Multisource Feedback to Assess Surgical Practice: A Systematic Review

Khalid Al Khalifa, FRCSI,* Ahmed Al Ansari, MBBCh,* Claudio Violato ,† and Tyrone Donnon†

*Department of General Surgery, Bahrain Defense Force Hospital, Bahrain; and †Department of Community Health Sciences, Faculty of Medicine, University of Calgary, Calgary, Canada

BACKGROUND: The assessment, maintenance of competence, and recertification for surgeons have recently received increased attention from many health organizations. Assessment of physicians' competencies with multisource feedback (MSF) has become widespread in recent years. The aim of the present study was to investigate further the use of MSF for assessing surgical practice by conducting a systematic review of the published research.

METHODS: A systematic literature review was conducted to identify the use of MSF in surgical settings. The search was conducted using the electronic databases EMBASE, PsycINFO, MEDLINE, PubMed, and CINAHL for articles in English up to August 2012. Studies were included if they reported information about at least 1 out of feasibility, reliability, generalizability, and validity of the MSF.

RESULTS: A total of 780 articles were identified with the initial search and 772 articles were excluded based on the exclusion criteria. Eight studies met the inclusion criteria for this systematic review. Reliability (Cronbach $\alpha \geq 0.90$) was reported in 4 studies and generalizability ($Ep^2 \geq 0.70$) was reported in 4 studies. Evidence for content, criterion-related, and construct validity was reported in all 8 studies.

CONCLUSION: MSF is a feasible, reliable, and valid method to assess surgical practice, particularly for nontechnical competencies such as communication skills, interpersonal skills, collegiality, humanism, and professionalism. Meanwhile, procedural competence needs to be assessed by different assessment methods. Further implementation for the use of MSF is desirable. (J Surg 70:475-486. © 2013 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: multisource feedback, assessment, competence, professionalism

COMPETENCIES: Systems-Based Practice, Practice-Based Learning and Improvement, Professionalism

The assessment and maintenance of competence of surgeons has received great interest from healthcare organizations in recent years.[1] This interest developed in response to concern about surgeons' perfomance,[2] patient safety,[3] and healthcare organization satisfaction. Surgeons have very little opportunity to receive systematic feedback about their practices. This is particularly true for nontechnical competencies like professionalism, communication skills, humanism, and interpersonal relationships.[4]

Multisource feedback (MSF) (also called 360° assessment) has emerged as a common method for assessing professional attitudes, behaviors, and competence in the workplace both in healthcare and industry.[5] MSF has gained widespread acceptance for both formative and summative assessment of professionals and can be a stimulus for reflecting on where change is required.[5] Research, in both industry and healthcare, has demonstrated that this method of assessment is practical, valid, and reliable when applied appropriately.[5]

MSF has been widely implemented in industry as a way of providing feedback to employees to guide self-directed learning and improve workplace performance.[6] The feedback in industrial settings differs from that in medical settings. MSF is used more frequently in industry where the employee works in a team or cannot be directly and easily supervised by managers or both.[7] In such settings supervisors, peers, and occasionally clients provide feedback. However, in medical settings, physicians complete a self-assessment instrument and receive feedback from medical colleagues (peers), nonmedical coworkers (e.g., office staff and secretaries), coworkers (e.g., nurses and physiotherapists), and patients.[8] This feedback system using questionnaires by different personnel (the assessed person as well as colleagues, peers, and clients) provides a more global perspective than can be provided by 1 or a few sources

Correspondence: Inquiries to Ahmed Al Ansari, MBBCh, MRCSI, MHPE, University Ambrosiana and University of Calgary, G15, Heritage Medical Research Centre, Faculty of Medicine, 3330 Hospital Drive NW, Calgary, AB, Canada T2N 1N4; e-mail: drahmedalansari@gmail.com

alone.[9] Certain characteristics of health professionals such as clinical skills, personal communication, and patient or client management combined with improved performance can be assessed by MSF.

MSF is gaining acceptance and credibility as a means of providing physicians and surgeons with the necessary information that helps them in monitoring and improving their performance and maintaining competence. Therefore, some postgraduate training programs and licensing bodies have made new efforts to implement MSF systems to recertify surgeons every 5 years.[1] Numerous studies have now been conducted on MSF in healthcare professionals generally and physicians in particular. Several studies of MSF have also been conducted with surgeons[1] but there is not yet clear evidence about its effectiveness for assessing various competencies such as professionalism, communication skills, medical knowledge, surgical skills, and interpersonal relationships. Accordingly, we wished to review and summarize the research in MSF for assessing surgical practice. The main purpose of the present study, therefore, was to conduct a systematic literature review to describe the use of MSF in surgical settings and to determine the psychometric characteristics and the evidence of its validity based on the published literature.

METHODS

The guidelines of the Preferred Reporting Items for Systematic reviews and Meta-Analyses were followed for this systematic review.[10]

Information Sources and Search

A systematic literature search was conducted for studies in English published from 1975 to 2012 for the following databases: MEDLINE, EMBASE, CINAHL, PubMed, and PsychINFO. The potential articles from the reference lists of selected articles were searched as well. The following terms were used in the search: MSF, MSF in surgical settings, 360° evaluation, and 360° evaluation in surgical settings.

