Academic Medicine

The Reliability, Validity and Feasibility of Multi-Source Feedback for Assessing Physicians: A Systematic Review

--Manuscript Draft--

Manuscript Number:	ACADMED-D-13-00337R1
Full Title:	The Reliability, Validity and Feasibility of Multi-Source Feedback for Assessing Physicians: A Systematic Review
Article Type:	Research Report
Corresponding Author:	Tyrone Donnon, PhD University of Calgary Calgary, AB CANADA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Calgary
Corresponding Author's Secondary Institution:	
First Author:	Tyrone Donnon, PhD
First Author Secondary Information:	
Order of Authors:	Tyrone Donnon, PhD
	Ahmed Al Ansari, MBBCh MRCSI PhD
	Samah Al Alawi, MD
	Claudio Violato, PhD
Order of Authors Secondary Information:	
Manuscript Region of Origin:	CANADA
Abstract:	Purpose The use of multisource feedback (MSF) or 360 degree evaluation has become a recognized method of assessing physician performance in practice. The purpose of the present systematic review was to investigate the reliability, generalizability, validity, and feasibility of MSF for the assessment of physicians. Method The authors searched the EMBASE, PsycINFO, MEDLINE, PUBMED, and CINAHL databases for peer-reviewed, English-language articles up to January, 2013. Studies were included if they met the following inclusion criteria: use one or more MSF instruments to assess physician performance in practice, reported psychometric evidence of the instrument(s) in the form of reliability, generalizability coefficients and construct or criterion-related validity, and provided information regarding the administration or feasibility of the process in collecting the feedback data. Results Of the 96 full-text articles assessed for eligibility, 43 articles were included in the final systematic review. The use of MSF has been shown to be an effective method for providing feedback to physicians from a multitude of specialties about their clinical and nonclinical (i.e., professionalism, communication, interpersonal relationship, management) performance. In general, assessment of physician performance was based on the completion of the MSF instruments by 8 medical colleagues, 8 coworkers and 25 patients to achieve adequate reliability and generalizability coefficients of $\alpha > 0.90$ and $Ep2 > 0.80$, respectively. Conclusions The use of multisource feedback employing medical colleagues, coworkers, and patients as a method to assess physicians in practice has been shown to have high reliability, validity and feasibility.
Response to Reviewers:	Reviewer Comments:

Reviewer #1
General comments: This is a systematic review of MSF studies, reporting upon their reliability, feasibility and validity. The systematic review appears to have been conducted according to protocol and provides a worthwhile overview of the MSF studies since 1975. However, the writing is unclear at times, and definitions of terms and explanations of concepts that would enhance transparency are not included. Suggestions are included in the comments below.

Intro - Para 2, p.4 - needs rewording as follows:

*First sentence, line 38 "MSF is frequently used in workplace settings where employees work in a team and cannot be directly or easily supervised by managers",
*Please reword the first part of this sentence to say -" MSF originated in industry... "
The first part of this sentence was modified to read “MSF originated in industry…”

*Please confirm the remainder of this sentence by going back and checking the references cited , " where employees work in a team and cannot be directly or easily supervised by managers". I'm not sure this was the main reason. I believe it was the realization that others working with an individual could assess particular domains quite readily. Please check this.
The references cited were checked and the remainder of this sentence was re-written to reflect the main reason for the growth in the use of MSF in industry:
“MSF originated in industry during a time when the search for competent employees and the reliance on a single supervisor’s evaluation was recognized as a restrictive approach to the assessment of a worker’s specific abilities 5,6”

Para 2, second sentence - a number of Canadian and US physicians still work mainly solo in private practice. Please reframe this sentence to reflect this.
To reflect the variability of persons that will work with a physician, and not necessarily in a team, we have rewritten this sentence to read:
“Similarly, physicians work with a variety of people (i.e., medical colleagues, consultants, therapists, nurses, and coworkers) that are able to provide a better assessment and contextually based understanding of physician performance than any single person.”

Para 2, third sentence - Not all MSF programs include a self-assessment. Some programs involving residents include supervisors
To reflect the fact that not all MSF process include a self-assessment and that some physicians in-training will be assessed by a supervisor or preceptor, we have rewritten this sentence to read:
“In MSF physicians may complete a self-assessment instrument and receive feedback from a number of medical colleagues (peers), in-training supervisors or preceptors, non-physician co-workers (e.g., nurses, psychologists, pharmacists), as well as their own patients.7”

Intro, para 3, p 5:
- second sentence - other domains such as professionalism also the focus of MSF
The word “professionalism” was included in this sentence to indicate that it is also a MSF domain that is assessed.

- last sentence, rationale for the study, does not seem to flow clearly from the preceding sentence
The last sentence was re-written to fit better with the preceding sentence and purpose of the study:
“Accordingly, the purpose of the present study was to conduct a systematic review of the published, peer-reviewed research on the different types of MSF instruments used to assess physicians’ clinical/nonclinical skills performance and to investigate the evidence for reliability, generalizability, validity and feasibility of this assessment approach.”

Please expand the Introduction and lit review to provide evidence as to the scope to which MSF has been implemented; e.g. for physicians, residents, students and in approximately how many countries. This too will contribute to the rationale for the study.

	To expand on the literature review in the Introduction section (staying just within the word count limits for the manuscript length) the following sentences were added to a revised second paragraph: "While early attempts at the development of MSF questionnaires in medicine focused on the assessment of residents in the late 1970s, today they are being used in North America (Canada, US) and Europe (Netherlands, UK) across a number of physician specialties.4 As a self-regulating profession, medicine is accountable for ensuring that physicians' are competent in the performance of their clinical roles and duties. Incumbent on regulatory bodies to monitor physician practice and patient safety, Canada was the first country to introduce a MSF process as a viable approach to providing an assessment of physician performance in the late 1990s…" Methods: Eligibility criteria, p. 5 - 6 - it appears from reading the results that studies of residents and students were also included. Please clarify this in the eligibility section. To clarify that studies that looked at MSF with residents were also eligible for inclusion, we modified inclusion criteria #1 in the "Eligibility criteria" to read: "…1) use one or more multi-source feedback instruments (e.g., self, colleague, coworker, and/or patient) to assess physician or resident performance in practice,.." And clarified that studies that looked at MSF with medical students were not eligible or excluded, we modified the exclusion criteria #1 to read: "…1) were used to assess other than physicians or residents (i.e., medical students) or non-physician health professionals…" Line 7, p. 6 - We excluded studies if they 1) were used to assess other than physicians or non-physician health professionals, - this is not clear. To clarified that studies that looked at MSF with non-physician health professionals were excluded, we modified the exclusion criteria #1 to include specific example groups: "…1) were used to assess other than physicians or residents (i.e., medical students) or non-physician health professionals (i.e., nurses, occupational or respiratory therapists, chiropractors, etc.),…" Study selection process appears clear and appropriate. Results: p.6-7 - clarity would be added to the tables by grouping the studies as described in this paragraph; i.e., - 1. Physician Assessment Review (Canada n = x, Netherlands = 1) 2. Sheffield Peer review Assessment Tool (UK n=x) 3. Other UK studies (n = X) 4. USA studies (n = x) 5. Studies from other countries (n = 4) To clarify the studies included in the systematic review as they are grouped in the Tables, the first paragraph in Results section was re-written as follows: "Although there are a variety of MSF instruments used in the studies, they include: the Physician Assessment Review (PAR) process (Canada, n = 13; Netherlands, n = 1), the Sheffield Peer Review Assessment Tool (SPRAT) process (UK, n = 6), multiple MSF instruments from the USA (n = 14), other UK related studies (n = 4), and three separate studies from other countries (China, Denmark and Taiwan)." Specialty - this para could be written more clearly, or perhaps use a table? To clarify the studies included in the systematic review as they are grouped by specialty, the first paragraph in the Specialty of Physicians Assessed Using MSF subsection was re-written as follows: "There were a number of MSF studies that assessed physicians across multiple specialties (n = 10). In a study of the psychometrics of the PAR MSF instruments, for example, Hall et al.13 evaluated the results from 308 physicians from multiple specialties in Alberta. With respect to specific physician practices there were MSF studies for each of the following specialties: family medicine (n = 5), pediatrics (n = 5), internal medicine (n = 5), surgery (n = 4), obstetrics/gynecology (n = 3), psychiatry (n = 3), anesthesia (n = 2), and single studies for emergency medicine, pathology/laboratory medicine, histopathology, radiology, and physical medicine and rehabilitation."

Types of MSF instruments used - this section might be better named- " Raters and length of questionnaires"
We modified the subsection title to read:
“MSF Assessors and Length of Questionnaires”

- re raters, did any include residents' supervisors or attendings?
To clarify that in some studies where the physicians (residents in-training) may have been evaluate by peers or medical colleagues that are their superiors, the following addition was added to the sentence:
“In MSF with physicians, information can come from a variety of sources (i.e., peers or medical colleagues including supervisors and preceptors,…”

- - para 2 in this section, first sentence - shorten to "The questionnaires used ranged in length from..."
To summarize the variability in length of the various MSF questionnaires used/included in Table 1, the first sentence of this 2nd paragraph now reads:
“The MSF questionnaires varied greatly in the number of items depending on the assessor: 4 to 57 items for self-assessment, 4 to 60 items for peer or medical colleague, 4 to 60 items for co-workers, and 3 to 49 items for patient questionnaires.”

Constructs/ domains assessed
- first sentence, suggest wording as " As shown in Table 1, a number of constructs were measured using MSF. "
The first and second sentence were combined to now read:
“As shown in Table 1, a number of constructs were measured using MSF: 1) professionalism, 2), clinical competence, 3) communication, 4) manager, and 5) interpersonal relationship.”

- identification of constructs: Please describe how you did this. E.g., Were they consistently identified by the authors, or did you have to interpret the authors' descriptions? How did communication differ from interpersonal relationships? Please define/ give examples of both of these. Also for manager. This will add clarity for the reader and also contribute to understanding of construct validity.
To clarify that there was consensus among the authors and to provide examples of how specific items from communication differ from interpersonal relationships and manager categories, we added the following sentences to the first paragraph
“Consensus for the five general category domains was achieved by three of the authors (TD, AA, SA) and were based on existing constructs or examples of items provided from the included studies”… “For example, items that were written “Communicates effectively with patients” or “Communicates effectively with other health care professionals” were clearly associated with the communication category, “Collaborates with medical colleagues” the interpersonal relationship category, and “Manages health care resources efficiently” the manager category.13”

Administration and feasibility - this section appears unclear.
It would help the reader to have definitions of "administration" and "feasibility". What criteria were looked for in descriptions of each of these? Eg, should "Administration" include how it was developed, administered and # of participants?
In Table 1, some data in the "Administration and feasibility" column appear to be psychometric in nature, and others, more admin or feasibility oriented. Consistency would help the reader.
To clarify that this section is as much about the general information about the process than just the administration and/or feasibility, the subsection heading was changed in the text (as well as on Table 1) and the first sentence in the first paragraph was modified and another sentence added.
General Information on Process, Administration and/or Feasibility
“Each of the 42 studies included in the MSF systematic review provided general information about the findings of their study with comments on the process, administration, and/or feasibility (Table 1). For example, general information comments emphasized how studies’ psychometric results provided support for the MSF process, was able to be administered to various participants in an efficient manner, and/or was a feasible method to collect multiple performance measures of physicians in practice.”

Reliablity and generalizabilty - please indicate which studies used each of these analyses.
At the top of the column in Table 2, the studies that reported either or both reliability and generalizability coefficients for each of the MSF questionnaires used are identified with the recognized, corresponding statistical symbols “α” and “Ep2” respectively. For example, in the first row/study for Violato et al., 1997 there are reliability coefficients reported for each of the MSF instruments (range from α = 0.89 to 0.95) and generalizability coefficients for the Medical Colleague (Ep2 = 0.77 for 8 raters) and Patient (Ep2= 0.80 for 25 raters) questionnaires.

Construct and Criterion-Related Validity -
- para 2: Please provide a sentence describing how each of these analyses indicates construct or criterion validity
To clarify each of the indications of construct validity outlined, a short explanation was given in parentheses as follows:
“Further evidence of construct validity was provided through analyses that showed: 1) measures of mean difference ratings between respondent groups (i.e., mean ratings from patients and coworkers are consistently higher than medical colleagues and are lowest on self-assessments), 2) improvement in performance ratings from Time 1 to Time 2 (i.e., increase in mean ratings are consistently higher from an earlier period, indicating an expected improvement in practice over time), 3) consistently higher ratings given to advanced trainees by year of program (i.e., increase in mean ratings as residents gain clinical experience from year to year of an in-training program), and 4) younger practitioners were rated higher than older ones (i.e., higher mean ratings are generally given to young practitioners that have been educated to be more conscious of MSF domain measures than practitioners that have been in practice for a greater number of years).”

- para 3 - this sentence is unclear: "Criterion-related validity was adduced in some studies where positive correlations: 1) were found between the MSF instruments/measures (concurrent validity)". Specifically, "adduced" is not a familiar word, and it's unclear what " between the MSF instruments" means.
The word “adduced” was replaced with the word “indicated”, and to clarify what is meant by between MSF instruments the following sentence was added:
“As shown in Risucci et al,33 there was strong concurrent validity for the medical colleague MSF questionnaire where supervisor and peer mean ratings on the same measures of physician performance correlated at $r = 0.92$, $p < 0.001$.”

Discussion , p. 11
- please comment upon the longitudinal and multi-study nature of the PAR and SPRAT programs, as compared to the others, and potential impact of this upon study rigour and program stability. This may lead to an important conclusion.
To emphasis the length of time that longitudinal and multi-studies of the PAR and SPRAT programs have been in place, the following sentence was modified to read:
“Most studies that provide evidence of reliability, generalizability, and validity (construct and criterion-related) are from the PAR process in Canada and the SPRAT instruments used in the UK where the longitudinal and multi-study nature of the MSF research on physician performance has been in progress for 16 and 8 years, respectively.”
In addition, the following sentence was included in the final paragraph:
“As indicated above, there exists a substantial body of rigorous and consistent research on the PAR and SPRAT programs that demonstrate the use of MSF will continue to play an important role in the formative and potentially summative assessment of physician performance in practice.”

- P.1 2 - Line 7 to the end of this paragraph about construct validity, other than the first bit about principal component analysis , is not transparent to the reader. Kindly explain the rationale for how these items relate to validity.
To clarify that there is a difference between physician discipline in what is being emphasized or measured with MSF questionnaires, the following sentence was modified and another sentence added to illustrate the variability.
“While the construct validity of MSF questionnaires may be found within a particular discipline (e.g., family medicine, internal medicine, surgery), many authors acknowledge that measures of various competencies or constructs are a function of

the specialization assessed (i.e., the percentage of variance associated with measures of patient management, clinical assessment, communication and/or professional development was found to vary across specialties).10,15,30,34
For example, Lockyer and Violato15 found in a principal component factor analysis of the medical colleague MSF questionnaire that the resulting four factor solution accounting for 73.4% of the variance for internal medicine physicians, 70% for psychiatrists and only 67.6% for pediatricians."

