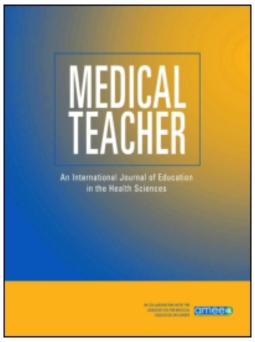

Taylor & Francis
Taylor & Francis Group

Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: http://www.tandfonline.com/loi/imte20

A BEME (Best Evidence in Medical Education) review of the use of workplace-based assessment in identifying and remediating underperformance among postgraduate medical trainees: BEME Guide No. 43

Aileen Barrett, Rose Galvin, Yvonne Steinert, Albert Scherpbier, Ann O'Shaughnessy, Mary Horgan & Tanya Horsley

To cite this article: Aileen Barrett, Rose Galvin, Yvonne Steinert, Albert Scherpbier, Ann O'Shaughnessy, Mary Horgan & Tanya Horsley (2016): A BEME (Best Evidence in Medical Education) review of the use of workplace-based assessment in identifying and remediating underperformance among postgraduate medical trainees: BEME Guide No. 43 , Medical Teacher, DOI: 10.1080/0142159X.2016.1215413

To link to this article: http://dx.doi.org/10.1080/0142159X.2016.1215413

View supplementary material

Published online: 14 Sep 2016.

Submit your article to this journal

View related articles

View Crossmark data

MEDICAL TEACHER, 2016
http://dx.doi.org/10.1080/0142159X.2016.1215413

BEME GUIDE

A BEME (Best Evidence in Medical Education) review of the use of workplace-based assessment in identifying and remediating underperformance among postgraduate medical trainees: BEME Guide No. 43

Aileen Barrett[a,b], Rose Galvin[c], Yvonne Steinert[d], Albert Scherpbier[e], Ann O'Shaughnessy[a], Mary Horgan[b] and Tanya Horsley[f,g]

[a]Education, Innovation and Research, Royal College of Physicians of Ireland, Dublin, Ireland; [b]School of Medicine, College of Medicine and Health Sciences, Brookfield Health Sciences Complex, University College Cork, Cork, Ireland; [c]Discipline of Physiotherapy, Department of Clinical Therapies, Faculty of Education and Health Sciences, University of Limerick, Limerick, Ireland; [d]Centre for Medical Education, Faculty of Medicine, McGill University, Montreal, Quebec, Canada; [e]Faculty of Health, Medicine and Life Sciences, University of Maastricht, Maastricht, The Netherlands; [f]Research Unit, Royal College of Physicians and Surgeons of Canada, Ottawa, Canada; [g]Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada

ABSTRACT

Introduction: The extent to which workplace-based assessment (WBA) can be used as a facilitator of change among trainee doctors has not been established; this is particularly important in the case of underperforming trainees. The aim of this review is to examine the use of WBA in identifying and remediating performance among this cohort.

Methods: Following publication of a review protocol a comprehensive search of eight databases took place to identify relevant articles published prior to November 2015. All screening, data extraction and analysis procedures were performed in duplicate or with quality checks and necessary consensus methods throughout. Given the study-level heterogeneity, a descriptive synthesis approach informed the study analysis.

Results: Twenty studies met the inclusion criteria. The use of WBA within the context of remediation is not supported within the existing literature. The identification of underperformance is not supported by the use of stand-alone, single-assessor WBA events although specific areas of underperformance may be identified. Multisource feedback (MSF) tools may facilitate identification of underperformance.

Conclusion: The extent to which WBA can be used to detect and manage underperformance in postgraduate trainees is unclear although evidence to date suggests that multirater assessments (i.e. MSF) may be of more use than single-rater judgments (e.g. mini-clinical evaluation exercise).

Introduction

The research problem

Progression to competence in postgraduate medical education is complex and the demand for better accountability in the assessment of performance standards and ensuring patient safety and quality of care continues to grow. One of the key challenges facing medical educators is the identification and remediation of underperformance. Almost two decades of research have sought to determine whether the implementation of workplace-based assessment (WBA) can provide accurate, informative, and learning-oriented judgments. While it appears that WBA does not appear to influence changes in practice among the general medical population (Overeem et al. 2009; Miller and Archer 2010; Saedon et al. 2012), we do not know whether or how these assessments can assist in identifying poor performance (diagnostic assessment) or in remediating or changing practice among the subgroup of underperforming trainees or those at a risk of underperforming. As the practice of medical education shifts toward an outcome-based paradigm, and increases its reliance on valid and meaningful work-based assessment methods and practices, we aim to add to the evidence for the use of WBA in the specific context of underperformance among postgraduate trainees.

Practice points

- Evidence for the use of WBA in detecting underperformance is limited, due in part to varied implementation processes, lack of ongoing longitudinal formative assessment and heterogeneity in study designs
- Evidence for the influence of WBA on remediating performance among underperforming trainees has not yet been established
- Multisource feedback, in which overall performance over time is evaluated by multiple raters, may provide better indication of performance than single-episode, single-rater WBA tools

Background

WBA was introduced with the aim of providing trainees with observation-based feedback on their performance in a

CONTACT Aileen Barrett aileenbarrett@rcpi.ie Education, Innovation and Research, Royal College of Physicians of Ireland, Frederick House, 19 South Frederick St, Dublin 2, Ireland

Supplemental data for this article can be accessed here.

real-time work-based setting (Norcini et al. 1995; Norcini and Burch 2007). The implementation of WBA in postgraduate medical education and training programs has consisted of various combinations of tools designed to address observation and feedback on, for example, practical, technical, communication, and judgment skills including the mini clinical evaluation exercise (mini-CEX), direct observation of procedural skills (DOPS), and case-based discussion (CbD) (Kogan et al. 2009). Over time, the implementation of these tools has been highly variable and a debate now exists as to their main purpose and role i.e. as an assessment of *performance*, or an assessment for *learning*.

This debate has been fuelled by a number of recent studies aiming to determine the best use of WBA. Firstly, evidence now suggests that WBA tools do not perform well as both summative and formative assessments (Hawkins et al. 2010; Hatala et al. 2015), in part due to confusion among users – both assessors and trainees – as to the primary aim of the assessments (Menon et al. 2012; Bok et al. 2013; Rees et al. 2014), potentially adding to educators' fears of making erroneous or inaccurate judgments while also ensuring that the patients are safe in the care of their trainees. Secondly, the number of WBAs required to make a reliable summative judgment is considerable; for example, it is estimated that the number of mini-CEX assessments required for such a judgment is between 8–10 (Alves de Lima et al. 2013) which, in busy clinical settings is becoming less and less feasible and acceptable. Finally, emerging research on rater variability in assessment – including the complex and multifaceted cognitive, social and psychological origins of this variability – has also raised questions as to whether WBA can reliably and validly be used to judge performance in a workplace setting (Govaerts et al. 2011, Yeates et al. 2013a, 2013b, Gingerich et al. 2014, Govaerts 2015).