Study Selection Criteria

Studies were included if they (1) described the instrument design, (2) identified factors measured by the instruments, (3) employed surgeons or surgical practice, (4) included information about at least 1 out of feasibility, reliability, generalizability, and validity of the MSF, and (5) were published in English. We excluded studies if they (1) were in nonsurgical specialties such as pediatrics, family medicine, obstetrics, and gynecology etc., (2) provided only general descriptions and information about MSF without empirical data, (3) reported only the process of MSF, and (4) only reported changes in performance after feedback.

Data Collection Process

Each article in this study was evaluated by 2 coders (K.A. and A.A.) independently, based on the title and abstract. Any disagreements about inclusion were solved by retrieving the full article and reviewed by a third coder (C.V.). Based on discussions among the 3 coders, we achieved 100% agreement on the studies to be included.

The initial search yielded 780 articles, as described in Figure 1. Of these, 461 articles were excluded based on the title, a further 265 articles were excluded based on the abstract, and another 47 were eliminated after reading the full articles. Finally, we agreed on 8 articles to be included in the present study.

RESULTS

As summarized in Figure 1, of the 786 initial articles, only 8 met the inclusion criteria and 778 were excluded. One

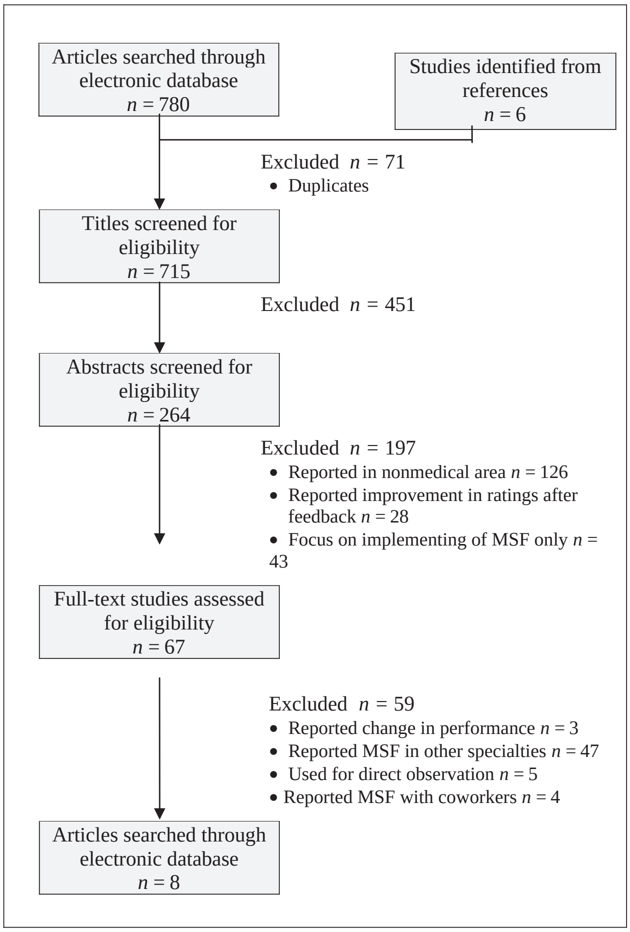
Articles searched through electronic database
n = 780
Studies identified from references
n = 6
Excluded n = 71
• Duplicates
Titles screened for eligibility
n = 715
Excluded n = 451
Abstracts screened for eligibility
n = 264
Excluded n = 197
• Reported in nonmedical area n = 126
• Reported improvement in ratings after feedback n = 28
• Focus on implementing of MSF only n = 43
Full-text studies assessed for eligibility
n = 67
Excluded n = 59
• Reported change in performance n = 3
• Reported MSF in other specialties n = 47
• Used for direct observation n = 5
• Reported MSF with coworkers n = 4
Articles searched through electronic database
n = 8

FIGURE 1. Selection of studies for the systematic review.

study was published prior to 2000 (in 1989). Four studies were published between the years of 2000 and 2010. Another 2 studies were published in 2011 and 1 was published in 2012. Three studies were conducted in the United States, another 3 studies in the United Kingdom, and 2 studies in Canada (Table 1).

Type of Assessment Instruments

Two studies used the Physician Achievement Review[1,11] instruments and another one used the Sheffield Patient Assessment Tool[12] to assess surgeons. The remaining 5 studies used single questionnaires with variable numbers of items ranging from 13 to 69 across the instruments. The details of the studies are summarized in Tables 1 and 2. The instruments have been designed to assess a range of competencies including patient relationships, diagnostic and treatment skills, collegiality, leadership, decision making, judgment, and the 6 competencies of the Accreditation Council For Graduates Medical Education (ACGME), patient care, medical knowledge, professionalism, system-based practice, practice-based learning and improvement, and interpersonal and communication skills (Table 1).