- as noted some claims made in the Results and other sections are unclear. Clarifying these may then require revising the Discussion and Conclusions to reflect changes made.
We have taken the revisions/additions into consideration and feel that they reflect the changes made.

Conclusions p. 12, 13 - please add references to substantiate these claims.
In the final paragraph, we added "In summary,…" at the beginning of the first sentence to indicated that we are generally summarizing the overall findings – adding the references that support this would ultimately include all of the primary studies in the systematic review.

Reviewer #2
Multi-Source Feedback is an important methodology used to provide information and assess learners and practitioners in health care. Analyzing the statistical properties of these tools is valuable. The authors are to be commended on identifying this timely topic for their review and on a clearly written paper. The abstract is well aligned and adequately summarizes the paper. The authors were in line with many of the published guidelines on conducting systematic reviews (1). The major deficit is in the lack of detailed description of the analytic processes used. Overall this is paper has merit but there are some issues that should be addressed.

1.The focus of the review is broad: "to investigate the evidence for reliability, generalizability, validity and feasibility". Given the various characteristics of each of those terms, a more detailed description of the analyses (see issue #5) conducted would help to focus the review parameters. There was no mention of other reviews done focused on MSF.
As far as we know there currently are not any other extensive MSF reviews published specific to the assessment of healthcare professions. The data were summarized within the categories identified as subheading within the text of the Results section and as headings at the top of the columns in Table 1 and 2. No statistical pooling or quantitative data analysis was conducted other than to compile by the number or percentage of studies that reported on any one specific area (i.e., country, specialty, MSF assessor types, etc.) Nevertheless, we have other revisions throughout based on some of the other reviewers' suggestions that we believe provide further clarification.

2.The qualifications of the review team are not mentioned. Was a medical librarian used to identify articles/keywords?
Two of the authors (TD and CV) have been involved and published meta-analyses/systematic reviews previously and publish extensively in the areas of educational/psychological assessment and evaluation. One of the other authors (AA) is a recent PhD graduate from our Medial Education Specialization program. A medical librarian was not required.

3.The timeframe for the population of studies included wasn't clearly justified. Given the relatively recent use of MSF in health sciences, why were studies from 1975 to the present included? What other studies/reviews were considered to help make this determination or to identify gaps?
To clarify that use of MSF is a relatively recent occurrence in physician assessment. The following sentences were added to identify when MSF with residents began and when a formal physician performance process was introduced later in the 1990s:
"While early attempts at the development of MSF questionnaires in medicine focused on the assessment of residents in the late 1970s, today they are being used in North America (Canada, US) and Europe (Netherlands, UK) across a number of physician specialties.4 As a self-regulating profession, medicine is accountable for ensuring that

physicians' are competent in the performance of their clinical roles and duties. Incumbent on regulatory bodies to monitor physician practice and patient safety, Canada was the first country to introduce a MSF process as a viable approach to providing an assessment of physician performance in the late 1990s."

4.What piloting was done for the search terms?
To clarify, we added in the Selection of studies subsection of the Methods section the following sentence:
"Initial identification of search terms to pilot were drawn from practical guides and a handbook on MSF.4,5"

5.How was the data analyzed? Was there any statistical pooling across studies? If so, what model was used? What qualitative approaches were used by the team to determine common constructs across studies (pg. 7 line 3)? Did the team look for variations across the studies? Without more transparency in the methods used, any threats to the validity of the review are difficult to ascertain and were not discussed in the study limitations.
The data were summarized within the categories identified as subheading within the text of the Results section and as headings at the top of the columns in Table 1 and 2. No statistical pooling or quantitative data analysis was conducted other than to compile by the number or percentage of studies that reported on any one specific area (i.e., country, specialty, MSF assessor types, etc.) In addition, these variations across studies (as related to variation of reported MSF validity measures was included as a separate study limitation (see #9 below).

6.Page 10, last line. Typo. After Time 2, it should read 3) consistently.
This was changed from a "2)" to a "3)".

7.It is not clear why "construct validity was provided" because "4) younger practitioners were rated higher than older ones". Depending on the factors assessed, age alone may not be an issue.
To clarify each of the indications of construct validity outlined, a short explanation was given in parentheses, and in regards to "4) young practitioners.." as follows:
"…, and 4) younger practitioners were rated higher than older ones (i.e., higher mean ratings are generally given to young practitioners that have been educated to be more conscious of MSF domain measures than practitioners that have been in practice for a greater number of years)."

8.Page 11. The paragraph on criterion-validity should be supported with summaries of the "positive correlations" found.
This paragraph was expanded to include an example of a strong positive correlation between MSF instruments in the following sentence added:
"As shown in Risucci et al,33 there was strong concurrent validity for the medical colleague MSF questionnaire where supervisor and peer mean ratings on the same measures of physician performance correlated at $r = 0.92$, $p < 0.001$."

9.Discussion. As noted in #5 above, the limitations do not address any potential threats to validity due to the team's analyses.
To clarify this as a limitation, this was acknowledged in a separate sentence as follows:
"Third, variability in the reporting of reliability (i.e., generalizability, intraclass correlation) and validity (i.e., construct and criterion-related) measures while supportive of the MSF process were difficult to combine consistently between studies."

10.Given that "each article focused on the use of a new MSF or a modified version of an existing instrument" (pg 11), the concluding statement (pg. 12) that MSF "is reliable, valid and feasible" seems a bit strong. Are all the instruments reviewed in this category?
To clarify, the sentence was re-worded to reflect that "In summary,…" this is the case and followed by a new sentence specific the PAR and SPRAT instruments that did fall into this category:
"In summary, MSF where various assessors (self, peers, coworkers, and patients) provide assessment of physicians' performance on various domains (clinical and nonclinical) is reliable, valid and feasible. As indicated above, there exists a substantial body of rigorous and consistent research on the PAR and SPRAT programs

	that demonstrate the use of MSF will continue to play an important role in the formative and potentially summative assessment of physician performance in practice." Reviewer #3 THE FOLLOWING REVIEW WAS PREPARED BY A MEMBER OF THE ACADEMIC MEDICINE EDITORIAL STAFF. ALL COMMENTS MUST BE ADDRESSED BEFORE RESUBMITTING YOUR MANUSCRIPT. 1.Please revise your abstract to be in the third person (e.g. "The authors searched EMBASE?" instead of "We searched EMBASE?"). The body of the paper, however, should use first person, active voice whenever possible. The abstract was revised in two places to be in the third person. 2.The Academic Medicine website offers a resource for preparing systematic reviews for publication: http://journals.lww.com/academicmedicine/Documents/AMSystematicReviewTips.pdf. In addition to addressing all external reviewer comments, I suggest you review this resource to make sure your manuscript contains all the required components of a systematic review. A few specific points: The AM Systematic Review Tips were reviewed to ensure the components were met. a.Be sure to comment on the level of agreement and how you resolved disagreement during the data abstraction process. To clarify the authors full agreement on included studies, the following sentence was included in the Data selection and abstraction subsection in the Methods section: "Review of all full-text articles was completed independently by the four authors until 100% agreement was achieved." b.Add details about how you addressed and minimized issues of publication, selection, and/or measurement bias during the data collection process. To clarify this issue, we added the following sentence at the beginning of the Data selection and abstraction subsection of the Methods section: "To address concerns of bias we conducted a comprehensive search using strict selection criteria based on rigorous interrater reliability." c.Comment on how you assessed the quality of the studies you included. To clarify that the quality of each of the studies included was determined to be 'high', we included the following sentence at the end of the Eligibility criteria subsection in the Methods section: "Although the studies included in this systematic review are based on the completion of MSF questionnaires by various assessors, the quality of the studies are considered to be 'high' for this type of research as each study needed to provide evidence of both reliability and construct (or criterion-related) validity." d.In the Results, be sure to cite all included studies at least once. At the minimum, this can be done by citing the whole set of articles when you note that the study included "a total of 43 peer-reviewed articles on physician MSF." The whole set of articles were cited at the end of this sentence. 3.Please remove table and figure placement notations from the text. Placement notations for Tables 1 & 2 and Figure 1 were removed from the text. 4.You are responsible for verifying that all the information in your reference list is present and correct. Please check citations against original publications for accuracy, check all links (if applicable) and update their access dates, and ensure that your references are formatted according to the AMA Manual of Style (see http://journals.lww.com/academicmedicine/Pages/references.aspx for more information about Academic Medicine's reference style). Citations were checked against the original publications for accuracy and formatted to reflect AMA style. Links were not referenced in this manuscript. 5.Please read the six disclosures statements below and add to your article the statements that are required and any others that may apply. The statements should be placed right after the end of your article. Acknowledgments: [This statement is optional. If you have no acknowledgments,

	please omit this statement. If you do have acknowledgments, please write them in the third person, e.g., "The authors thank?."] Not required Funding/Support: [This statement is required. If you have no sources of funding or support to list, please enter "None."] Added Other disclosures: [This statement is required. If you have no other disclosures to list, such as conflicts of interest, please enter "None."] Added Ethical approval: [This statement is required. If ethical approval was not needed, please enter "Not applicable." Otherwise, state the agency or group that granted approval, and make sure that this information is also in your report.] Added Disclaimer: [This statement is optional. If you do not wish to include a disclaimer, please omit this statement.] Not required Previous presentations: [This statement is optional. If you have no previous presentations to report (e.g., presenting the abstract; a poster; a speech), please omit this statement.] Not required

*Cover Letter (explaining revisions to manuscript)

Tyrone Donnon, PhD
Associate Professor
Medical Education and Research Unit
Department of Community Health Sciences
Faculty of Medicine, University of Calgary
3330 Hospital Drive NW, Calgary, Alberta, Canada T2N 4N1
Phone (403) 210-9682 Fax: (403) 270-7507
Email: tldonnon@ucalgary

May 07, 2013

David P. Sklar, MD
Editor-in-Chief
Academic Medicine

RE: ACADMED-D-13-00337, entitled "The Reliability, Validity and Feasibility of Multi-Source Feedback for Assessing Physicians: A Systematic Review"

Dear Dr. Sklar:

We have made the following point-by-point responses to the reviewers' comments by reiterate each comment verbatim and follow each one (in yellow highlighted text and bolded typeface and/or quotations) with our response to that comment in this cover letter, and have indicated where/how the manuscript has been revised to address the comments using the 'track changes' feature in Word.

Thank you for your consideration of this manuscript for publication in Academic Medicine.

Yours sincerely,

Tyrone Donnon, PhD (Corresponding Author)
Ahmed Al Ansari, MBBCh MRCSI, PhD
Samah Al Alawi, MD
Claudio Violato, PhD

Reviewer Comments:

Reviewer #1

General comments: This is a systematic review of MSF studies, reporting upon their reliability, feasibility and validity. The systematic review appears to have been conducted according to protocol and provides a worthwhile overview of the MSF studies since 1975. However, the writing is unclear at times, and definitions of terms and explanations of concepts that would enhance transparency are not included. Suggestions are included in the comments below.

Intro - Para 2, p.4 - needs rewording as follows:

* First sentence, line 38 "MSF is frequently used in workplace settings where employees work in a team and cannot be directly or easily supervised by managers",
* Please reword the first part of this sentence to say -" MSF originated in industry... "

The first part of this sentence was modified to read "MSF originated in industry…"

* Please confirm the remainder of this sentence by going back and checking the references cited , " where employees work in a team and cannot be directly or easily supervised by managers". I'm not sure this was the main reason. I believe it was the realization that others working with an individual could assess particular domains quite readily. Please check this.

The references cited were checked and the remainder of this sentence was re-written to reflect the main reason for the growth in the use of MSF in industry:

"MSF originated in industry during a time when the search for competent employees and the reliance on a single supervisor's evaluation was recognized as a restrictive approach to the assessment of a worker's specific abilities [5,6]"

Para 2, second sentence - a number of Canadian and US physicians still work mainly solo in private practice. Please reframe this sentence to reflect this.

To reflect the variability of persons that will work with a physician, and not necessarily in a team, we have rewritten this sentence to read:

"Similarly, physicians work with a variety of people (i.e., medical colleagues, consultants, therapists, nurses, and coworkers) that are able to provide a better assessment and contextually based understanding of physician performance than any single person."

Para 2, third sentence - Not all MSF programs include a self-assessment. Some programs involving residents include supervisors

To reflect the fact that not all MSF process include a self-assessment and that some physicians in-training will be assessed by a supervisor or preceptor, we have rewritten this sentence to read:

"In MSF physicians may complete a self-assessment instrument and receive feedback from a number of medical colleagues (peers), in-training supervisors or preceptors, non-physician co-workers (e.g., nurses, psychologists, pharmacists), as well as their own patients.[7]"

Intro, para 3, p 5:

- second sentence - other domains such as professionalism also the focus of MSF

The word "professionalism" was included in this sentence to indicate that it is also a MSF domain that is assessed.

- last sentence, rationale for the study, does not seem to flow clearly from the preceding sentence

The last sentence was re-written to fit better with the preceding sentence and purpose of the study:

"Accordingly, the purpose of the present study was to conduct a systematic review of the published, peer-reviewed research on the different types of MSF instruments used to assess physicians' clinical/nonclinical skills performance and to investigate the evidence for reliability, generalizability, validity and feasibility of this assessment approach."

Please expand the Introduction and lit review to provide evidence as to the scope to which MSF has been implemented; e.g. for physicians, residents, students and in approximately how many countries. This too will contribute to the rationale for the study.

To expand on the literature review in the Introduction section (staying just within the word count limits for the manuscript length) the following sentences were added to a revised second paragraph:

"While early attempts at the development of MSF questionnaires in medicine focused on the assessment of residents in the late 1970s, today they are being used in North America (Canada, US) and Europe (Netherlands, UK) across a number of physician specialties.[4] As a self-regulating profession, medicine is accountable for ensuring that physicians' are competent in the performance of their clinical roles and duties. Incumbent on regulatory bodies to monitor physician practice and patient safety, Canada was the first country to introduce a MSF process as a viable approach to providing an assessment of physician performance in the late 1990s…"

Methods:

Eligibility criteria, p. 5 - 6 - it appears from reading the results that studies of residents and students were also included. Please clarify this in the eligibility section.

To clarify that studies that looked at MSF with residents were also eligible for inclusion, we modified inclusion criteria #1 in the "Eligibility criteria" to read:

"…1) use one or more multi-source feedback instruments (e.g., self, colleague, coworker, and/or patient) to assess physician or resident performance in practice,.."

And clarified that studies that looked at MSF with medical students were not eligible or excluded, we modified the exclusion criteria #1 to read:

"…1) were used to assess other than physicians or residents (i.e., medical students) or non-physician health professionals…"

Line 7, p. 6 - We excluded studies if they 1) were used to assess other than physicians or non-physician health professionals, - this is not clear.

To clarified that studies that looked at MSF with non-physician health professionals were excluded, we modified the exclusion criteria #1 to include specific example groups:

"…1) were used to assess other than physicians or residents (i.e., medical students) or non-physician health professionals (i.e., nurses, occupational or respiratory therapists, chiropractors, etc.),…"

Study selection process appears clear and appropriate.