The literature continues to provide consistent evidence that the delivery of negative feedback is a significant concern even for experienced educators (Kogan et al. 2012) and that the concept of "failure to fail" is complex and multifaceted (Dudek et al. 2005). To date, the evidence suggests that the greatest impact of WBA lies in providing observation-based feedback but the impact of these tools on identifying areas for change, and subsequently on changing behavior appears to be limited. Conclusions from a number of systematic reviews (Overeem et al. 2009; Miller and Archer 2010; Pelgrim et al. 2011; Saedon et al. 2012; Ferguson et al. 2014) suggest that multiple factors may need to be considered. Firstly, studies that use changes in performance as the main outcome measures have interpreted this change in practice as "evidence" of learning; while this functional definition is useful to an extent, it is limited and does not include the possibility of a learning effect from affirmation of good practice (De Houwer et al. 2013). Secondly, the reviews we found did not look at whether or not a change in practice was required prior to the WBA episode or intervention; doctors performing at expected levels may not require as many changes to their practice as those deemed to be underperforming.

The impact – or lack thereof – of WBA on changing practice also needs to be considered in light of contemporary feedback literature, specifically studies addressing how trainees perceive and decide to act upon feedback. A recent series of studies addressing this issue provides some interesting insights into how trainees process feedback; using a regulatory focus theory to explore this question, Watling et al. (2012) attempted to understand the complexity and influence of a "promotion" or "prevention" focus on the acceptance or denial of feedback. A key feature of other studies in this program of research also highlighted the importance of feedback culture (Watling et al. 2013a, 2013b; Watling 2014) and the credibility judgments that a learner makes about the feedback provider before deciding on the usefulness or relevance of that feedback (Watling et al. 2012).

With the emergence and adoption of competency-based education (CBME) models in postgraduate medical education comes a dependence on robust, longitudinal and continuous low-stakes assessment tools and methods that will aim to assist in tailoring learning and development to an individual trainee's needs and achievement of pre-defined program outcomes. Early identification – and/or remediation – of underperformance is one key goal of these contemporary medical education models; the question for our review is thus to determine the extent to which WBA tools and the methods by which they are implemented assist in identifying underperforming trainees and whether or how WBA may also assist in remediation of this underperformance.

Review objectives

The aim of this review is to comprehensively review the existing WBA literature to answer two overarching research questions:

1. How has WBA been used to identify and/or remediate underperforming postgraduate medical trainees?
2. What features or implementation conditions of WBA tools specifically contribute to identifying or remediating underperformance among postgraduate medical trainees?

In order to ensure the team shared a consistent understanding of terms and interpretation a series of definitions for the purposes of the review were determined *a priori* (Table 1).

Methods

The study methods followed the BEME-approved study protocol (Barrett et al. 2015). The review is reported here in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) (Moher et al. 2009) (Appendix 1) and the standards for reporting literature searches (STARLITE) guidelines (Booth 2006) (Appendix 2, both are available online on the Journal website as Supplemental Material).

Search strategy

A detailed MEDLINE search strategy was developed in collaboration with an information scientist during the development of the study protocol (Appendix 3, available online on the Journal website as Supplemental Material). Following the acceptance and publication of the study

Table 1. Agreed terms and definitions.

Term	Definition for the purposes of the review
Underperformance	*Underperformance* within a clinical context is inconsistent within the literature and terms are often used interchangeably. The most contemporary (2013) definition provided within a UK-based study defines the underperforming trainee as "requiring intervention beyond the normal level of supervisor–trainee interaction" (Mitchell et al. 2013). While this definition does not classify the root cause of the trainee's difficulties it provides an overarching articulation of a trainee who is not currently meeting the expectations of their training level and we decided to use it as a reference to terms including "the trainee in difficulty," "the difficult/problem trainee" and "the trainee in trouble"
WBA	Any assessment tool or method designed to provide feedback on performance and inform improvement in a practice setting and included (but was not limited to) tools such as: a. Mini-CEX b. DOPS c. CbD d. OSATS (Objective structured assessment of technical skills) e. MSF was used to refer to various tools designed to collect evaluations of the performance by multiple assessors, which is then collated and discussed with the trainee by a single facilitator. The tools in use include the Mini-PAT (mini-peer assessment tool) and TAB (team assessment of behavior) and other formats referred to as 360^0 feedback
Postgraduate medical trainees	Post-qualification doctors pursuing further clinical training in order to register as a specialist, for example, in medicine or surgery (e.g. resident, trainee, doctor-in-training, nonconsultant hospital doctors
Remediation	"The act or process of correcting a deficiency" as described by Cleland et al. (2010, p. e185). This particular definition was chosen for our review as it links closely with the purpose of formative assessment, which is to provide information on the performance strengths and deficiencies, to provide a structure for feedback and guidance on improving performance
WBA tool features or factors	The *features* or *factors* of WBA tools we expected would be described included: • The WBA rating systems and feedback structures • WBA methods of use including such considerations as whether they are used routinely or in the case of suspected underperformance, if multiple tools are used and if one or many encounters were used in identifying or remediating performance-related issues

Table 2 Inclusion/exclusion criteria.

	Inclusion criteria	Exclusion criteria
Population	• Postgraduate medical/surgical trainees	• Nonmedical trainees • Medical students (undergraduate and graduate-entry programs) • Studies not involving humans • Studies in medical areas not related to humans (e.g. veterinary studies)
Intervention	• WBA tools e.g. Mini-CEX, DOPS, case-based discussion, OSATS	
Outcomes	• Studies that described/reported outcomes distinctly related to identification/remediation	
Research Design	• No restriction for study design was applied	• Studies that did not report primary data • Reports published only in dissertation format
Context	• Routine or targeted use of WBA • Trainee-led or trainer-led WBA • Single or multiple WBA events • Use of WBA as part of a wider program of assessment or in the context of a range of assessment evidence • Management or remediation of underperformance for knowledge, skills, and attitudes • Presence/absence of facilitation and/or written or verbal feedback	

protocol the search strategy was independently reviewed by a second information scientist experienced in BEME reviews.

A third university-based information scientist collaborated with the lead author (AB) on the adaptation of the search strategy for the eight bibliographic databases used for the study (MEDLINE, Science Direct, PsycInfo, Australian Education Index, British Education Index, Education Resource Information Centre (ERIC), Cumulative Index to Nursing and Allied Health Literature (CINAHL) and Excerpta Medica database (EMBASE)) and also provided a "tracking tool" as a search log for each database. The searches were updated in November 2015.

The Cochrane database of systematic reviews, the BEME published reviews, conference abstracts for AMEE (Association of Medical Education in Europe) and ASME (Association for the Study of Medical Education) and a number of medical education journals (*Medical Teacher*, *Medical Education*, *The Clinical Teacher,* and *Advances in Health Sciences Education*) were hand-searched for the years 2010–2015.

