Evaluating the Evaluators: Implementation of a Multi-Source Evaluation Program for Graduate Medical Education Program Directors

Mary Ellen J. Goldhamer, MD, MPH
Keith Baker, MD, PhD
Amy P. Cohen, EdM
Debra F. Weinstein, MD

ABSTRACT

Background Multi-source evaluation has demonstrated value for trainees, but is not generally provided to residency or fellowship program directors (PDs).

Objective To develop, implement, and evaluate a PD multi-source evaluation process.

Methods Tools were developed for PD evaluation by trainees, department chairs, and graduate medical education (GME) leadership. Evaluation questions were based on PD responsibilities, including Accreditation Council for Graduate Medical Education (ACGME) requirements. A follow-up survey assessed the process.

Results Evaluation completion rates were as follows: trainees in academic year 2012–2013, 53% (958 of 1824), and in academic year 2013–2014, 42% (800 of 1898); GME directors in 2013–2014, 100% (95 of 95); and chairs/chiefs in 2013–2014, 92% (109 of 118). Results of a follow-up survey of PDs (66%, 59 of 90) and chairs (74%, 48 of 65) supports the evaluations' value, with 45% of responding PDs (25 of 56) and 50% of responding chairs (21 of 42) characterizing them as "extremely" or "quite" useful. Most indicated this was the first written evaluation they had received (PDs 78%, 46 of 59) or provided (chairs 69%, 33 of 48) regarding the PD role. More than 60% of PD (30 of 49) and chair respondents (24 of 40) indicated trainee feedback was "extremely" or "quite" useful, and nearly 50% of PDs (29 of 59) and 21% of chairs (10 of 48) planned changes based on the results. Trainee response rates improved in 2014–2015 (52%, 971 of 1872) and 2015–2016 (69%, 1276 of 1837).

Conclusions In our institution, multi-source evaluation of PDs was sustained over 4 years with acceptable and improving evaluation completion rates. The process and assessment tools are potentially transferrable to other institutions.

Introduction

Residency and fellowship program directors (PDs) play a critical role in the design, operation, and success of physician training programs. To date, limited attention has been focused on optimizing PD effectiveness, and reports regarding assessments related to the PD role are absent from the literature. This is particularly striking because PDs are required to implement a system of multi-source evaluation of residents without participating in such systems themselves. Moreover, the lack of emphasis on evaluation represents a missed opportunity to help PDs maximize their positive impact: multi-source feedback is a valid and reliable method for evaluating physicians and other professionals,[1,2] and is considered an effective development tool for leaders across professions.[1(p293)]

DOI: http://dx.doi.org/10.4300/JGME-D-15-00543.1

Editor's Note: The online version of this article contains an evaluation of a program director by trainee, by graduate medical education director, and by chair/chief; a program director follow-up survey; and a chair/chief follow-up survey.

Partners HealthCare implemented a centralized process for multi-source evaluation of directors of more than 100 programs accredited through the Accreditation Council for Graduate Medical Education (ACGME). Implementation spanned 5 sponsoring institutions: 2 large academic medical centers, 2 community hospitals, and 1 rehabilitation hospital. This effort was part of a broader initiative to enhance multi-source evaluation of trainees and faculty across our institutions.[3] The goal was to develop a process for formative, multi-source assessment aimed at enhancing PD performance, supporting career development, highlighting best practices, stimulating program improvements, and identifying areas for expanded institutional training and support of PDs.

Methods

Development of Evaluation Tools

Evaluation tools for PDs were not found in the published literature, so new evaluation tools were developed for 3 groups expected to have valuable perspectives on PD performance: residents and fellows, department chairs, and graduate medical

education (GME) directors (vice president for GME, designated institutional official, and 2 associate directors of GME). Items on the evaluation forms were based on PD responsibilities, including those outlined in the ACGME Common Program Requirements.[4] Draft evaluation forms were revised based on pilot testing, using "think aloud" technique and focusing on understandability, length, and content. Pilot testing included past and newly appointed chairs, PDs (who were not eligible to participate in the evaluation process), and a 16-member evaluation and feedback subcommittee, which included current PDs and trainee representatives.

Pilot testing supported construct validity[5] along with fitness for purpose, meaningfulness, and acceptability.[6,7] Traditional components of validity (eg, Cronbach's alpha and/or factor analysis) were neither possible nor relevant since the evaluations were designed for brevity (to improve response rate) with nonoverlapping questions. Test-retest reliability could not be assessed because (1) the trainee survey was anonymous; (2) the cohort of trainees evaluating PDs changes each year; and (3) a variation of responses would be expected given the dynamic nature of programs and leadership.