Results:
p.6-7 - clarity would be added to the tables by grouping the studies as described in this paragraph; i.e., -
1. Physician Assessment Review (Canada n = x, Netherlands = 1)
2. Sheffield Peer review Assessment Tool (UK n=x)
3. Other UK studies (n = X)
4. USA studies (n = x)
5. Studies from other countries (n = 4)

To clarify the studies included in the systematic review as they are grouped in the Tables, the first paragraph in Results section was re-written as follows:

"Although there are a variety of MSF instruments used in the studies, they include: the Physician Assessment Review (PAR) process (Canada, *n* = 13; Netherlands, *n* = 1), the Sheffield Peer Review Assessment Tool (SPRAT) process (UK, *n* = 6), multiple MSF instruments from the USA (*n* = 14), other UK related studies (*n* = 4), and three separate studies from other countries (China, Denmark and Taiwan)."

Specialty - this para could be written more clearly, or perhaps use a table?

To clarify the studies included in the systematic review as they are grouped by specialty, the first paragraph in the Specialty of Physicians Assessed Using MSF subsection was re-written as follows:

"There were a number of MSF studies that assessed physicians across multiple specialties (*n* = 10). In a study of the psychometrics of the PAR MSF instruments, for example, Hall et al.[13] evaluated the results from 308 physicians from multiple specialties in Alberta. With respect to specific physician practices there were MSF studies for each of the following specialties: family medicine (*n* = 5), pediatrics (*n* = 5), internal medicine (*n* = 5), surgery (*n* = 4), obstetrics/gynecology (*n* = 3), psychiatry (*n* = 3), anesthesia (*n* = 2), and single studies for emergency medicine, pathology/laboratory medicine, histopathology, radiology, and physical medicine and rehabilitation."

Types of MSF instruments used - this section might be better named- " Raters and length of questionnaires"

We modified the subsection title to read:

"MSF Assessors and Length of Questionnaires"

- re raters, did any include residents' supervisors or attendings?

To clarify that in some studies where the physicians (residents in-training) may have been evaluate by peers or medical colleagues that are their superiors, the following addition was added to the sentence:

"In MSF with physicians, information can come from a variety of sources (i.e., peers or medical colleagues including supervisors and preceptors,…"

- - para 2 in this section, first sentence - shorten to "The questionnaires used ranged in length from..."

To summarize the variability in length of the various MSF questionnaires used/included in Table 1, the first sentence of this 2nd paragraph now reads:
"The MSF questionnaires varied greatly in the number of items depending on the assessor: 4 to 57 items for self-assessment, 4 to 60 items for peer or medical colleague, 4 to 60 items for co-workers, and 3 to 49 items for patient questionnaires."

Constructs/ domains assessed
- first sentence, suggest wording as " As shown in Table 1, a number of constructs were measured using MSF. "
The first and second sentence were combined to now read:
"As shown in Table 1, a number of constructs were measured using MSF: 1) professionalism, 2), clinical competence, 3) communication, 4) manager, and 5) interpersonal relationship."

- identification of constructs: Please describe how you did this. E.g., Were they consistently identified by the authors, or did you have to interpret the authors' descriptions? How did communication differ from interpersonal relationships? Please define/ give examples of both of these. Also for manager. This will add clarity for the reader and also contribute to understanding of construct validity.
To clarify that there was consensus among the authors and to provide examples of how specific items from communication differ from interpersonal relationships and manager categories, we added the following sentences to the first paragraph
"Consensus for the five general category domains was achieved by three of the authors (TD, AA, SA) and were based on existing constructs or examples of items provided from the included studies"… "For example, items that were written "Communicates effectively with patients" or "Communicates effectively with other health care professionals" were clearly associated with the communication category, "Collaborates with medical colleagues" the interpersonal relationship category, and "Manages health care resources efficiently" the manager category.[13]"

Administration and feasibility - this section appears unclear.
It would help the reader to have definitions of "administration" and "feasibility". What criteria were looked for in descriptions of each of these? Eg, should "Administration" include how it was developed, administered and # of participants?
In Table 1, some data in the "Administration and feasibility" column appear to be psychometric in nature, and others, more admin or feasibility oriented. Consistency would help the reader.
To clarify that this section is as much about the general information about the process than just the administration and/or feasibility, the subsection heading was changed in the text (as well as on Table 1) and the first sentence in the first paragraph was modified and another sentence added.
General Information on Process, Administration and/or Feasibility
"Each of the 42 studies included in the MSF systematic review provided general information about the findings of their study with comments on the process, administration, and/or feasibility (Table 1). For example, general information comments emphasized how studies' psychometric results provided support for the MSF process, was able to be administered to various participants in an efficient manner, and/or was a feasible method to collect multiple performance measures of physicians in practice."

Reliablity and generalizabilty - please indicate which studies used each of these analyses.
At the top of the column in Table 2, the studies that reported either or both reliability and generalizability coefficients for each of the MSF questionnaires used are identified with the recognized, corresponding statistical symbols "α" and "Ep^2" respectively. For example, in the first row/study for Violato et al., 1997 there are reliability coefficients reported for each of the MSF instruments (range from α = 0.89 to 0.95) and generalizability coefficients for the Medical Colleague (Ep^2 = 0.77 for 8 raters) and Patient (Ep^2= 0.80 for 25 raters) questionnaires.

Construct and Criterion-Related Validity -
- para 2: Please provide a sentence describing how each of these analyses indicates construct or criterion validity
To clarify each of the indications of construct validity outlined, a short explanation was given in parentheses as follows:
"Further evidence of construct validity was provided through analyses that showed: 1) measures of mean difference ratings between respondent groups (i.e., mean ratings from patients and coworkers are consistently higher than medical colleagues and are lowest on self-assessments), 2) improvement in performance ratings from Time 1 to Time 2 (i.e., increase in mean ratings are consistently higher from an earlier period, indicating an expected improvement in practice over time), 3) consistently higher ratings given to advanced trainees by year of program (i.e., increase in mean ratings as residents gain clinical experience from year to year of an in-training program), and 4) younger practitioners were rated higher than older ones (i.e., higher mean ratings are generally given to young practitioners that have been educated to be more conscious of MSF domain measures than practitioners that have been in practice for a greater number of years)."

- para 3 - this sentence is unclear: "Criterion-related validity was adduced in some studies where positive correlations: 1) were found between the MSF instruments/measures (concurrent validity)". Specifically, "adduced" is not a familiar word, and it's unclear what " between the MSF instruments" means.
The word "adduced" was replaced with the word "indicated", and to clarify what is meant by between MSF instruments the following sentence was added:
"As shown in Risucci et al,[33] there was strong concurrent validity for the medical colleague MSF questionnaire where supervisor and peer mean ratings on the same measures of physician performance correlated at $r = 0.92$, $p < 0.001$."

Discussion , p. 11
- please comment upon the longitudinal and multi-study nature of the PAR and SPRAT programs, as compared to the others, and potential impact of this upon study rigour and program stability. This may lead to an important conclusion.
To emphasis the length of time that longitudinal and multi-studies of the PAR and SPRAT programs have been in place, the following sentence was modified to read:
"Most studies that provide evidence of reliability, generalizability, and validity (construct and criterion-related) are from the PAR process in Canada and the SPRAT instruments used in the UK where the longitudinal and multi-study nature of the MSF research on physician performance has been in progress for 16 and 8 years, respectively."
In addition, the following sentence was included in the final paragraph:

"As indicated above, there exists a substantial body of rigorous and consistent research on the PAR and SPRAT programs that demonstrate the use of MSF will continue to play an important role in the formative and potentially summative assessment of physician performance in practice."

- P.1 2 - Line 7 to the end of this paragraph about construct validity, other than the first bit about principal component analysis , is not transparent to the reader. Kindly explain the rationale for how these items relate to validity.

To clarify that there is a difference between physician discipline in what is being emphasized or measured with MSF questionnaires, the following sentence was modified and another sentence added to illustrate the variability.

"While the construct validity of MSF questionnaires may be found within a particular discipline (e.g., family medicine, internal medicine, surgery), many authors acknowledge that measures of various competencies or constructs are a function of the specialization assessed (i.e., the percentage of variance associated with measures of patient management, clinical assessment, communication and/or professional development was found to vary across specialties).[10,15,30,34] For example, Lockyer and Violato[15] found in a principal component factor analysis of the medical colleague MSF questionnaire that the resulting four factor solution accounting for 73.4% of the variance for internal medicine physicians, 70% for psychiatrists and only 67.6% for pediatricians."

- as noted some claims made in the Results and other sections are unclear. Clarifying these may then require revising the Discussion and Conclusions to reflect changes made.

We have taken the revisions/additions into consideration and feel that they reflect the changes made.

Conclusions p. 12, 13 - please add references to substantiate these claims.

In the final paragraph, we added "In summary,…" at the beginning of the first sentence to indicated that we are generally summarizing the overall findings – adding the references that support this would ultimately include all of the primary studies in the systematic review.

Reviewer #2

Multi-Source Feedback is an important methodology used to provide information and assess learners and practitioners in health care. Analyzing the statistical properties of these tools is valuable. The authors are to be commended on identifying this timely topic for their review and on a clearly written paper. The abstract is well aligned and adequately summarizes the paper. The authors were in line with many of the published guidelines on conducting systematic reviews (1). The major deficit is in the lack of detailed description of the analytic processes used. Overall this is paper has merit but there are some issues that should be addressed.

1.The focus of the review is broad: "to investigate the evidence for reliability, generalizability, validity and feasibility". Given the various characteristics of each of those terms, a more detailed description of the analyses (see issue #5) conducted would help to focus the review parameters. There was no mention of other reviews done focused on MSF.

As far as we know there currently are not any other extensive MSF reviews published specific to the assessment of healthcare professions. The data were summarized within the categories identified as subheading within the text of the Results section and as headings at the top of the columns in Table 1 and 2. No statistical pooling or quantitative data analysis was conducted other than to compile by the number or percentage of studies that reported on any one specific area (i.e., country, specialty, MSF assessor types, etc.) Nevertheless, we have other revisions throughout based on some of the other reviewers' suggestions that we believe provide further clarification.

2.The qualifications of the review team are not mentioned. Was a medical librarian used to identify articles/keywords?
Two of the authors (TD and CV) have been involved and published meta-analyses/systematic reviews previously and publish extensively in the areas of educational/psychological assessment and evaluation. One of the other authors (AA) is a recent PhD graduate from our Medial Education Specialization program. A medical librarian was not required.

3.The timeframe for the population of studies included wasn't clearly justified. Given the relatively recent use of MSF in health sciences, why were studies from 1975 to the present included? What other studies/reviews were considered to help make this determination or to identify gaps?
To clarify that use of MSF is a relatively recent occurrence in physician assessment. The following sentences were added to identify when MSF with residents began and when a formal physician performance process was introduced later in the 1990s:
"While early attempts at the development of MSF questionnaires in medicine focused on the assessment of residents in the late 1970s, today they are being used in North America (Canada, US) and Europe (Netherlands, UK) across a number of physician specialties.[4] As a self-regulating profession, medicine is accountable for ensuring that physicians' are competent in the performance of their clinical roles and duties. Incumbent on regulatory bodies to monitor physician practice and patient safety, Canada was the first country to introduce a MSF process as a viable approach to providing an assessment of physician performance in the late 1990s."

4.What piloting was done for the search terms?
To clarify, we added in the Selection of studies subsection of the Methods section the following sentence:
"Initial identification of search terms to pilot were drawn from practical guides and a handbook on MSF.[4,5]"

5.How was the data analyzed? Was there any statistical pooling across studies? If so, what model was used? What qualitative approaches were used by the team to determine common constructs across studies (pg. 7 line 3)? Did the team look for variations across the studies? Without more transparency in the methods used, any threats to the validity of the review are difficult to ascertain and were not discussed in the study limitations.
The data were summarized within the categories identified as subheading within the text of the Results section and as headings at the top of the columns in Table 1 and 2. No statistical pooling or quantitative data analysis was conducted other than to compile by the

number or percentage of studies that reported on any one specific area (i.e., country, specialty, MSF assessor types, etc.) In addition, these variations across studies (as related to variation of reported MSF validity measures was included as a separate study limitation (see #9 below).

6.Page 10, last line. Typo. After Time 2, it should read 3) consistently.
This was changed from a "2)" to a "3)".

7.It is not clear why "construct validity was provided" because "4) younger practitioners were rated higher than older ones". Depending on the factors assessed, age alone may not be an issue.
To clarify each of the indications of construct validity outlined, a short explanation was given in parentheses, and in regards to "4) young practitioners.." as follows:
"…, and 4) younger practitioners were rated higher than older ones (i.e., higher mean ratings are generally given to young practitioners that have been educated to be more conscious of MSF domain measures than practitioners that have been in practice for a greater number of years)."

8.Page 11. The paragraph on criterion-validity should be supported with summaries of the "positive correlations" found.
This paragraph was expanded to include an example of a strong positive correlation between MSF instruments in the following sentence added:
"As shown in Risucci et al,[33] there was strong concurrent validity for the medical colleague MSF questionnaire where supervisor and peer mean ratings on the same measures of physician performance correlated at $r = 0.92$, $p < 0.001$."

9.Discussion. As noted in #5 above, the limitations do not address any potential threats to validity due to the team's analyses.
To clarify this as a limitation, this was acknowledged in a separate sentence as follows:
"Third, variability in the reporting of reliability (i.e., generalizability, intraclass correlation) and validity (i.e., construct and criterion-related) measures while supportive of the MSF process were difficult to combine consistently between studies."

10.Given that "each article focused on the use of a new MSF or a modified version of an existing instrument" (pg 11), the concluding statement (pg. 12) that MSF "is reliable, valid and feasible" seems a bit strong. Are all the instruments reviewed in this category?
To clarify, the sentence was re-worded to reflect that "In summary,…" this is the case and followed by a new sentence specific the PAR and SPRAT instruments that did fall into this category:
"In summary, MSF where various assessors (self, peers, coworkers, and patients) provide assessment of physicians' performance on various domains (clinical and nonclinical) is reliable, valid and feasible. As indicated above, there exists a substantial body of rigorous and consistent research on the PAR and SPRAT programs that demonstrate the use of MSF will continue to play an important role in the formative and potentially summative assessment of physician performance in practice."

Reviewer #3

THE FOLLOWING REVIEW WAS PREPARED BY A MEMBER OF THE ACADEMIC MEDICINE EDITORIAL STAFF. ALL COMMENTS MUST BE ADDRESSED BEFORE RESUBMITTING YOUR MANUSCRIPT.

1. Please revise your abstract to be in the third person (e.g. "The authors searched EMBASE?" instead of "We searched EMBASE?"). The body of the paper, however, should use first person, active voice whenever possible.

The abstract was revised in two places to be in the third person.

2. The Academic Medicine website offers a resource for preparing systematic reviews for publication: http://journals.lww.com/academicmedicine/Documents/AMSystematicReviewTips.pdf. In addition to addressing all external reviewer comments, I suggest you review this resource to make sure your manuscript contains all the required components of a systematic review. A few specific points: **The AM Systematic Review Tips were reviewed to ensure the components were met.**

a. Be sure to comment on the level of agreement and how you resolved disagreement during the data abstraction process.