Inclusion and exclusion criteria

The inclusion and exclusion criteria are presented in Table 2.

Following the initial screening process by the two authors, we devised a novel "voting" system for the articles requiring further consideration; a "voting spreadsheet" was compiled to identify (non)agreement on inclusion

Table 3. Study outcomes.

Outcome	Example
Individual-level	• Number of trainees identified as poorly performing through the use (either routine or targeted) of a WBA process • Progression/remediation statistics • Changes in trainee performance (knowledge, skills, attitudes, etc.) • Trainee satisfaction • Kirkpatrick's educational outcomes
Practice-level	• Changes in implementation methods, e.g. nonroutine to routine • Implementation of new/differing WBA tools
System-level	• Changes in system-wide implementation of WBA tools or methods e.g. throughout a deanery

(Appendix 4, available online on the Journal website as Supplemental Material

Evaluation of study outcomes

The primary outcomes of the review were those perceived to be resultant from the use of a WBA intervention at three conceptual levels: (1) the individual (trainee) level; (2) practice level (e.g. change from nonroutine to routine use of WBA by the training body or institution), and (3) system-level (e.g. deanery-wide implementation of a new tool) (Table 3). Other outcomes included the conditions under which the use of WBA is most useful for identifying or remediating underperformance and, where possible, the features of WBA tools, or factors in using WBA, may directly contribute to successful remediation of underperformance.

Educational outcomes were organized using Kirkpatrick's framework of educational outcomes, using Barr's adaptations for medical education research (Barr et al. 2000) and further adaptations by Steinert et al. (2006) that subdivided the original Level 3 into self-reported (3a) and observable (3b) changes in behavior (Appendix 5, available online on the Journal website as Supplemental Material)

Assessment of methodological quality

Observational studies

We chose to use the methodological quality assessment approach described by Buckley et al. (2009) for observational studies. Each criterion was independently rated as "met," "unmet," or "unclear." This framework suggests that in order to be deemed of high quality, studies should meet a minimum of seven of these 11 quality indicators. Recent guidelines suggest that reporting of ethical issues should consider both ethical approval for a study and issues of informed consent separately (Lo 2012). Where this is not reported, the guidelines suggest that this be deemed "unclear." We, therefore, modified the BEME criterion ("are all ethical issues articulated and managed appropriately?") to these specifications.

Qualitative studies

Of the two included qualitative studies in this review, one used mixed methods and it was agreed that the Critical Appraisal Skills Programme (CASP 2013) guidelines for the reporting of all qualitative studies would be appropriate for the purposes of this review.

Data extraction and coding

A modified BEME coding sheet (Appendix 6, available online on the Journal website as Supplemental Material) was developed by the review team and approved by the study's appointed BEME International Collaborating Centre (BICC). The modifications related to specific data required for the review included

- Study design, presence/absence of a conceptual framework
- Population and setting including training program, year in training
- WBA "intervention" tool characteristics (including rating scales) and method of implementation
- Conditions of use i.e. routine/targeted and any specified implementation factors
- Evaluation and outcomes of the study including educational impact, identification and/or remediation outcomes

All included articles were independently coded by two of the investigators (AB and RG). A coding pilot was performed in which both coders independently reviewed a set of two papers; a third reviewer (unrelated to the project) independently performed a quality check of the data and reported no major discrepancies in data extraction between the two coders.

Results

Selection of papers

Original searches identified 7067 papers. Following deduplication this resulted in 6261 papers for screening. All searches were imported into EndNote X7. A flow diagram of the selection process is detailed in Figure 1, available online on the Journal website as Supplemental Material. A search log of all records retrieved by each database (Supplementary Appendix 7) is also available online.

The initial screening was completed by AB; at the start of the process, a quality check was performed in which AB and RG independently screened the titles and abstracts of the first 836 citations and performed an inter-rater reliability calculation. Current practice suggests that Cohen's weighted kappa and percentage agreement both have strengths and limitations and more than one determination of agreement should be used (McHugh 2012); we therefore performed a percentage agreement (99.04%) along with a weighted kappa (Appendix 8, available online on the Journal website as Supplemental Material), which was 0.641 indicating moderate-to-good agreement but also cannot rule out the role of chance in this agreement statistic. All disagreements at this stage (8 of 836 citations) were resolved by discussion.

Of the 6261 papers screened, 6059 were excluded on the basis of title and abstract. The full text of 202 papers was retrieved. AB performed a second screening of this set, compiling those papers into three files: *exclude*, *include*, and *for team discussion*. At this stage 169 papers were excluded, 16 were included and 17 were identified as requiring further discussion by the review team.

RG performed a second quality check of all three sets and there was complete agreement.

Using the voting system described earlier in the methods section, there was complete team agreement on the exclusions of nine papers. For the remaining eight papers, team discussion took place via email. No study was excluded on the basis of quality throughout this process; methodological quality was evaluated following agreement on the final included set of articles. Twenty studies were included in the final review.

Review findings

Overview

All reports were published as journal articles from 2000–2015. The majority of studies (13) took place in the UK and four studies were carried out by the same research team (Wood et al. 2006; Whitehouse et al. 2007; Bullock et al. 2009; Hassell et al. 2012;) and using different data sets. The study designs included evaluative and retrospective studies, with the majority taking place in hospital settings.

Of the 18 quantitative studies included, multisource feedback (MSF) was the WBA "intervention" in 13 studies (Appendix 9, available online on the Journal website as Supplemental Material); six of these studies specifically used the mini-team assessment of behavior (mini-TAB) format (Appendix 10, available online on the Journal website as Supplemental Material). Two cross-sectional studies surveyed program directors about the methods used to identify trainees in difficulty (Yao and Wright 2000; Brown et al. 2008) and one study attempted to determine whether scores in a number of WBAs could predict underperformance specifically. One study considered OSATS only (Hiemstra et al. 2011) and the minicard direct observation tool was evaluated by Donato et al. (2015) A new pharmacotherapy-structured clinical observation (P-SCO) tool was implemented and evaluated by Young et al. (2011).

Two prospective studies (Whitehouse et al. 2007; Hassell et al. 2012) and three cross sectional studies (Yao and Wright 2000; Brown et al. 2008; Burford et al. 2010) were identified for our review. Of the 13 remaining studies, seven were retrospective while the timeline of the intervention (vs the research study) was unclear in the other six studies. A number of studies reported conclusions that were not fully supported by the data emerging from the studies; in these studies the data appeared to be retrospectively reported and would have been enhanced by pre-intervention consideration of study outcomes.

The qualitative studies included an evaluation of anesthetic trainer and trainee perceptions of the mini-CEX (Weller et al. 2009) and participant experiences of a new model of facilitation of MSF (Sargeant et al. 2011) (Appendix 11, available online on the Journal website as Supplemental Material).

