2018-10

Validation of Multisource Feedback in Assessing Medical Performance: A Systematic Review

Stevens, SAG

10.1097/ceh.0000000000000219
Journal of Continuing Education in the Health Professions
Lippincott, Williams & Wilkins

Validation of Multisource Feedback in Assessing Medical Performance: A Systematic Review

Stevens, S, Read, J. Baines, R. Chatterjee, A. & Archer, J.

Abstract

Introduction

Over the past ten years, a number of systematic reviews have evaluated the validity of multisource feedback (MSF) to assess and quality assure medical practice. The purpose of this study is to synthesise the results from existing reviews to provide a holistic overview of the validity evidence.

Methods

This review identified eight systematic reviews evaluating the validity of MSF published between January 2006 and October 2016. Using a standardised data extraction form, two independent reviewers extracted study characteristics. A framework of validation developed by the American Psychological Association (APA) was used to appraise the validity evidence within each systematic review.

Results

In terms of validity evidence, each of the eight reviews demonstrated evidence across at least one domain of the APA validity framework. Evidence of assessment validity within the domains of 'internal structure' and 'relationship to other variables' has been well established. However, the domains of content validity (i.e. ensuring MSF tools measure what they are intended to measure); consequential validity (i.e. evidence of the intended or unintended consequences MSF assessments may have on participants or wider society) and response process validity (i.e. the process of standardisation and quality control in the delivery and completion of assessments) remain limited.

Discussion

Evidence for the validity of MSF has, across a number of domains, been well-established. However, the size and quality of the existing evidence remains variable. In order to determine the extent to which MSF is considered a valid instrument to assess medical performance, future research is required to determine: 1) how best to design and deliver MSF assessments that address the identified limitations of existing tools, and 2) how to ensure involvement within MSF supports positive changes in practice. Such research is integral if MSF is to continue to inform medical performance and subsequent improvements in the quality and safety of patient care.

Key words: multisource feedback, MSF, systematic review, medical education, validity, physician

Introduction

Multisource feedback (MSF) is a method of workplace based assessment (WBA) used to facilitate the collection of feedback from colleagues, and at times patients, in order to inform on-going performance. This method of assessment has a long history of use outside of healthcare [1], however more recently it has been adopted internationally within healthcare environments, particularly within medicine, as an instrument to assess and quality-assure clinical practice [2,3].

Within medicine, MSF increasingly forms a key component of regulatory processes worldwide [4-7]. The use of MSF in regulatory settings requires an assurance for stakeholders (including patients, physicians and the regulators), that instruments have substantial validity evidence. This issue is particularly pertinent when the outcome of such assessments could have career affecting consequences for physicians (e.g. remediation or license withdrawal), and potential implications for care quality and patient safety.

The wide adoption of MSF across a multitude of medical disciplines internationally predicates an inherent need to critically evaluate evidence to support or refute its validity. Whilst a number of systematic reviews demonstrate MSF to be a valid, reliable and feasible method of performance assessment [8-11], critics regularly cite concerns around important issues that may undermine assessment validity [12-14]. Past reviews have largely focussed on specific areas of assessment validity (e.g. statistical and psychometric properties) or have explored the validity of feedback instruments in regards to particular medical specialities [8-10,15-19]. This review critically evaluated existing reviews on the validity evidence of MSF in assessing medical performance among qualified physicians in the healthcare settings, through answering the following question:

- To what extent is MSF a valid instrument to assess medical performance?

For the purpose of this review, MSF was inclusive of all colleague feedback instruments that include ratings from peers, colleagues and co-workers [20]. While patient feedback can be encapsulated within MSF assessments[21], it was not the focus of this review. In order to maintain research integrity, data included in studies that discuss patient feedback are not extracted, synthesised or reported.

Methods

We conducted a systematic review of reviews and narrative synthesis, adopting a configurative approach to the review design[22]. One author carried out a systematic search of MEDLINE, PubMed, PsycINFO, CINAHL and Cochrane Library for systematic reviews published in the English Language,

between January 2006 and October 2016. Search terms listed in Table 1. were reviewed using the Peer Review of Electronic Search Strategies (PRESS) initiative [23]. Electronic searches were supplemented with reference list searches to ensure sufficient coverage.

