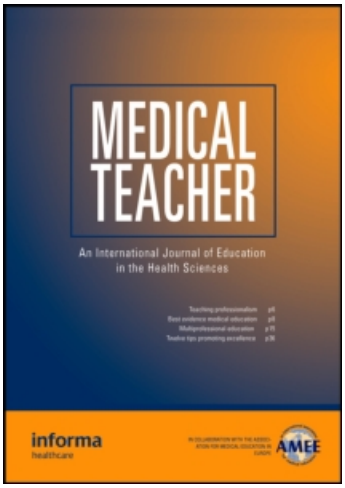

Medical Teacher
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713438241

Using the 360° multisource feedback model to evaluate teaching and professionalism

Ronald A. Berk [a]
[a] The Johns Hopkins University, USA

Online publication date: 07 December 2009

To cite this Article Berk, Ronald A.(2009) 'Using the 360° multisource feedback model to evaluate teaching and professionalism', Medical Teacher, 31: 12, 1073 — 1080
To link to this Article: DOI: 10.3109/01421590802572775
URL: http://dx.doi.org/10.3109/01421590802572775

2009; 31: 1073–1080 MEDICAL TEACHER

Using the 360° multisource feedback model to evaluate teaching and professionalism

RONALD A. BERK
The Johns Hopkins University, USA

Abstract

Background: Student ratings have dominated as the primary and, frequently, only measure of teaching performance at colleges and universities for the past 50 years. Recently, there has been a trend toward augmenting those ratings with other data sources to broaden and deepen the evidence base. The 360° multisource feedback (MSF) model used in management and industry for half a century and in clinical medicine for the last decade seemed like a best fit to evaluate teaching performance and professionalism.

Aim: To adapt the 360° MSF model to the assessment of teaching performance and professionalism of medical school faculty.

Methods: The salient characteristics of the MSF models in industry and medicine were extracted from the literature. These characteristics along with 14 sources of evidence from eight possible raters, including students, self, peers, outside experts, mentors, alumni, employers, and administrators, based on the research in higher education were adapted to formative and summative decisions.

Results: Three 360° MSF models were generated for three different decisions: (1) formative decisions and feedback about teaching improvement; (2) summative decisions and feedback for merit pay and contract renewal; and (3) formative decisions and feedback about professional behaviors in the academic setting. The characteristics of each model were listed. Finally, a top-10 list of the most persistent and, perhaps, intractable psychometric issues in executing these models was suggested to guide future research.

Conclusions: The 360° MSF model appears to be a useful framework for implementing a multisource evaluation of faculty teaching performance and professionalism in medical schools. This model can provide more accurate, reliable, fair, and equitable decisions than the one based on just a single source.

Introduction

More than 15,000 studies have been published on the topic of teaching effectiveness. It is the major criterion (98%) in assessing overall faculty performance in US colleges when compared to research (41%) and publications (31%) (Seldin 1999). Unfortunately, student ratings have dominated as the primary and, frequently, only measure of teaching performance for the last 50 years. In surveys over the past decade, it was found that 86% of US liberal arts college deans and 97% of department chairs (US Department of Education 1991) use student ratings for summative decisions. It is likely that the percentages in medical schools may be even higher.

Recently, there has been a trend toward using multiple sources of evidence for formative and summative decisions about the teaching behaviors of faculty. This article will examine this trend and proffer the 360° multisource feedback (MSF) model to evaluate teaching and professionalism. These models will be grounded in the characteristics, research, and practices of the MSF models in management/industry and clinical medicine. The first section presents the rationale for this approach to evaluate teaching.

Practice points

- The 360° MSF model can be configured for medical school professors to obtain feedback for formative and summative decisions on teaching performance.
- The characteristics of the formative and summative models are distinctly different based on the specific decisions being made.
- The MSF model can be adapted to provide feedback on professional behaviors of faculty in the academic as well as clinical setting.
- Several psychometric issues related to the construction, reliability, and validity of the scales still require research attention.
- Timely standardized administration, processing, interpretation, and feedback are the most significant practical concerns.

Rationale for multisource assessment

Several arguments have been given for considering multiple sources of evidence in making personnel decisions about

Correspondence: R. A. Berk, 10971 Swansfield Rd, Columbia, MD 21044-2726, USA. Tel: 410 730 9339, 410-940-7118; fax: 888 410 8089; email: rberk@son.jhmi.edu

ISSN 0142–159X print/ISSN 1466–187X online/09/121073–8 © 2009 Informa Healthcare Ltd.
DOI: 10.1080/01421590802572775

employees. There are three in particular that are most pertinent to evaluating faculty: (1) psychometric limitations of top-down model, (2) fallibility of all sources of evidence, and (3) expert recommendations.