The resulting assessment tools (provided as online supplemental material) were endorsed by the Partners Education Committee, which includes elected and appointed educators, program directors, administrators, and trainees from the 5 teaching hospitals in our health system. The "Evaluation of Program Director by Trainee" included 9 items rated on a 5-point Likert scale, and solicits text comments about PD strengths, along with suggestions for improvement. Evaluations by the GME director and the department chair (or division chief) included 7 and 23 items, respectively; each were rated according to an expectation scale, with a prompt for text comments.

PD Evaluation Process

Trainee evaluation of PDs was piloted in academic year (AY) 2012–2013, including 102 ACGME PDs who had been in the role at least 6 months. Trainees were assured anonymity, and that results would be provided to PDs and chairs/chiefs only in aggregate and if 4 or more trainee evaluations were received. The evaluations were conducted utilizing REDCap, a secure web-based application designed to support research.[8] In AY 2013–2014, a multi-source evaluation of 108 PDs was conducted, incorporating evaluation by trainees, GME directors, and chairs/chiefs. The GME directors completed evaluations on those PDs for whom they routinely provided GME oversight. For 13 GME programs utilizing more than 1 hospital, multiple chairs were asked to provide an evaluation. For simplicity, department chairs and/or division chiefs—representing the person to whom a program director reports—will be referred to as "chairs." Of note, some PDs report to more than 1 chair and some chairs oversee more than 1 PD.

What was known and gap
Multi-source evaluation generally is not provided to residency or fellowship program directors.

What is new
A program director evaluation by residents, fellows, department chairs, and institutional graduate medical education leadership.

Limitations
Survey instruments lack formal validity evidence.

Bottom line
Multi-source evaluation can guide program director improvement efforts; nearly 50% of responding program directors reported planned changes based on the results, and the process and assessment tools are transferrable to other institutions.

In 2013, multi-source evaluation results were sent to PDs and the chairs to whom they report (see an example in TABLE 1). Trainee evaluation results included mean ratings for each item, provided along with system-wide mean ratings for all PDs of the same category (residency or fellowship) to use as benchmarking data; aggregated, anonymous, unedited text comments from trainees were also included. To protect anonymity, trainee evaluations were not provided to PDs or chairs when fewer than 4 evaluations were received from trainees. Evaluations from GME directors and chairs were provided in full and were not anonymous. PDs and chairs were encouraged to meet and discuss findings and action plans.

Postevaluation Survey

We conducted an anonymous paired follow-up survey of PDs and chairs in AY 2013–2014 to determine (1) the perceived usefulness of multi-source evaluations; (2) the perceived relative value of information obtained from different evaluator groups; and (3) whether the process stimulated meetings and/or specific actions (surveys are provided as online supplemental material).

The Partners HealthCare Institutional Review Board determined that the study was exempt (Protocol #32014P000247).

Results

Evaluation completion rates were as follows: AY 2012–2013 trainees had a 53% (958 of 1824)

TABLE 1
Evaluation of Fellowship Program Director by Trainee—Sample Report (Academic Year 2012)[a]

Evaluation Question	Individual Program Director Rating	Overall Average[b]	Difference
Is available, approachable, and supportive	5.0	4.7	+0.3
Actively pursues improvements in academic programs	5.0	4.6	+0.4
Articulates a clear vision for my education	4.9	4.5	+0.4
Provides a system for explicit, effective feedback	5.0	4.5	+0.5
Has effective communication and interpersonal skills	4.9	4.5	+0.4
Seeks fair resolution of conflicts	4.9	4.6	+0.3
Provides trainees with an effective advisor/mentor	4.8	4.3	+0.5
Addresses resident concerns regarding the work environment	5.0	4.6	+0.4
Addresses my documentation requirements and requests	5.0	4.8	+0.2

[a] Overall survey response rate was 53% (958 of 1824); program response rate 53%. Survey rating scale: 1, strongly disagree; 2, disagree; 3, neutral; 4, agree; and 5, strongly agree.
[b] Overall average is the mean rating by program type (eg, residency or fellowship) for all training programs.

completion rate; AY 2013–2014 trainees had a 42% (800 of 1898) completion rate; GME directors completed 100% (95 of 95) of the evaluations; and chairs/chiefs completed 92% (109 of 118) of the evaluations. Of note, in AY 2013, 44 PDs and their corresponding chairs were not provided with trainee evaluation data because there were fewer than 4 respondents (generally in small fellowship programs). Cumulative trainee data were provided to these programs in subsequent years once evaluation results totaled 4 responses.