Validity

Out of the 8 studies included in the present review (Table 1), 1 reported evidence of content validity by determining if the content that the instrument contained was an adequate sample of the domain it was supposed to represent.[1] Enhancing content validity of instruments (sampling of appropriate content and skills) can be achieved by using a table of specifications based on a list of core competency areas and methods to assess them and having experts systematically review items to ensure that each competency is adequately assessed. Applying this procedure, Violato et al.[1] constructed instruments to assess a surgeon in practice in communication skills, interpersonal skills, collegiality, professionalism, and ability to continuously improve. These researchers developed a committee of experts (i.e., surgeons and psychometric experts) to construct questionnaires of 34 items for medical colleague, 19 items for coworker, 33 for self-assessment, and 39 items for a patient questionnaire. The questionnaires were subsequently sent to surgeons to provide systematic feedback (a modified Delphi procedure). Questionnaires were edited following the feedback to enhance the content validity of the instruments.[1]

Two studies (Table 1) reported concurrent, criterion-related validity by comparing the results of MSF with the results obtained using another assessment method.[13,14] Criterion-related validity refers to the relationship between scores obtained using the MSF instruments and scores obtained using 1 or more other instruments or measures. Risucci et al. examined the predictive validity by comparing MSF with American Board of Surgery in Training Examination (ABSITE). They found a significant correlation between MSF and ABSITE ($r = 0.58$, $p < 0.01$). This relationship suggests that as surgeons received higher ratings in MSF, they also received higher rating scores in the ABSITE.[14]

Crossley et al.[13] compared the MSF assessment in the form of Non-Technical Skills for Surgeons (NOTSS) with the Procedure-Based Assessment (PBA) global summary, and Objective Structured Assessment of Technical Skills. They found that the NOTSS scores were positively correlated with PBA global summary scores ($r = 0.48$, $p < 0.001$). Also, MSF in the form of NOTSS was positively correlated with the generic part of the Objective Structured Assessment of Technical Skills score; ($r = 0.51$, $p < 0.001$).

Evidence for construct validity, which refers to the nature of the psychological construct or characteristic being measured by the instrument, was reported in all the studies.[1,11-17] Violato et al.[1] conducted principal component factor analysis to derive a 5-factor solution for the medical colleague questionnaire accounting for 69% of the variance, 3 factors for the coworker questionnaire accounting for 70.9%, 5 factors for the patient questionnaire accounting for 73.5%, and 4 factors for the self-assessment questionnaire accounting for 65.1%. In addition, the mean score was calculated between self-assessment and medical colleague. Surgeons rated themselves lower than medical colleague with self $M = 4.07$ (0.73) and medical colleague $M = 4.5$ (0.64). Crossley et al.[13] derived 4 factors with principal component factor analyses in their MSF instruments with 6 surgical specialties.

Risucci et al.[14] also investigated construct validity of their MSF. Principal component factor analysis was conducted to derive a 1-factor solution accounting for 85.3 % of the variance. In addition, the mean score was calculated between self-assessment and medical colleague. Surgical residents rated themselves higher than medical colleague with self $M = 3.89$ (0.59) and medical colleague $M = 3.53$ (0.67). As well, the mean score was calculated between self-assessment and supervisors' assessment. Surgical residents rated themselves higher than supervisors with self $M = 3.89$ (0.59) and supervisors' $M = 3.73$ (0.91). Moreover, the mean score for ratings of surgical residents was calculated between medical colleague and supervisors. Medical colleague rated surgical residents lower than did supervisors with medical colleague $M = 3.53$ (0.67), and supervisors $M = 3.73$ (0.91).[14]

Chipp et al.,[15] employing plastic surgeons, found that consultants rated trainees more stringently than trainees, nurses, and patients. Sinclair et al.,[12] employing urologists in the UK, addressed construct validity by testing the instrument in different settings and on different occasions.

Consultants had an average of 6 free-text comments (range 3-10) on the assessments. Of the 60 free-text

TABLE 1. Specialty, Instruments, Factors Assessed, and Validity of MSF Studies for Surgical Practice

Study Name	Specialty and Participant	MSF Instrument Personnel and No. of Items	Factors Assessed by MSF	Validity and Findings
Violato et al.[1] (Canada)	Surgery (n = 252) 25 surgeons from each subspecialty	*PAR* – Medical colleague instrument consists of 34 items – Coworker instrument consists of 19 items – Patient instrument consists of 39 items, and – Self-instrument consists of 34 items	– MC instrument examined communication, diagnostic and treatment skills, medical records transfer, coordination of care, respect for patients, collaboration, professionalism, ability to assess medical literature, continuing learning, and stress management. – CW instrument focused on communication, collaboration, respect for patients and colleagues, accessibility and support for colleagues, and coworker learning. – PT instrument focused on communication, respects, the office staff, and information received. – Self-assessment instrument is identical for MC.	*Construct validity*: Principal component factor analysis was conducted to derive a 5-factor solution for MC, accounting for 69% of the variance; 3 factors for CW, accounting for 70.9%; 5 factors for Pt, accounting for 73.5%; and 4 factors for self, accounting for 65.1% of the variance. *Construct validity*: The mean score was calculated between self-assessment and MC. Surgeons rated themselves lower than MC with self $M = 4.07$ (0.73), MC $M = 4.5$ (0.64). *Findings*: Using PAR questionnaire data from patients, medical colleagues, and coworkers is gaining acceptance and credibility as a means of providing primary care physicians with quality improvement data as part of overall strategy of maintaining competence and certification.
Lockyer et al.[11] (Canada)	Surgery (n = 216) Surgeons from different specialties	*PAR* – Medical colleague instrument consists of 34 items – Coworker instrument consists of 19 items – Patient instrument consists of 39 items, and – Self-instrument consists of 34 items	– MC instrument examined communication, professionalism, medical expert, scholar, and manager. – CW instrument focused on oral communication, and written communication. – PT instrument focused on communication, manager, follow-up, and management. – Self-assessment instrument is identical for MC.	*Construct validity*: Principal component factor analysis was conducted to derive a 4-factor solution for MC, accounting for 75% of the variance; a 2-factor solution for CW, accounting for 72%; a 4-factor solution for Pt, accounting for 77%. *Construct validity*: The mean score was calculated between self-assessment and MC. Surgeons rated themselves lower than MC with self $M = 4.03$ (0.77), MC $M = 4.68$ (0.30). *Findings*: The comparison of the aggregate mean scores and mean factors scores showed that there were no differences by school for any of the assessments or factors within questionnaires. This suggests an equivalency of performance for graduates of the University of Calgary and those from 4-y medical schools.