To clarify the authors full agreement on included studies, the following sentence was included in the Data selection and abstraction subsection in the Methods section:
"Review of all full-text articles was completed independently by the four authors until 100% agreement was achieved."

b. Add details about how you addressed and minimized issues of publication, selection, and/or measurement bias during the data collection process.

To clarify this issue, we added the following sentence at the beginning of the Data selection and abstraction subsection of the Methods section:
"To address concerns of bias we conducted a comprehensive search using strict selection criteria based on rigorous interrater reliability."

c. Comment on how you assessed the quality of the studies you included.

To clarify that the quality of each of the studies included was determined to be 'high', we included the following sentence at the end of the Eligibility criteria subsection in the Methods section:
"Although the studies included in this systematic review are based on the completion of MSF questionnaires by various assessors, the quality of the studies are considered to be 'high' for this type of research as each study needed to provide evidence of both reliability and construct (or criterion-related) validity."

d. In the Results, be sure to cite all included studies at least once. At the minimum, this can be done by citing the whole set of articles when you note that the study included "a total of 43 peer-reviewed articles on physician MSF."

The whole set of articles were cited at the end of this sentence.

3. Please remove table and figure placement notations from the text.

Placement notations for Tables 1 & 2 and Figure 1 were removed from the text.

4. You are responsible for verifying that all the information in your reference list is present and correct. Please check citations against original publications for accuracy, check all links (if applicable) and update their access dates, and ensure that your references are formatted according to the AMA Manual of Style (see http://journals.lww.com/academicmedicine/Pages/references.aspx for more information about Academic Medicine's reference style).
Citations were checked against the original publications for accuracy and formatted to reflect AMA style. Links were not referenced in this manuscript.

5. Please read the six disclosures statements below and add to your article the statements that are required and any others that may apply. The statements should be placed right after the end of your article.

Acknowledgments: [This statement is optional. If you have no acknowledgments, please omit this statement. If you do have acknowledgments, please write them in the third person, e.g., "The authors thank?."]
Not required

Funding/Support: [This statement is required. If you have no sources of funding or support to list, please enter "None."]
Added

Other disclosures: [This statement is required. If you have no other disclosures to list, such as conflicts of interest, please enter "None."]
Added

Ethical approval: [This statement is required. If ethical approval was not needed, please enter "Not applicable." Otherwise, state the agency or group that granted approval, and make sure that this information is also in your report.]
Added

Disclaimer: [This statement is optional. If you do not wish to include a disclaimer, please omit this statement.]
Not required

Previous presentations: [This statement is optional. If you have no previous presentations to report (e.g., presenting the abstract; a poster; a speech), please omit this statement.]
Not required

Research Report

Title:

The Reliability, Validity and Feasibility of Multi-Source Feedback for Assessing Physicians: A Systematic Review

Authors:

Tyrone Donnon, PhD, Ahmed Al Ansari, MBBCh MRCSI~~,~~ PhD, Samah Al Alawi, MD, Claudio Violato, PhD

Bios:

Dr. Tyrone Donnon is an associate professor with the Medical Education and Research Unit, Department of Community Health Sciences, Faculty of Medicine, University of Calgary, Calgary, Canada.

Dr. Ahmed Al Ansari is ~~senior resident~~the director of training and development in the Department of ~~Surgery~~Medical Education, Faculty of Medicine, Bahrain Defense Force Hospital, Riffa, Bahrain.

Dr. Samah Al Alawi is a faculty member in the Department of Family Medicine, Faculty of Medicine, Bahrain Defense Force Hospital, Riffa, Bahrain.

Dr. Claudio Violato is a professor with the Medical Education and Research Unit, Department of Community Health Sciences, Faculty of Medicine, University of Calgary, Calgary, Canada.

Correspondence to:

Tyrone Donnon, Medical Education and Research Unit, G13 Health Medical Research Bldg, Faculty of Medicine, University of Calgary, 3330 Hospital Drive, NW, Calgary, AB Canada, T2N 4N1. Tel: 403-210-9682; Fax: 403-210-7507; E-mail: tldonnon@ucalgary.ca

Abstract

The Reliability, Validity and Feasibility of Multi-Source Feedback for Assessing Physicians: A Systematic Review

Purpose

The use of multisource feedback (MSF) or 360 degree evaluation has become a recognized method of assessing physician performance in practice. The purpose of the present systematic review was to investigate the reliability, generalizability, validity, and feasibility of MSF for the assessment of physicians.

Method

The authors~~We~~ searched the EMBASE, PsycINFO, MEDLINE, PUBMED, and CINAHL databases for peer-reviewed, English-language articles up to January, 2013. Studies were included if they met the following inclusion criteria: use one or more MSF instruments to assess physician performance in practice, reported psychometric evidence of the instrument(s) in the form of reliability, generalizability coefficients and construct or criterion-related validity, and provided information regarding the administration or feasibility of the process in collecting the feedback data.

Results

Of the 96 full-text articles assessed for eligibility, ~~we include~~ 43 articles were included in the final systematic review. The use of MSF has been shown to be an effective method for providing feedback to physicians from a multitude of specialties about their clinical and nonclinical (i.e., professionalism, communication, interpersonal relationship, management) performance. In general, assessment of physician performance was based on the completion of

the MSF instruments by 8 medical colleagues, 8 coworkers and 25 patients to achieve adequate reliability and generalizability coefficients of $\alpha \geq 0.90$ and $Ep^2 \geq 0.80$, respectively.

Conclusions

The use of multisource feedback employing medical colleagues, coworkers, and patients as a method to assess physicians in practice has been shown to have high reliability, validity and feasibility.

Keywords: Multisource feedback, systematic review, physician performance, reliability, construct validity

Assessment and maintenance of physician competence is of a great importance to healthcare physician organizations. This is particularly important with growing concerns for patient safety[1] and an understanding of the importance that professional roles and responsibilities, including interpersonal skills and professionalism, should be integrated into physicians' clinical practice.[2] Thus, the view of competence has changed from a focus on the ability to conduct specific medical procedures to a more comprehensive framework for the assessment of physician performance.[3] Multi-source feedback (MSF), also referred to as 360 degree evaluation, has emerged as an important approach for assessing professional competence, behaviours, and attitudes in the workplace.[4]

While early attempts at the development of MSF questionnaires in medicine focused on the assessment of residents in the late 1970s, today they are being used in North America (Canada, US) and Europe (Netherlands, UK) across a number of physician specialties.[4] As a self-regulating profession, medicine is accountable for ensuring that physicians' are competent in the performance of their clinical roles and duties. Incumbent on regulatory bodies to monitor physician practice and patient safety, Canada was the first country to introduce a MSF process as a viable approach to providing an assessment of physician performance in the late 1990s.

Typically, this feedback is collected using surveys or questionnaires designed to elicit responses from various respondents (e.g., peers, coworkers, patients) and, in some cases, from the physicians themselves through a corresponding self-assessment version of the measurement instrument. MSF has gained widespread acceptance for evaluation of professionals and is seen as a catalyst for the practitioner to reflect on where change may be required.

~~MSF is frequently used in workplace settings where employees work in a team and cannot be directly or easily supervised by managers.~~MSF originated in industry during a time

when the search for competent employees and the reliance on a single supervisor's evaluation was recognized as a restrictive approach to the assessment of a worker's specific abilities.[5,6] Similarly, physicians work ~~in teams~~ with a variety of people (i.e., medical colleagues, consultants, therapists, nurses, and coworkers) that are able to provide a better assessment and contextually based understanding of physician performance than any single person. In MSF physicians may complete a self-assessment instrument and receive feedback from a number of medical colleagues (peers), in-training supervisors or preceptors, non-physician co-workers (e.g., nurses, psychologists, pharmacists), as well as their own patients.[7] Different respondents focus on characteristics of the physician that they can assess (e.g., patients are not expected to assess a physician's clinical expertise) and provide a more comprehensive evaluation than what could be derived by any one source alone.[8]

MSF is gaining acceptance and credibility as a means of providing doctors with relevant information about their practice to help them monitor, develop, maintain and improve their competence. MSF has focused on clinical skills, communication, collaboration with other health care professionals, professionalism and patient management.[9] Accordingly, the purpose of the present study was to conduct a systematic review of the published, peer-reviewed research on the different types of MSF instruments used to assess physicians' clinical/nonclinical skills performance and to investigate the evidence for reliability, generalizability, validity and feasibility of this assessment approach.

Method

Selection of studies

A systematic review of the research on MSF published from the 1975 to January 2013 was conducted using the following databases: MEDLINE, PubMed, EMBASE, CINAHL, and

PsycINFO and the Cochrane Database of Systematic Reviews. Initial identification of search terms to pilot were drawn from practical guides and a handbook on MSF.[4,5] The search was limited to English language, peer-reviewed journals, using the following terms: "multisource-feedback" and "360 degree evaluation" to identify MSF related studies and combined them with them with physician related assessments with the terms "assessment of physician competencies," "assessment of physician professionalism," "assessment of physician in practice." We also manually searched from the reference lists of relevant studies.

Eligibility criteria

Studies were included if they: 1) use one or more multi-source feedback instruments (e.g., self, colleague, coworker, and/or patient) to assess physician or resident performance in practice, 2) describe the MSF instrument or its' design, 3) reported psychometric evidence of the instrument(s) in the form of reliability, generalizability and/or feasibility (administration) of collecting the feedback data, 4) provided evidence of either construct and/or criterion-related validity (predictive/concurrent), and 5) published in an English language, peer-reviewed journal. We excluded studies if they 1) were used to assess other than physicians or residents (i.e., medical students) or non-physician health professionals (i.e., nurses, occupational or respiratory therapists, chiropractors, etc.), and 2) studies failed to provide adequate information about the psychometrics of the MSF instrument (reliability and validity). For example, Violato and Lockyer[10] compared mean self and peer MSF ratings between three different specialties, Sinclair et al.[11] focused on the issue of patient reliability using the SHEFFPAT questionnaire, and Noonan et al.[12] provided information on the test-retest reliability, but all three of these studies failed to provide an analysis on the validity of the MSF instruments. Although the studies included in this systematic review are based on the completion of MSF questionnaires by various

assessors, the quality of the studies are considered to be 'high' for this type of research as each study needed to provide evidence of both reliability and construct (or criterion-related) validity.

Data selection and abstraction

To address concerns of bias we conducted a comprehensive search using strict selection criteria based on rigorous interrater reliability. Each article in the present study was reviewed and coded by two authors (TD and AA) independently; initially titles and abstracts were screened before full-text articles were assessed for eligibility (Figure 1). Review of all full-text articles was completed independently by the four authors until 100% agreement was achieved. Once articles were identified for inclusion, the following information was extracted: the name of the MSF instrument (unless a specific name was provided for the MSF instrument, the generic terms '360 degree evaluation' or 'multi-source feedback' were used), physician specialty, number of participants, assessor type, construct/factors assessed by the MSF instrument, administration/feasibility issues, mean number of raters per assessor type (response percentage), reliability/ generalizability/ intra-class correlation coefficients, and analysis of construct and criterion-related validity.

~~[Insert Figure 1]~~

Results

As shown in Figure 1, the review of 96 full-text studies resulted in a total of 43 peer-reviewed articles on physician MSF (Table 1).[7,13-54] Although there are a variety of MSF instruments used in the studies, they include: the Physician Assessment Review (PAR) process (Canada, *n* = 13; Netherlands, *n* = 1), the Sheffield Peer Review Assessment Tool (SPRAT) process (UK, *n* = 6), multiple MSF instruments from the USA (*n* = 14), other UK related studies (*n* = 4), and three separate studies from other countries (China, Denmark and Taiwan). ~~Thirteen~~

~~(30%) of the articles were studies from Canada and focused on the use of the Physician Assessment Review (PAR) MSF instruments, six (14%) articles were from the UK and used the Sheffield Peer Review Assessment Tool (SPRAT), and 14 (33%) studies from the US used a variety of MSF instruments (although common constructs/factors were assessed across these studies, each MSF instrument was different). The UK studies included in the systematic review also reported on the use of a variety of other MSF instruments (*n* = 6, 14%). There were four (9%) articles included that were from other countries (i.e., China, Denmark, Netherlands, Taiwan. The study from the Netherlands used a modified version of the PAR MSF instruments from Canada.~~

~~[Insert Table 1]~~

Specialty of Physicians Assessed Using MSF

There were ~~10 (23%)~~a number of MSF studies that assessed physicians across multiple specialties (*n* = 10). In a study of the psychometrics of the PAR MSF instruments, for example, Hall et al.[13] evaluated the results from 308 physicians from multiple specialties in Alberta. With respect to specific physician practices , ~~there were five (12%)~~MSF studies for each of the following specialties: family medicine (*n* = 5), pediatrics (*n* = 5), ~~and~~ internal medicine (*n* = 5), ~~for a total of 15 (35%) articles. Other specialties that were used in MSF articles included~~ surgery (*n* = 4~~, 9%~~), obstetrics/gynecology (*n* = 3~~, 7%~~), psychiatry (*n* = 3~~, 7%~~), anesthesia (*n* = 2~~, 5%~~), and ~~in~~ single studies ~~of~~for emergency medicine, pathology/laboratory medicine, histopathology, radiology, and physical medicine and rehabilitation.

MSF Assessors and Length of Questionnaires ~~Types of MSF Instruments Identified~~

In MSF with physicians, information can come from a variety of sources (i.e., ~~medical colleagues or~~ peers or medical colleagues including supervisors and preceptors, co-workers such

as nurses and other allied health professionals, patients and their families, and a self-assessment). In 38 (91%) of the studies, the use of a MSF instrument was completed by the physicians' peers or medical colleagues. In most studies, however, assessment were also obtained from coworkers (*n* = 32, 74%), patients and/or their families (*n* = 23, 53%), and from self-assessments (*n* = 22, 51%).