The AY 2013–2014 follow-up survey of PDs (66%, 59 of 90) and chairs (74%, 48 of 65) supports the value of multi-source evaluation (TABLE 2), with a majority of PD respondents (78%, 46 of 59) indicating that this was the first written evaluation they had received relating to the role, and 69% of responding chairs (33 of 48) acknowledged that it was the first evaluation they had provided. The majority of PD (86%, 48 of 56) and chair (90%, 38 of 42) respondents considered the overall evaluation process to be at least somewhat useful, and 45% of PDs (25 of 56) and 50% of chairs (21 of 42) characterized it as "quite" or "extremely" useful. Evaluations from trainees were considered most helpful, followed by those from chairs and then GME directors. For this reason our education committee determined that trainee evaluations would continue annually, while chairs and GME directors would evaluate PDs every 3 to 4 years.

Providing PDs and chairs with evaluations appears to have stimulated desirable activities, with 39% of PDs (23 of 59) and 42% of chairs (20 of 48) reporting they met to discuss the PD multi-source evaluation. Additionally, 19% (7 of 36) of PDs and 31% (8 of 26) of chairs indicated that a meeting was scheduled for the future. Among PDs who met with their chair(s), 30% (7 of 23) indicated that this was their first meeting to discuss their performance in the PD role. Nearly 50% of PDs (29 of 59) and 21% of chairs (10 of 48) planned changes based on feedback.

The trainee evaluation has continued annually with improving response rates: AY 2014–2015, 52% (971 of 1872) and AY 2015–2016, 69% (1276 of 1837). In

TABLE 2
Perceptions of Program Directors and Chiefs/Chairs on the Usefulness of Evaluating Program Directors Based on Source of Evaluation[a]

Evaluation Type	Program Directors (n = 59)			Chiefs/Chairs (n = 48)		
	Total Responses	% Quite or Extremely Useful[b]	% At Least Somewhat Useful[c]	Total Responses	% Quite or Extremely Useful[b]	% At Least Somewhat Useful[c]
Overall	56	45	86	42	50	90
Trainees	49	61	90	39	62	90
Chair/chief	49	47	88	33	52	85
GME director	52	40	81	37	51	84

[a] Denominators refer to respondents who reported receiving the information from different evaluator groups. Rating options: not at all useful, minimally useful, somewhat useful, quite useful, and extremely useful/essential.
[b] Sum of 2 highest ratings: quite useful, extremely useful/essential.
[c] Sum of 3 highest ratings: somewhat useful, quite useful, extremely useful/essential.

AY 2014–2015, mean ratings of PDs were highest in the areas "Addresses my documentation requirements/requests," "Pursues improvements in academic programs," "Is available, approachable, and supportive," and "Addresses concerns regarding the work environment." The lowest-rated items were "Provides trainees with an effective advisor/mentor" and "Provides a system for trainees to receive explicit, effective feedback." Evidence that the evaluation tool offers some degree of discrimination is that in AY 2014–2015, 59% of PDs (61 of 103) received at least 1 rating of a 3 (neutral) or lower, and of the 11 048 responses to individual questions, 1067 (9.7%) received a rating of less than or equal to 3.

The development and initial implementation of this PD evaluation process was considerably more time consuming than its annual continuation because of the need to create survey tools, develop an approach for distributing reports, and cultivate an understanding of the process and institutional buy-in. Currently, the evaluation process requires the equivalent of 2 weeks of full-time effort from a member of the GME office staff.

Discussion

This study represents an important first step in using multi-source evaluation to support PD development and improve performance in this critical role. We demonstrated that multi-source evaluation of PDs can be implemented successfully, and we anticipate that other institutions can do so with less effort by importing and/or adapting the tools, methods, and report formats we have developed.

Our survey suggests that many PDs and chairs consider PD evaluation to be useful, with evaluations from trainees being the most valuable component. This probably relates to the nature of the PD-trainee relationship, which makes it difficult for trainees to provide candid assessments of their supervisor without the benefits of a structured and anonymous process. Also, trainee perceptions would seem to be of particular interest since PD activities are generally for their benefit. Finally, text comments from trainees were more detailed than those of chairs and GME directors, and the availability of multiple responses makes it possible to identify themes that rose above the level of individual opinion.

Beyond benefit to PDs and programs, aggregated institutional data can be helpful in identifying areas for improvement. For example, the lower mean ratings for "Providing a system for trainees to receive explicit, effective feedback" correlated with results of our ACGME resident survey, and the PD evaluations allowed us to assess areas, such as mentoring, not addressed on the ACGME survey.