Sinclair et al.[12] (UK)	Urologists consultant (n = 10)	*SHEFFPAT* Patients (single instrument with 13 items)	Assess the 7 domains of GMC "Good Medical Practice" 1. Good medical care 2. Maintaining good medical practice 3. Teaching training and assessing 4. Relationship with patients 5. Working with colleagues 6. Probity 7. Health	*Construct validity*: Construct validity achieved by testing the instrument in different sittings and different occasions. The instrument was tested before with pediatrician. However, testing the same instrument with different specialty supports the validity of that instrument. *Findings*: Consultants had an average of 6 free-text comments (range 3-10). Of the 60 free-text comments, 86.7% were positive with only 13.3% commenting on a negative aspect. All of these 8 negative comments were constructive criticism about the department and organization rather than the specific consultant. *Findings*: The SHEFFPAT questionnaire appears to provide reliable, valid, and unbiased feedback from the patients for urologists.
Crossley et al.[13] (UK)	Six specialties in surgery (cardiac surgery, colorectal, gastrointestinal, orthopedics, vascular, and obstetrics and gynecology) (n = 85)	*NOTSS* MC, CW, and independent assessors using (single instrument with 16 items)	Four main factors 1. Situation awareness 2. Decision making 3. Communication and team work 4. Leadership	*Construct validity*: Principal component factor analysis was conducted to derive a 4-factor solution. *Construct validity*: The assessment using Non-Technical Skills for Surgeons (NOTSS) were positively correlated with Procedure-based assessment (PBA) global summary scoring. The Pearson correlation was 0.48 ($p < 0.001$). *Construct validity*: The assessment using NOTSS were positively correlated with the generic part of the Objective Structured assessment of Technical skills (OSATS) score. The Pearson correlation was 0.51 ($p < 0.001$). *Findings*: Thirty of the 56 anesthetists and 26 of the 39 scrub nurses who completed the validity, feasibility, and acceptability of NOTSS reported the following: only 5 agreed that NOTSS added too much time to the operating list, whereas the majority perceived NOTSS to be useful for the supporting insight and for providing feedback. Most regarded NOTSS as an important adjunct to surgical skills–assessment methods. Twenty-five felt that the routine use of NOTSS would enhance patient safety in the operating theater.
Risucci et al.[14] (USA)	Surgical residents (n = 32)	*NA* MC + self-assessment (single instrument with 10 items)	– Technical ability – Basic science knowledge – Clinical knowledge – Judgment – Relations with patients – Relations with peers	*Construct validity*: Principal component factor analysis was conducted to derive a 1-factor solution accounting for 85.3% of the variance. *Construct validity*: The mean score was calculated between self-assessment and MC. Surgical residents rated themselves higher than MC with self $M = 3.89$ (0.59), and MC $M = 3.53$ (0.67).

TABLE 1 *(continued)*

Study Name	Specialty and Participant	MSF Instrument Personnel and No. of Items	Factors Assessed by MSF	Validity and Findings
			– Reliability – Industry – Personal appearance – Reaction to pressure	*Construct validity*: The mean score was calculated between self-assessment and supervisors. Surgical residents rated themselves higher than supervisors with self $M = 3.89$ (0.59), and supervisors $M = 3.73$ (0.91). *Construct validity*: The mean score for ratings of surgical residents was calculated between MC and supervisors. MC rated surgical residents lower than supervisors with MC $M = 3.53$ (0.67), and supervisors $M = 3.73$ (0.91). *Predictive validity*: The average of overall ratings by peer and supervisors correlated moderately with the total raw score on American Board of Surgery In training Examination (ABSITE), $r = 0.58$, $p < 0.01$.
Chipp et al.[15] (UK)	Plastic surgery (30 trainees have experience with the format of MSF revision course but the results of the last 9 candidates were reported in this study)	*NA* Consultants, trainees, patients, and nurses. It is station-based assessment where each lasts for 30 min. Each station consisted of viva style–structured interview based around photographs of clinical conditions. One station consisted of long cases and the final 2 stations were each made up of 5 short cases.	Four main factors 1. overall professional capability, 2. knowledge and judgment, 3. communication and responses, 4. bedside manner.	*Findings*: Scores were obtained from consultants, trainees, patients, and nurses for each candidate and used to calculate an average score for every station. An overall average score of 6 or more is required to pass the exam. Differences in scores between different groups were as follows: Consultants = 5.9, Trainees = 6.3, Nurses = 6.7, and Patients = 6.9. *Construct validity*: Consultants rated trainees more stringently than trainees, nurses, and patients. *Findings (predictive validity)*: There were 9 candidates who had taken the FRCS (plastic) exam at the next available sitting after the revision course. The exam course accurately predicted actual exam results in 6 of the 9 candidates. The remaining 3 candidates passed the exam despite scoring less than 6 on the exam preparation course; this may be due to the feedback from the course which allowed intensive and focused revision in certain areas before the exam.
Higgins et al.[16] (USA)	Cardiothoracic surgery	*NA*	Six general competencies of *ACGME* 1 Patient care, 2. Medical knowledge, 3. Professionalism, 4. System-based practice, 5. Practice-based learning and	*Construct*: Residents demonstrated improved scores in every domain of the 6 categories when comparing the first and second administrations of the survey with a mean improvement of 4.46 on every scale. The 2 assessments were performed with an 8-month interval.