The MSF questionnaires varied greatly in the number of items depending on the assessor: 4 to 57 items for self-assessment, 4 to 60 items for peer or medical colleague, 4 to 60 items for co-workers, and 3 to 49 items for patient questionnaires. The PAR studies use a variety of MSF instruments for each of the assessors with the number of items (depending on specialty) ranging from 11 to 40 items for the patient, 12 to 22 for the coworker, 22 to 39 for the medical colleague, and 21 to 39 for the self-assessment instrument. The SPRAT uses the same 24 item MSF instrument for medical colleagues and coworkers, although modified versions for histopathology (21 item PATH-SPRAT),[27] junior residents (16 item mini-PAT),[28] and patients (13 item SHEFFPAT)[29] have been introduced. In two studies, medical students were also involved in the MSF process and completed the same 10 or 12 item instrument that medical colleagues, coworkers and patients used.[39,45]

Constructs/Domains Assessed

As shown in Table 1, ~~there are~~ a number of constructs ~~that can be~~were measured using MSF~~. We identified five constructs~~: 1) professionalism, 2), clinical competence, 3) communication, 4) manager, and 5) interpersonal relationship. Consensus for the five general category domains was achieved by three of the authors (TD, AA, SA) and were based on existing constructs or examples of items provided from the included studies. Professionalism, for example, consisted of a variety of measures of psychosocial skills, professional management/

responsibilities, humanistic qualities, compassion, attitude, teaching and professional development. Clinical Competence included items that assessed clinical care, good medical practice, patient care, safe practice, clinical performance, clinical knowledge, critical thinking, diagnosis, and management of complex problems. Items connected to the 'communication,' 'interpersonal relationship, and 'manager' constructs were group and categorized similarly. For example, items that were written "Communicates effectively with patients" or "Communicates effectively with other health care professionals" were clearly associated with the communication category, "Collaborates with medical colleagues" the interpersonal relationship category, and "Manages health care resources efficiently" the manager category.[13]

General Information on Process, Administration and/or Feasibility

Each of the 42 studies included in the MSF systematic review provided general information about the findings of their study with comments on the process, administration, and/or feasibility ~~or both~~ (Table 1). For example, general information comments emphasized how studies' psychometric results provided support for the MSF process, was able to be administered to the various participants in an efficient manner, and/or was a feasible method to collect multiple performance measures of physicians in practice. Researchers have acknowledged that the MSF instruments are effective when used in triangulation with patients, coworkers and medical colleagues in conjunction with the physician's self-assessment.[7] The authors of some studies recognize that the feedback provided to physicians regarding their performance on key competencies have the potential to initiate changes in practice.[14] There was an initial PAR study that considered MSF to be feasible as a function of the estimated cost per physician, but it was suggested that the MSF on the physician be re-administered every five years.[13] In a subsequent PAR study, family medicine physicians were re-assessed over a five

year period (i.e., Time 1 and Time 2) providing evidence of measurement stability but the incorporation of feedback by the physicians was limited.[20,21] In PAR related studies, the administration of the MSF process was found to be feasible and adaptable for a variety of specialties (e.g., paediatrics,[19] surgery,[14] emergency medicine,[17] family medicine,[20] psychiatry,[22] etc.) and potentially for use in other countries.[24] Although the SPRAT originated with the use of a common 24-item MSF instrument for medical colleagues and coworkers in paediatrics, modified versions of the peer review assessment instruments has also been used with multiple specialities.[26-31] In 2008, the study by Crossley et al.[29] introduced a 13-item patient MSF instrument (SHEFFPAT) that in a subsequent study by Archer and McAvoy[31] failed to show that patients were able to identify doctors in potential difficulty.

Reliability and Generalizability of MSF Instruments

The reliability of the various MSF instruments was reported in 26 (62%) of the studies included in this systematic review. Reliability coefficients are reported typically as Cronbach's alpha (α) and reflect the internal consistency of the items. MSF instruments should have an $\alpha \geq 0.90$, which is typically achieved in PAR related studies for the medical colleague (0.89 to 0.99), coworker (0.91 to 0.96), and patient (0.93 to 0.99) instruments. Although only one of the SPRAT studies included a combined medical colleague and coworker reliability coefficient ($\alpha = 0.98$),[28] the standard error of measurement (SEM) was calculated for 5 of the 6 included studies. In general, to achieve a SEM of $\pm$ 0.40 with the combined SPRAT a minimum of 8 raters are required.

Using generalizability analyses, generalizability coefficients (Ep^2) were derived in 17 studies (40%). Ep^2 provides a measure of the dependability of the MSF instruments as a function of the various factors that can influence the physicians' ratings. The coefficients for the

medical colleague instrument ranged from Ep^2 = 0.61 to 0.88, coworker from 0.56 to 0.87, and patient from 0.65 to 0.85. In four studies, the intraclass correlation coefficient (ICC) was calculated as a way to determine the consistency in ratings across the evaluators and were found to range from 0.45 to 0.90 (suggesting that the ratings obtained from the various evaluators was moderate to highly consistent).

~~[Insert Table 2]~~

Construct and Criterion-Related Validity

To be included in this systematic review, a study had to provide evidence of either construct and/or criterion-related validity (predictive/concurrent). In 28 (67%) of the studies, evidence for the construct validity of the MSF instrument used was provided through exploratory factor analyses (principal component). As we have seen, each of the MSF instruments were found to assess a variety of constructs based on the particular instrument used (i.e., PAR, SPRAT, other) or the respondent (i.e., medical colleague, coworker, patient).

Further evidence of construct validity was provided through analyses that showed: 1) measures of mean difference ratings between respondent groups (i.e., mean ratings from patients and coworkers are consistently higher than medical colleagues and are lowest on self-assessments), 2) improvement in performance ratings from Time 1 to Time 2 (i.e., increase in mean ratings are consistently higher from an earlier period, indicating an expected improvement in practice over time), ~~2~~3) consistently higher ratings given to advanced trainees by year of program (i.e., increase in mean ratings as residents gain clinical experience from year to year of an in-training program), and 4) younger practitioners were rated higher than older ones (i.e., higher mean ratings are generally given to young practitioners that have been educated to be more conscious of MSF domain measures than practitioners that have been in practice for a

greater number of years). In 30 (71%) of the studies evidence of construct validity was supported with findings that patients, followed by coworkers, tended to rate physicians more positively than did residents who were more positive than faculty and consultant raters.

Criterion-related validity was ~~adduced~~ indicated in some studies where positive correlations: 1) were found between the MSF instruments/measures (concurrent validity), and 2) between MSF ratings and other assessment instruments/measures (predictive or concurrent validity). As reported in Risucci et al,[33] there was strong concurrent validity for the medical colleague MSF questionnaire where supervisor and peer mean ratings on the same measures of physician performance correlated at $r = 0.92$, $p < 0.001$. The PATH-SPRAT total aggregated score, for example, was found to correlate at $r = 0.48$ ($p < 0.001$) with histopathology residents performance on an Objective Structured Practice Examination.[27]

Interpretation

In a review of the MSF instruments included in this systematic review, there appears to be agreement that the administration of a 360 degree evaluation of physicians in practice from a variety of specialties are feasible from a self-assessment, medical colleague, coworker and patient perspectives. Most studies that provide evidence of reliability, generalizability, and validity (construct and criterion-related) are from the PAR process in Canada and the SPRAT instruments used in the UK where the longitudinal and multi-study nature of the MSF research on physician performance has been in progress for 16 and 8 years, respectively. Although there are a number of American MSF studies (14), each article focused on the use of a new MSF instrument or a modified version of an existing instrument/evaluation guideline (e.g., American Board of Internal Medicine Guide to the Evaluation of Residents in Internal Medicine).

In general, physician performance assessment with MSF instruments employed a minimum of 8 medical colleagues, 8 coworkers and 25 patients to achieve reliability and generalizability coefficients of $\alpha \geq 0.90$ and $Ep^2 \geq 0.80$, respectively. Although a variety of constructs are assessed, there are five key domains identified across the MSF instruments: 1) professionalism, 2), clinical competence, 3) communication, 4) manager, and 5) interpersonal relationships. The majority of the studies provided evidence of the construct validity of the MSF instruments used by conducting a principal component factor analysis or comparing mean rating scores between rater groups (patients tend to rate most positively followed by coworkers, resident peers, faculty and consultant evaluators). Interestingly, in a reversed finding Lockyer et al.[16] found that self assessments were higher than peers in a general practice sample of international medical graduates. While the construct validity of MSF questionnaires may be found within a particular discipline (e.g., family medicine, internal medicine, surgery), ~~M~~many authors acknowledge that measures of various competencies or constructs are a function of the specialization assessed (i.e., the percentage of variance associated with measures of patient management, clinical assessment, communication and/or professional development was found to vary across specialties).[10,15,30,34] For example, Lockyer and Violato[15] found in a principal component factor analysis of the medical colleague MSF questionnaire that the resulting four factor solution accounting for 73.4% of the variance for internal medicine physicians, 70% for psychiatrists and only 67.6% for pediatricians.

Although the present systematic review was rigorous, there are limitations to the present study. First, there is heterogeneity in the MSF instruments used and the number of items employed to measure the various constructs identified. Accordingly, the identification of a single best MSF instrument is difficult and context/specialty specific. Second, the feasibility of

using MSF is based primarily on the reported response rate percentages but does not typically include costs and administration concerns in the assessment of physician performance. Third, variability in the reporting of reliability (i.e., generalizability, intraclass correlation) and validity (i.e., construct and criterion-related) measures while supportive of the MSF process were difficult to combine consistently between studies. Finally, our search was limited to English peer-review journal articles and may not reflect MSF processes in other countries or currently in use but not published.

In summary, MSF where various assessors (self, peers, coworkers, and patients) provide assessment of physicians' performance on various domains (clinical and nonclinical) is reliable, valid and feasible. As indicated above, there exists a substantial body of rigorous and consistent research on the PAR and SPRAT programs that demonstrate the use of MSF will continue to play an important role in the formative and potentially summative assessment of physician performance in practice. Future research should focus on consolidating measures of competence domains between and within physician specialties, while taking into consideration issues related to the establishment of a MSF process at local and national levels.

Funding/Support: None

Other disclosures: None

Ethical approval: Not applicable

References

1. Kohn LT, Corrigan JM, Donaldson MS (eds). Too Err is Human: Building a Safer Health System. Washington, DC: National Academy Press; 1999.
2. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA. 2002;287:226-235.
3. Bandiera G, Sherbino J, Frank JR. The CanMEDS Assessment Tool Handbook. An Introductory Guide to Assessment of the CanMEDS Competencies. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2006.
4. Lockyer J, Clyman S. Multisource feedback (360-degree evaluation). In: Holmboe ES, Hawkins RE, eds. Practical Guide to the Evaluation of Clinical Competence. Philadelphia, PA: Mosby; 2008.
5. Bracken DW, Timmreck CW, Church AH. Introduction: A multisource feedback process model. In: Bracken DW, Timmreck CW, Church AH, eds. The Handbook of Multisource Feedback: The Comprehensive Resource for Designing and Implementing MSF Processes. San Francisco, Jossey-Bass; 2001:3-14.
6. Bracken DW, Church AH~~, Bracken DW~~. Advancing the state of the art of 360-degree feedback: guest editors' comments on the research and practice of multi rater assessment methods. Group Org Manag. 1997;22(2):149-161.
7. 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.
8. Sala F, Dwight S. Predicting executive performance with muti-rater surveys: Whom you ask makes a difference. J Consult Psych Res Prac. 2002;54(3):166-172.

9. Fidler H, Lockyer J, Violato C. Changing physicians practice: The effect of individual feedback. Acad Med. 1999:74:702-714.

10. Violato C, Lockyer J. Self and peer assessment of pediatricians, psychiatrists and medicine specialists: implications for self-directed learning. Adv Health Sc Educ. 2006;11:235-244.

11. Sinclair AM, Gunendran T, Archer J, et al. Re-certification for urologists: is the SHEFFPAT questionnaire valid for assessing clinicians' 'relationships with patients'? Brit J Med Surg Urol. 2009;2:100-104.

12. Noonan CLF, Monagle J, Castanelli D. Development of a multisource feedback tool for consultant anaesthetist performance. Aust Health Rev. 2011;35:141-145.

13. Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: the Physician Achievement Review. CMAJ. 1999;161:52-57.

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

15. Lockyer J, Violato C. An examination of the appropriateness of using a common peer assessment instrument to assess physician skills across specialties. Acad Med. 2004;79:S5-S8.

16. Lockyer J, Blackmore D, Fidler H, et al. A study of multi-source feedback system for international medical graduates holding defined licences. Med Educ. 2006;40:340-347.

17. Lockyer J, Violato C, Fidler H. The assessment of Emergency Physicians by a regulatory authority. Acad Emerg Med. 2006;13:1296-1303.

18. Lockyer J, Violato C, Fidler H. A multi source feedback program for anesthesiology. Can J Anesth. 2006;53(1):33-39.

19. Violato C, Lockyer J, Fidler H. Assessment of pediatricians by a regulatory authority. Pediatrics. 117(3):796-802.

20. Lockyer J, Violato C, Fidler H. What multisource feedback factors influence physicians' self-assessments? A five-year longitudinal study. Acad Med. 2007;82(10):S77-S80.

21. 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.

22. Violato C, Lockyer J, Fidler H. Assessment of psychiatrists in practice through multisource feedback. Can J Psychiatry. 2008;53(8):525-533.

23. Lockyer J, Violato C, Fidler H, et al. The assessment of pathologists/laboratory medicine physicians through a multisource feedback tool. Arch Pathol Lab Med. 2009;133:1301-1308.

24. Overeem K, Wollersheim H, Arah OA, et al. Evaluation of physicians' professional performance: an iterative development and validation study of multisource feedback instruments. BMC Health Serv Res. 2012;12(80):1-11.

25. Lockyer J, Violato C, Wright B, et al. Long-term outcomes for surgeons from 3- and 4-year medical school curricula. Can J Surg. 2012;55:S1-5.

26. Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. BMJ. 2005;330(7502):1251-1253.

27. Davies H, Archer J, Bateman A, et al. Specialty-specific multi-source feedback: assuring validity, information training. Med Educ. 2008;42:1014-1020.

28. Archer J, Norcini J, Southgate L, et al. Mini-PAT (Peer Assessment Tool): a valid component of a national assessment programme in the UK? Adv Health Sc Educ. 2008;13:181-192.

29. Crossley J, McDonnell J, Cooper C, et al. Can a district hospital assess its doctors for re-licensure? Med Educ. 2008;42:359-363.

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

31. Archer J, McAvoy P. Factors that might undermine the validity of patient and multi-source feedback. Med Educ. 2011;45:886-893.

32. DiMatteo MR, DiNicola DD. Sources of assessment of physician performance: a study of comparative reliability and patterns of intercorrelation. Med Care. 1981;19:829-842.

33. Risucci DA, Tortolani AJ, Ward RJ. Ratings of surgical residents by self, supervisors and peers. Surg Gyn Obs. 1989;169:519-526.

34. Ramsey PG, Wenrich MD, Carline JD, et al. Use of peer ratings to evaluate physician performance. JAMA. 1993;269:1655-1660.

35. Wenrich MD, Carline JD, Giles LM, et al. Ratings of the performances of practicing internists by hospital-based registered nurses. Acad Med. 1993;68(9):680-687.

36. Thomas PA, Gebo KA, Hellmann DB. A pilot study of peer review in residency training. J Gen Intern Med. 1999;14:551-554.

37. Lipner RS, Blank LL, Leas BF, et al. The value of patient and peer ratings in recertification. Acad Med. 2002;77(10):S64-66.

38. Davis JD. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. Obstet Gynecol. 2002;99:647-651.

39. 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(5):458-463.

40. 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.

41. Wood L, Wall D, Bullock A, et al. '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(7):e177-184.

42. 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.

43. Allerup P, Aspegren K, Ejlersen E, et al. Use of 360-degree assessment of residents in internal medicine in a Danish setting: a feasibility study. Med Teach. 2007;29:166-170.

