CI 0.40–0.69) for manager skills to d=1.78 (95% CI 1.20–2.30) for communication skills.

When differences between physician/surgeon performance were investigated on two different occasions, we found four studies (group B) that showed differences in clinical performance across the five domain scores of MSF. In particular, Brinkman et al[19] compared ratings for 36 pediatric residents on two occasions with regard to the professionalism and communication skills domains, and their results showed that there were consistently large effect size differences between time 1 and time 2. The ratings on these MSF items ranged from d=1.31 for the professionalism domain to d=2.00 for the communication skills domain. Correspondingly, Lockyer et al[22] found a range of MSF scores that varied from d=0.01 for physicians over a 5-year period on the professionalism, communication skills, and management domains for self-rating assessment to d=0.66 with the same physicians over the professionalism, communication skills, and interpersonal relationship domains as rated by medical colleagues. Violato et al[9] reported a small effect size of d=0.46 when the performance of 250 family physicians was compared after a 5-year interval between MSF assessments.

Criterion (predictive/concurrent) validity of the MSF system

In group C, we combined the outcomes in 21 (60%) studies that investigated the differences in MSF scores provided by different raters (eg, residents, self, medical colleague, coworker, patients) across the five domains identified. Effect size differences in performance between the different

Table 1 Characteristics of MSF studies with construct and criterion (concurrent/predictive) validity effect size measures

Study source	Group	Contrast†	MSF domain*	Effect size difference (d_{UWM}‡)
Archer et al[20]	A	SPRS (MC)/SHO (MC)	2, and 5	1.22
Sample size, 112 pediatrics				
(20 specialist registrars, 92 senior house officers)				
Total forms =921				
Brinkman et al[19]	A	Feedback (MC)/No-feedback (MC)	1, 2, and 3	1.8
Sample size, 36 pediatric residents				
(16 with feedback and 16 with no feedback)				
Total forms =1,263				
Massagli and Carline[21]	A	PGY2/PGY3	1, 2, 4, and 5	0.05
Sample size, 56 rehabilitation residents		PGY2/PGY4	1, 2, 4, and 5	0.17
(nine PGY2, nine PGY3, nine PGY4)		PGY3/PGY4	1, 2, 4, and 5	0.23
Total forms =930				
Archer et al[8]	A	Foundation year 1 (MC)/Foundation year 2 (MC)	2, and 5	0.34
Sample size, 553 multiple specialties residents				
(219 Foundation year 1,334 Foundation year 2)				
Total forms =5,544				
Archer et al[15]	A	SPRS year 2 (MC)/SPRS year 4 (MC)	2, and 5	0.29
Sample size, 577 pediatric (343 SPRS year 2,				
201 SPRS year 4, 10 pediatricians in years 1, 3, 5, 6)				
Total forms =4,770				
Wood et al[18]	B	ObGyn time 1/ObGyn time 2	4, and 5	2.41
Sample size, 67 obstetrics and gynecology residents				
Total forms =578				
Lockyer et al[22]	B	Phys time 1/Phys time 2 (Self)	1, 2, 3, and 4	0.46
Sample size, 250 family physicians				
Total forms =500				
Brinkman et al[19]	B	Nurse time 1 (CW)/Nurse time 2	1, and 2	1.31
Sample size, 36 pediatric residents		(CW)	1, and 2	2.00
Total forms =1,263		(Parents) time 1/(Parents) time 2		
Violato et al[9]	B	Phys time 1/Phys time 2 (MC)	1, 2, and 5	0.66
Sample size, 250 family physicians		Phys time 1/Phys time 2 (CW)	1, and 3	0.22
Total forms =20,500		Phys time 1/Phys time 2 (Patients)	1, 3, and 4	0.01
Risucci et al[7]	C	Self/Peer (MC)	1, 2, and 5	0.56
Sample size, 32 surgical residents		Self/Supervisors (MC)	1, 2, and 5	0.21
Total forms =1,024		Peer (MC)/Supervisors (MC)	1, 2, and 5	0.25
Wenrich et al[41]	C	Nurse (CW)/Phys (MC) medical	2, and 5	0.51
Sample size, 318 internal medicine physicians		knowledge	2, and 5	–0.46
Total forms =1,877		Nurse (CW)/Phys (MC) humanistic		
Lelliott et al[42]	C	Self/MC	2, 3, and 5	0.47
Sample size, 347 psychiatrists		Patients/MC	2, 3, and 5	0.85
Total forms =11,426				
Violato et al[43]	C	Self/MC	1, 2, 4, and 5	0.58
Sample size, 28 family physicians		Self/Patients	1, 2, 4, and 5	0.95
Total forms =170		Self/CW	1, 2, 3, and 5	0.77
Hall et al[3]	C	Self/Patients	1, 2, 3, 4, and 5	1.30
Sample size, 295 multiple specialties		Self/MC	1, 2, and 5	0.37
Physicians		Self/Consultant (MC)	1, 2, and 5	0.80
Total forms =11,665		Self-Referring physicians (MC)	1, 2, and 5	1.18
		Self/CW	1, 2, 3, and 5	0.76
		Consultant (MC)/MC	1, 2, and 5	0.46
		Consultant (MC)/CW	1, 2, 3, and 5	0.18
Thomas et al[44]	C	MC (Intern)/MC	2, and 5	0.41
Sample size, 16 internal medicine residents		MC (Intern)/CW	2, and 5	1.06
Total forms =177		MC/CW	2, and 5	0.65
Lipner et al[45]	C	MC/Patients	1, 2, and 3	2.60
Sample size, 356 internal medicine physicians				
Total forms =12,460				