Psychometric limitations of top-down model

Traditionally, employees in most organizations have been evaluated using the 'top-down' supervision model. The immediate supervisor rates the performance of the employee at the time of the scheduled performance appraisal. That's it. There is one evaluation by one person. There are three issues that have plagued this approach: (1) the potential unfairness and bias of one rater; (2) the invalidity of a single set of ratings which may be based on an incomplete assessment of all relevant job skills, no direct observation of job performance, and/or a reliance on secondary information; and (3) the unreliability of a single source of evidence. This approach and those issues would apply to the case where a department chair, dean, or associate dean alone conducted the evaluation of a professor.

Fallibility of all sources of evidence

The bias, invalidity, and unreliability issues of the single source can also be applied to other sources of evidence. Put simply: there is no perfect source or combination of sources of evidence. In fact, almost all sources are based on human judgment in the form of either the individual or collective opinions of other people. 'There is no known objective method for measuring teaching performance'.

Each source of evidence can supply unique information, but also is fallible in one or more ways, usually different from the other sources. For example, the unreliability and biases of peer ratings are not the same as those of student ratings; student ratings have other weaknesses. By triangulating three or more different sources of evidence, the 'strengths of each source can compensate for weaknesses of the other sources', thereby converging on a decision about teaching effectiveness that is more accurate, reliable, fair, and equitable than the one based on any single source (Appling et al. 2001). This strategy is derived from a compensatory model of decision making.

Given the complexity of measuring the act of teaching in a real-time classroom environment, the use of multiple sources of 'informed judgment' is a sound, defensible strategy. The decision maker should integrate the information from only those sources for which validity evidence is available. The quality of the sources chosen should be beyond reproach.

Expert recommendations

Historically, student ratings have dominated as the primary measure of teaching effectiveness for the past five decades (Seldin 1999). In fact, the evaluation of teaching has been in a metaphorical cul-de-sac with student ratings as the universal barometer. Only recently has there been a trend toward augmenting those ratings with other data sources to broaden and deepen the evidence base (Centra 1993; Braskamp & Ory 1994; Knapper & Cranton 2001; Berk 2006; Seldin 2006; Arreola 2007).

Multisource feedback models

Rather than adopt an existing model or propose a new one, I have chosen to adapt a time-tested, industry standard with which many medical educators may already be familiar: the 360° MSF model. It was developed in 'management and industry' more than a half-century ago and applied to clinical medicine a decade ago. These applications are described next. The extension of the MSF model to 'medical faculty evaluation' will follow.

Management/industry

Description. The 360° MSF method has its origin and most frequent applications in the corporate world. An employee's 'job behaviors' and 'outcomes' are rated anonymously by persons who are most knowledgeable about his or her work – those hierarchically above, below, and on the same level as the employee – to furnish different perspectives (Edwards & Ewen 1996). This approach taps their collective wisdom to provide a more balanced, complete, accurate, and fair assessment than the traditional single-source, top-down, supervisor-only method. The 'ratings are compared to self-ratings to give precise feedback to the employee' so he or she can plan specific improvements in his or her job performance (formative) in order to meet career goals. In some cases, the results may be used by the supervisor for promotion and pay-raise decisions (summative). The ratings supplement the available information for supervisory decisions.

The 'employee is the hub' of the ratings and the raters may consist of his or her supervisor (above), co-workers (same level), internal and external customers (below), self, and others, as displayed in Figure 1.

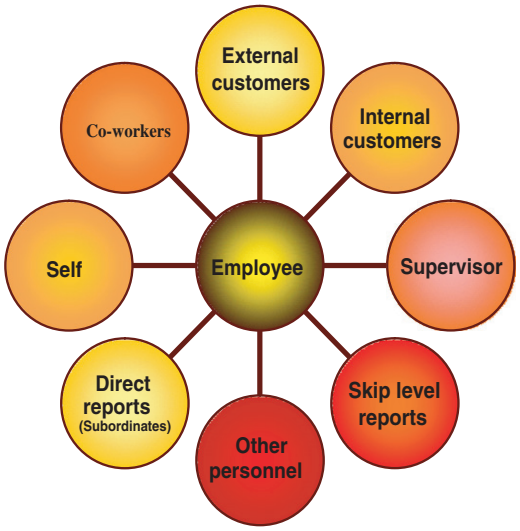

Figure 1. A 360° MSF assessment of an employee (adapted from Edwards and Ewen 1996, Figure 1-1, p. 8).

History. The 360° assessment is not a recent development. It has its roots in the military assessment centers developed during World War II and the US military academies' use of a multisource process called 'peer grease', which was designed to evaluate the leadership skills of students in the 1950s and 1960s (Edwards & Ewen 1996; Fleenor & Prince 1997). Private industry began to experiment with the multisource approach in the 1960s and 1970s. Corporations such as Federal Express, Bank of America, Bell Labs, and Disney World used the method in job evaluations, hiring and internal selection panels, promotion boards, and talent selection (Boyd 2005). By the mid-1980s the 360° assessment was being employed for performance appraisals in organizations like Fidelity Bank, Monsanto, and Dupont.