Two reviewers independently examined titles and abstracts (facilitated through the online systematic review application, Rayyan[24]). Inter-reviewer agreement was sought through consensus, with any disagreements resolved by a third reviewer. The criteria for inclusion were systematic reviews exploring the use of MSF for qualified physicians in any healthcare setting. Previous systematic reviews focussed on medical students were excluded due to the differing nature of performance assessment in undergraduate medical education. Due to the nature of the review of reviews methodology, non-systematic literature reviews were excluded and the grey literature was not searched. In order to standardise the inclusion process, an inclusion criteria form was developed and piloted (Table 2).

Quality assessment was conducted using a modified version of the AMSTAR checklist as adapted by SIGN [25-27]. One author conducted a full quality appraisal of all the potentially relevant reviews , after a high level of interrater reliability was reached on appraisal of a sub sample of the reviews by two authors (100%). Methodological quality of the included studies was not the main focus of this systematic review, therefore content relevance took precedence over methodological rigour [28]. However, sensitivity analyses were conducted to assess the impact of study quality on review findings (29). Sensitivity analyses test the effect of including (or excluding) review findings of differing quality on the review synthesis. In this instance, any low-quality studies that contributed no new themes to the findings were removed from the analysis.

Using a standardised data extraction form, two reviewers extracted the study characteristics from the included articles. The findings and author conclusions of articles reviewed were extracted, themed and reported in a systematic format (full study characteristics available in Supplementary Data 1). Previous studies and reviews, exploring the validity of MSF assessments within medical education, have often categorised evidence in terms of construct, criterion and/or content validity. For this review however, the themed findings were mapped against a validity framework developed by the American Psychological Association (APA), the American Educational Research Association (AERA), the and the National Council on Measurement in Education (NCME), as described by Downing [29]. The "APA framework", with its five domains of validity evidence, has been described as the '*the current standard of assessment validation'* [30]. Table 3 provides a summary of the framework adapted from Downing [29].

The analysis was synthesised using a modified narrative synthesis technique grounded in Popay et al's guidance [31]. A modified narrative synthesis relies on three non-sequential framework elements: i) developing a preliminary synthesis of findings of included studies; ii) exploring relationships within and between studies; and iii) assessing the robustness of the synthesis by ''relying primarily on the use of words and text to summarise and explain findings from a synthesis'' [31].

Table 1. Search terms

Setting: "health" OR "healthcare" OR "medic*" OR "care" AND **Perspective**: "doctor*" OR "physician*" OR "surgeon*" AND **Intervention**: "multisource feedback" OR "multi-source feedback" OR "peer feedback" OR "colleague feedback" OR "360-degree feedback" OR "360-degree evaluation" OR "MSF" OR "performance feedback" AND **Evaluation:** "reliability" OR "feasibility" OR "accuracy" OR "validity" OR "effectiveness" OR "strength*" OR "weakness*" OR "limitation*" AND "systematic review" OR "review"

Table 2. Inclusion criteria form

1. Is the study published post 2006?
    a. Yes (proceed) b. No (reject)
2. Is the study available in English?
    a. Yes (proceed) b. No (reject)
3. Is the study type a systematic review?
    a. Yes (proceed) b. No (reject)
4. Is the context of the study healthcare?
    a. Yes (proceed) b. No (reject)
5. Does the study discuss MSF in healthcare?
    a. Yes (proceed) b. No (reject)
6. Are qualified medical doctors the target population?
    a. Yes (include) b. No (exclude)

Table 3 Summary of the APA validity framework, adapted from Downing[29]

Domains of Validity Evidence	Description
Content	*Content validity is concerned with ensuring that the content of the test is sufficiently similar to, and representative of, the task that it is intending to measure.*
Response Process	*Response process validity is concerned with ensuring that all sources of error associated with the administration of the test are recognised and limited to the full extent possible*
Internal Structure	*Internal structure validity is concerned with the statistical and psychometrics characteristics of the questions or prompts, and the psychometric properties of the model used to score/scale the assessment. This aspect of validity is involved in determining the generalisability and reproducibility of the assessment.*
Relationship to other variables	*This type of validity evidence is concerned with the correlational or relationship of assessment results with other previous or existing measures of performance*
Consequences	*Consequential validity is concerned with the impact that the assessment has on both the examinees, as well as on the health service, patients and wider society.*