(*Continued*)

Table 1 (*Continued*)

Study source	Group	Contrast†	MSF domain*	Effect size difference (d_{UWM}‡)
Violato et al[5]	C	Self/MC	1, 2, 3, and 5	0.62
Sample size, 252 surgeons		Self/CW	1, 2, 3, and 5	0.61
Total forms =7,237		Self/Patients	1, 2, 3, 4, and 5	0.58
		MC/CW	1, 2, 3, and 5	0.00
		MC/Patients	1, 2, 3, 4, and 5	0.00
		CW/Patients	3, 4, and 5	0.00
Wood et al[27]	C	Patients/MC	1, and 3	0.98
Sample size, 7 radiology residents		Patients/CW	1, and 3	1.31
Total forms =57		MC/CW	1, and 3	0.04
Joshi et al[46]	C	MC/CW	3, and 5	1.34
Sample size, 8 obstetrics/gynecology residents		MC/Patients	3, and 5	0.43
Total forms =512		CW/Patients	3, and 5	0.97
Lockyer et al[47]	C	MC/Patients	1, 2, and 3	0.06
Sample size, 197 anesthesiology physicians				
Total forms =5,957				
Violato et al[48]	C	Self/MC	1, 2, and 3	0.04
Sample size, 100 pediatric physicians		Self/CW	1, 2, 3, and 5	0.18
Total forms =3,963		Self/Patients	1, 2, 3, and 4	0.07
		MC/CW	1, 2, 3, and 5	0.97
		MC/Patients	1, 2, 3, and 4	0.79
		CW/Patients	1, 3, 4, and 5	0.26
Violato et al[32]	C	Self/MC	1, 2, and 4	0.83
Sample size, 101 psychiatry physicians		Self/CW	1, 2, 3, 4, and 5	1.52
Total forms =4,069		Self/Patients	1, 2, 3, and 4	1.13
		MC/CW	1, 2, 3, and 5	0.68
		MC/Patients	1, 2, 3, and 4	0.28
		CW/Patients	1, 3, 4, and 5	0.40
Archer et al[8]	C	(Consultant) MC/(Resident) MC	2, and 5	0.37
Sample size, 553 multiple specialties residents				
Total forms =5,544				
Pollock et al[14]	C	CW/MC	1, 2, 3, 4, and 5	0.87
Sample size, 6 plastic surgery residents				
Total forms =240				
Davies et al[40]	C	Consultant (MC)/CW	2, and 4	0.98
Sample size, 92 histopathology residents				
Total forms =1,012				
Campbell et al[33]	C	Patients/MC	1, 2, 3, and 5	0.19
Sample size, 291 multiple specialties physicians				
Total forms =18,023				
Meng et al[34]	C	Nurse (CW)/Secretaries (CW)	1, 3, and 5	0.16
Sample size, 15 anesthesiology residents		Nurse (CW)/Nurse aids (CW)	1, 3, and 5	0.64
Total forms =429		Nurse (CW)/Technicians (CW)	1, 3, and 5	0.65
		Secretaries (CW)/Nurse aids (CW)	1, 3, and 5	0.16
		Secretaries (CW)/Technicians (CW)	1, 3, and 5	0.46
		Nurse aids (CW)/Technicians (CW)	1, 3, and 5	0.00
Lockyer et al[35]	C	Self/MC	1, 2, and 5	0.22
Samples size, 101 pathologists/laboratory physicians		Self/Referring physicians (MC)	1, 2, 4, and 5	0.58
Total forms =808		Self/CW	1, 2, 3, and 5	0.18
		MC/Referring physicians (MC)	1, 2, 4, and 5	0.38
		MC/CW	1, 2, 3, and 5	0.03
		Referring physicians (MC)/CW	1, 2, 3, and 4	0.40

(*Continued*)

Table 1 (*Continued*)