44. Pollock RA, Donnelly MB, Plymale MA, 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(2):639-649.

45. Massagli TL, Carline JD. Reliability of a 360-degree evaluation to assess resident competence. Am J Phys Med Rehabil. 2007;86(10):845-852.

46. Lelliott P, Williams R, Mears A, et al. Questionnaires for 360-degree assessment of consultant psychiatrists: development and psychometric properties. Brit J Psych. 2008;193:156-160.

47. Campbell JL, Richards SH, Dickens A, et al. 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.

48. 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.

49. Campbell J, Narayanan A, Burford B, et al. Validation of a multi-source feedback tool for use in general practice. Educ Prim Care. 2010;21:165-179.

50. Chandler N, Henderson G, Park B, et al. 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.

51. Yang YY, Lee FY, Hsu HC, et al. Assessment of first-year post-graduate residents: usefulness of multiple tools. J Chine Med Assoc. 2011;74:531-538.

52. Wall D, Singh D, Whitehouse A, et al. Self-assessment by trainees using self-TAB as part of the team assessment of behavior multisource feedback tool. Med Teach. 2012;34:165-167.

53. Qu B, Zhao YH, Sun BZ. Assessment of resident physicians in professionalism, interpersonal and communication skills: a multisource feedback. Internat J Med Sci. 2012;9(3):228-236.

54. Wright C, Richards SH, Hill JJ, et al. Multisource feedback in evaluating the performance of doctors: the example of the UK General Medical Council patient and colleague questionnaires. Acad Med. 2012;87(12):1668-1678.

Figure

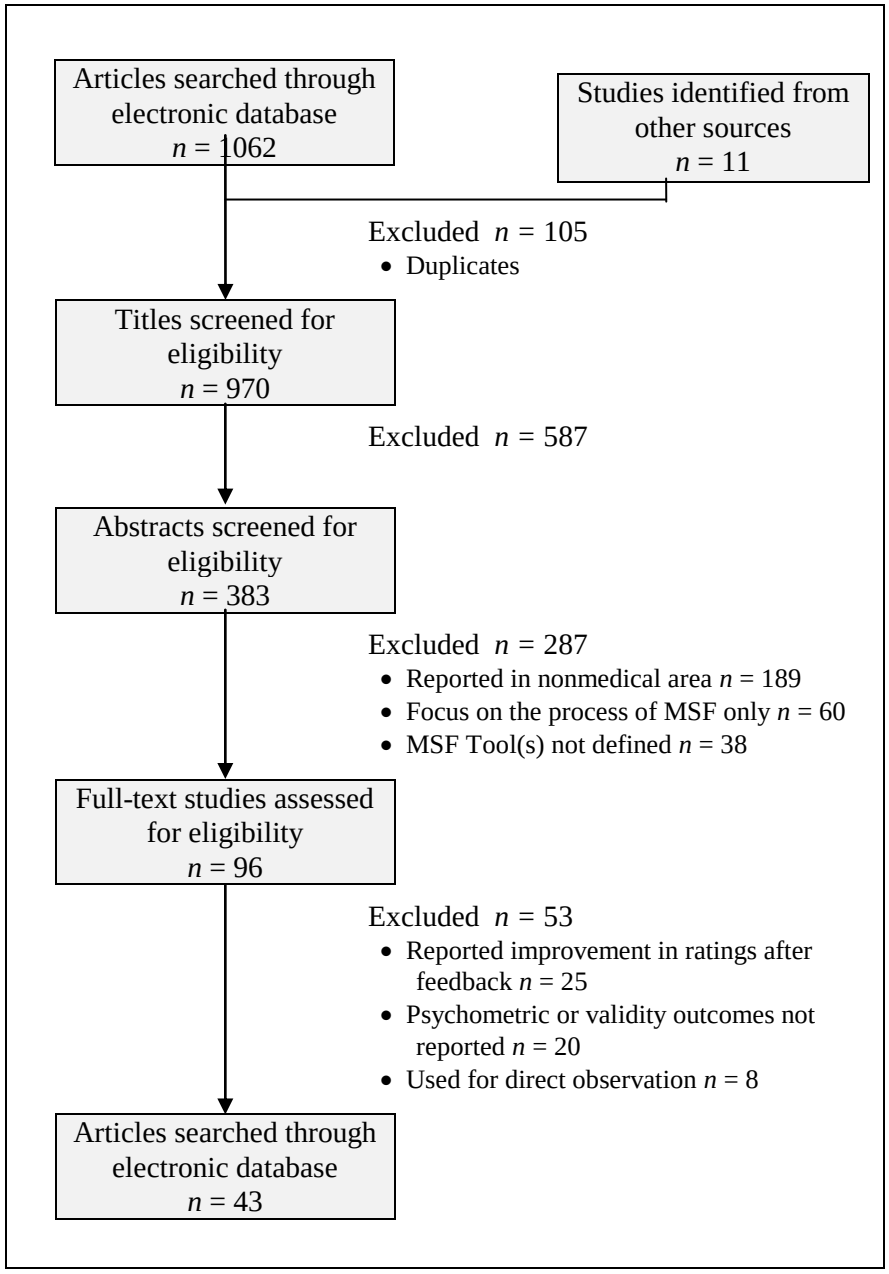

Figure 1: Selection of studies for the systematic review.

Table

Table 1: Description of the 43 studies on physician multisource feedback included in the systematic analysis

Study (Origin)	Specialty (*n*, participants)	MSF Instrument Personnel type (No. items)	Constructs/Factors assessed	General Information on Process, Administration and/or Feasibility
Physician Assessment Review (PAR)				
Violato et al., 1997[7] (Canada)	Family Physicians (n = 17), Internal Medicine and Surgery (*n* = 11) (*n* = 28, physicians)	***PAQ*** Medical Colleague (34 items) ***SAQ*** Self (34 items) ***PS*** Patient (49 items) ***CAQ*** Co-Worker (18 items) ***APCQ*** MC (39 items) ***ACRPQ*** MC (34 items)	Prof,Clin comp, Inter Per Prof, Clin comp, Inter Per Prof, Mngr Prof, Inter Per, Comm Prof, Clin comp, Inter Per Prof, Clin comp, Inter Per	The results of this study provided evidence of reliable and validity for four of the six (PAQ, SAQ, PS and CAQ) multi-source feedback questionnaires used to triangulated measures of professionalism, interpersonal skills, and clinical competencies between peers or medical colleagues (MC), coworkers (CW), and patients (Pt) with a physician's self (Self) assessment. A precursor to the PAR instruments, the authors concluded that the findings provide evidence that patients, peers, coworkers and medical colleagues can provide reliable and multidimensional theoretically meaningful assess of physicians in practice.
Hall et al., 1999[13] (Canada)	Multiple Specialties (*n* = 308, physicians)	***PAR (Generic)*** Self (26 Items) MC (26 Items) CW (17 Items) Pt (44Items) Consultant (23 Items) Referring (21 Items)	 Prof, Clin comp,Inter Per Prof, Clin comp,Inter Per Prof, Comm, Inter Per Prof, Comm, Mager Prof, Clin comp, Inter Per Prof, Clin comp	In this pilot study of registered physicians with the College of Physicians and Surgeons of Alberta (CPSA) the Physician Review Assessment (PAR) program was initially introduced. This PAR project was found to be feasible at an estimated cost of $200 per physician and based on these findings was implemented in the province where all physicians are required to participate every 5 years.
Violato et al., 2003[14] (Canada)	Surgery (*n* = 201, surgeons)	***PAR (Surgery)*** Self (34Items) MC (34 Items) CW (19 Items) Pt (39Items)	 Prof, Clin comp, Comm, Inter Per Prof, Clin comp, Comm, Inter Per Comm, Inter Per Comm, Inter Per,Mngr	As part of the CPSA PAR process, modified versions of the instruments were developed to be used with surgeons. The authors concluded that a multisource feedback system is feasible, reliable and valid in assessing key competencies and, moreover, provide feedback to initiate change in surgeons' practice.
Lockyer & Violato, 2004[15] (Canada)	Psychiatry (*n* = 101),Pediatrics (*n* = 100)and Internal Medicine (*n* = 103) (*n* = 304, physicians)	***PAR (Specialty Generic)*** MC(36 Items)	Prof, Clin comp, Comm	The reliability and generalizability coefficients provide support for the use of the Physician Achievement Review (PAR) program in Alberta across three different specialties. Although consistency is found in the number of factors measured, percentage of variance accounted for any one factor reflects differences in competencies assessed between the specialties.
Lockyer et al., 2006[16] (Canada)	General Practice (*n* = 37, physicians)	***PAR modified (IMG)*** Self (21 Items) MC (22 Items) CW (12 Items) Pt (13 Items)	 Prof, Clin Comp Prof, Clin Comp Prof, Comm, Prof, Comm, Mngr	The findings indicate that the modified PAR tools have acceptable psychometric properties for the assessment of international medical graduates (IMG) whose knowledge and skills have not been formally assessed through national examination processes. The authors suggest that further research comparing IMG with a benchmark

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

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

		MC (9 Items) Pt (3 Items)	Clin comp, Inter Per	measures of physician's performance. The reliabilities of measures from four sources were found to be substation, suggesting the usefulness of these sources for physician evaluation.
Risucci et al., 1989[33] (USA)	Surgery (n = 32, residents)	**360 degree evaluation** Self, MC (same 10 Items)	Prof, Clin comp, Inter Per	The authors concluded that the use of the found that the use of 360 degree evaluation was valid in relation to peer and supervisor ratings of surgical residents. Discrepancies found on the self assessment with those of the peers and supervisors are suggested to reflect the need for residents to address concerns related to professional, interpersonal and clinical skill performance.
Ramsey et al., 1993[34] (USA)	Internal Medicine (*n* = 314, physicians)	**peer physician assessment** MC (11 Items)	Clin comp, Inter Per	The findings suggest that it is feasible to use peer-assessment from professional associates to assess practicing physicians in domains such as clinical skills and interpersonal or humanistic qualities that are difficult to measure using other sources.
Wenrich et al., 1993[35] (USA)	Internal Medicine (*n* = 232, physicians)	**360 degree evaluation** MC (10 Items) CW (13 Items)	Clin comp, Inter Per	The author concluded that nurses' ratings appear to provide a feasible and reliable method of evaluating internists' communication skills and humanistic qualities; however, suggested that this be used in conjunction with ratings provided by peer physicians.
Thomas et al. 1999[36] (USA)	Internal Medicine (*n* = 16, residents)	**peer physician assessment** MC (10 Items)	Clin Comp, Inter Per	The authors concluded that the use of peer review was reliable and feasible when completed by residents, but less so by faculty members. In addition, the authors reported that the residents gave high ratings to the value of the feedback provided by their peers in an end of year survey.
Lipner et al.,2002[37] (USA)	Internal Medicine (*n* = 356, physicians)	**peer/patient assessment** MC (11 Items) Pt (10 Items)	Prof, Clin comp Prof, Clin comp, Comm	The patient and peer assessment module was introduced to evaluate the value of multisource feedback in a recertification professional development program for practicing physicians. Participants reported that the module provided feedback that was beneficial for use in improving their practices.
Davis, 2002[38] (USA)	Obstetrics/ Gynecology (*n* = 16, residents)	**multi-source feedback** Self, MC and CW (same 16 Items)	Clin comp, Inter Per	This evaluation form found support for the use of multi-source feedback when used with other medical colleagues (i.e., faculty members and peers), however, showed discrepancies when compared with the ratings given by self and coworker (nurses) assessments. Suggested that residents may benefit from doing the self-assessment to improve their ability to honestly appraise their clinical and interpersonal skills.
Joshi et al., 2004[39] (USA)	Obstetrics/ Gynecology (*n* = 8, residents)	**360 degree evaluation** Self, MC, CW, Pt and Medical Students (same 10 Items)	Comm, Inter Per	The authors concluded that the 360-degree evaluation questionnaire appear to be reliable in evaluating residents' competencies in interpersonal and communication skills. Further research on the determining the reliability between evaluator categories and throughout the 4 years of the residency program is suggested.
Wood et al.,2004[40] (USA)	Radiology (*n* = 7, residents)	**360 degree evaluation** Self, MC, CW, Pt (same 10 Items)	Prof, Comm	This study shows that the **360 degree evaluation** form was a reliable measurement of radiology residents' professionalism and interpersonal/communication skills. Although the time to complete was feasible, there were organizational and analysis challenges.
Wood et al., 2006[41]	Obstetrics/	***Team Observation*** tool		The **Team Observation** tool has become mandatory in Obs and Gyn

(UK)	Gynecology (*n* = 113, residents)	MC (4 items)	Mngr, Inter Per	training for the past 6 years. The aim was to assist in the facilitation and assessment of the implementation of 'Calman's Structured Training' program.
Brinkman et al., 2007[42] (USA)	Paediatrics (*n* = 36, residents)	**multi-source feedback** Parents (10 Items) CW (14 Item)	 Prof, Comm Prof, Clin Comp, Comm	Adapted from the American Board of IM surveys, the Parent Satisfaction Questionnaire consists of 10 communication and humanistic related questions and the nurse evaluation consists of 14 items related to professionalism, communication and clinical competence. These questionnaires were shown to enhance standard feedback on resident performance with and improved pediatric resident communication skills and professionalism.
Allerup et al., 2007[43] (Denmark)	Internal Medicine (*n* = 42, residents)	**360 degree evaluation** MC and CW (same 15 Items)	Prof, Clin comp, Comm, InterPer	The purpose of this study was to explore the feasibility of 360 degree assessment in an internal medicine residency program in a Danish setting. Although the feasibility and reliability was found to be acceptable, the construct validity of the multisource feedback tool was not determined or verified based on the domains identified in this study.
Pollock et al., 2007[44] (USA)	Plastic Surgery (*n* = 6, residents)	**360 degree evaluation** MC, CW (same 60 Items)	Prof, Clin comp, Comm, Mngr, Inter Per	In this study, plastic surgery residents' performance was rated differently by health care professionals. Nevertheless, the resident found the 360 degree evaluation to be beneficial as they received two independent, formative assessments over a number of years of integrated training.
Massagli & Carline., 2007[45] (USA)	Physical Medicine & Rehabilitation (*n* =56, residents)	**360 degree evaluation** CW, Rehab Staff, Medical Students (same12 Items)	Prof, Clin comp, Comm, Inter Per	The authors concluded that the use of a Web-based 360 degree evaluation tool is a feasible way to obtain reliable ratings from rehabilitation staff about resident behaviors. This instrument showed adequate reliability and validity in assessing residents in the physical and rehabilitation program.
Lelliott et al., 2008[46] (UK)	Psychiatry (*n* = 347, physicians)	***ACP 360*** Self, MC (same 57 Items) Pt (17 Items)	 Clin comp, Comm, InterPer Clin comp, Comm, InterPer	The 360 degree Assessment of Consultant Psychiatrists (ACP 360) service was implemented by the Royal College of Psychiatrists in the UK since 2005 to provide feedback for individual consultants for performance improvement. The author reported that the use of the ACP 360 is considered to be a reliable and feasible service in assessing psychiatrists who work in large multi professional teams.
Campbell et al., 2008[47] (UK)	Multiple Specialties (*n* = 291, physicians)	***GMC Survey*** MC (17 Items) Pt (9 Items)	 Prof, Clin comp, Comm, InterPer Prof, Clin comp, Comm, InterPer	The authors concluded that the General Medical Council (GMC) patient and colleague questionnaires were reliable and provided a basis for the assessment of professionalism among UK doctors. It is suggested that further research is need to explore the validity of the questionnaires as reliable indicators of acceptable professional performance, especially for revalidation of physicians' registration.
Meng et al.,2009[48] (USA)	Anesthesia (*n* = 15, residents)	**360 degree evaluation** CW (13 Items)	Prof, Comm, Inter Per	This **360 evaluation form** may be useful for post anesthetic care unit rotations. It appears to correlate well with traditional global ratings, although coefficients were not provided, was feasible and provided formative feedback to the residents.
Campbell et al., 2010[49] (UK)	Family Physicians	***CFET/DISQ* (*CFEP360)*** MC (CFET: 18 Items)	 Prof, Clin comp, Comm,	The authors concluded that physician performance, as assessed using the Colleague Feedback Evaluation Tool (CFET) and Doctor's