Results

Review characteristics and Study Quality

Eight studies were ultimately included in our qualitative synthesis (see PRISMA flow diagram; Figure 1). Key characteristics of the included reviews are summarised in Table 4. Of the reviews included in

the final analysis, the majority of studies were observational in design with no control group. With the validity evidence for MSF based largely on observational studies, the risk of bias associated with the findings may be viewed as high. However, descriptive and observational studies can still provide useful information of validity evidence, particularly in relation to consequential validity [17].Of the eight reviews included, five included articles exploring only MSF to assess medical performance [8-10,15,16]. The remaining three included studies exploring MSF alongside other WBA methods[17-19]. All reviews included qualified physicians as the target population, with six including studies from multiple specialities [10,15-19], one including paediatric physicians only [8], and one including surgical specialities only [9]. In terms of validity evidence, each of the eight reviews included demonstrated evidence across at least one domain of the APA validity framework: content validity (n=5); response process (n=5); internal structure (n=5) [8-10,15,18]; relationship to other variables (n=3)[9,10,18]; and consequential validity (n=6) [10,15-19], of which three provided evidence of this domain only [16,17,19].

The methodological quality of included reviews was mixed (Supplementary Data 2). Three studies were considered high in quality [16-18], five were considered acceptable [8-11,19] and one study was considered low [32]. The most common methodological weaknesses were reviews excluding studies based on their status (e.g. excluding grey literature/non-peer reviewed articles) and not listing the studies excluded at full text.

Content

Of the literature included in the analysis, content validity evidence for MSF instruments can be categorised into two themes. Firstly, validity is discussed in terms of the technical and non-technical competencies that can effectively and feasibly be assessed through MSF assessments [8-10,15]. Donnon et al (2014) [15] identify five key domains of; professionalism, clinical competence, communication, management, and interpersonal relationships across which MSF can be a valid means of assessing medical performance. However, other reviews highlighted further competencies that can be successfully assessed including treatment skills, patient relationships, collegiality, leadership, decision making, system based practice, probity, and knowledge and judgment [8-10]. One study did however demonstrate that in terms of surgical specialties, MSF appears to adequately assess non-technical skills but fails to adequately assess areas of clinical procedural competence[9]. Secondly, reviews discuss evidence of how the content validity for MSF instruments can be maximised [8,10,18]. A number of reviews conclude that in order to enhance content validity, experts should review MSF question items, as part of the development stage, to ensure desired competencies are adequately assessed by the instrument. It is proposed that this can best be carried out systematically using a modified Delphi to ensure consensus across a number of experts [8,10]. Further validity evidence is required within the

content validity domain in order to provide a consistent understanding of the areas of clinical practice that can be suitably assessed using MSF, as well demonstrating how to best to design MSF questionnaires used within specific medical specialities or for differing purposes of assessment (e.g. regulatory vs. professional development).

Response Process

In the context of MSF, response process validity is concerned with the process of standardisation and quality control in the delivery and completion of assessments. A number of reviews presented validity evidence within this domain[8-10,15,18], however this evidence focussed solely on the feasibility of assessment implementation. Overall, the feasibility of MSF was considered to be high as; 1) assessments often take a short period of time to complete[8,9,18]; 2) assessments are cost effective[10,18] and 3) physicians often receive high rates of response to requests for feedback[8-10]. One review did demonstrate however that physicians of certain medical specialties have issues in finding adequate numbers of suitable assesors[10]. Although the feasibility of MSF is well established within the reviews, no alternative factors that may impact the standardisation and quality control in assessment delivery were reported.

Internal Structure

Evidence exploring the reliability, replicability and generalisability of MSF assessments falls within the domain of structural validity. Five reviews reported evidence supporting the internal structure of MSF assessments [8-10,15,18], using statistical analyses to explore the psychometric properties of assessments. Firstly, overall reliability of MSF was high. Cronbach's Alpha (a) scores for MSF instruments generally were reported to be >=.90 [8,10,15], or standard error measurements (SEM) <=.40 for a number of studies that evaluated the SPRAT tool [8,15]. Secondly, generalisability of test scores was widely evidenced with high generalisability coefficients (>=.80) for instruments involving six to eight (or more) assessors [8-10,15,18]. Thirdly, the expectation of progression was observed through increased feedback scores over time with consistently higher ratings given to advanced trainees by year of programme [8,9,15]. Finally, the consistency of feedback scores between different assessor groups was generally moderate to high with interclass correlations (ICC) ranging between .45 to .90. These scores were often >.70 [8,10,15], demonstrating a good level of consistency between different assessor groups [15]. Despite the high ICC results, a number of reviews did report that clear differences were present in the mean feedback scores between different assessor groups (e.g. senior physicians rated more stringently than junior