It was not until the early 1990s that the strategy gained wide acceptance for formative feedback and summative appraisal decisions. Now it is estimated that over 90% of Fortune 1000 firms use MSF systems to evaluate employees (Boyd 2005). The response in the public sector has been much more limited (Ghorpade 2000).

Characteristics. A few of the most important characteristics of the 360° MSF reported in the literature are as follows:

(1) Employee should be involved in the selection of raters;
(2) Raters should be credible and knowledgeable of employee's behaviors;
(3) Behaviors and outcomes rated should relate to actual job tasks and expectations;
(4) Moderate size sample of raters (4–12) should be used to preserve anonymity and increase reliability;
(5) A common scale should be completed by all raters;
(6) Likert-type scales with 4–7 options, such as 'strongly agree–strongly disagree' or 'always–never', should be constructed properly;
(7) Scales should be administered online rather than on paper to preserve anonymity, increase response rate, and increase quality and quantity of comments;
(8) Feedback should be sensitive, timely, face-to-face, and regular; and
(9) Improvements in performance should be documented over time.

Resources and research. There are several books (Bracken et al. 2001; Edwards & Ewen 1996; Jones 1996; Lepsinger & Lucia 1997; Tornow & London 1998; Waldman & Atwater 1998) on how to design and implement the 360° approach, evaluate feedback instruments (van Velsor et al. 1997), and guide best practices (Gray et al. 2003). There are also research reviews in narrative (McCarthy & Garavan 2001; Seifert et al. 2003; Smither et al. 2005) and meta-analysis (Conway & Huffcutt 1997; Smither et al. 2005) forms that can guide the development and implementation of a multisource assessment in public and private industrial settings. The research continues to accumulate on several psychometric issues such as the equivalence of multisource ratings across a variety of subordinate positions (Diefendorff et al. 2005; Gillespie 2005) and factors that influence employees' intentions to provide honest upward feedback (Smith & Fortunato 2008).

If the primary purposes of the 360° MSF model are to provide meaningful feedback to increase self-awareness and to motivate the employee toward self-improvement and growth, the latest research indicates that some recipients of that feedback will be more likely to improve than others. The magnitude of those improvements may be very small and across-the-board performance improvement is unrealistic (Smither et al. 2005). Further, it is critical to account for the type of organization and the culture of the organization before introducing the 360° process (Brutus et al. 1998). Openness, mutual trust, and honesty, plus a genuine interest in and desire for performance improvement must exist for the 360° MSF model to be successful.

Clinical medicine

Description. The 360° MSF approach has been applied differently in medicine. It has been used by physician-licensing boards, medical schools, and hospitals for quality control and improvement of healthcare delivery, and to identify poorly performing physicians, beginning with medical students (Tyler 2006). Most of these applications have occurred within the past decade.

The model has been used to assess the performance of residents (Woolliscroft et al. 1994; Johnson & Cujec 1998; Allerup et al. 2007; Davis 2002; Joshi et al. 2004; Wood et al. 2004) and senior licensed physicians (Wenrich et al. 1993; Violato et al. 1997; Hall et al. 1999; Lipner et al. 2002). It has also been applied to most specialties, including anesthesiology (Lockyer & Violato 2006), emergency medicine (Lockyer et al. 2006), radiology (Wood et al. 2004), obstetrics and gynecology (Risucci et al. 1989; Davis 2002), surgery (Risucci et al. 1989; Violato et al. 2003), general practice (Griffin et al. 2000; Murphy et al. 2008), family practice (Sargeant et al. 2005), internal medicine (Wenrich et al. 1993; Woolliscroft et al. 1994; Allerup et al. 2007), and pediatrics (Archer et al. 2005).

Figure 2 depicts the medical analog to the management model shown previously. Here the 'physician is the hub'.

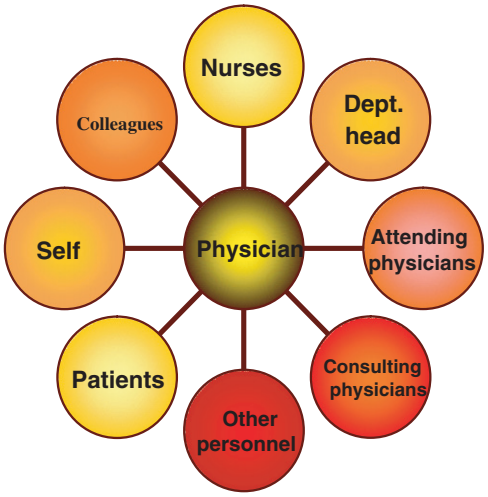

Figure 2. A 360° MSF assessment of a physician (adapted from Berk 2006, Figure 1.2, p. 41).

You can easily guess into which category (above, same level, and below) each bubble rater would fall.