	(*n* = 179, physicians)	Pt (DISQ: 12 Items)	Mngr Inter Per Prof, Clin comp, Comm, Inter Per	Interpersonal Skills Questionnaire (DISQ) or CFEP360 system, should be able to identify physicians who are underperforming, while still being of use to for the majority of physicians for revalidation purposes.
Chandler et al., 2010[50] (USA)	Paediatrics (*n* = 66, residents)	**360 degree evaluation** Self, MC, CW and Pt (same10 Items)	 Comm, Inter Per Comm, Inter Per Comm, Inter Per Comm, Inter Per	Overall, the 360 degree evaluation ratings for the paediatric residents were high and provided guidance to them their interpersonal and communication skills. The authors indicated that the results provide evidence for the use of multiple evaluator feedback in a residency program that can feasibly be replicated annually.
Yang et al., 2011[51] (Taiwan)	Multiple Specialties (*n* = 245, residents)	**360 degree evaluation** MC, CW (same 12 Items)	Prof, Clin comp, Comm	The authors conclude that the use of 360 degree evaluation as formative method in assessment helped the residents to understand how other members of their team view their knowledge and attitudes. Subsequently, this helped the residents to develop an action plan and improve their behavior.
Wall et al., 2012[52] (UK)	Multiple Specialties (*n* = 834, residents)	***TAB*** Self: (4 Items) MC, CW (same 4 Items)	 Prof, Comm Prof, Comm	The authors concluded that the use of the 4 item TAB assessment tool can help some physicians to identify concerns with professional or communication performance. The use of Self-TAB in comparison with the TAB, however, demonstrates physicians limited ability to self assess.
Qu et al., 2012[53] (China)	Multiple Specialties (*n* = 258, residents)	***EOS Group Tools*** Self (21 Items) MC (21 items) Attending (21 items) CW (26 items) Office staff (15 items) Pt (25 items)	 Prof, Comm Prof, Comm Prof, Comm Prof, Comm Prof, Comm, Prof, Clin comp, Mngr Inter Per	The author concluded that the 360 degree evaluation tools developed by the Education Outcomes Service (EOS) group from the Arizona Medical Education Consortium are reliable and valid in assessing resident professionalism and interpersonal communication skills in China. It was suggested that further studies are required to determine how the residents used their data to produce changes in their professional and interpersonal communication skills.
Wright et al., 2012[54] (UK)	Multiple Specialties (*n* = 1,065, physicians)	***GMC Survey*** MC (18 Items) Pt (9 Items)	 Prof, Clin comp, Comm, InterPer Prof, Clin comp, Comm, InterPer	The General Medical Council (GMC) has introduced a five-year cycle whereby all licensed doctors must be 'revalidation', in part, through the use of feedback on the Colleague and Patient Questionnaires. Although found to be feasible for formative purposes, concerns about the utility of the Pt and MC feedback as a stand-alone assessment of physician practice are expressed.

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

Table

Table 2: Reliability and validity characteristics of the 43 studies on physician multisource feedback

Study (Origin)	Mean no. raters (Response %)	Reliability (α), Generalizability (Ep^2) and/or Intra-Class Correlation (ICC)	Validity
Physician Assessment Review (PAR)			
Violato et al., 1997[7] (Canada)	Self (SAQ): 1 (100%) MC (PAQ): 7.8 (76.8%) Pt (PS): 26.2 (87.4%) CW (CAQ): 8.5 (85.4%) MC (APCQ): 7.4 (73.5%) MC (ACRPQ): 8.6 (85.5%)	Self (SAQ): α = 0.95 MC (PAQ): α = 0.95, for 8 raters Ep^2 = 0.77 Pt (PS): α = 0.95, for 25 raters Ep^2= 0.80 MC (CAQ): α = 0.95 MC (APCQ): α = 0.92 MC (ACRPQ): α = 0.89	***Construct:*** Principal component factor analysis was conducted for the *PAQ* (four factor solution), *PS* (seven factor solution), and *CAQ* (three factor solution) questionnaires accounting for 73.1%, 70.0%, and 72.8% of the variance, respectively. The mean rating scores were shown to be higher for medical colleagues (MC) or peers ($p < 0.05$), co-workers and patients when compared with physicians' self assessments.
Hall et al., 1999[13] (Canada)	Self: 1 (95.8%) MC: Consultant and Referring: 6.4 (79.7%) CW: 5.2 (86.7%) Pt: 22.1 (88.6%)	Self: α = 0.95 MC: α = 0.95 Consultant: α = 0.93 Referring: α = 0.91 CW: α = 0.95 Pt: α = 0.95	***Construct:*** The mean ratings showed that self assessments were consistently lower than reported by peers (MC, Consultants and Referring), coworkers (CW) and patients (Pt).
Violato et al., 2003[14] (Canada)	Self: 1 (96.5%) MC: 7.3 (89.6%) CW: 7.2 (88.2%) Pt: 22.6 (83.2%)	Self: α = 0.97 MC: α = 0.98 CW: α = 0.95 Pt: α = 0.93	***Construct:*** A principal component factor analysis showed a five factor solution for peers (MC) accounting for 69.0% of the variance, three factor for coworker (CW) accounting for 70.9%, five factors for patients (Pt) accounting for 73.5%, and four factors for self accounting for 65.1%.The mean ratings showed that self assessments were consistently lower than reported by peers, coworkers and patients.
Lockyer & Violato, 2004[15] (Canada)	MC (Psych): 7.6 (94.6%) MC (Peds): 7.6 (95.5%) MC (IM): 7.6 (94.4%)	MC (Psych): α= 0.98, for 7.6 raters Ep^2 = 0.81 MC (Peds): α= 0.98, for 7.6 raters Ep^2 = 0.88 MC (IM): α= 0.99, for 7.6 raters Ep^2 = 0.82	***Construct:*** Principal component factor analysis was conducted to derive a four factors solution for MC (psychiatrists) accounting for 70% of the variance, four factors for MC (pediatricians) accounting for 67.6%, and four factors for MC (internal medicine) accountings for 73.4%.
Lockyer et al., 2006[16] (Canada)	Self: 1 (91.8%) MC: 5.7 (71.8%) CW: 6.9 (86.1%) Pt: 17.5 (69.9%)	Self: α = 0.83 MC: α = 0.98, for 5.7 raters Ep^2= 0.67 CW: α = 0.91, for 6.9 raters Ep^2= 0.59 Pt: α = 0.95, for 17.5 raters Ep^2= 0.71	***Construct:*** Principal component factor analysis showed a two factor solution for medical colleague (MC) accounting for 71.5 % of the variance, two factors for coworker (CW) accounting for 59.5%, and two factors for patient (Pt) accounting for 74.9% of the variance. Unlike other findings, mean ratings for self assessment were higher than reported by medical colleague (MC) and near identical to mean ratings that were reported by their patients.
Lockyer et al., 2006[17] (Canada)	Self: 1 (100%) MC: 7.7 (95.5%) CW: 7.6 (94.9%) Pt: 21.6 (86.3%)	Self: α = 0.97 MC: α = 0.97, for 7.7 raters Ep^2= 0.84 CW:α = 0.94, for 7.6 raters Ep^2= 0.85 Pt: α = 0.97, for 21.6 raters Ep^2= 0.68	***Construct:*** An exploratory factor analysis showed a four factor solution for the peer (MC), two for the coworker (CW) and two for the patient (PT) instruments that accounted for 71.9%, 62.5%, and 80.0% of the variance, respectively. The mean ratings showed that self assessments were consistently lower than reported by peers, coworkers and patients.
Lockyer et al., 2006[18] (Canada)	Self: 1 (100%) MC: 7.8 (94.6%) CW : 7.8 (95.1%) Pt: 17.7 (56.2%)	Self:α = 0.97 MC:α = 0.97, for 7.8 raters Ep^2= 0.69 CW:α = 0.95, for 7.8 raters Ep^2= 0.56 Pt:α = 0.93, for17.7 raters Ep^2= 0.65	***Construct:*** An exploratory factor analysis showed a three factor solution for the peer (MC), two for the coworker (CW) and two for the patient (PT) instruments that accounted for 74.5%, 67.5%, and 77.6% of the variance, respectively. The mean ratings showed that self assessments

			were consistently lower than reported by peers, coworkers and patients.
Violato et al., 2006[19] (Canada)	Self: 1 (100%) MC: 7.6 (95.5%) CW: 7.6 (94.8%) Pt: 23.4 (93.6%)	Self: α = 0.98 MC: α = 0.98, for 7.6 raters Ep^2= 0.78 CW: α = 0.95, for 7.6 raters Ep^2= 0.87 Pt: α = 0.99, for 23.4 raters Ep^2= 0.85	***Construct:*** Principal component factor analysis showed a four factor solution for peers (MC) accounting for 67.6% of the variance, three factor for coworkers (CW) accounting for 63.8%, and four factor for patients (Pt) accounting for 77.6 %. Self-instrument is identical to co-worker instrument. The mean ratings showed that self assessments were consistently lower than reported by peers, coworkers and patients.
Lockyer et al., 2007[20] (Canada)	Self: 1 (100%)	Self, α = 0.96	***Construct:*** Principal component factor analysis was conducted to derive a three factor solution accounting for 71% of the variance. ***Predictive:*** The sum of the mean scores calculated for self-ratings between Time 1 and 2 (5 year interval) showed that physicians rated themselves higher in the second iteration, $p<0.05$.
Violato et al., 2008[21] (Canada)	MC: 7.19 (93%) CW: 7.34 (94%) Pt: 24.09 (97%)	MC: α = 0.96, for 8 raters Ep^2= 0.78 CW: α = 0.96, for 8 raters Ep^2 = 0.83 Pt: α = 0.98, for 23 raters Ep^2 = 0.80	***Construct:*** Confirmatory factor analyses were conduct on the MC (CFI = 0.91), CW (CFI = 0.87) and Pt (CFI = 0.83) instruments. ***Predictive:*** From Time 1 to Time 2 (5 year interval) on both the MC and CW total, there was found to be a significant improvement, $p < 0.001$. From Time 1 to Time 2 (5 year interval) on the Pt total, however, not significant difference was shown.
Violato et al., 2008[22] (Canada)	Self: 1 (100%) MC: 7.6 (94.6%) CW: 7.4 (92.1%) Pt: 24.3 (97.3%)	Self: α = 0.96 MC: α = 0.98, for 7.6 raters Ep^2= 0.81 CW: α = 0.96,for 7.4 raters Ep^2 = 0.82 Pt: α = 0.98, for 24.3 raters Ep^2= 0.78	***Construct:*** Principal component factor analysis showed a four factor solutions for peers (MC) accounting for 66.8%, three factor solution for coworker (CW) accounting for 68.8%, and five factor solution for patients (Pt) accounting for 73.7% of the variance. The mean ratings showed that self assessments were consistently lower than reported by peers, coworkers and patients.
Lockyer et al., 2009[23] (Canada)	Self: 1 (100%) MC: 7.6 (91.3%) CW: 7.6 (91.8%) Referring: 7.4 (90.3%)	MC: α = 0.98, for 7.6 raters Ep^2 = 0.78 CW: α = 0.95, for 7.6 raters Ep^2 = 0.80 Referring: α = 0.98, for 7.4 raters Ep^2 =0.81	***Construct:*** Principal component factor analysis showed a five factor solution for peers (MC) accounting for 68.8% of the variance, three factor for referring physicians (Referring)accounting for 66.9%, and two factors for coworkers (CW) accounting for 59.9%.The mean ratings showed that self assessments were consistently lower than reported by peers, coworkers and referring physicians.
Overeem et al., 2012[24] (Netherland)	MC: 6.5 (81.3%) CW: 6.7 (83.8) Pt: 15 (51.8%)	MC: α = 0.95 CW: α = 0.95 Pt: α = 0.94	***Construct:*** Principal component factor analysis showed a six factor solution for peers (MC) accounting for 67 % of the variance, three factor solution for coworker (CW) accounting for 70%, and a single factor solution for patient (Pt) accounting for 60%. Physicians with more work experience were rated lower by MC and CW; $p < 0.05$. MC ratings showed a medium correlation with CW ratings ($r = 0.35$, $p < 0.01$), a small correlation with Pt ratings ($r = 0.21$, $p < 0.01$), and CW ratings showed a small correlation with Pt rating ($r = 0.22$, $p < 0.01$).
Lockyer et al., 2012[25] (Canada)	Self: 1 MC: 7.67 CW: 7.60 Pt: 24	Self: α = 0.97 MC: α = 0.98, for 7.27 raters Ep^2 = 0.61 CW: α = 0.95, for 7.20 raters Ep^2 = 0.70 Pt: α = 0.98, raters 22.63 raters Ep^2 = 0.81	***Construct validity***: Principal component factor analysis showed a four factors solution for medical colleague (MC) accounting for 75% of the variance, two factor solution for coworker (CW) accounting for 72%, four factor solution for patient (Pt) accounting for 77% of the variance. The mean ratings showed that self assessments were consistently lower than reported by peers, coworkers and patients.