Study source	Group	Contrast†	MSF domain*	Effect size difference (d_{UWM}‡)
Lockyer et al[36]	C	Self/MC	1, 2, and 4	0.78
Sample size, 187 emergency medicine physicians		Self/CW	1, 2, 4, and 5	0.93
Total forms =6,889		Self/Patients	1, 2, 3, 4, and 5	1.13
		MC/CW	1, 2, 4, and 5	0.43
		MC/Patients	1, 2, 3, 4, and 5	0.63
		CW/Patients	1, 2, 3, and 5	0.17
Archer et al[15]	C	Consultant (MC)/Resident (MC)	2, and 5	0.64
Sample size, 577 pediatric residents				
Total forms =4,770				
Chandler et al[16]	C	Self/Attending (MC)	3, and 5	0.87
Sample size, 66 pediatrics residents		Self/CW	3, and 5	1.10
Total forms =823		Self/Patients	3, and 5	0.08
		Attending (MC)/CW	3, and 5	0.26
		Attending (MC)/Patients	3, and 5	0.30
		CW/Patients	3, and 5	0.45
Campbell et al[17]	C	Patients/MC	1, 2, 3, and 5	0.02
Sample size, 179 family physicians				
Total forms =10,895				
Archer and McAvoy[37]	C	Patients/MC	2, and 5	1.90
Sample size, 68 different specialties physicians		Assessor nominated by physicians/ assessors nominated by referring body	2, and 5	1.91
Total forms =2,365				
Overeem et al[38]	C	MC/Patients	1, 2, 3, 4, and 5	0.44
Sample size, 146 multiple specialties		MC/CW	1, 2, 3, and 4	0.75
Physicians		CW/Patients	1, 2, 3, and 5	0.45
Total forms =3,648				
Lockyer et al[39]	C	Self/MC	1, 2, 3, and 4	1.11
Sample size, 216 surgeons		Self/CW	1, 2, and 3	0.86
Total forms =9,072		Self/Patients	1, 2, 3, 4, and 5	1.00
		MC/CW	1, 2, 3, and 4	0.44
		MC/Patients	1, 2, 3, 4, and 5	0.30
		CW/Patients	3, 4, and 5	0.21
Qu et al[23]	C	Self/Attending (MC)	1, and 3	0.30
Sample size, 258 multiple specialties residents		Self/MC	1, and 3	0.13
Total forms =4,128		Self/CW	1, and 3	–0.55
		Self/Patients	1, 2, 3, 4, and 5	0.19
		Self/Office staff (CW)	1, and 3	1.78
		Attending (MC)/MC	1, and 3	0.08
		Attending (MC)/CW	1, and 3	0.82
		Attending (MC)/Patients	1, 2, 3, 4, and 5	0.38
		Attending (MC)/Office staff (CW)	1, and 3	2.31
		Patients/Office staff (CW)	1, 2, 3, 4, and 5	1.87
		Patients/MC	1, 2, 3, 4, and 5	0.37
		Patients/CW	1, 2, 3, 4, and 5	0.42
Lockyer et al[49]	C	Self/MC	1, and 2	0.22
Sample size, 37 general practice physicians		Self/CW	1, 2, and 3	0.05
Total forms =1,130		Self/Patients	1, 2, 3, and 4	0.04
		MC/CW	1, 2, and 3	0.22
		MC/Patients	1, 2, 3, and 4	0.21
		CW/Patients	1, 3, and 4	0.00
Risucci et al[7]	D	MSF/ABSITE	1, 2, and 5	1.45
Sample size, 32 surgical residents				
Total forms =1,024				
Wood et al[27]	D	MSF (PT)/global examination	1, and 3	1.96
Sample size, 7 radiology residents		MSF (MC)/global examination	1, and 3	1.02
Total forms =57		MSF (CW)/global examination	1, and 3	1.60

(*Continued*)

Table 1 (*Continued*)

Study source	Group	Contrast†	MSF domain*	Effect size difference (d_{UWM}‡)
Davies et al[40] Sample size, 92 histopathology residents Total forms =1,012	D	MSF (PATH-SPRAT)/OSPE	2, and 3	1.09
Yang et al[24] Sample size, 245 multiple specialties residents Total forms =1,053	D	MSF/small scale OSCE MSF/small scale OSCE + DOPS	1, 2, and 3 1, 2, and 3	0.79 2.07

Notes: †A, predictive validity (physicians in different years level); B, predictive validity (physicians performance on MSF in two occasions separated with time); C, concurrent validity (differences in personnel ratings); D, construct validity (comparing MSF with standardized measures). *MSF domains consist of the following: 1= professionalism, covering psychosocial skills, psychosocial management, humanistic qualities, compassion, attitude, professional development, teaching, and professional responsibilities and professional managements; 2= clinical competence covering clinical care, good medical practice, patient care, safe practice, clinical performance, knowledge, critical thinking, diagnosis, and management of complex problem; 3= communication, covering communication with staff and interpersonal communication skills; 4= management, covering reporting, self-management, administrative skills, office personal, access to doctor, practice process, physical office, and physical space; and 5= interpersonal relationships, covering relationships with patients, colleagues, family members, collegiality, collaboration, patient education, information provision, and patient interaction. Two of the authors (AA, TD) agreed on the names of the main five domains and agreed on the items included. *dUWM*‡ refers to the unweighted mean effect size difference as defined by Cohen's d.
Abbreviations: CW, coworkers; MC, medical colleagues; MSF, multi-source feedback; PGY, postgraduate year; SPRS, specialist registrar; Phys, family physician; ObGyn, obstetrics and gynecology; CW, coworkers; ABSITE, American Board Of Surgery In-Training Examination; PATH-SPRAT, Pathology-Sheffield Peer Review Assessment Tool; OSPE, Objective Structured Practical Examination; OSCE, Objective Structured Clinical Examination; DOPS, Direct Observation of Procedural Skills; SHO, senior house officer; PT, patients.