Methodological quality

The methodological quality of all included studies was independently evaluated by two reviewers (AB and RG) and minor differences were resolved by discussion. However, it is worth noting that two specific methodological evaluation judgments included in the BEME criteria and CASP guidelines should be considered in the context of the time at which the majority of studies took place and which may impact on the overall impression of methodological quality:

1. None of the studies included in the review provided a pre-identified conceptual or theoretical framework. The inclusion of such a critical "lens" through which a study's results can be interpreted or analyzed is a relatively recent development in medical education research
2. We also attempted to identify ethical issues associated with both informed consent and ethical approval. Until relatively recently, medical education research was generally considered exempt from institutional approval as it was not perceived to involve "risk" to the participants. Complete reporting of ethical issues was a limitation of all 18 quantitative studies. While a number of studies reported prospectively obtaining research ethics committee approval, informed consent was not evident in 10 studies where trainee data or participant information were analyzed

Methodological quality of quantitative studies

While study designs and data collection methods were generally appropriate to the stated research question, the strength of the findings was limited by the absence of before-and-after interventional studies. Two of the included cross-sectional studies (Yao and Wright 2000; Brown et al. 2008) were reliant on participant recall of the number of trainees identified as being "in difficulty." Unfortunately, these data were not triangulated with any documentary evidence which contributed to the limitations of the study conclusions. Lack of triangulation in general was a limitation of 12 of the 18 quantitative studies (Appendix 12, available online on the journal website as Supplemental Material).

Although none of the included studies were randomized trials, we included a specific quality indicator addressing whether the study authors had attempted a risk-of-bias assessment *and/or* included a statement of researcher positionality (modification of the BEME quality indicator No.5 (Buckley et al. 2009)). Only a single evaluation study (Young et al. 2011) explicitly addressed risk-of-bias and stated that the raters involved in the study had not been blinded to the study hypothesis. A large-scale study by Archer and McEvoy (2011) would also have been enhanced by an assessment of risk-of-bias inherent in the study design in which MSF raters were not blinded to the fact that the doctors under assessment had already been referred to the National Clinical Assessment Service (NCAS) with suspected issues of underperformance.

Methodological quality of qualitative studies

Two qualitative studies were identified for inclusion in our review. The methodological approaches to both qualitative studies included in the review and their study designs were generally appropriate to the articulated research questions (Appendix 13, available online on the Journal website as Supplemental Material).

The methodological quality of the study by Sargeant et al. (2011) was good, with the study meeting 8 of the 10

CASP criteria. The main issue of concern to the quality of this study was the lack of consideration of the influence of the researchers on the data collection and analysis process, given that they had also acted as facilitators of the workshops around which the study was based. The interview topic guide was not provided and therefore it is unclear as to whether the study methods were entirely matched to the research question.

The study by Weller et al. (2009) also met 7 of the 10 CASP criteria; the main issues limiting the quality of this study involved lack of clarity around recruitment and what, if any, influence the researchers had on the data collection and/or analysis process. While the interviewer-participant relationship was articulated, the relationship between the study gatekeeper and the invitees is unclear.

Descriptive summary of review outcomes

Trainee-level outcomes

Seven studies in our review reported trainee-level outcomes. Using the modified Kirkpatrick's hierarchy of educational outcomes, we determined that five of the studies reported outcomes at *Level 1a* i.e. learner reaction to the educational intervention (Hesketh et al. 2005; Whitehouse et al. 2007; Weller et al. 2009; Burford et al. 2010; Chipp et al. 2011). Outcomes at *Level 2a* (i.e. evidence of change in skills) were reported by Hiemstra et al. (2011) by developing learning curves to plot the trajectory of OSATS ratings and achievement of pre-defined levels of "good" practice. Self-reported changes in trainees' clinical practice (*Level 3a*) were reported by Sargeant et al. (2011) in which GP trainees were interviewed about the impact of a new model of feedback delivery in the context of MSF; however, as is the limitation of self-reported behaviors, this outcome was not corroborated by triangulation with any other performance data.

We also explored the data for evidence of any other impact of the WBA intervention on the individual trainee. Where outcome data were reported, trainees generally improved performance or progressed throughout training. Black and Welch (2009) reported outcomes for trainees identified as "in difficulty" for the study period. Identification mechanisms included use of the mini-peer assessment tool (mini-PAT). The outcomes, however, were reported for all trainees in difficulty, regardless of the mechanism of identification and did not differentiate those identified as underperforming using this tool.

Brown et al. (2008) reported on a survey of program directors in which the presence or absence of a remediation mechanism was explored. They concluded that programs with an established remediation program were more likely to report the identification of trainees requiring remediation, however, the outcomes (and descriptions) of these remediation interventions was not provided.

In another survey of program directors, Yao and Wright (2000) reported that participants "estimated" the rate of completion of residency among "problem residents" as 57%, with 18% requiring additional time but completing the program, 9% moving to another similar residency program, 10% moving to a difference residency program, and 4% leaving medicine. These study findings are, however, limited by reliance on recall and were not triangulated by any documentary evidence to support the findings. It was also unclear as to how many of these "problem residents" had been identified using the WBA (mini-CEX) or if WBA was a feature of any of the remediation interventions.

The development and implementation of an assessment system incorporating MSF was described by Hesketh et al. (2005). They reported that, of the trainees given four or fewer ratings of "requires help/attention" (Rh/A) on any part of the assessment system (including a 360° feedback), all trainees (100%) received a satisfactory overall evaluation. In the case of trainees with 4–10 Rh/As, educational supervisors differed in their management, with some requiring a repeat evaluation and others progressing, while trainees with up to 10 Rh/As were all required to repeat their evaluation. The overall outcomes for this group, along with any remediation interventions, were not reported.

Practice-level and system-level outcomes

None of the included studies reported changes in the practice of using or implementing WBA or in system-level outcomes. While seven studies described the development and/or implementation of a WBA tool or methods (Hesketh et al. 2005; Wood et al. 2006; Warm et al. 2010; Chipp et al. 2011; Sargeant et al. 2011; Young et al. 2011; Donato et al. 2015), it was unclear as to whether these studies and their findings resulted in changes at either of these levels.

Synthesis of findings

In attempting to answer our original research questions, we have synthesized the review findings to discuss firstly, the use of WBA in identification of underperformance and secondly, its use in the remediation of underperformance among postgraduate medical trainees. Although the variation in study purposes, designs, outcomes and implementation among the studies precluded a qualitative or quantitative meta-analysis, we have instead attempted to provide a descriptive synthesis of themes emerging from the included papers in the context of these original research questions.

How has WBA been used to identify underperforming postgraduate medical trainees?

While the optimal mechanisms and conditions of implementation of WBA to identify underperformance are not yet clear, due to the small number of empirical studies included in this review, the emerging themes presented here tend to support the longitudinal integration of regular, continuous low-stakes WBA as important determinant of overall performance among this cohort.

Routine or targeted use of WBA

Of the 12 studies that described the routine (nontargeted) use of WBA, nine referred to the routine use of MSF. The majority of the MSF events occurred once per year or once per six-month training post.