	(n = 6) Rotating in year 3	MC, CW, people were selected by program director using single instrument with 45 items.	improvement, and 6. Interpersonal and communication skills.	*Findings*: In the first administration of the survey, the residents as a group scored highest in the ACGME competencies of medical knowledge, patient care, and professionalism. However, residents scored lowest in the system-based practice, interpersonal and communication skills, and practice-based learning and improvement.
Pollock et al.[17] (UK)	Plastic surgery (n = 6)	*NA* MC+ CW (single instrument with 4 parts consists of 60 items).	Part 1, 6 general competencies of ACGME 1. Patient care, 2. Medical knowledge, 3. Professionalism, 4. System-based practice, 5. Practice-based learning and improvement, and 6. Interpersonal and communication skills. Part 2, the raters were asked if they will choose the same surgeon (2 items) Part 3, the raters asked to mark items on checklist of 25 performance characteristics that need improvement (30 items) Part 4, the same 25 performance characteristics offered in part 3; however, the raters asked whether the items were achieved (30 items)	*Construct validity*: The correlation between MC and CW was calculated. $r = 0.42$. $p = 0.35$. However, CW rated residents significantly higher than the MC all over for the 4 competencies. *Findings*: Raters in ambulatory surgery sittings tend to check more negative characteristics than do other nurses and clinical staff. *Construct validity*: Surgeons rated trainees more stringently than nurses. The mean rating of surgeons was $M = 3.24$ and the mean rating of nurses was $M = 3.6$.

PAR, Physician Achievement Review; MC, medical colleague; CW, coworker; Pt, patient; SHEFFPAT, the Sheffield Patient Assessment Tool; GMC, General Medical Council; (OSATS), Objective Structured assessment of Technical skills; ACGME, Accreditation Council for Graduate Medical Education.

TABLE 2. Feasibility, Reliability, and Generalizability Evidence for MSF Studies for Surgical Practice

Study Name	Mean No. of Raters (% Response)	Reliability Coefficient (α) or (95% CI)	Administration/ Feasibility	Generalizability (Ep^2) or IntraClass Correlation (ICC)
Violato et al.[1] (Canada)	MC, 7.27 (89.6%) CW, 7.20 (88.2%) Pt, 22.63 (83.2%) Self, 1 (96.5%)	MC, $\alpha = 0.98$ CW, $\alpha = 0.95$ Pt, $\alpha = 0.93$ Self, $\alpha = 0.97$	The College of Physicians and Surgeons of Alberta adopted a performance appraisal or MSF system for all physicians/surgeons in its jurisdiction. As a part of its overall goal of ensuring that all physicians/surgeons in the province participate in a multisource feedback process every 5 years, the college implemented this evaluation system for different specialties as well.	7.27 MC, Ep$^2 > 0.70$ 7.20 CW, Ep$^2 > 0.70$ 22.63 Pt, Ep$^2 > 0.70$
Lockyer et al.[11] (Canada)	MC, 7.67 CW, 7.60 Pt, 24 Self, 1	MC, $\alpha = 0.98$ CW, $\alpha = 0.96$ Pt, $\alpha = 0.98$ self, $\alpha = 0.97$	The purpose of this study was to compare the performance of practicing surgeons in Alberta who graduated from the University of Calgary (a 3-y school) with matched samples from other 4-y Canadian medical schools and to determine the reliability and validity of PAR instrument in assessing surgeons.	7.27 MC, Ep$^2 = 0.61$ 7.20 CW, Ep$^2 = 0.70$ 22.63 Pt, Ep$^2 = 0.81$
Sinclair et al.[12] (UK)	Twenty-three patients for each consultant	Not reported	The aim of this study was to implement a validated and objective way to measure the relationship with patients with urologists. In addition, to evaluate the feasibility, reliability of the SHEFFPAT questionnaire in urology.	With 23 patients, Ep$^2 = 0.70$ (95% CI = 0.21)
Crossley et al.[13] (UK)	Fifty-six anesthetists, 39 scrub nurses, 2 surgical care, and 3 independent assessors. 8.4 Raters for each candidate. Pt response rates (67.1%).	With a total of 6 raters in assessing trainee over (2 different cases) the reliability, $\alpha = 0.88$	The nontechnical skills for surgeons can affect patient safety and clinical effectiveness. Therefore, the aim of this study was to develop a reliable and valid tool to assess the nontechnical skills of individual surgeons in the operating room.	With 6 raters, Ep$^2 = 0.80$.