Characteristics. There are several striking differences between the management and medical applications:

(1) Large sample of raters (10–30) is often clustered into groups, such as five nurses, four patients, three colleagues, plus self;
(2) Rating scales focus on a variety of competencies, including medical knowledge, teaching, clinical care, communication skills, and management/administrative skills, with a particular emphasis on interpersonal skills with patients and colleagues and professionalism (Joshi et al. 2004; Wood et al. 2004);
(3) Number and type of anchor responses vary from 3 to 7 options, such as 'major concern—no concern', 'excellent—unacceptable', 'above expectations—below expectations', and 'among the best—among the worst';
(4) Sometimes a common scale may be used with minor modifications for all sources, but more frequently, different scales are given to different categories of raters (Wenrich et al. 1993; Woolliscroft et al. 1994; Hall et al. 1999; Lockyer & Violato 2004; Allerup et al. 2007);
(5) Online scale administrations have encountered a range of response rates by institution and country compared to paper-based administrations; and
(6) Proper and timely feedback is critical to the formative assessment process (Norcini & Birch 2007).

Research. A corpus of research is accumulating with a wide range of applications in the US, Canada, the UK, Denmark, and numerous other countries. Lockyer and Clyman (2008) provide an overview of MSF and descriptions of various applications and uses in clinical settings. Reviews have concentrated on the medical uses of the MSF (Lockyer 2003), peer ratings (Ramsey et al. 1993), peer-rating instruments (Evans et al. 2004), and the effects of feedback on performance (Veloski et al. 2006). Studies have compared the different sources for rating residents (Woolliscroft et al. 1993; Johnson & Cujec 1998; Risucci et al. 1989; Davis 2002; Wood et al. 2004) and physicians (Ramsey et al. 1993; Violato et al. 1997; Lipner et al. 2002; Lockyer et al. 2007) and the influence of emotions upon feedback acceptance and use (Sargeant et al. 2006).

The research indicates significant progress in these areas as experience with different applications of the 360° MSF increases. However, there is considerable room for improvement in the definition of the behaviors measured, the quality of the rating scales, the systematic and standardized administration of the ratings by professionals with different levels of skill and motivation, the process by which raters assess the knowledge, skills, and abilities (KSAs) and, especially, the interpersonal skills, humanistic qualities, and professionalism of physicians, and the feedback process. These problems are attributable, in part, to the complexity of operating a 360° assessment with multiple raters using different scales in uncontrolled, real-time environments.

Medical faculty evaluation

Critical reviews of strategies to evaluate teaching behaviors in the higher education literature suggest a variety of possible raters, including students, self, peers, outside experts, mentors, alumni, employers, and administrators, and 14 different potential sources of evidence (Berk 2005, 2006). They are all applicable to medical faculty evaluation. These sources are as follows: (1) student ratings, (2) peer ratings, (3) external expert ratings, (4) self-ratings, (5) videos, (6) student interviews, (7) alumni ratings, (8) employer ratings, (9) mentor's advice, (10) administrator ratings, (11) teaching scholarship, (12) teaching awards, (13) learning outcome measures, and (14) teaching portfolio.

Description. The 360° MSF model hinges on the specific decisions to be made. In faculty evaluation, there are three applications of the model that can be created for formative and summative decisions about teaching behaviors and formative decisions about professional behaviors. Let's consider how the raters and sources above can fit into these different models.

The 'professor is the hub' in all applications. The raters/sources vary for each decision. These may be different for each department and institution. To illustrate the models here, a 'best practices' combination of raters/sources, based on the previous reviews, is recommended for each decision. They may not be the most appropriate choices for every institution.

Formative decisions about teaching behaviors. Among the 14 strategies identified previously, which ones would you select to improve your teaching? Which ones provide the most accurate information to pinpoint your strengths, weaknesses, and suggestions on how to improve?

Assuming the scales are properly constructed with adequate evidence of reliability and validity, five of the best sources you could use are teacher mentor (above), peer ratings and video with self/peer (same level), student ratings and student interviews (below), plus self-ratings. Different scales would be given to the mentor, peer, and students. The professor would also complete those scales. Discrepancies between his or her (self) ratings and those of the other three raters can yield a profile of strengths and weaknesses to pinpoint specific classroom behaviors needing attention. The 360° MSF model with these six sources of evidence is shown in Figure 3.