In a single study reporting the targeted or purposeful use of WBA, Archer and McAvoy (2011) described the use of MSF for doctors referred to the UK NCAS with suspected issues of underperformance and the ability of the tool to discriminate between doctors previously identified as underperforming and a "normative reference group" of pilot participants. While this study did reveal that the MSF scores for the underperforming group were significantly lower than those of the control group, two additional study design issues need to be considered in the interpretation of these findings.

Firstly, the assessors who completed the MSF were aware that the doctor being assessed had been referred to the NCAS. From the vast literature on assessor rating variability (Gingerich et al. 2011; Yeates et al. 2013a, 2013b; Govaerts 2015) and evidence that assessors are reluctant to provide negative feedback in face-to-face situations (Kogan et al. 2012; Menon et al. 2012) this may have swayed assessors' ratings and potentially facilitated the delivery of negative feedback if they were reassured that the underperformance had already been identified by others.

Secondly, the NCAS provides assessments for all doctors, including trainees, but this study did not distinguish between trainees and nontrainees. It is therefore not possible to extrapolate their findings purely to the group of underperforming trainees. It would be interesting to further explore this study's data to ascertain differences in the discriminatory ability of the MSF tool between these two cohorts.

Trainee- or trainer-led WBA

This aspect of WBA implementation was not clearly described in the majority of studies included in this review, with the exception of studies related to MSF. As described previously, MSF is characterized by the collection of feedback from multiple assessors. In general, these assessors are identified by the trainee and supervisor together. In our review, MSF was implemented in this way by all authors with the exception of Wood et al. (2006); these authors described the identification of assessors by supervisors only which they determined was justified in the context of the pilot phase in which this was implemented. In general, implementation of the routine MSF was either voluntary or mandatory but coordinated by the local educational supervisor or deanery lead.

None of the included studies reported or described trainee-led implementation processes and methods; therefore, a comparison of the impact of either method on the identification or remediation use of WBA could not be established.

Single or multiple WBA events

Studies included in the review generally reported on single WBA events and therefore on immediately related outcomes. This is in direct contrast with the premise of WBA that supports the use of multiple low-stakes assessments and with current evidence that suggestsa single WBA event is not, by itself, a reliable judgment of overall trainee performance (Hatala et al. 2015). However, the use of a single WBA event to identify and provide feedback on specific *areas* of underperformance has not yet been established.

Of the 13 MSF studies identified in this review, there were no comparison studies of the use of single-vs-multiple events and their impact on outcomes. It was also not possible to determine from the study by Mitchell et al. (2011), whether the programs using mini-CEX, DOPS, and case-based discussion had implemented these according to research guidelines for the number required for good reliability e.g. in the case of mini-CEX, whether a minimum of 8–10 assessments were recorded (Alves de Lima et al. 2013). This retrospective study looked at "mean" scores for each assessment type (mini-CEX, DOPS, CbD, mini-PAT) for each trainee by converting the narrative ratings (expectations-based) into scores of 1–6 and retrospectively compared these means for two groups – trainees already "tagged" as in difficulty and those who had not been tagged. While some associations were noted between lower mean scores on mini-CEX and CbD in trainees already "tagged" as in difficulty, there was little evidence of predictive ability of WBA to identify trainees in difficulty using receiver operator characteristic (ROC) curves.

The study did not provide information as to how underperformance had been identified among this group and did not evaluate how many assessments were performed per trainee to generate this mean. The use of a single "mean" score also limited the ability of the tool to identify underperformance as this may not pick up on subtle "dips" in the performance as opposed to trends seen over time.

Use of WBA as part of a wider program of assessment or in the context of a range of assessment evidence

The use of WBA as part of an overall system or program of assessment was not a key feature of the studies included in this review. However, a single study by Hesketh et al. (2005) described the use of MSF as part of a PHAST (pre-registration house officer appraisal and assessment system) system of assessment for pre-registration house officers (this study took place prior to the establishment of the 2-year Foundation Programme in the UK). The 360° (or MSF) assessment tool was completed by four raters twice in the year using a four-point narrative scale ("excellent – good – satisfactory – requires help/attention"). The paper provided details of the implementation of this 360^0 assessment alone and did not compare the results to the entire portfolio of evidence generated within the PHAST system and therefore could not be compared to using the MSF tool alone.

Internal medicine residents participating in a year-long ambulatory clinical attachment were assessed at two points in the year using MSF along with clinical quality data, patient ratings and knowledge-based test scores and ranked relative to peers for each component (Warm et al. 2010). The authors were able to identify poor performance relative to peers, and compared to quality data, however, it is not clear from the study whether this ranking system was more or less effective than using MSF alone to identify underperformance, although there were a number of highlighted areas of inconsistency between MSF scores and other values.

Presence/absence of facilitation and/or written or verbal feedback

Three studies included in our review examined whether or not the addition of verbal or written feedback had an impact on the quality of WBA; in general, the impact of that feedback on detection of underperformance appears to depend to a large extent on the quality and or specificity of that feedback.

A qualitative study by Sargeant et al. (2011) explored the impact of an "ECO" (emotion, content, outcomes) model of facilitated feedback on trainer and trainee perceptions of the MSF process. As described above, the impact of the intervention was only evaluated at the level of self-reported changes in practice among trainees; nevertheless all participants (GP trainers and trainees) were positive in their evaluation of the model. The study did not, however, explore whether this model was more or less useful than usual MSF processes in identifying or remediating underperformance.

The "learner-centeredness" of written MSF feedback was analyzed by Vivekananda-Schmidt et al. (2013). Of a total sample of 11,483 MSF forms, only 4777 (42%) contained any free-text comments. Using a content analysis approach, the authors determined that where feedback was provided, this generally tended to be "rater-centered," with an emphasis on the trainee's impact on the assessor's working life rather than goal-oriented feedback for the trainee's development e.g. whether the trainee was a 'good colleague' and contributed to the team.

The authors also specifically analyzed the 513 forms containing a "below average" rating; of these, only 56% contained free-text feedback despite explicit instructions that all such ratings should be accompanied by feedback. Given the lack of trainee-centered free-text feedback in general, and specifically in trainees at risk of underperformance, the authors concluded that MSF may be of limited use in identifying or remediating underperformance where the feedback is not informative.

Young et al. (2011) explored the implementation of a new tool, the Pharmacotherapy Structured Clinical Observation Tool (P-SCO) for third-year psychiatry trainees and compared the written comments on the tool to those provided on the comparator tool, a global rating scale. Their results showed that assessors were more likely to provide specific comments on the P-SCO compared to the global rating scale, providing 2.6 times more affirmatory feedback and 5.3 times more corrective feedback. The P-SCO also identified more ratings of "below expectation" than the global rating scale. The form specifically requested "key feedback points, including what was done well and at least one task to work on" and all assessors had participated in a faculty development workshop prior to the implementation.