Risucci et al.[14] (USA)	MC (peers): 27 MC (supervisors): 4 and self-assessment	Not reported	The aim of this study was to examine the validity of ratings through comparison ratings among raters and to analyze the extent to which they obtained ratings could differentiate attending surgeons from surgical residents.	Not reported
Chipp et al.[15] (UK)	Eight consultants, 2-3 trainees, 11 patients, and 11 nurses (station based)	Not reported	The aim of this study was to establish a new clinically based exam preparation course, utilizing multisource feedback, to identify candidates at risk of failure and improve pass rates.	Not reported
Higgins et al.[16] (USA)	Supervisors, peers, nurses, self, and administrative personnel People (12-15 raters for each candidate)	Not reported	The aim of this study was to develop and implement an evaluative tool that would provide data to residents and program leadership regarding their performance and to provide the training program in cardiothoracic surgery with a reliable way to assess this component of the program.	Not reported
Pollock et al.[17] (USA)	Twelve medical colleagues and 28 coworkers	Not reported	The aim of this study was develop methods to evaluate resident performance using competencies essential for outcomes, and to determine whether ratings of resident performance varied systematically among healthcare professional.	Not reported

comments, 86.7% were positive with only 13.3% commenting on a negative aspect. All of these 8 negative comments were constructive criticism about the department and organization rather than the specific consultant. Pollock et al. found that the correlation between medical colleague and coworker was correlated ($r = 0.42$, $p = 0.35$) with plastic surgery. The coworker rated residents significantly higher than the medical colleagues overall rating for the 4 competencies.[23]

Higgins et al.[16] studied American cardiothoracic surgeons. They found that residents demonstrated improved scores in every domain of the 6 categories when comparing the first and second administrations of the survey with a mean improvement of 4.46 on every scale at an 8-month interval. Moreover, in the first administration of the survey, the residents scored highest in the Accreditation Council for Graduate Medical Education (ACGME) competencies of medical knowledge, patient care, and professionalism. Conversely, they scored lowest for system-based practice, interpersonal and communication skills, and practice-based learning and improvement.

Internal Structure, Reliability, and Generalizability

Reliability refers to the consistency of the scores obtained or the consistency of measurement. The internal consistency reliability using Cronbach coefficient alpha (α) was reported for most MSF instruments, both for subscales and the total scale. Violato et al.[1] in assessing surgeons reported Cronbach α of 0.98, 0.97, 0.95, and 0.93, for medical colleague, self, coworker, and patient instruments, respectively. Crossley et al.[13] reported Cronbach α of 0.88 for their 16-item instrument. Similarly, Lockyer et al.[11] reported α coefficients = 0.90, 0.96, 0.98, and 0.97, respectively, for colleague, coworker, patient, and self.

In addition to the internal consistency of the questionnaires, several researchers investigated the number of raters and the number of items that are sufficient to provide stable data to the individual being assessed. They thus employed generalizability theory deriving generalizability coefficients (Ep^2).[18] In this work, studies showed that it is possible to achieve $Ep^2 > 0.70$ with moderate number of observers.[19] For example, Sinclair et al. achieved $Ep^2 = 0.70$ with a 13-item instrument and 23 raters.[12] Violato et al. found adequate generalizability coefficients ($Ep^2 > 0.70$) for groups of 8 assessors (medical colleague and coworkers) and 25 patients.[1] Crossley et al. achieved $Ep^2 = 0.80$ with 6 raters.[21] Lockyer et al.[11] achieved $Ep^2 = 0.61$ for 8 medical colleagues, $Ep^2 = 0.70$ with 8 coworkers, and $Ep^2 = 0.81$ with 25 patients.

Generalizability was reported in only these 4 studies and it ranged from $Ep^2 = 0.70$ to 0.80.[1,11-13] The other 4 studies in Table 2 did not report any generalizability analyses.

Feasibility

Several researchers concluded that the feasibility of using MSF is good (Table 2). Some of the studies used the response rates as indication of feasibility. Violato et al.[1] reported high response rates for patients (83.2%), coworkers (88.2%), medical colleague (89.6%), and self (96.5%). Lockyer et al.[11] found similar response rates as did others. Other researchers identified the feasibility of the MSF by the time needed to complete the forms which generally took between 6 and 15 minutes.

In several of the studies (especially the Canadian and UK ones), participation in the MSF is mandated by the regulatory or licensing authorities and surgeons must therefore participate (Table 2). In other studies (e.g., in the US) MSF has been developed to assess surgical residents in technical and nontechnical skills. It appears feasible, therefore, to employ MSF for both residents and practicing surgeons.