The characteristics of this model are as follows:

(1) Professor selects raters and sources of evidence;
(2) Raters should be knowledgeable of professor's teaching behaviors;
(3) Teaching behaviors defined for each source may be different and complementary, although there may be intentional overlap between subscales of the student and peer scales;
(4) Sample of raters may be large, including students, one or more peers, and a mentor;
(5) Different Likert-type scales for the different types of raters that measure different faculty behaviors are developed, such as 'student rating scales', which concentrate on teaching behaviors related to content and organization, learning outcomes, instructional

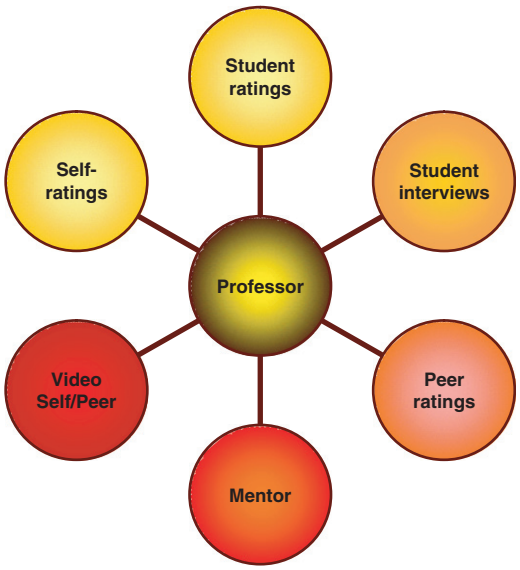

Figure 3. A 360° MSF assessment of a professor (formative decisions about teaching).

methods, learning atmosphere, and evaluation methods; 'clinical scales', which measure instructional methods and clinical experiences; and 'peer observation scales', which focus on content and organization, communication style, questioning skills, critical thinking skills, rapport with students, learning environment, and teaching methods;

(6) Quality of many homegrown scales varies from very good to poor, but commercial student rating scales are better;

(7) Number of points and anchors vary on every scale, for example, 4–7-point 'strongly disagree–strongly agree' and 'always–never' scales or 4-point 'excellent–needs improvement' scale;

(8) Administration of the student rating scales are online and paper-and-pencil, with an increase in the former at many institutions with U.S. response rates from 30% to 90%, but all other scales are typically administered in paper-and-pencil form; the data from the different sources may be collected at different times during the semester;

(9) Feedback to faculty member on student ratings can be less than 2 weeks (online) or more than a month (paper-and-pencil), immediate from student interviews, or within days and face-to-face by peers and mentor; and

(10) Professor tracks changes and progress in teaching improvement from the different sources of evidence and documents improvements.

Summative decisions about teaching behaviors. Drawing on the same pool of 14 strategies listed earlier, which ones would you pick for your department chair, dean, or associate dean to use to determine your annual merit pay or whether your contract should be renewed (for non-tenured full- and part-time faculty)? Again which sources provide accurate information on teaching effectiveness, but also collectively converge on a decision that is fair and equitable? After all, your teaching career is on the line.

Interestingly, the 'best' sources in this application are almost identical to the ones chosen for formative decisions. They include department chair and mentor ratings (above), peer ratings (optional) and video (optional) (same level), student ratings (below), plus self-ratings. The use of peer ratings, video performance and feedback, and mentor ratings should be determined at the discretion of the faculty member; otherwise, it would be a breach of confidentiality (Berk et al. 2004). That evidence was originally collected for formative use in confidence with the respective peer or mentor. Summative decisions are different. Either the intent of the information for both types of decisions is agreed upon in advance or, if not, the final use of any of those sources should rest with the faculty member.

The characteristics of this model are very different from the formative model, although the sources may be the same:

(1) Faculty and administrator(s) determine the raters and sources of evidence;

(2) Multiple raters should be chosen for their expertise and to minimize several types of rating bias (Berk 2006);

(3) Teaching behaviours defined for each source may be different;

(4) Large sample of raters may include all students, one or more peers agreed upon by the professor and the administrator, the mentor, the professor (self), and the administrator;

(5) Separate peer observation reports, different in content and scope from the one shared with the professor, may be prepared by the raters and submitted to the department chair (Berk et al. 2004);

(6) Feedback to the professor by department chair on overall performance drawn from all of the data sources occurs face-to-face at the end of the year; and

(7) Department chair tracks changes and progress in teaching longitudinally and documents improvements; determines merit pay and contract renewal; evidence may also be included in teaching portfolio (dossier) for promotion and tenure review.

Formative decisions about professional behaviors. Professionalism has been included in many of the 360° MSF clinical assessments of medical students, interns, residents, and licensed physicians (Wood et al. 2004). The professional behaviors measured in those models relate primarily to the clinical environments in which the medical personnel work (Berk 2009). In fact, the US National Board of Medical Examiners has developed a list of 59 behaviors (http://professionalbehaviors.nbme.org/2008list of behaviors.pdf) and is currently conducting field trials in collaboration with medical schools and residency programs to test their instrument in the context of an MSF program (http://professionalbehaviors.nbme.org/guide.pdf). That model can be extended to professional behaviors as a faculty member in the school of medicine.

In this formative decision application, the professor is again the hub of the rating wheel. A large sample of raters is chosen

by the faculty and department chair, which, in this case, is different and more diverse than in the preceding models. In addition to the previous sources of students (below), peers (same level), and department chair (above), this model might also include colleagues and students in the roles of mentees, research associates (RAs), teaching assistants (TAs), and lab assistants (LAs), plus administrative assistants (AAs), IT staff, admissions personnel, and other staff.