Rating scale variation

While MSF rating scales varied slightly throughout all 13 studies, one paper specifically addressed whether scale length impacted on the number of "below expectation" ratings identified by assessors (Hassell et al. 2012). Using four versions of the team assessment of behavior (TAB) MSF tool, the authors reported trends toward fewer underperformance ratings using the longer versions of the scale. However, the study design meant that the four versions of the form were used in four different training locations; there was no direct comparison of rating scales among a single group, therefore, the findings are limited in their generalizability.

Rater variation

It is widely accepted that rater variability is an important factor in any assessment of performance and this has been extensively studied in the context of WBA (Govaerts et al. 2011; Yeates et al. 2013a, 2013b; Govaerts 2015). This BEME review included one such study which looked at whether different rater groups were more or less likely to identify a trainee in difficulty (Bullock et al. 2009); while these authors determined that some assessors were more likely to be more lenient than others and the study reported concern ratings, the impact of this rater variability specifically on detection or remediation of underperformance was not fully explored.

How has WBA been used to remediate underperforming postgraduate medical trainees?

Only one study included in this review attempted to determine if a relationship between remediation processes and trainee outcomes exits. Brown et al. (2008) surveyed 100 otolaryngology program directors to determine which assessment tools were in place across all program and found some weak correlations between the provision of formative feedback and identification of underperformance. Where program had a remediation mechanism in place, they were more likely to identify underperformance but the study could not identify whether having the mechanisms in place first allowed for better identification *or* if issues related to identification of underperformance necessitated the development of remediation mechanisms.

However, the usefulness of these findings were limited by the study design; this was a survey of program directors in which data collection relied upon on recall and was therefore subject to recall bias as there was no triangulation with documentary evidence of trainees in difficulty. The self-reported feedback mechanisms may or may not reflect actual practice.

Discussion

Over the past twenty years, since the introduction of the mini-CEX (Norcini et al. 1995) a vast body of literature has emerged on the implementation of workplace-based assessment. While a number of previous systematic reviews have failed to unearth definitive effectiveness of WBA in changing practice (Overeem et al. 2009; Miller and Archer 2010; Saedon et al. 2012), we recognized that in a group of well-performing doctors, change may not always be necessary. We therefore focussed our attention on the use of WBA within the context of trainees in difficulty.

Our review has allowed us to examine this previously unexplored research topic and to understand the extent to which WBA is currently used in identifying trainees in difficulty and its use in remediation practices. The review also allowed us to identify and describe the limitations of current research in contributing to this important conversation in postgraduate medical education including methodological and study design limitations.

While it appears that the routine integration of some WBA methods and tools may assist in identifying *areas* of underperformance, its use in identifying trainees who are generally underperforming is not yet clear. Although this is due to multiple factors, including the implementation challenges and variations encountered by training bodies and institutions, the absence of well-designed interventional studies also limits our ability to answer this question definitively.

WBA and underperformance

Our search strategy uncovered a number of studies reporting underperformance data; in many cases the authors proposed that this in itself provided evidence of the ability of the tool to identify underperformance, but these data were generally not supported by any other sources. In two studies that compared WBA ratings of performing and underperforming trainees (Archer and McAvoy 2011; Mitchell et al. 2011) the strength of the association provided in the results was limited by the study design. In both cases, the group of underperforming trainees had been identified or flagged as underperforming by other means which were unclear, possibly by "expert opinion". In particular, the study by Archer and McAvoy (2011), while providing statistically significant differences in ratings between the two groups, was limited by a design bias in which the MSF assessors for the underperforming doctors were not blinded to the fact that the doctor under assessment had already been identified as underperforming and had been referred to the NCAS for assessment.

On reviewing the included studies as a whole, it is not possible to definitively articulate how WBA may be of use in identifying trainees in difficulty, or the implementation conditions that may contribute to this detection. The majority of studies reported outcomes such as numbers of concern ratings seen on trainee assessments, but the implications and outcomes of this detection were in general not provided and in very few cases triangulated by other performance markers. Nevertheless, we can make some observations on themes we found throughout the studies.

MSF, which assesses general aspects of trainee performance including communication skills and ability to work as part of team, was the most commonly used WBA tool among studies included in this review. While we cannot directly compare its features to those of other WBA tools, which focus on specific aspects of clinical performance (e.g. mini-CEX), it is worth noting a number of features of that tool – and its implementation – that may influence its ability to identify underperformance.

Firstly, the MSF process requires input from a number of assessors. In this review, the majority of MSF interventions involved more than six assessors and trainees in all except one study (Wood et al. 2006) were involved in choosing their assessors. The responsibility of providing ratings is cushioned by the fact that the feedback is collated from all assessors and delivered to the trainee anonymously, therefore in contrast to, for example, the mini-CEX, an assessor may be more willing to provide below average or concern ratings. This may also mitigate against the bias effects of a single-assessor WBA in which rater variability may provide a threat to reliability and validity (Gingerich et al. 2011; Govaerts et al. 2011; Yeates et al. 2013; Gingerich et al. 2014). In our review, Bullock et al. (2009) noted differences in MSF ratings among different professional groups which are line with other studies of rater variability. The collation of feedback among 10 assessors may, therefore, provide a more reliable overall assessment of performance than a single-rater judgment.

Secondly, the majority of studies reported the routine use of the MSF once a year, or twice at most. Again in contrast to other WBA tools that document a single WBA event, the MSF assessment and ratings provide for general impressions of the trainee over time and do not focus on a single interaction. The supervisor or trainer therefore delivers collated feedback which may assist in delivering negative feedback if issues of concern have been identified by more than one rater.

Our review suggests that the shorter versions of the TAB form (using 3- or 4-point scales) appeared to have slightly better detection rates of underperformance than longer versions (Hassell et al. 2012). The limitation of this finding, however, lies in the fact that the variations of the forms were used in different cohorts of trainees and therefore we cannot imply that for the same group, one type of rating is better than another. This finding, however, contrasts with those of a 2008 study of mini-CEX rating scales in which both five- and nine-point rating scales had similar interrater reliability, but the longer version was deemed to be more "accurate" in determining competence (Cook and Beckman 2008).

It appears that MSF is generally implemented in the same manner across all training programs, however, the various MSF tools have not been compared to each other (generic MSF, mini-PAT, and TAB) and it would be worth considering whether one tool may be superior to others in its ability to detect or remediate underperformance.

WBA and remediation

It appears from this review that there is little published literature on the use of WBA in remediation of underperforming trainees. This is particularly important in the context of programmatic approaches to assessment – and outcome-based education models – in which it is evidence of multiple types of assessments, under multiple conditions and stakes, and using multiple assessors that will create the overall picture of the performing or underperforming trainee. Remediation mechanisms will require in-depth evidence-based approaches to be effective and to ensure that graduates are competent and fit to practice.