DISCUSSION

The main findings of the present study are: (1) MSF can be applied to surgical practice both in residency and subsequent independent practice, (2) a range of competencies such as diagnostic and treatment skills, patient relationships, collegiality, leadership, decision making, system-based practice, probity, professionalism, and knowledge and judgment, and communications can be assessed, (3) various raters such as medical colleagues, non-MD coworkers, supervisors, patients and self-assessment can be employed, (4) high internal consistency reliability of the instruments can be achieved, (5) as few as 8 raters and 23 patient surveys can achieve an Ep^2 coefficient ≥ 0.70, and (6) there is evidence of validity (content, criterion-related, and construct) for the use of MSF in the assessment of surgical practice.

A number of nontechnical competencies can effectively and feasibly be assessed using MSF for both surgical residents and independently practicing surgeons. A full MSF model should include data from a self-assessment, medical colleagues (e.g., other surgeons, referring physicians, and anesthesiologists), nonmedical coworkers (e.g., office staff and secretaries), coworkers (e.g., nurses and physiotherapists), and patients. As we have seen, this range of data can be employed to assess leadership, decision making, system-based practice, probity, professionalism, and knowledge and judgment, and communications, and so forth.[1,13] The MSF system is feasible with typically high response rates of questionnaires which require only a brief period of time to complete.

Across the several studies reviewed, the internal consistency reliability was high (≥ 0.85) and typically in excess of 0.90. Similar results were reported with the use of MSF in other specialties. Lockyer et al.[2] reported high internal consistency reliability ($\alpha = 0.94$), using the MSF questionnaires to assess emergency room physicians.

In the UK, Archer et al.[18] reported high internal consistency reliability ($\alpha = 0.98$) with the MSF process using an instrument which was modified from Sheffield Peer Review Assessment Tool across different specialties. It consists of 16 questions (mini-PAT) rated on a 6-point scale. With anesthesiology, Lockyer et al.[23] developed a survey with 11, 19, 29, and 29 items for patients, coworkers, medical colleague, and self-assessment, respectively, using a 5-point scale to assess 186 anesthesiologists. The internal consistency reliability was high in the patient survey ($\alpha = 0.93$), coworker survey ($\alpha = 0.95$), medical colleagues survey $\alpha = 0.97$, and self-assessment survey ($\alpha = 0.97$). Additionally, the number of raters required to assess a surgeon is around 6 to 8. With questionnaires in excess of about 17 items, e.g., the Ep^2 coefficient is generally ≥ 0.70, the accepted standard. Approximately, 23 patients achieve a similar Ep^2 coefficient. These results correspond with the findings from other studies. Ramsey et al.[24] with 313 family physicians achieved Ep^2 coefficient $= 0.70$ with an 11-item global instrument and 10 to 11 peer physician raters. Violato et al.[25] with 100 pediatricians reported adequate generalizability coefficients $Ep^2 > 0.78$ for groups of 8 assessors (medical colleague and coworkers) and 25 patients.

Our systematic review of the 8 MSF studies has revealed several sources of validity evidence for use with surgeons. These include evidence of content, criterion-related, and construct validity. Most of the construct validity evidence comes from factor analytic studies that identify the basic factors of latent variables (e.g., communication skills and professionalism) in the questionnaires. These findings correspond to the results reported by others who have applied the MSF process to other specialties. Archer et al.[26] examined validity by comparing MSF scores between year 2 and year 4 pediatrician trainees. Year 4 trainees scored significantly higher than year 2. In another study, Archer et al.[27] examined construct validity by comparing MSF scores between senior house officers and specialist registrar trainees who scored significantly higher than the senior house officers. Consistently higher ratings given to advanced trainees by year of program support the construct validity of the MSF instruments.

Wood et al.[28] examined the construct validity of MSF over a period of 6 years in Obstetrics and Gynecology training in the UK. They found a correlation between first assessments and second assessments for 67 doctors having 2 sets of assessments (usually separated by 6-7 months; $r = 0.77$, $p < 0.001$). Similarly, Violato et al.[29] examined the evidence of construct validity of MSF instruments for general physicians. Researchers investigated changes in performance for doctors from the College of Physicians and Surgeons of Alberta who participated twice, 5 years apart, and determined the associations between change in performance and initial assessment and sociodemographic characteristics. The paired sample *t*-test used to compare the sum of the mean aggregate score for the 2 times indicated significant differences ($p < 0.001$). Confirmatory factor analysis provided evidence for the validity of factors that were theoretically expected, meaningful, and cohesive.

The present systematic review, although comprehensive, is based on a relatively modest number of studies (8) that were published in refereed journals in English. Although MSF appears adequate to assess nontechnical skills, this approach fails to assess aspects of clinical competence reflecting surgeon's knowledge and skills; these may be more accurately obtained through other methods such as the PBA[13] or objective structured performance-related examination.[21]

In addition, MSF assessments are entirely questionnaire-based and rely on judgment and inference by the assessors and respondents, which are known to be subject to a variety of influences and heuristics.[22] Therefore, generalizability theory should be applied in further studies to determine the potential sources of error that can occur due to different assessors and respondents.

Future research should be done to replicate and extend some of the empirical findings, especially validity evidence. Criterion-related validity studies of correlations between direct observations of behavior or performance and MSF scores are required to add further evidence of validity. Future research may well include confirmatory factor analysis which provides stronger construct validity evidence than do the principal component factor analyses conducted to date.[20] Meanwhile the current empirical evidence is promising.