raters (eg, comparison of patients with self assessment, medical colleagues to coworkers) ranged from d=0.50 (95% CI 0.47–0.52) for interpersonal relationships to d=0.57 (95% CI 0.55–0.60) for both professionalism and clinical competence. Most of the studies in group C showed that physicians consistently rated themselves lower than did other assessor groups. However, in a study of 258 residents within different specialties reported by Qu et al, residents on self-assessments rated themselves higher than did other raters.[23] As shown in the forest plot (Figure 2), the combined random-effects size calculation for the professionalism domain was "medium" (d=0.66, 95% CI 0.44–0.69).

In group D (Table 3), of the 35 studies included in the meta-analysis, four reported data on physician/surgeon performance on MSF in comparison with other criterion measures (eg, OSPE, OSCE). The mean effect size differences were found to be "medium" to "high" across each of the five domains identified on MSF. Effect size differences in performance between domain scores and other examination measurement scores ranged from d=1.28 (95% CI 1.15–1.41) for clinical competence to d=1.43 (95% CI 0.87–2.00) for interpersonal relationships. Yang et al[24] found a range of MSF scores that varied from d=0.79 for residents on the domains of professionalism, clinical competence, and communication skills to d=2.07 with the same physicians on the same domains when their MSF scores were compared with other clinical performance measures such as the OSCE.

Although the Cochran Q test shows significant heterogeneity between the studies included in the four groups, a subgroup analysis to determine the potential differences as a result of moderator variables such as physician/surgeon sex or age was limited by the data reported across the primary studies included in the meta-analysis. Nevertheless, the studies were weighted by their respective sample sizes, and the random-effects model analysis (with greater than 95% CIs) provide a more conservative estimate of the combined effect sizes as illustrated by a forest plot (Figure 2).

Table 2 Random effects model (Cohen's *d*) of the MSF domains with different physician years (group A)/different physician performance in two occasions (group B)

MSF domain measure	Studies included (number of outcomes)	Sample size	MSF with different physician years*	Studies included (number of outcomes)	Sample size	Difference between physicians' performance on two occasions**
Professional	2 (4)	126	0.56 (0.39–1.59)	3 (6)	1,054	0.65 (0.30–1.00)
Clinical competence	5 (7)	1,335	0.62 (0.25–1.00)	3 (4)	554	0.99 (0.53–1.45)
Communication	1 (1)	72	1.78 (1.22–2.34)	2 (3)	750	0.23 (0.02–048)
Manager	1 (3)	54	0.14 (0.40–0.69)	3 (3)	567	0.92 (0.01–1.84)
Interpersonal relationships	4 (6)	1,263	0.42 (0.16–0.67)	2 (2)	317	1.50 (0.19–3.22)

Notes: *Effect sizes combined for physicians in different year levels (different PGY level, eg, year 1, year 2, senior house officer, specialist registrar);[8,15,19–21] **effect sizes combined for physicians' performance on two occasions separated by time (eg, 5 years, 7 months, 7 years).[9,18,19,22]
Abbreviations: MSF, multi-source feedback; PGY, post graduate year.