physicians) potentially affecting the validity of the instrument [8,15]. Overall, the size and quality of the evidence underpinning the structural validity of MSF within the reviews was high.

Relationship to other variables

To verify if results of MSF assessments are providing a valid representation of physicians' performance, feedback scores can be correlated against scores of other WBA methods to explore consistency of findings. This provides validity evidence, as one might postulate that those who do well in other WBA assessments should also do well in MSF. Three reviews demonstrate validity evidence within this domain[9,10,18], with each reporting significant correlations between feedback scores for MSF assessments with other WBA methods. Significant correlations were observed across a number of medical specialties between MSF assessment results and the results of other examinations including: 1) Procedures Based Assessments (PBA), 2) Objective Structured Assessment of Technical Skills (OSATS), 3) American Board of Surgery in Training Examinations (ABSITE), 4) Patient Satisfaction Questionaires (PSQ), 5) Significant Event Analysis (SEA), plus many others[9,10,18]. Although the evidence base underpinning this domain of assessment validity is small compared with others, findings are consistent and demonstrate that in comparison with other methods of workplace based assessment, MSF can provide a valid representation of physician performance.

Consequences

Consequential validity is concerned with evidence of the intended or unintended consequences MSF assessments may have on participants or wider society. Six reviews demonstrate validity evidence within this domain with much of the evidence focussed around the likelihood of positive change in physicians' attitudes or behaviours as a result of receiving feedback.[10,15-19] In order to stimulate modifications to behaviours and attitudes, reviews identified a number of factors influencing the likelihood of change. In terms of the source of feedback, participants must perceive assessors as credible and familiar with their work.[10,16] In terms of assessment delivery, feedback should be facilitated[16,17] and narrative comments should be employed alongside quantitative questionnaire results.[16,19] As for the content of the feedback, mixed conclusions are drawn about the likelihood of change as a result of negative comments. Two reviews concluding negative feedback reduces the likelihood of change,[10,17] however a further review concluded that negative comments may not stimulate changes in performance where feedback is inconsistent with a physicians' own perceptions of their performance.[16] Repetitive comments about the same behaviour is understood to increase likelihood of change,[16] as does providing the participant time to reflect on the feedback[16] and ensuring that the feedback is specific and action based, avoiding global judgements of performance.[10,19] One review noted variability in the likelihood of change as a result of MSF assessments by seniority and

medical speciality, with some junior physicians and most surgeons displaying little willingness to change. This variability may however be due to individual differences.[17] A general consensus within reviews suggests that well designed, delivered and evaluated colleague feedback (MSF) instruments can lead to modifications in attitudes and changes in behaviour. However, inconsistencies in the findings predicate the need for further research to ensure that MSF can reliably support positive changes in physician performance.