Implications for practice

Although the findings of this review have few definitive or immediate implications for practice, we have attempted to comprehensively review the existing body of literature to bring together emergent themes and trends that support the use of WBA in identifying or remediating underperformance among postgraduate trainees. Further research is required to determine whether certain tools (and/or their implementation methods) are better than others in detecting underperformance and this will require robust comparison-based study designs along with consideration of the interpretation of single-episode concern ratings as opposed to ongoing underperformance issues. It appears that of the WBA tools identified in this review, MSF may provide a method to detect underperformance more than other single-rater tools where the number and range of assessors is adequate to do so but in an already crowded schedule for educators this may prove operationally difficult, particularly as institutions gradually progress to outcome-based education models, including CBME, which are heavily reliant on assessment approaches.

Recommendations for future research

One of the larger gaps in the literature we identified included a consistent definition and description of the *underperforming trainee* versus indicators of concern or specific *areas of underperformance*. It is important to determine how many concern ratings, or what patterns of WBA ratings may indicate trainees who are in difficulty. Validity studies, informed by contemporary understanding of the definition of the concept of validity (St-Onge and Young 2015) and the use of newer validity frameworks (Cook et al. 2014; Cook et al. 2015; Hatala et al. 2015) should also be considered in determining the value of WBA in identifying and/or remediating underperformance.

Strengths and limitations of the review

The main strengths of this review included the breadth of expertise and experience of our research team. We were fortunate to have a number of experienced BEME and Cochrane reviewers who brought significant methodological strength and rigor to the development and execution of the review process.

Consultation with three information scientists at various stages of the review also allowed us to ensure that our search strategy and searches were systematic, complete, thorough and rigorously documented. We performed quality checks at all appropriate stages of the review and given the international make-up of our team, the inclusion of our email "voting" and discussion process allowed us to ensure that the final set of review articles met our inclusion criteria.

The primary limitations of the review relate to the lack of empirical studies on the use of WBA in remediation of underperformance. While we can make some assertions as to the use of WBA in the detection of underperformance, evidence specifically related to remediation was a considerable gap that we identified. The variability of study designs and methodological approaches also limits our ability to provide definitive statements on how WBA is (or could be) used to maximize its impact on the detection of underperformance.

Conclusion

Evidence for the use of WBA in identifying and/or remediating underperformance among postgraduate medical trainees has not yet been established. While this is partly due to the quality and focus of studies already published, it appears that the question of how useful WBA is for this group of postgraduate medical trainees has not been addressed in general. We hope, however, that this review will be of use in designing focussed programs of research aiming to definitively determine the role and value of WBA for this specific postgraduate medical education group.

Acknowledgements

We wish to acknowledge the invaluable expertise of the information scientists who contributed to the review; Ms Elise Cogo, Mr Cathal Kerrigan and Ms Jane Burns. We also wish to thank our BEME Institutional Collaborating Centre and the BEME team for their guidance and advice throughout the review process.

Disclosure statement

The authors report no declarations of interest. No funding was sought for the conduct of this BEME review

References

Alves de Lima A, Conde D, Costabel J, Corso J, Vleuten C. 2013. A laboratory study on the reliability estimations of the mini-CEX. Adv Health Sci Educ Theory Pract. 18:5–13.

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

Barr H, D, Freeth M Hammick I. Koppel S, Reeves. 2000. Evaluations of interprofessional education; a United Kingdom review of health and social care. London: CAIPE/BERA.

Barrett A, Galvin R, Steinert Y, Scherpbier A, O'Shaughnessy A, Horgan M, Horsley T. 2015. A BEME (Best Evidence in Medical Education) systematic review of the use of workplace-based assessment in identifying and remediating poor performance among postgraduate medical trainees. Syst Rev. 4:1–6.

Black D, Welch J. 2009. The under-performing trainee – concerns and challenges for medical educators. ClinTeach. 6:79–82.

Bok H, Teunissen P, Favier R, Rietbroek N, Theyse L, Brommer H, Haarhuis J, van Beukelen P, van der Vleuten C, Jaarsma D. 2013. Programmatic assessment of competency-based workplace learning: when theory meets practice. BMC Med Educ.13:123.

Booth A. 2006. "Brimful of STARLITE": toward standards for reporting literature searches. J Med Libr Assoc. 94:421–e205.

Brown DJ, Thompson RE, Bhatti NI. 2008. Assessment of operative competency in otolaryngology residency: survey of US program directors. Laryngoscope. 118:1761–1764.

Buckley S, Coleman J, Davison I, Khan KS, Zamora J, Malick S, Morley D, Pollard D, Ashcroft T, Popovic C, et al. 2009. The educational effects of portfolios on undergraduate student learning: a Best Evidence Medical Education (BEME) systematic review. BEME Guide No. 11. Med Teach. 31:282–298.

Bullock AD, Hassell A, Markham WA, Wall DW, Whitehouse AB. 2009. How ratings vary by staff group in multi-source feedback assessment of junior doctors. Med Educ. 43:516–520.

Burford B, Illing J, Kergon C, Morrow G, Livingston M. 2010. User perceptions of multi-source feedback tools for junior doctors. Med Educ. 44:165–176.

CASP: Critical Appraisal Skills Programme [Internet]. 2013. [cited 2016 Jan 14]. Available from: http://www.casp-uk.net/#!checklists/cb36.

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

Cleland J, Mackenzie RK, Ross S, Sinclair HK, Lee AJ. 2010. A remedial intervention linked to a formative assessment is effective in terms of improving student performance in subsequent degree examinations. Med Teach. 32:e185–e190.

Cook D, Zendejas B, Hamstra S, Hatala R, Brydges R. 2014. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. Adv Health Sci Educ. 19:233–250.

Cook DA, Beckman TJ. 2008. Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX. Adv Health Sci Educ Theory Pract. 14:655–664.

Cook DA, Brydges R, Ginsburg S, Hatala R. 2015. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ. 49:560–575.

De Houwer J, Barnes-Holmes D, Moors A. 2013. What is learning? On the nature and merits of a functional definition of learning. Psychon Bull Rev. 20:631–642.

Donato AA, Park YS, George DL, Schwartz A, Yudkowsky R. 2015. Validity and feasibility of the Minicard Direct Observation Tool in 1 Training Program. J Grad Med Educ. 7:225–229.

Dudek NL, Marks MB, Regehr G. 2005. Failure to fail: the perspectives of clinical supervisors. Acad Med. 80:S84–S87.

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. BMC Med Educ. 14:76–76.

Gingerich A, Regehr G, Eva KW. 2011. Rater-based assessments as social judgments: rethinking the etiology of rater errors. Acad Med. 86:S1–S7.

Gingerich A, van der Vleuten CPM, Eva KW, Regehr G. 2014. More consensus than idiosyncrasy: categorizing social judgments to examine variability in mini-CEX ratings. Acad Med. 89:1510–1519.

Govaerts M. 2015. Workplace-based assessment and assessment for learning: threats to validity. J Grad Med Educ. 7:265–267.