You might have noticed that in the academic food chain most of these raters are below the professor. These individuals often consider themselves most vulnerable and, in fact, may be the object of relational problems, such as shouting, harsh words, rudeness, ridicule, mean and nasty comments, or underhanded, passive aggressive, or bullying behaviors by faculty members. Unfortunately, such behaviors in the classroom and work environment are on the rise in the US under the rubric of 'faculty incivility' (Twale & DeLuca 2008). None of the aforementioned persons deserves to be humiliated, embarrassed, undermined, insulted, belittled, put down, shunned, or marginalized by a professor; yet many are.

As universities throughout the US are adopting codes of civility (Forni 2002) for the classroom behaviors of students, it now seems necessary to set guidelines for faculty and to hold them accountable so students have appropriate role models to emulate.

Categories of professional behaviors of professors would include the following: emotional intelligences of intrapersonal and interpersonal skills, team work, communication, accessibility, responsibility, altruism, honor, integrity, honesty, trust, respect, caring, patience, and compassion.

Character dimensions or attributes would also fit under the domain of professionalism: leadership, excellence, creativity, motivation, diversity, values, aspirations, self-confidence, and initiative. None of these behaviors is usually measured by student ratings or any other source. All of these traits can also be assessed for administrators and other academic personnel.

A persuasive case could be made for any combination of the preceding behaviors to be part of the formative 360° MSF assessment of faculty. Although I am unaware of any such assessments of faculty or administrators in the US, the potential is clear and the need is justified. Medical school faculties and administrators should give serious consideration to measuring these behaviors as part of a comprehensive assessment program.

Appropriate scales could be developed for the different raters, but a scale with a common core of behaviors, which multiple raters are able to observe, may be part of the measurement process. A separate subscale on the student rating scale can address these behaviors too. To preserve anonymity and confidentiality, these scales should be administered online twice a year (or once a semester) at standardized times. The professor (self) should also complete the scale.

Once all analyses have been conducted and reported to the department chair, meetings with individual faculty members should be scheduled to provide prompt face-to-face formative feedback on the positive behaviors and trends across the multiple raters on negative behaviors. These ratings can also be contrasted with self-ratings and the department chair's ratings to identify discrepancies. An action plan should then be developed to address the negative behaviors and track improvements over the months that follow until the next semi-annual administration.

Conclusions

It seems as though we have come full circle several times to extend the 360° MSF model to evaluating teaching and professionalism. I am getting dizzy. This article provided an overview of the salient characteristics, research, and practices of the 360° MSF models in management/industry and clinical medicine. Drawing on that foundation, the model was adapted to the specific decisions rendered to evaluate medical school faculty teaching performance and professional behaviors.

What remains unchanged in every application is the original spirit of the model and its primary function:

Multisource ratings → Quality feedback → Action plan to improve→Improved performance

Although the ratings were intended for formative decisions, in many cases they have also ended up being used for summative decisions.

All of these applications of the 360° MSF model have advantages and disadvantages. In fact, it is possible to distill several persistent and, perhaps, intractable psychometric issues in executing these models. Here are the top 10 issues that deserve attention:

(1) There is no objective measure of teaching performance; all sources of evidence are fallible;
(2) Almost all evidence of performance is derived from the 'informed' judgments of students and those persons in the department with whom a professor works;
(3) There is difficulty in defining the domains of 'job' and 'professional' behaviors and in obtaining consensus on what each scale should measure;
(4) The quality of many homegrown scales is inadequate in terms of reliability and validity evidence for the decisions for which they are used;
(5) Peer observation scales and protocols are particularly weak, even though inter-rater reliabilities are moderate to high in many applications;
(6) Nine sources of rater bias are challenging to eliminate or even minimize;
(7) Standardization of scale administration in online or paper-and-pencil mode is often inadequate to ensure correct score interpretations;
(8) Response rates for online scale administrations are variable and unacceptably low at some institutions (e.g., 30–50%);
(9) Clear, understandable, meaningful, and appropriate report forms for rating results are needed for faculty and administrators; and
(10) Guidelines and training to interpret the results from multiple raters of performance accurately, fairly, and equitably and to provide sensitive and appropriate feedback are essential for faculty and administrators.

All of these areas require continued research. Although much has been learned during the 80-year history of scaling,

60-year history of faculty evaluation, and 50-year history of the 360° MSF model in management/industry, a lot of work is still necessary to realize the true meaning of 'best practices' in evaluating teaching and professionalism.

Declaration of interest: The author reports no conflicts of interest. The author alone is responsible for the content and writing of this article.

Notes on contributors

RONALD A. BERK, PhD, is professor emeritus, biostatistics and measurement, and former assistant dean for teaching at The Jons Hopkins University, where he served for 30 years. He has published 11 books and 130 journal articles/book chapters and has presented 80 keynotes and 200 workshops on teaching and faculty evaluation.