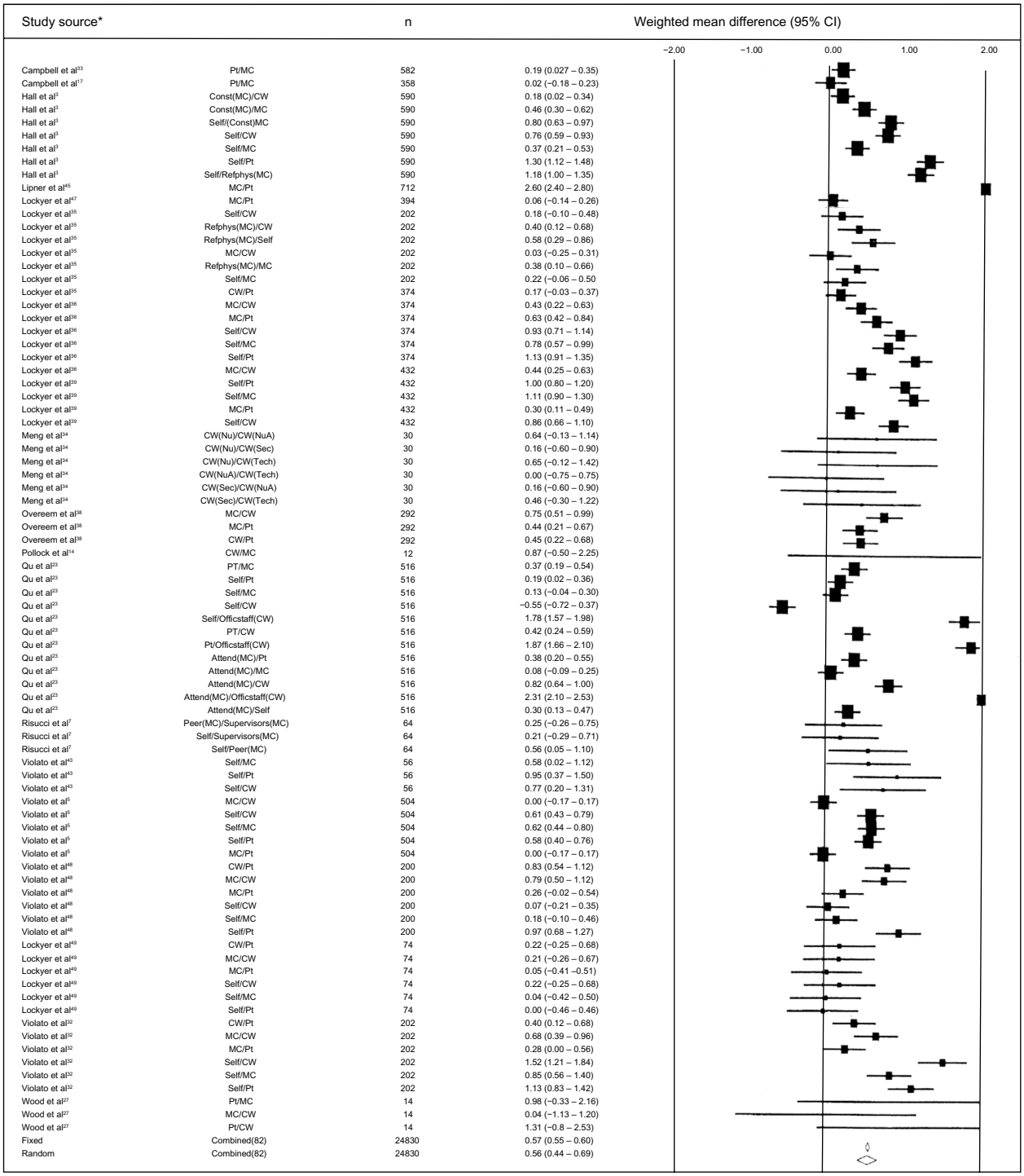

Figure 2 Random and fixed effects model forrest plots for the MSF "personnel rating differences" for professional measures.

Notes: *The effect size values are taken from the raw data reported for the outcomes in studies group C. The Cochran *Q*-test for heterogeneity shows significant overall heterogeneity between studies.

Abbreviations: MSF, multi-source feedback; Pt, patients; MC, medical colleagues; Const, consultant; CW, co-workers; RefPhys, referring physicians; Nu, nursing; NuA, nursing aid; Sec, secretary; Tech, technicians; Officstaff, office staff; Attend, attending.

Discussion

In this meta-analysis, the MSF demonstrates evidence of construct validity when used with physicians and surgeons across the years of a residency program or a number of years of practice. Physician/surgeon performance on the MSF domains across a single year of practice showed "small" to "large" effect size differences, with effect sizes ranging from d=0.14 (95% CI 0.40–0.69) in the manager skills domain to d=1.78 (95% CI 1.20–2.30) in the communication skills domain.

Table 3 Random effects model (Cohen's *d*) of the MSF domains with personnel ratings/academic performance (groups C and D)

MSF domain measure	Studies included (number of outcomes)	Sample size	Personnel rating differences*	Studies included (number of outcomes)	Sample size	MSF with different global measurement**
Professional	19 (82)	12,415	0.56 (0.44–0.67)	3 (6)	543	1.42 (0.72–2.12)
Clinical competence	24 (75)	12,720	0.60 (0.49–0.72)	3 (4)	614	1.34 (0.65–2.05)
Communication	20 (76)	11,280	0.56 (0.42–0.67)	3 (6)	603	1.35 (0.71–1.99)
Manager	13 (38)	6,089	0.60 (0.45–0.74)	–	–	
Interpersonal relationships	23 (74)	11,660	0.54 (0.44–0.64)	1 (1)	32	1.43 (0.87–2.00)

Notes: *Effect size combined between differences in personnel ratings (ie, resident versus faculty, specialist versus consultant);[3,5,7,8,14–17,23,27,32–39,41–49] **effect sizes combined between MSF with standardized measures (eg, global ratings, OSPE).[7,24,27,40]
Abbreviation: MSF, multi-source feedback; OSPE, Objective Structured Practical Examination.