Govaerts M, Schuwirth L, Van der Vleuten C, Muijtjens A. 2011. Workplace-based assessment: effects of rater expertise. Adv Health Sci Educ Theory Pract. 16:151–165.

Hassell A, Bullock A, Whitehouse A, Wood L, Jones P, Wall D. 2012. Effect of rating scales on scores given to junior doctors in multisource feedback. Postgrad Med J. 88:10–14.

Hatala R, Cook D, Brydges R, Hawkins R. 2015. Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): a systematic review of validity evidence. Adv Health Sci Educ:20:1149–1175.

Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. 2010. Constructing a validity argument for the mini-clinical evaluation exercise: a review of the research. Acad Med. 85:1453–1461.

Hesketh EA, Anderson F, Bagnall GM, Driver CP, Johnston DA, Marshall D, Needham G, Orr G, Walker K. 2005. Using a 360 degrees diagnostic screening tool to provide an evidence trail of junior doctor performance throughout their first postgraduate year. Med Teach. 27:219–233.

Hiemstra E, Kolkman W, Wolterbeek R, Trimbos B, Jansen FW. 2011. Value of an objective assessment tool in the operating room. Can J Surg. 54:116–122.

Kogan JR, Conforti LN, Bernabeo EC, Durning SJ, Hauer KE, Holmboe ES. 2012. Faculty staff perceptions of feedback to residents after direct observation of clinical skills. Med Educ. 46:201–215.

Kogan JR, Holmboe ES, Hauer KE. 2009. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA: JAm Med Assoc. 302:1316–1326.

Lo B. 2012. Informed consent. Ethical issues in clinical research: a practical guide. Philadelphia, PA: Lippincott Williams & Wilkins.

McHugh M. 2012. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 22:276–282.

Menon S, Winston M, Sullivan G. 2012. Workplace based assessment: attitudes and perceptions among consultant trainers and comparison with those of trainees. Psychiatrist. 36:16–24.

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

Mitchell C, Bhat S, Herbert A, Baker P. 2011. Workplace-based assessments of junior doctors: do scores predict training difficulties? Med Educ. 45:1190–1198.

Mitchell C, Bhat S, Herbert A, Baker P. 2013. Workplace-based assessments in Foundation Programme training: do trainees in difficulty use them differently? Med Educ. 47:292–300.

Moher D, Liberati A, Tetzlaff J, Altman DG. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. AnnIntern Med. 151:264–269.

Norcini J, Burch V. 2007. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Med Teach. 29:855–871.

Norcini JJ, Blank LL, Arnold GK, Kimball HR. 1995. The mini-CEX (Clinical Evaluation Exercise): a preliminary investigation. Ann InternMed. 123:795–799.

Overeem K, Wollersheim H, Driessen E, Lombarts K, Van De Ven G, Grol R, Arah O. 2009. Doctors' perceptions of why 360-degree feedback does (not) work: a qualitative study. Med Educ. 43:874–882.

Pelgrim EAM, Kramer AWM, Mokkink HGA, den Elsen L, Grol RPTM, Vleuten CPM. 2011. In-training assessment using direct observation of single-patient encounters: a literature review. Adv Health Sci Educ. 16:131–142.

Rees CE, Cleland JA, Dennis A, Kelly N, Mattick K, Monrouxe LV. 2014. Supervised learning events in the Foundation Programme: a UK-wide narrative interview study. BMJ Open. 4:e005980.

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. BMC Med Educ. 12:25.

Sargeant J, McNaughton E, Mercer S, Murphy D, Sullivan P, Bruce DA. 2011. Providing feedback: exploring a model (emotion, content, outcomes) for facilitating multisource feedback. Med Teach. 33:744–749.

St-Onge C, Young M. 2015. Evolving conceptualisations of validity: impact on the process and outcome of assessment. Med Educ. 49:548–550.

Steinert Y, Mann K, Centeno A, Dolmans D, Spencer J, Gelula M, Prideaux D. 2006. A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME Guide No. 8. Med Teach. 28:497–526.

Tong A, Palmer S, Craig JC, Strippoli GFM. 2014. A guide to reading and using systematic reviews of qualitative research. Nephrol Dial Transpl. doi: 10.1093/ndt/gfu354.

Tong A, Sainsbury P, Craig J. 2007. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care. 19:349–357.

Vivekananda-Schmidt P, MacKillop L, Crossley J, Wade W. 2013. Do assessor comments on a multi-source feedback instrument provide learner-centred feedback? Med Educ.47:1080–1088.

Warm EJ, Schauer D, Revis B, Boex JR. 2010. Multisource feedback in the ambulatory setting. J Grad Med Educ. 2:269–277.

Watling C. 2014. Cognition, culture, and credibility: deconstructing feedback in medical education. Perspect Med Educ. 3:124–128.

Watling C, Driessen E, van der Vleuten CPM, Lingard L. 2012. Learning from clinical work: the roles of learning cues and credibility judgements. Med Educ. 46:192–200.

Watling C, Driessen E, van der Vleuten CPM, Vanstone M, Lingard L. 2012. Understanding responses to feedback: the potential and limitations of regulatory focus theory. Med Educ. 46:593–603.

Watling C, Driessen E, van der Vleuten CPM, Vanstone M, Lingard L. 2013a. Beyond individualism: professional culture and its influence on feedback. Med Educ. 47:585–594.

Watling C, Driessen E, van der Vleuten CPM, Vanstone M, Lingard L. 2013b. Music lessons: revealing medicine's learning culture through a comparison with that of music. Med Educ. 47:842–850.

Weller JM, Jones A, Merry AF, Jolly B, Saunders D. 2009. Investigation of trainee and specialist reactions to the mini-clinical evaluation exercise in anaesthesia: implications for implementation. BJA: British J Anaesth. 103:524–530.

Whitehouse A, Hassell A, Bullock A, Wood L, Wall D. 2007. 360 degree assessment (multisource feedback) of UK trainee doctors: field testing of team assessment of behaviours (TAB). Med Teach. 29:171–176.

Wood L, Wall D, Bullock A, Hassell A, Whitehouse A, Campbell I. 2006. "Team observation": a six-year study of the development and use of multi-source feedback (360-degree assessment) in obstetrics and gynaecology training in the UK. Med Teach. 28:e177–e184.

Yao DC, Wright SM. 2000. National survey of internal medicine residency program directors regarding problem residents. JAMA. 284:1099–1104.

Yeates P, O'Neill P, Mann K, Eva KW. 2013a. Effect of exposure to good vs poor medical trainee performance on attending physician ratings of subsequent performances. JAMA. 30:2226–2232. [Erratum appears in JAMA. 2013 Jan 16;309(3):237].

Yeates P, O'Neill P, Mann K, Eva K. 2013b. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Health Sci Educ Theory Pract. 18:325–341.

Young JQ, Lieu S, O'Sullivan P, Tong L. 2011. Development and initial testing of a structured clinical observation tool to assess pharmacotherapy competence. Acad Psychiatr. 35(1):27–34.