References

Allerup P, Aspegren K, Ejlersen E, Jørgesen G, Malchow-Møller A, Møller MK, Pedersen KK, Rasmussen OB, Rohold A, Sørensen B. 2007. Use of 360-degree assessment of residents in internal medicine in a Danish setting: A feasibility study. Med Teach 29:166–170.

Appling SE, Naumann PL, Berk RA. 2001. Using a faculty evaluation triad to achieve evidence-based teaching. Nurs Health Care Perspect 22:247–251.

Archer JC, Norcini J, Davies H. 2005. Use of SPRAT for peer review of paediatricians in training. Brit Med J 330:1251–1253.

Arreola RA. 2007. Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system. 3rd ed. Bolton, MA: Anker.

Berk RA. 2005. Survey of 12 strategies to measure teaching effectiveness. Int J Teach Learn Higher Educ 17(1):48–62. (http:/www.isetl.org/ijtlthe).

Berk RA. 2006. Thirteen strategies to measure college teaching: A consumer's guide to rating scale construction, assessment, and decision making for faculty, administrators, and clinicians. Sterling, VA: Stylus.

Berk RA. 2009. Derogatory and cynical humor in clinical teaching: A case for professionalism. Med Educ 43:7–9.

Berk RA, Naumann PL, Appling SE. 2004. Beyond student ratings: Peer observation of classroom and clinical teaching. Int J Nurs Educ Scholarsh 1(1):1–26.

Boyd NM. 2005. 360-Degree performance appraisal systems. In: Rabin J, editor. Encyclopedia of public administration and public policy. Oxford, England: Taylor & Francis.

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

Braskamp LA, Ory JC. 1994. Assessing faculty work. San Francisco: Jossey-Bass.

Brutus S, Fleenor JW, London M. 1998. Does 360-degree feedback work in different industries? A between-industry comparison of the reliability and validity of multi-source performance ratings. J Manage Dev 17(3):177–190.

Centra JA. 1993. Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness. San Francisco: Jossey-Bass.

Conway JM, Huffcutt AI. 1997. Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. Hum Perform 10:331–360.

Davis J. 2002. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. Obstet Gynecol 99:647–651.

Diefendorff JM, Silverman SB, Greguras GJ. 2005. Measurement equivalence and multisource ratings for non-managerial positions: Recommendations for research and practice. J Bus Psychol 19(3):399–425.

Edwards MR, Ewen AJ. 1996. 360° Feedback: The powerful new model for employee assessment and performance improvement. New York: American Management Association (AMACOM).

Evans R, Elwyn G, Edwards A. 2004. Review of instruments for peer assessment of physicians. Brit Med J 328:1240–1245.

Forni PM. 2002. Choosing civility: The 25 rules of considerate conduct. New York: St. Martin's Press.

Ghorpade J. 2000. Managing five paradoxes of 360-degree feedback. Acad Manage Exec 14(1):140–150.

Gillespie TL. 2005. Internationalizing 360-degree feedback: Are subordinate ratings comparable? J Bus Psychol 19(3):361–382.

Gray A, Lewis A, Fletcher C, Burke E, Mackay J, Kubelius E, Lindley P. 2003. 360 degree feedback: Best practices guidelines (http://www.psychtesting.org.uk/).

Griffin E, Sanders G, Craven D, King J. 2000. A computerized 360° feedback tool for personal and organizational development in general practice. Health Informatics J 6:71–80.

Hall W, Violato C, Lewkonia R, Lockyer JM, Fidler H, Toews J, Jennett P, Donoff M, Moores D. 1999. Assessment of physician performance in Alberta: The physician achievement review. Can Med Assoc J 161:52–57.

Johnson D, Cujec B. 1998. Comparison of self, nurse and physician assessment of residents rotating through an intensive care unit. Crit Care Med 26:1811–1816.

Jones JE. 1996. 360-Degree feedback: Strategies, tactics, and techniques for developing leaders. Amherst, MA: Human Resource Development Press.

Joshi R, Ling FW, Jaeger J. 2004. Assessment of a 360-degree instrument to evaluate residents' competency in interpersonal and communication skills. Acad Med 79:458–463.

Knapper C, Cranton P, editors. 2001. Fresh approaches to the evaluation of teaching (New Directions for Teaching and Learning, No. 88). Jossey-Bass: San Francisco.

Lepsinger R, Lucia AD. 1997. The art and science of 360 degree feedback. San Francisco: Pfeiffer.

Lipner RS, Blank LL, Leas BF, Fortna GS. 2002. The value of patient and peer ratings in recertification. Acad Med 77(10S):64S–66S.

Lockyer JM. 2003. Multi source feedback in the assessment of physician competencies. J Cont Educ Health Prof 23(1):4–12.