Figure 1. Flow diagram of study selection

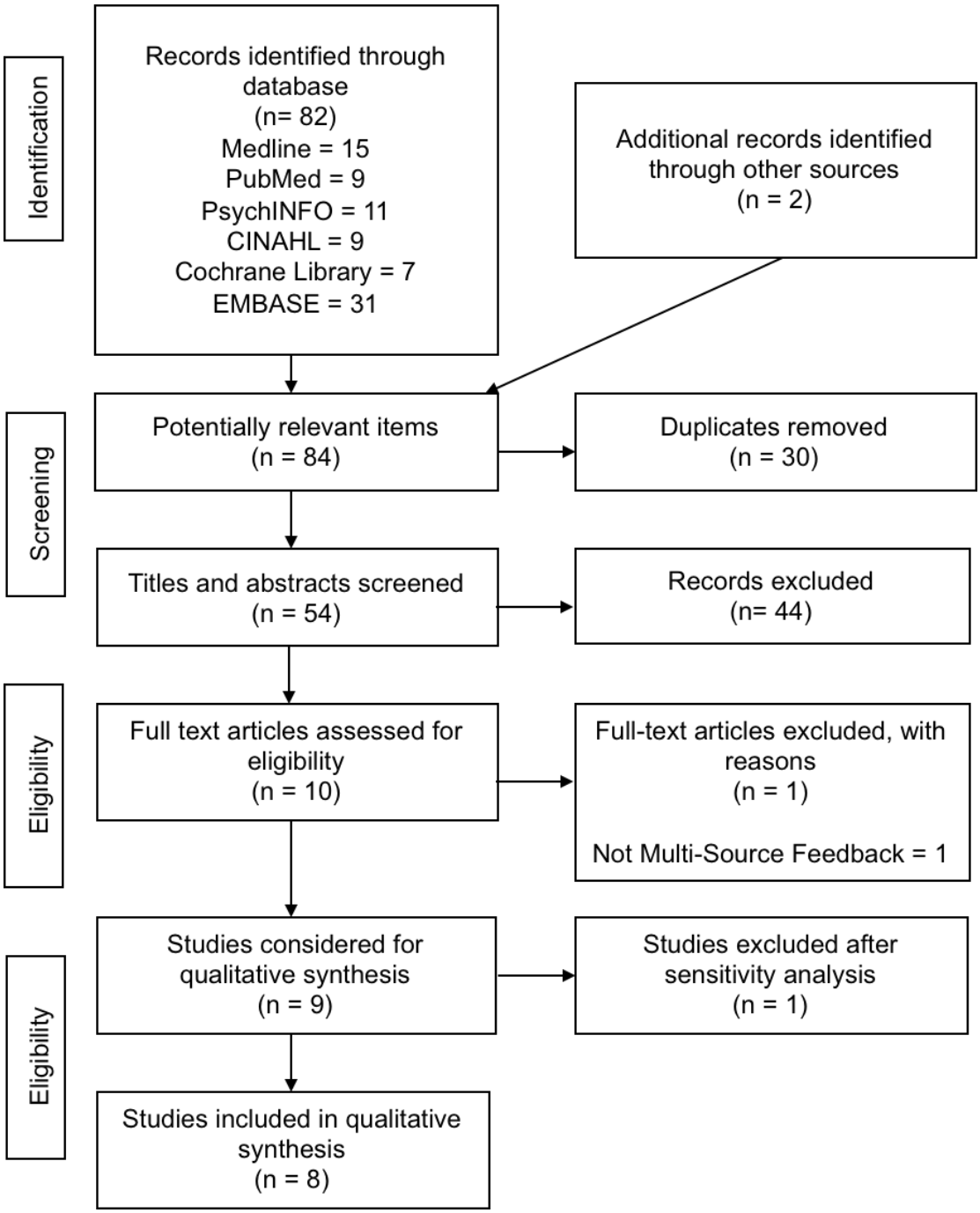

Table 4. Characteristics of included reviews

Authors	Date	Title	Aim	Perspective	Studies	WBA Methods
Al Alawi, S. Al Ansari, A. Raees, A. & Al Khalifa, S.	2013	Multisource feedback to assess paediatric practice: a systematic review	Describe the use of MSF in pediatric settings and to determine its psychometric characteristics and evidence of its validity based on the published literature.	Doctors (paediatrics only)	6	Multi-source feedback (MSF) = 100%
Al Khalifa, K. Al Ansari, A. Violato, C. & Donnon, T	2013	Multisource feedback to assess surgical practice: a systematic review	Describe the use of MSF in surgical settings and to determine the psychometric characteristics and the evidence of its validity based on the published literature.	Doctors (surgical specialties only)	8	Multi-source feedback (MSF) = 100%
Andrews, J. Violato, C. Al Ansari, A. Donnon, T & Pugliese, G.	2013	Assessing psychologists in practice: Lessons from the health professions using multisource feedback.	Review of the use of MSF in healthcare and to summarize the evidence of its feasibility, reliability, generalizability, validity, and other psychometric characteristics.	Doctors (multiple specialties) (n=46), Occupation Therapy (n=1), Medical Radiation Technology (n=1)	48	Multi-source feedback (MSF) = 100%
Donnon, T. Al Ansari, A. Al Alawi, S. & Violato, C.	2014	The reliability, validity, and feasibility of multisource feedback physician assessment: A systematic review	Review research on the different types of MSF instruments used to assess physicians' performance on clinical and nonclinical skills and to investigate the evidence for reliability, generalizability, validity, and feasibility of this assessment approach.	Doctors (multiple specialties)	43	Multi-source feedback (MSF) = 100%
Ferguson, J. Wakeling, J. & Bowie, P	2014	Factors influencing the effectiveness of multisource feedback in improving the professional practice of medical doctors: a systematic review	Identify the key factors that influence the effectiveness of multisource feed- back in improving the professional practice of medical doctors	Doctors (multiple specialties)	16	Multi-source feedback (MSF) = 100%
Miller, A & Archer, J	2010	Impact of workplace based assessment on doctors' education and performance: a systematic review	Investigate whether workplace based assessment affects doctors' education and performance.	Doctors (multiple specialties)	16	Multi-source feedback (MSF) = 50.0% Mini-clinical examination exercise (mini-CEX) = 25.0% Direct observation of procedural skills (DOPS) = 6.25% Multiple assessment methods = 18.75%
Saedon, H. Salleh, S. Balakrishnan, A. Imray, C. & Saedon, M	2012	The role of feedback in improving the effectiveness of workplace based assessments: a systematic review	Elucidate the impact of feedback on the effectiveness of WBAs in postgraduate medical training.	Doctors (multiple specialties)	15	Multi-source feedback (MSF) = 46% Mini-clinical examination exercise (mini-CEX) = 20% Procedure based assessment (PBA) = 14% General workplace based assessments = 7% Multiple assessment methods = 13%