The effect size differences between physician/surgeon performance on two occasions (time 1/time 2) ranged from *d*=0.23 (95% CI 0.13–0.33) for the communication skills domain to *d*=0.90 (95% CI 0.74–1.10) for the interpersonal relationship domain measure.

The differences in rating for physician/surgeon performance on MSF between different assessor groups (self-assessments, medical colleagues, consultants, patients, and coworkers) showed "medium" effect size differences that ranged from *d*=0.50 (95% CI 0.47–0.52) for the interpersonal relationship domain to *d*=0.57 (95% CI 0.55–0.60) for the professionalism and clinical competence domains. In particular, these results were supported by the findings from other assessment methods such as the mini-clinical evaluation exercise (mini-CEX). Ratings with different raters in the mini-CEX have showed that in comparison with faculty evaluator ratings, residents tend to be more lenient and score trainees higher on in-training evaluation checklists.[25,26] In our study of the MSF, we found that physicians and surgeons consistently rated themselves lower than did other assessor groups.[23] In addition, patients and coworkers typically rated physicians/surgeons more leniently than did other raters, such as medical colleagues or consultants.

The MSF showed evidence of criterion-related validity when compared with other performance examination measures (eg, global examination, OSPE, OSCE). We found a "large" correlation coefficient, with combined effect sizes ranging from *d*=1.28 (95% CI 1.15–1.41) for the communication skills domain to *d*=1.43 (95% CI 0.87–2.00) for the interpersonal relationship domain.

The construct-related and criterion-related validity of MSF was supported by the findings outlined within the studies included in one or more of the four group comparisons. As illustrated in the forrest plots for the professionalism domain in group C, not all of the reported differences between personnel ratings were found to be statistically significant. When combined with the outcomes from 19 different studies, however, we found that there was a significant combined random-effects size of *d*=0.65 (95% CI, 0.44–0.69).

In general, the findings of this meta-analysis shows "medium" combined effect sizes for the construct-related and criterion-related validity of the five main MSF domains identified. Although different questionnaires and different numbers of items were used in MSF across different specialties, they were found to consistently measure similar domains of physician/surgeon performance.[15] This feedback process using multiple questionnaires in different type of raters provides a more comprehensive evaluation of clinical practice than can typically be provided by one or few sources.[1]

Strengths and weaknesses of the study

There are limitations to this meta-analysis. Because we were interested in determining the construct-related and criterion-related validity of MSF as a method for physician/surgeon evaluation, consistency in the use of the evaluation tool varied from a research design perspective. In addition, there was variability in the performance domains measured and in the number of items used to measure each domain depending on the MSF instrument used (ie, ranging from four items to 60 items), the raters used (ie, self, patients, medical colleague, coworker), and whether or not the MSF was being compared with other clinical skill measures (ie, OSCE). To overcome this limitation, the more conservative random-effects size analysis was performed to accommodate for the heterogeneity between the studies as indicated by the significant values obtained using the Cochran *Q* test. Nevertheless, we were unable to undertake subsequent subgroup analyses to determine where there may have been between-study differences because these data (eg, sex, age of participant) were rarely reported. Although some of

the studies had small sample sizes such as six[14] and seven participants,[27] this was in part compensated by the 40 and eight raters who completed the questionnaire, respectively, on each of the participants in these studies. To achieve some control over the quality of the studies that were included in this meta-analysis, only papers that had been published in refereed journals were selected.

Implications for clinicians and policymakers

Certain characteristics of health professionals, such as clinical skills, personal communication, and client management, combined with improved performance can be assessed using MSF.[8] MSF is a unique form of assessment that has been shown to have both construct-related and criterion-related validity in assessing a multitude of clinical and nonclinical performance domains. In addition, MSF has been shown to enhance changes in clinical performance,[15] communication skills,[7] professionalism,[7] teamwork[28], productivity,[29] and building trusting relationship with patients.[30]

Consequently, MSF has been adopted and used extensively as a method for assessment of a variety of domains identified in medical education programs and licensing bodies in the UK, Canada, Europe, and other countries as well. Although MSF has gained widespread acceptance, the literature has raised a number of concerns about its implementation and its validity. Therefore, the availability of evidence to support the validity of the process and the instruments used to date is of crucial importance to enable policymakers to make the decision to implement MSF within their own programs or organizations.

Conclusion and future research

Although MSF appears to be adequate for assessment of a variety of nontechnical skills, this approach is limited to feedback from peers or medical colleagues abilities to assess aspects of clinical skills competence that reflect physicians'/surgeons' knowledge and non-cognitive behavior. In particular, as part of the process of assessing clinical performance, other methods such as procedures-based assessment or the OSCE should be used in conjunction with the peer MSF questionnaire to ensure accurate assessment of these specific skills.