Lockyer JM, Clyman S. 2008. Multi source feedback. In: ES Holmboe, RE Hawkins, editors. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby/Elsevier. pp 75–85.

Lockyer JM, Violato C. 2004. An examination of the appropriateness of using a common peer assessment instrument to assess physician skills across specialties. Acad Med 79:5S–8S.

Lockyer JM, Violato C. 2006. A multi source feedback program for anesthesiologists. Can J Anesth 53:33–39.

Lockyer JM, Violato C, Fidler H. 2006. The assessment of emergency physicians by a regulatory authority. Acad Emerg Med 13:1296–1303.

Lockyer JM, Violato C, Fidler H. 2007. What multisource feedback factors influence physician self-assessments? A five-year longitudinal study. Acad Med 82:577–580.

McCarthy AM, Garavan TN. 2001. 360 degree feedback processes: Performance improvement and employee career development. J Eur Ind Training 25(1):3–32.

Murphy DJ, Bruce DA, Mercer SW, Eva KW. 2008. The reliability of workplace-based assessment in postgraduate medical education and training: A national evaluation in general practice in the United Kingdom. Adv Health Sci Educ 10:1007/s10459-008-9104-8.

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

Ramsey P, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. 1993. Use of peer ratings to evaluate physician performance. J Am Med Assoc 269:1655–1660.

Risucci DA, Tortolani AJ, Ward RJ. 1989. Ratings of surgical residents by self, supervisors and peers. Surg Gynecol Obstet 169(5):519–526.

Sargeant JM, Mann KV, Ferrier S. 2005. Understanding family physicians' reactions to MSF performance assessment: Perceptions of credibility and usefulness. Med Educ 39:497–504.

Sargeant JM, Mann KV, Sinclair D, van der Vleuten C, Metsemakers J. 2006. Understanding the influence of emotions and reflections upon multi-source feedback acceptance and use. Adv Health Sci Educ 10:1007/s10459-006-9039-x.

Seifert CF, Yukl G, McDonald RA. 2003. Effects of multisource feedback and a feedback facilitator on the influence behavior of managers toward subordinates. J Appl Psychol 88:561–569.

Seldin P. 1999. Current practices – good and bad – nationally. In: Seldin P, editor. Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions. Bolton, MA: Anker. pp 1–24.

Seldin P, editor. 2006. Evaluating faculty performance: A practical guide to assessing teaching, research, and service. Bolton, MA: Anker.

Smith AFR, Fortunato VJ. 2008. Factors influencing employee intentions to provide honest upward feedback ratings. J Bus Psychol 22(3):191–207.

Smither JW, London M, Reilly RR. 2005. Does performance improve following multisource feedback? A theoretical model, meta-analysis, and review of empirical findings. Personnel Psychol 58:33–66.

Tornow WW, London M. 1998. Maximizing the value of 360-degree feedback: A process for successful individual and organizational development. San Francisco: Jossey-Bass.

Twale DJ, DeLuca BM. 2008. Faculty incivility: The rise of the academic bully culture and what to do about it. San Francisco: Jossey-Bass.

Tyler KM. 2006. Peer-level multiple source feedback for fitness to practice. Med Educ 40:459–489.

US Department of Education. 1991. Assessing teaching performance. The Department Chair: A Newsletter for Academic Administrators 2(3):2, winter.

van Velsor E, Leslie JB, Fleenor JW. 1997. Choosing 360: A guide to evaluating multi-rater feedback instruments for management development. Greensboro, NC: Center for Creative Leadership.

Veloski J, Boex JR, Grasberger MJ, Evans A, Wolfson DB. 2006. Systematic review of the literature on assessment, feedback, and physicians' clinical performance 1: BEME Guide No. 7. Med Teach 28:117–128.

Vinson MN. 1996. The pros and cons of 360-degree feedback: Making it work. Train Dev 50(4):11–12.

Violato C, Lockyer JM, Fidler H. 2003. Multi source feedback: A method of assessing surgical practice. Brit Med J 326:546–548.

Violato C, Marini A, Toews J, Lockyer JM, Fidler H. 1997. Feasibility and psychometric properties of using peers, consulting physicians, co-workers, and patients to assess physicians. Acad Med 72:82S–84S.

Waldman D, Atwater LE. 1998. Power of 360 degree feedback: How to leverage performance evaluations for top productivity. Burlington, MA: Gulf Professional Publishing.

Wenrich MD, Carline JD, Giles LM, Ramsey PG. 1993. Ratings of the performances of practicing internists by hospital-based registered nurses. Acad Med 68:680–687.

Wood J, Collins J, Burnside ES, Albanese MA, Propeck PA, Kelcz F, Spilde JM, Schmaltz LM. 2004. Patient, faculty, and self-assessment of radiology resident performance: A 360-degree method of measuring professionalism and interpersonal/communication skills. Acad Radiol 11:931–939.