Discussion

This review has systematically collected, synthesised and categorised the evidence underpinning the validity of MSF as an assessment tool to assess the ongoing performance of qualified physicians. No review to date has drawn together all of the evidence supporting or refuting the validity of MSF, to provide an up-to-date and holistic analysis of MSF validity. Using the APA framework to map the current validity evidence for the use of MSF in medicine,[29] it is clear that the size and strength of evidence across the different domains of validity is variable.

This review has demonstrated that the evidence base supporting the statistical and psychometric properties of MSF is sufficient. The internal structural validity of MSF has been repeatdely tested, with feedback instruments often demonstrated to be statistically reliable methods of performance assessment. What is also apparent, although the size of the evidence base is smaller, is that results of MSF assessments often correlate highly with other WBA methods., sufficient evidence exists to demonstrate that MSF is a feasible method of assessing medical performance in terms of cost, time and response rates.[8-10,18] We have also shown however that validity evidence is currently lacking in to order to determine 1) how best to ensure MSF tools measure what they intend to measure (content validity); 2) how best to maximise positive impact on practice (consequential validity); and 3) how to ensure the process of assessment delivery is rigorous, robust, and free from bias (response process validity).

Ensuring the MSF can provide a valid assessment of physician performance is a central component of current debate within the UK. Adopted within a recent process of medical relicensure for physicians,[6] MSF has recently been criticised by Sir Keith Pearson for not being able to "consistently identify physicians…whose behaviours are 'disruptive'", which may impact on "the quality and safety of care provided to patients".[33] Physicians choosing their own assessors and the potential for this to undermine the validity of feedback where "colleagues sometimes lacks the necessary objectivity, honesty and candour" has also been raised as a continuing concern for the validity of MSF.[33] Early work by Ramsey et al suggested that the self-selection of assessors had no significant impact on MSF results.[34] However, the issue of bias within the selection of assessors has previously been brought into question, with one study demonstrating that the 'practice of choosing one's own raters is likely to lead to more favourable results'.[14] A number of studies have suggested that interpersonal relationships may play a part in a physician's selection of assessors,[12-14,35] however more research is required to understand this threat to assessment validity.