We are faced with the challenge of ensuring that use of MSF for assessment of physicians and surgeons in practice is reliable and valid. As shown above, MSF has proved to be a useful method for assessing the clinical and nonclinical skills of physicians/surgeons in practice with clear evidence of construct and criterion-related validity. Although MSF is considered to be a useful assessment method, it should not be the only measure used to assess physicians and surgeons in practice. Other reliable and valid methods should be used in conjunction with MSF, in particular to assess procedural skills performance and to overcome the limitation of using a single measure.

Future research should be considered by researchers in order to replicate and extend some of the empirical findings, especially the evidence for criterion-related validity. Criterion-related validity studies looking at correlations between direct observations of behavior or performance and MSF scores are required to add further evidence of validity. Future research on the various MSF instruments available may well include confirmatory factor analysis, which provides stronger construct validity evidence than the principal component factor analyses conducted currently.[31] In addition, MSF assessments are entirely questionnaire-based and rely on the judgment of and inference by the assessors and respondents, which are subject to a variety of biases and heuristics. Therefore, generalizability theory should be used in future studies to determine potential sources of error measurement that can occur due to use of different assessors and specialties, as well as the characteristics of the respondents themselves.

Author contributions

All authors contributed to the conception and design of the study. AA and TD acquired the data. All authors analyzed and interpreted the data. KA, AD, and CV provided administrative, technical, or material support and supervised the study. AA and TD drafted the manuscript. All of the authors critically revised the manuscript for important intellectual content and approved the final version submitted for publication.

Disclosure

All authors have no competing interests in this work.

References

1. Bracken DW, Timmreck CW, Church AH. Introduction: a multisource feedback process model. In: Bracken DW, Timmreck CW, Church AH, editors. *The Handbook of Multisource Feedback: The Comprehensive Resource for Designing and Implementing MSF Processes*. San Francisco, CA, USA; Jossey-Bass: 2001.
2. Lockyer J, Clyman S. Multisource feedback (360-degree evaluation). In: Holmboe ES, Hawkins RE, editors. *Practical Guide to the Evaluation of Clinical Competence*. Philadelphia, PA, USA; Mosby: 2008.
3. Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: the Physician Achievement Review. *CMAJ*. 1999;161:52–57.