CONCLUSION

The present systematic literature review has shown that MSF is feasible, reliable, and valid in assessing surgeons in practice. The results indicate that MSF systems can be used to assess key competencies such as communication skills, interpersonal skills, collegiality, and medical expertise. In addition, further implementing of MSF system in surgical settings has promising possibilities. This feedback system can provide information beyond that which can be provided by 1 or few sources alone.[9] Although reliability and validity challenges remain, MSF shows a promising, feasible, reliable, and valid means of assessing surgeons across a broad range of competences such as professionalism, leadership, interpersonal skills, collegiality, and communication skills.

REFERENCES

1. Violato C, Lockyer J, Fidler H. Multisource feedback: a method of assessing surgical practice. *BMJ*. 2003;326:546-548.

2. Lockyer J, Violato C, Fidler H. The assessment of emergency physicians by a regulatory authority. *Acad Emerg Med*. 2006;13:1296-1303.

3. Wilson RM, Harrison BT, Gibberd RW, Hamilton JW. An analysis of adverse events from the quality in Australian health care study. *Med J Aust*. 1999;170:411-415.

4. Lockyer J, Violato C, Fidler H. A study assessing surgeon use of multisource feedback data. *Teach Learn Med*. 2003;15:168-174.

5. Lockyer J, Clyman S. Multi source feedback. Holmboe E, Hawkins R, editors. A Practical Guide to the Assessment of Clinical Competence. Mosby/Elsevier; 2008.

6. Sala F, Dwight S. Predicting executive performance with multi-rater surveys: whom you ask makes a difference. *Consult Psychol J Pract Res*. 2002;54:166-172.

7. Church AH, Bracken DW. Advancing the state of the art of 360 feedback: guest editors' comments on the research and practice of multi rater assessment methods. *Group Organ Manag*. 1997;22:149-161.

8. Violato C, Lockyer J, Fidler H. Assessment of psychiatrists with multisource feedback. *Can J Psychiatry*. 2007;53:525-533.

9. Bracken DW, Church AH. The Handbook of Multisource Feedback: The Comprehensive Resource for Designing and Implementing MSF Processes. San Francisco: Jossey-Bass; 2001.

10. Moher D, Liberati A, Tetzlaff J, Altman DG. The PRISMA Group. Preferred Reporting Items for Systematic reviews and Meta-Analyses: the PRISMA statement. *PLoS Med*. 2008. Available at: 10.1371/journal.pmed.1000097 [e1000097].

11. Lockyer J, Violato C, Wright B, Fidler HM, Chan R. An analysis of long term outcomes for surgeons from three and four year medical school curricula. *Can J Surg*. 2012;55:1-11.

12. Sinclair A, Gunendran T, Archer J, Bridgewater B, O'Flynn K, Pearce I. Re-certification for urologists: is the SHEFFPAT questionnaire valid for assessing clinicians' relationships with patients? *Br J Med Surg Urol*. 2009;2:100-104.

13. Crossley J, Marriott J, Purdie H, Beard J. Prospective observational study to evaluate NOTSS (Non-Technical Skills for Surgeons) for assessing trainees' non-technical performance in the operating theatre. *Br J Surg*. 2011;98:1010-1020.

14. Risucci D, Tortolani A, Ward R. Ratings of surgical residents by self, supervisors and peers. *Surg Gynecol Obstet*. 1989;169:519-526.

15. Chipp E, Srinivasan K, Khan M, Rayatt S. Incorporating multi-source feedback into a new clinically based revision course for the FRCS (Plast) exam. *Med Teach*. 2011;33:e263-e266.

16. Higgins RSD, Bridges J, Burke JM, et al. Implementing the ACGME general competencies in a cardiothoracic surgery residency program using 360-degree feedback. *Ann Thorac Surg*. 2004;77:12-17.

17. Pollock R, Donnely M, Plymale M, et al. 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.

18. 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*. 2008;13:181-192.

19. Archer J, McGraw M, Davies H. Assuring validity of multisource feedback in national program. *Postgrad Med J*. 2010;86:526-531.

20. Violato C, Hecker K. How to use structural equation modeling in medical education research: a brief guide. *Teach Learn Med*. 2008;19:362-371.

21. Ponton-Carss A, Hutchinson C, Violato C. Assessing surgical skills, professionalism and communications in surgeons. *Am J Surg*. 2011;202:433-440.

22. Kahneman D. Thinking Fast and Slow. Toronto: Doubleday, Canada; 2011.

23. Lockyer J, Violato C, Fidler H. A multi source feedback program for anesthesiology. *Can J Anaesth*. 2006;53:33-39.

24. Ramsey PG, Wenrich MD, Carline JD, Inui T, Laeson E, LoGerfo J. Use of peer ratings to evaluate physicain performance. *JAMA*. 1993;269:1655-1660.

25. Violato C, Lockyer J, Fidler H. Assessment of pediatricians a regulatory authority. *Pediatrics*. 2006;117:796-802.

26. Archer J, McGraw M, Davies H. Assuring validity of multisource feedback in national program. *Arch Dis Child*. 2010;95:330-335.

27. Archer J, Norcini J, Davies A. Use of SPRAT for peer review of paediatricians in training. *BMJ*. 2005;330:1251-1253.

28. 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.

29. 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.