Finally, central to the validity debate for MSF is that the priority for different aspects of validity varies depending on its proposed purpose. Reliability and other components of internal structure are paramount if the purpose is to identify poor practice as part of a patient safety agenda. Whereas content, response process and consequences validity come to the fore if the focus is more formative; with the hypothesis that feedback will drive up standards, thereby supporting better patient care. In order to "review different approaches and determine which works best, drawing upon learning from other sectors",[33] the purpose of which MSF is being used must be clearly articulated. As van der Vleuten concludes in his seminal paper there is always a "trade off"; when decisions need to be made about prioritising different aspects of validity.[36] When used within high stakes/regulatory processes, MSF instruments require validity with more evidence in the tool's statistical and psychometric properties (internal structure validity). However, utilising MSF within low stakes/formative processes focusses on the personal development of physicians and requires more evidence in how to facilitate positive changes to practice (consequence validity). While not mutually exclusive, the use of MSF within both high stakes or formative processes has a direct impact on resource allocation and requires a focus on different implementation approaches in how data is collected and analysed. This factors subsequently shapes the validity evaluation for MSF tools in order to understand "which works best".[33]

Limitations

The present review has some limitations. Although comprehensive, the review is based on a relatively modest number of prior reviews that were published in peer reviewed English language journals only. The grey literature was not searched and experts in the area were not contacted. Publication bias therefore cannot be ruled out. The methodological quality of the included reviews varied and the results should therefore be treated with some caution. Variability in the reporting of reliability (i.e., generalisability, intraclass correlation) and validity (i.e., construct and criterion related) measures, while supportive of the MSF process, were difficult to combine consistently between studies. There is also a potential for over reporting of results with four of the reviews using similar search terms and data sources as well as overlap of included studies.[8-10,15] Finally, the absence of evidence synthesis relating to patient feedback, an aspect of performance feedback which many view as part of MSF, is a recognised limitation of this review.

Conclusion

MSF is increasingly adopted within continuing professional development and regulatory frameworks worldwide as a method to assess medical performance and quality assure clinical practice. The validity

evidence for MSF used within medicine has, in many domains, been well-established; however, the size and quality of the evidence base is variable. To ensure that MSF can support improvements in medical performance and subsequently the quality and safety of patient care, further validity evidence is required to determine: 1) how best to design and deliver MSF assessments that address the identified limitations of existing tools, and 2) how to ensure participation in MSF supports positive changes in practice. Further validity evidence will be particularly important if the purpose of using MSF is to support improvements in medical performance and therefore the quality and safety of patient care.

Lessons for Practice

MSF is increasingly adopted within continuing professional development and regulatory frameworks worldwide as a method to assess medical performance and quality assure clinical practice.

The use of MSF within such settings requires assurances for physicians, patients and regulators that instruments contain substantial validity evidence.

More validity evidence is required to determine how best to design and deliver MSF assessments that address the identified limitations of existing tools, and how to ensure participation in MSF supports positive changes in practice.

References

1. Bracken DW, Timmreck CW, Church AH. *The handbook of multisource feedback: The comprehensive resource for designing and implementing MSF processes.* San Francisco, CA, US: Jossey-Bass; 2001.

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

3. Campbell JL, Richards SH, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires. *Qual Saf Health Care.* 2008;17(3):187-193.

4. Levinson W. Revalidation of physicians in Canada: Are we passing the test? *CMAJ : Canadian Medical Association Journal.* 2008;179(10):979-980.

5. Medical Board of Australia. *Registration Standard: Continuing Professional Development.* 2016.

6. General Medical Council. Supporting Information for appraisal and revalidation. 2012.

7. American Board of Medical Specialties. *Standards for the ABMS Program for Maintenance of Certification (MOC).* 2014.

8. Al Alawi S, Al Ansari A, Raees A, Al Khalifa S. Multisource feedback to assess pediatric practice: a systematic review. *Canadian medical education journal.* 2013;4(1):e86-95.

9. Al Khalifa K, Al Ansari A, Violato C, Donnon T. Multisource feedback to assess surgical practice: a systematic review. *Journal of surgical education.* 2013;70(4):475-486.

10. Andrews JJW, Violato C, Al Ansari A, Donnon T, Pugliese G. Assessing psychologists in practice: Lessons from the health professions using multisource feedback. *Professional Psychology: Research and Practice.* 2013;44(4):193-207.

11. Donnon T, Al Ansari A, Al Alawi S, Violato C. The reliability, validity, and feasibility of multisource feedback physician assessment: A systematic review. *Academic Medicine.* 2014;89(3):511-516.