4. Fidler H, Lockyer J, Violato C. Changing physicians practice: the effect of individual feedback. *Acad Med*. 1999;74:702–714.
5. Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. *BMJ*. 2003;326:546–548.
6. Miller A, Archer J. Impact of workplace based assessment on doctors education and performance: a systematic review. *BMJ*. 2010; 341:1–6.
7. Risucci DA, Tortolani AJ, Ward RJ. Ratings of surgical residents by self, supervisors and peers. *Surg Gynecol Obstet*. 1989;169:519–526.
8. Archer J, Norcini J, Southgate L, Heard S, Davies H. Mini-PAT (Peer Assessment Tool): a valid component of a national assessment programme in the UK? *Adv Health Sci Educ Theory Pract*. 2008;13:181–192.
9. Violato C, Lockyer J, Fidler H. Change in performance: a 5-year longitudinal study of participants in a multi-source feedback programme. *Med Educ*. 2008;42:1007–1013.
10. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. *JAMA*. 2000;238:2008–2012.
11. Rosenthal R, Rubin D. A simple general purpose display of magnitude of experimental effect. *J Educ Psychol*. 1982;74:166–169.
12. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials*. 1986;7:177–188.
13. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ, USA: Lawrence Earlbaum Associates; 1988.
14. Pollock RA, Donnelly MB, Plymale MA, Stewart DH, Vasconez HC. 360-degree evaluations of plastic surgery resident accreditation council for graduate medical education competencies: experience using a short form. *Plast Reconstr Surg*. 2008;122:639–649.
15. Archer J, McGraw M, Davies H. Assuring validity of multisource feedback in a national programme. *Postgrad Med J*. 2010;86:526–531.
16. Chandler N, Henderson G, Park B, Byerley J, Brown WD, Steiner MJ. Use of a 360-degree evaluation in the outpatient settings: the usefulness of nurse, faculty, patient/family, and resident self-evaluation. *J Grad Med Educ*. 2010;10:430–434.
17. Campbell J, Narayanan A, Burford B, Greco M. Validation of a multisource feedback tool for use in general practice. *Educ Prim Care*. 2010;21:165–179.
18. Wood L, Wall D, Bullock A, Hassell A, Whitehouse A, Campbell I. 'Team observation': a six-year study of the development and use of multi-source feedback (360-degree assessment) in obstetrics and gynecology training in the UK. *Med Teach*. 2006;28:e177–e184.
19. Brinkman WB, Geraghty SR, Lanpher BP, et al. Effect of multisource feedback on resident communication skills and professionalism. *Arch Pediatr Adolesc Med*. 2007;161:44–49.
20. Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. *BMJ*. 2005;330:1251–1253.
21. Massagli TL, Carline JD. Reliability of a 360-degree evaluation to assess resident competence. *Am J Phys Med Rehabil*. 2007;86:845–852.
22. Lockyer J, Violato C, Fidler H. What multisource feedback factors influence physicians' self-assessments? A five-year longitudinal study. *Acad Med*. 2007;82:77–80.
23. Qu B, Zhao YH, Sun BZ. Assessment of resident physicians in professionalism, interpersonal and communication skills: a multisource feedback. *Int J Med Sci*. 2012;9:228–236.
24. Yang YY, Lee FY, Hsu HC, et al. Assessment of first-year postgraduate residents: usefulness of multiple tools. *J Chin Med Assoc*. 2011;74:531–538.
25. Hatala R, Norman GR. In-training evaluation during an internal medicine clerkship. *Acad Med*. 1999;74(Suppl 10):118S–120S.
26. Hill F, Kendall K, Galbraith K, Crossley J. Implementing the undergraduate mini CEX: a tailored approach at Southampton University. *Med Educ*. 2009;43:326–334.
27. Wood J, Collins J, Burnside ES, et al. Patient, faculty, and self-assessment of radiology resident performance: a 360-degree method of measuring professionalism and interpersonal/communication skills. *Acad Radiol*. 2004;11:931–939.
28. Dominick P, Reilly R, McGourty J. The effects of peer feedback on team member Behaviour. *Group and Organization Management*. 1997;22:508–520.
29. Edwards M, Ewen A. *360 Feedback: The Powerful New Model for Employee Assessment and Performance Improvement*. New York, NY, USA: Amacom; 1996.
30. Waldman D. Predictors of employee preferences for multi-rater and group-based performance appraisal. *Group and Organization Management*. 1997;22:264–287.
31. Violato C, Hecker K. How to use structural equation modeling in medical education research: a brief guide. *Teach Learn Med*. 2008;19: 362–371.
32. Violato C, Lockyer J, Fidler H. Assessment of psychiatrists in practice through multisource feedback. *Can J Psychiatry*. 2008;53:525–533.
33. Campbell JL, Richards SH, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires. *Qual Saf Health Care*. 2008;17: 187–193.
34. Meng L, Metro DG, Patel RM. Evaluating professionalism and interpersonal and communication skills: implementing a 360-degree evaluation instrument in an anesthesiology residency program. *J Grad Med Educ*. 2009;10:216–220.
35. Lockyer J, Violato C, Fidler H, Alakija P. The assessment of pathologists/laboratory medicine physicians through a multisource feedback tool. *Arch Pathol Lab Med*. 2009;133:1301–1308.
36. Lockyer J, Violato C, Fidler H. The assessment of emergency physicians by a regulatory authority. *Acad Emerg Med*. 2006;13:1296–1303.
37. Archer J, McAvoy P. Factors that might undermine the validity of patient and multi-source feedback. *Med Educ*. 2011;45:886–893.
38. Overeem K, Wollersheim HC, Arah OA, Cruijsberg JK, Grol RP, Lombarts KM. Evaluation of physicians' professional performance: an iterative development and validation study of multisource feedback instruments. *BMC Health Serv Res*. 2012;12:1–11.
39. Lockyer J, Violato C, Wright B, Fidler H, Chan R. Long-term outcomes for surgeons from 3- and 4-year medical school curricula. *Can J Surg*. 2012;55:1–5.
40. Davies H, Archer J, Bateman A, et al. Specialty-specific multisource feedback: assuring validity, information training. *Med Educ*. 2008;42:1014–1020.
41. Wenrich MD, Carline JD, Giles LM, Ramsey PG. Ratings of the performances of practicing internists by hospital-based registered nurses. *Acad Med*. 1993;68:680–687.
42. Lelliott P, Williams R, Mears A, et al. Questionnaires for 360-degree assessment of consultant psychiatrists: development and psychometric properties. *Br J Psychiatry*. 2008;193:156–160.
43. Violato C, Marini A, Towes J, et al. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. *Acad Med*. 1997;72:82–84.
44. Thomas PA, Gebo KA, Hellmann DB. A pilot study of peer review in residency training. *J Gen Intern Med*. 1999;14:551–554.
45. Lipner RS, Blank LL, Leas BF, Fortna GS. The value of patient and peer ratings in recertification. *Acad Med*. 2002;77 Suppl 10:64–66.
46. Joshi R, Ling FW, Jaeger J. Assessment of a 360-degree instrument to evaluate residents' competency in interpersonal and communication skills. *Acad Med*. 2004;79:458–463.
47. Lockyer J, Violato C, Fidler H. A multi-source feedback program for anesthesiology. *Can J Anesth*. 2006;53:33–39.
48. Violato C, Lockyer J, Fidler H. Assessment of pediatricians by a regulatory authority. *Pediatrics*. 2006;117:796–802.
49. Lockyer JL, Blackmore D, Fidler H, et al. A study of a multi-source feedback system for international medical graduates holding defined licences. *Med Educ*. 2006;40:340–347.