Certain limitations should be noted. Compound questions were utilized in the evaluations for purposes of brevity and maximizing response rate; if more specific information about PD performance is required, questions can be reformulated accordingly. When the follow-up survey of PDs and chairs was conducted in AY 2013–2014, 44 PDs and their corresponding chairs (39% of the 112 pairs surveyed) had not received trainee evaluation data because there were fewer than 4 respondents. Since trainee evaluation was considered the most useful component (and the 44 pairs were surveyed based on their receipt of other evaluation components) the usefulness of the multi-source evaluation process may have been undervalued. Finally, a 66% response rate on the PD follow-up survey is not ideal, given that PDs are the focus of the feedback process. The multi-source evaluation did not include program coordinators because anonymity could not be provided. Faculty also were not included in the initial phase, but may be asked to evaluate PDs in the future. Enhancements to the PD evaluation process are being implemented. To improve trainee participation, we now highlight the PD evaluation process during trainee orientation.

The multi-source evaluation process will be continued, with evaluation by residents conducted annually, and by GME directors and chiefs/chairs every 3 to 4 years. The PD self-evaluation that was piloted in AY 2014 continues to be offered annually as an optional component.

Conclusion

Multi-source evaluation of PDs was successfully implemented in our health system, and was considered "somewhat," "quite," or "extremely" useful by the majority of PD and chair respondents. It can guide individual PD improvement efforts and, based on themes noted across specialties, institutional faculty development programs for PDs. The process and assessment tools are potentially transferable to other institutions.

References

1. Atwater LE, Brett JF, Charles AC. Multisource feedback: lessons learned and implications for practice. *Hum Resource Manag*. 2007;46(2):285–307.
2. Donnon T, Al Ansari A, Al Alawi S, et al. The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. *Acad Med*. 2014;89(3):511–516.
3. MedEdPORTAL Publications. Goldhamer M, Baker K, Rigg A, et al. Development and implementation of multi-

source assessment tools for ACGME residents and fellows. 2014;10:9839. https://www.mededportal.org/publication/9839. Accessed July 18, 2016.
4. Accreditation Council for Graduate Medical Education. Common program requirements. http://www.acgme.org/Portals/0/PDFs/Common_Program_Requirements_07012011%5B2%5D.pdf. Accessed May 13, 2016.
5. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;119(2):166.e7–166.e16.
6. Baartman LK, Bastiaens TJ, Kirschner PA, et al. The wheel of competency assessment: presenting quality criteria for competency assessment programs. *Stud Educ Eval*. 2006;32(2):153–170.
7. Swing SR, Beeson MS, Carraccio C, et al. Educational milestone development in the first 7 specialties to enter the next accreditation system. *J Grad Med Educ*. 2013;5(1):98–106.
8. Harris PA, Taylor R, Thielke R, et al. Research Electronic Data Capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. 2009;42(2):377–381.

Mary Ellen J. Goldhamer, MD, MPH, is Education Specialist, Partners Office of Graduate Medical Education, Partners HealthCare, and Instructor in Medicine, Massachusetts General Hospital and Harvard Medical School; **Keith Baker, MD, PhD,** is Program Director, Anesthesiology Residency, and Vice Chair for Education, Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, and Associate Professor of Anesthesiology, Massachusetts General Hospital and Harvard Medical School; **Amy P. Cohen, EdM,** is Assistant Fellowship Director, Harvard Medical School Academy Fellowship in Medical Education Research, and Director of Data Analytics and Instructor, Department of Health Policy and Management, Harvard T.H. Chan School of Public Health; and **Debra F. Weinstein, MD,** is Vice President, Graduate Medical Education, Partners HealthCare, and Associate Professor of Medicine, Massachusetts General Hospital and Harvard Medical School.

Funding: The authors report no external funding source for this study.

Conflict of interest: The authors declare they have no competing interests.

The authors would like to thank John Patrick T. Co, MD, MPH, Director of Graduate Medical Education, Partners HealthCare, and Associate Directors Eric Nadel, MD, and Lori Berkowitz, MD, for their assistance in implementing this multisource evaluation effort. The authors would also like to thank Michelle Brooks, MS, RD, who, at the time of the study, was at the Enterprise Research Infrastructure and Services, Partners HealthCare, for her contribution of REDCap database design and management.

Corresponding author: Mary Ellen J. Goldhamer, MD, MPH, Partners Office of Graduate Medical Education, Bulfinch 230-E, 55 Fruit Street, Boston, MA 02114, 617.726.5440, mgoldhamer@mgh.harvard.edu

Received October 21, 2015; revisions received February 24, 2016, and April 1, 2016; accepted May 2, 2016.