12. Bullock AD, Hassell A, Markham WA, Wall DW, Whitehouse AB. How ratings vary by staff group in multi-source feedback assessment of junior doctors. *Medical education.* 2009;43(6):516-520.

13. Burford B, Illing J, Kergon C, Morrow G, Livingston M. User perceptions of multi-source feedback tools for junior doctors. *Medical education.* 2010;44(2):165-176.

14. Archer JC, McAvoy P. Factors that might undermine the validity of patient and multi-source feedback. *Medical education.* 2011;45(9):886-893.

15. Donnon T, Al Ansari A, Al Alawi S, Violato C. The Reliability, Validity, and Feasibility of Multisource Feedback Physician Assessment: A Systematic Review. *Academic Medicine.* 2014;89(3):511-516 510.1097/ACM.0000000000000147.

16. Ferguson J, Wakeling J, Bowie P. Factors influencing the effectiveness of multisource feedback in improving the professional practice of medical doctors: a systematic review. *BMC Medical Education.* 2014;14:76-76.

17. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ.* 2010;341:c5064.

18. Overeem K, Faber MJ, Arah OA, et al. Doctor performance assessment in daily practise: does it help doctors or not? A systematic review. *Med Educ.* 2007;41(11):1039-1049.

19. Saedon H, Salleh S, Balakrishnan A, Imray CHE, Saedon M. The role of feedback in improving the effectiveness of workplace based assessments: a systematic review. *BMC medical education.* 2012;12:25-25.

20. Overeem K, Lombarts MJ, Arah OA, Klazinga NS, Grol RP, Wollersheim HC. Three methods of multi-source feedback compared: a plea for narrative comments and coworkers' perspectives. *Medical teacher.* 2010;32(2):141-147.

21. Edwards A, Evans R, White P, Elwyn G. Experiencing patient-experience surveys: a qualitative study of the accounts of GPs. *Br J Gen Pract.* 2011;61(585):e157-e166.

22. Gough D, Thomas J, Oliver S. Clarifying differences between review designs and methods. *Systematic Reviews.* 2012;1:28-28.

23. Sampson M. An evidence-based practice guideline for the peer review of electronic search strategies. *J Clin Epidemiol.* 2009;62.

24. Elmagarmid A, Fedorowicz Z, Hammady H, Ilyas I, Khabsa M, Ouzzani M. Rayyan: a systematic reviews web app for exploring and filtering searches for eligible studies for Cochrane Reviews. 2014, 2014.

25. Shea BJ. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J Clin Epidemiol.* 2009;62.

26. Shea BJ, Bouter LM, Peterson J, et al. External validation of a measurement tool to assess systematic reviews (AMSTAR). *PloS one.* 2007;2(12):e1350.

27. Scottish Intercollegiate Guidelines Network (SIGN). Systematic Reviews and Meta-Analyses Methodology Checklist 1 http://www.sign.ac.uk/methodology/checklists.html.

28. Dixon-Woods M, Cavers D, Agarwal S, et al. Conducting a critical interpretive synthesis of the literature on access to healthcare by vulnerable groups. *BMC medical research methodology.* 2006;6:35.

29. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ.* 2003;37(9):830-837.

30. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education.* 2014;19(2):233-250.

31. Popay J, Roberts H, Sowden A, et al. Guidance on the conduct of narrative synthesis in systematic reviews. *A product from the ESRC methods programme Version.* 2006;1:b92.

32. Wilkinson TJ, Wade WB, Knock LD. A blueprint to assess professionalism: Results of a systematic review. *Academic Medicine.* 2009;84(5):551-558.

33. Pearson K. *Taking revalidation forward: Improving the process of relicensing for doctors. Sir Keith Pearson's review of medical revalidation.* http://www.gmc-uk.org/Taking_revalidation_forward___Improving_the_process_of_relicensing_for_doctors.pdf_68683704.pdf: General Medical Council;2017.

34. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of Peer Ratings to Evaluate Physician Performance. *JAMA.* 1993;269(13):1655-1660.

35. Hill JJ, Asprey A, Richards SH, Campbell JL. Multisource feedback questionnaires in appraisal and for revalidation: a qualitative study in UK general practice. *Br J Gen Pract.* 2012;62(598):e314-321.

36. van der Vleuten C. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Education.* 1996;1:41-67.