Sheffield Peer Review Assessment Tool (SPRAT)			
Archer et al., (2005)[26] (UK)	Combined MC and CW: 8.2 (82.0%)	SEM for 4 raters ± 0.50 (95% CI)	***Construct:*** The mean ratings for specialist registrars were significantly higher than for senior house officers, $p < 0.001$. In a hierarchical regression, the rating of the residents by the peers (MC) accounted for 7.6% of the variation in the mean ratings.
Davies et al., 2008[27] (UK)	Self: 1 (100%) Combined MC and CW: 9.2 (92%)	SEM for 8 raters ± 0.37 (95% CI)	***Construct:*** Principal component factor analysis was conducted to derive a two factor solution accounting for 78% of the variance. Pearson's correlation for self-versus assessor ratings was shown to be negative ($r = –0.13$, $p > 0.05$). Consultants marked trainees lower than other occupational groups, $p < 0.001$. ***Predictive:*** A medium correlation was found between the trainees' PATH-SPRAT aggregated and Objective Structure Practical Examination scores; $r = 0.48$, $p < 0.001$.
Archer et al., 2008[28] (UK)	Combined MC and CW: 6.7 (67%)	Combined MC and CW: α = 0.98 SEM for 8 raters ± 0.45 (95% CI)	***Construct:*** Principal component factor analysis was conducted to derive a two factor solution accounting for 81% of the variance. Consultants scored trainees significant lower than other assessors; $p< 0.001$.The mean scores showed that year one (F1) trainees were rated significantly lower than year two (F2) trainees; $p< 0.001$.
Crossley et al., 2008[29] (UK)	Combined MC and CW: 14 (100%) Pt: 9.7 (27.4%)	Combined MC and CW: SEM for 9 raters ± 0.37 (95% CI) Pt: SEM for 15 raters ± 0.29 (95% CI)	***Construct:*** Patients (Pt) rated female physicians significantly higher than male physicians for their relational skills than male doctors, $p < 0.05$. The least stringent professional group (foundation doctors/pre-registration house officers) rated the residents higher on average than the most stringent professional group (allied health professionals), $p < 0.05$.
Archer et al., 2010[30] (UK)	Combined MC and CW: 8.26 (83%)	SEM for 8 raters ± 0.40 (95% CI)	***Construct:*** Principal component factor analysis was conducted to derive a two factor solution accounting for 76.5% of the variance. Consultants marked trainees significantly lower than all groups of raters ($p< 0.05$), whereas senior house officers and foundation doctors scored trainees significantly higher than consultants ($p < 0.05$). ***Predictive:*** The mean scores for Year 4 were significantly higher than for Year 2, $p < 0.01$.
Archer & McAvoy, 2011[31] (UK)	Combined MC and CW: 12.0 Pt: 22.8	NR	***Construct:*** The mean ratings showed that the assessors identified by the physicians were rated significantly higher than those that were identified by the referring body; $p < 0.001$. Nevertheless, patients scored the physicians higher than all assessors; $p < 0.001$. The mean ratings showed that these physicians in difficulty when compared to a normative reference group scored significantly lower; $p < 0.001$.
Multisource feedback or 360 degree evaluation			
DiMatteo et al., 1981[32] (USA)	Self: 1 Attending : 15 MC: 15 Pt: 15	Self: α = 0.56 (Clin comp) and 0.78 (Inter Per) Attending: α = 0.90 (Clin comp and Inter Per) MC: α = 0.67 (Clin comp) and 0.92 (Inter Per) Pt: α = 0.79 (Inter Per)	***Construct:*** Principal component factor analysis for Internal Medicine (IM) I group showed a two factor solution for Attending accounting for 68.7 % of the variance, a two factor solution for peers (MC) accounting for 87.5%, and a two factor solution for self (Self) accounting for 57.2% of the variance. Results from similar forms used with the IM II, surgery and family medicine residents found similar factor solution results. ***Concurrent:*** Correlations on the two factors between self (Self) with

			Attending (r = 0.08 to 0.31), peers (MC) (r = 0.06 to 0.38) and patients (Pt) (r = -0.07 to 0.44) are negative to moderate.
Risucci et al., 1989[33] (USA)	Self: 1 (84.4%) MC (peers): 27 MC: (supervisors): 4	NR	***Construct:*** Principal component factor analysis showed a three factor solution for self (Self) accounting for 68.7 % of the variance, two factor solution for supervisors (MC) accounting for 80.3%, and a single factor solution for peers (MC) accounting for 85.3 % of the variance. The mean ratings showed that self assessments were consistently higher than reported by peers and supervisors, and supervisors mean ratings were higher than peers. ***Concurrent:*** Supervisor and peer ratings strongly correlated (r = 0.92, p < 0.001). ***Predictive***: The peer and supervisor (MC) 360 degree evaluation showed large correlations with the American Board of Surgery In-Training Examination, r = 0.52 and r =0.55 (p < 0.01), respectively.
Ramsey et al.,1993[34] (USA)	MC: 8.7 (51.6%)	MC: For 11 raters Ep^2 = 0.70	***Construct:*** Principal component factor analysis showed a two factor solution accounting for 88.7 % of the variance.
Wenrich et al., 1993[35] (USA)	CW: 8.01 (68.2%)	CW: Based on a range of 6.6 to 13.9 raters (depending on item) Ep^2= 0.70	***Construct:*** Principal component factor analysis showed a two factor solution for the combined nurse (CW) and peer (MC) evaluation forms based on the 10 common items. The mean ratings showed that nurses scored the physicians lower on humanistic qualities (p < 0.01) but higher on medical knowledge (p < 0.001) than the peer (MC) raters.
Thomas et al. 1999[36] (USA)	MC: 11.1 (49.2%)	MC: α = 0.94	***Construct validity***: Principal component factor analysis showed a two factor solution for medical colleague (MC) accounting for between 84.4% (senior residents) to 88.2% (junior residents) of the variance, The mean ratings showed that faculty members scored the junior residents consistently lower than senior residents or peers.
Lipner et al., 2002[37] (USA)	MC: 10 (100%) Pt: 25 (100%)	MC: For 10 raters Ep^2 = 0.61 (95% CI $\pm$ 0.41) Pt: For 25 raters Ep^2=0.67 (95% CI $\pm$ 0.14)	***Construct:*** The mean rating of patients (Pt) was found to be higher than the ratings received from peer (MC) assessments.
Davis, 2002[38] (USA)	Self: 1 (93.7%) MC (Peers): 16 (100%) MC (Faculty): 16 (92.9%) CW (Nurses): 16 (83.3%)	MC (Faculty): ICC = 0.66 to 0.84 MC (Peers): ICC = 0.78 to 0.90 CW (Nurses): ICC = 0.23 to 0.45	***Concurrent:*** Pearson correlation coefficients between the MC faculty members and MC peers showed moderate to large correlations on both factors (r = 0.72 and 0.80, p < 0.01) and on the overall clinical assessment item (r = 0.86, p < 0.001). In comparison with MC faculty members ratings, however, the correlations with the Self and CW (Nurses) were non-significant and ranged between r = -0.12 to 0.36 and r = 0.04 to 0.24, respectively.
Joshi et al., 2004[39] (USA)	Self: 1 (100%) MC: 16 (100%) CW: 25 (100%) Pt: 10 (100%) Medical Students: 12 (100%)	MC: For 16 raters ICC = 0.72 CW: For 25 raters ICC = 0.86 Pt: For 10 raters ICC = 0.54 Medical Students: For 12 raters ICC = 0.82	Authors recognize that validity of the question was achieved by 'expert opinion' only. ***Concurrent:*** Faculty (MC) ratings showed a large correlation with nurse coworkers (CW) ratings (r = 0.55, p = 0.16), a small correlation with Pt ratings (r = 0.21, p = 0.61), and CW ratings showed a medium correlation with Pt rating (r = 0.43, p = 0.29).
Wood et al., 2004[40] (USA)	Combined MC, CW and Pt: 8.14 (57%)	MC: α = 0.85 CW: α = 0.87	***Construct:*** In an analysis of variance, it was found that the Pt mean score ratings of the trainees were significantly higher when compared with MC

		Pt: α = 0.86	and CW, $p < 0.001$. ***Concurrent:*** The correlation coefficients were calculated between a 5 item global ratings form (used as a gold standard) and the 1) Pt 360 degree evaluation ($r = 0.70$, $p = 0.08$), 2) MC ($r = 0.46$, $p = 0.30$), and 3) CW ($r = 0.62$, $p = 0.14$) were medium to large, however, not significant.
Wood et al., 2006[41] (UK)	MC: 12.52	MC: For 8 raters ICC = 0.80	***Construct:*** Principal component factor analysis was conducted on the Team Observation tool to derive a one factor solution accounting for 76% of the variance. ***Predictive:*** Spearman's correlation coefficients were calculated between Time 1 to Time 2 (6-7 month interval); $r = 0.77$, $p < 0.001$.
Brinkman et al., 2007[42] (USA)	Parents: 19.3 CW: 15.8	Parents: α= 0.95 CW: α = 0.96	***Construct:*** Although statistical results between groups at Time 1 and 2 were not reported at both Time 1 and 2, the multisource feedback group achieved higher ratings from parents and nurses on average than the control group at Time 2.
Allerup et al., 2007[43] (Denmark)	Self: 1 (97.6%) MC: 4.7 (94.0%) CW: 2.8 (55.0%)	Combined MC and CW, α = 0.46 to 0.89	***Construct:*** The mean correlation ratings between self and coworkers (CW) indicated that nurses on average rate the residents (Self) higher. The mean correlation ratings between self and peers (MC), however, show that other physicians (MC) on average rate the residents (Self) lower. Note that the construct validity of the measures used was not provided and, therefore, the domains identified were not confirmed.
Pollock et al., 2007[44] (USA)	MC: 12 CW: 28	NR	***Construct:*** The mean ratings by peers (MC) was significantly lower than the nurse coworkers (CW) across all competencies areas identified.
Massagli & Carline., 2007[45] (USA)	CW: 3.7 Rehab Staff: 9.9 Medical Students: 3.0	Combined CW, Rehab Staff and Medical Students: α = 0.89 CW: For 5 raters $Ep^2 = 0.80$ Rehab Staff: For 4 raters $Ep^2 = 0.80$ Medical Students: For 23 raters $Ep^2 = 0.80$	***Construct:*** Principal component factor analysis showed a single factor solution accounting for 84.0% of the variance. The mean scores for post graduate year 4 residents (Self) were shown to be higher than for year 2 and 3 residents.
Lelliott et al., 2008[46] (UK)	Self: 1 (100%) MC: 12.7 (85.0%) Pt: 19.2 (63.9%)	Self: α = 0.98 MC: α = 0.98, for 13 raters $Ep^2 \geq 0.75$, ICC = 0.75 Pt: α = 0.97, for 25 raters $Ep^2 \geq 0.75$, ICC = 0.70	***Construct:*** Principal component factor analysis showed a seven factor solution for peers (MC) accounting for 70.2 % of the variance and a single factor solution for the patient (Pt) tool accounting for 66.8 % of the variance. The mean ratings showed that self assessments were consistently lower than reported by peers and patients; $p < 0.001$.
Campbell et al., 2008[47] (UK)	MC:13.8 (69.1%) Pt: 36.2 (92.1%)	MC: α = 0.95, for 12 raters $Ep^2 = 0.76$ Pt: α = 0.96, for 36 raters $Ep^2 = 0.75$	***Construct:*** Principal component factor analysis showed a three factor solution for peers (MC) for the 17 performance-based items accounting for 61.0 % of the variance, and two factor for patients (Pt) for the 9 performance-based items accounting for 76.8 % of the variance. On mean ratings patients (Pt) scored the physicians higher than peers (MC), and younger physicians were rated higher than older physicians by both their peers and patients; $p < 0.05$.
Meng et al., 2009[48] (USA)	CW: 28.6 (88%)	CW: (Nurses) ICC = 0.87 CW: (Secretaries) ICC = 0.79 CW: (Nurse Aids) ICC = 0.83 CW: (Technicians) ICC = 0.86	***Construct:*** The average mean ratings across all items from post anesthetic care unit nurses were higher than secretarial staff. ***Concurrent:*** Although the authors indicated that residents who ranked high by global ratings were also ranked high by the 4 categories of 360

			degree evaluation ratings, no correlations were provided.
Campbell et al., 2010[49] (UK)	MC: 13.9 Pt: 47.3	MC: α = 0.84, for 14 raters Ep^2= 0.82 Pt: α = 0.95, for 25 raters Ep^2= 0.81	***Construct:*** Principal component factor analysis showed a two factor solution for medical colleague (MC) CFET form accounting for 66.0% of the variance, and a single factor solution for the patient (Pt) DISQ form accounting for 94.0% of the variance. The mean ratings for patients were slightly higher on average than reported by peers (MC).
Chandler et al., 2010[50] (USA)	Self: 1 (100%) MC: 2.6 CW: 7.4 Pt: 1.2	NR	***Construct:*** The mean ratings showed that self assessments were consistently lower than reported by peers (MC) and nurse coworkers (CW); $p < 0.001$. Self mean ratings were, however, not significantly different from the patients (Pt).
Yang et al., 2011[51] (Taiwan)	Combined MC and CW: 4.3 (85.3%)	Combined MC and CW: α = 0.86	***Predictive***: The 360 degree evaluation show a medium correlation with the small scale OSCE ($r = 0.37$, $p < 0.05$). Moreover, adding the DOPS score to small-scale OSCE scores increased it to a large correlation at $r = 0.72$ ($p < 0.05$), and adding the IM in-training examination increased it to $r = 0.85$, $p < 0.05$.
Wall et al., 2012[52] (UK)	Self: 1 (100%) Combined MC and CW: 11.6	NR	***Concurrent:*** The self ratings compared with combined peer (MC) and coworker (CW) ratings showed a small correlation on minor concerns ($r = 0.20$, $p< 0.001$) and major concerns ($r = 0.26$, $p< 0.001$).
Qu et al., 2012[53] (China)	Self: 1 (100%) MC: 2 (100%) Attending 1(100%) CW: 3 (100%) Office staff: 2 (100%) Pt: 7 (100%)	Self: α = 0.92 MC: α = 0.93 Attending: α = 0.91 CW: α = 0.92 Office staff: α = 0.90 Pt: α = 0.93	***Construct:*** Principal component factor analysis showed a two factor solution for self (Self) accounting for 71.0% of the variance, a two factor solution for the attending (Attending) accounting for 70.9 % of the variance, a two factor solution for peers (MC) accounting for 70.7%, a two factor solution for nurses (CW) accounting for 75.5%, a two factor solution for Office staff accounting for 74.6%, and a four factor solution for patients (Pt) accounting for 72.7% of the variance. The mean ratings showed that self assessments were consistently lower than reported by MC and Pt, but were higher when compared with the CW (nurses).
Wright et al., 2012[54] (UK)	MC:13.8 (69.1%) Pt: 36.2 (92.1%)	MC: α = 0.94, ICC = 0.85, for $\geq$ 15 raters $Ep^2 \geq 0.70$ Pt: α = 0.87, ICC = 0.83, for $\geq$ 34 raters $Ep^2 \geq 0.70$	***Construct:*** Principal component factor analysis showed a three factor solution for peers (MC) for the 18 performance-based items accounting for 58% of the variance, and two factor for patients (Pt) for the 9 performance-based items accounting for 79% of the variance. Convergent validity was shown with correlations between the Pt and Doctor's Interpersonal Skills Questionnaire (DISQ), rho = 0.63, p <0.001; and between the MC and Colleague Feedback Evaluation Tool (CFET), rho = 0.81, $p < 0.01$.

MC = Medical colleague, CW = Co-Worker, Pt =Patient, SEM = Standard Error of Measurement, NR = Not Reported, CFI = Confirmatory Fit Index, ICC = Intraclass correlation coefficient, Ep^2 = Generalizability Coefficient, CFET = Colleague Feedback Evaluation Tool, DISQ = Doctor's Interpersonal Skills Questionnaire