ORIGINAL ARTICLE

Resident Self-Other Assessor Agreement

Influence of Assessor, Competency, and Performance Level

Pamela A. Lipsett, MD, MHPE; Ilene Harris, PhD; Steven Downing, PhD

Objectives: To review the literature on self-assessment in the context of resident performance and to determine the correlation between self-assessment across competencies in high- and low-performing residents and assessments performed by raters from a variety of professional roles (peers, nurses, and faculty).

Design: Retrospective analysis of prospectively collected anonymous self-assessment and multiprofessional (360) performance assessments by competency and overall.

Setting: University-based academic general surgical program.

Participants: Sixty-two residents rotating in general surgery.

Main Outcome Measures: Mean difference for each self-assessment dyad (self-peer, self-nurse, and self–attending physician) by resident performance quartile, adjusted for measurement error, correlation coefficients, and summed differences across all competencies.

Results: Irrespective of self-other dyad, residents asked to rate their global performance overestimated their skills. Residents in the upper quartile underestimated their specific skills while those in the lowest-performing quartile overestimated their abilities when compared with faculty, peers, and especially nurse raters. Moreover, overestimation was greatest in competencies related to interpersonal skills, communication, teamwork, and professionalism.

Conclusions: Rater, level of performance, and the competency being assessed all influence the comparison of the resident's self-assessment and those of other raters. Self-assessment of competencies related to behavior may be inaccurate when compared with raters from various professions. Residents in the lowest-performing quartile are least able to identify their weakness. These data have important implications for residents, program directors, and the public and suggest that strategies that help the lowest-performing residents recognize areas in need of improvement are needed.

Arch Surg. 2011;146(8):901-906

PHYSICIANS ARE EXPECTED TO participate in lifelong learning and professional development. Integral to the process of training and improving skills is the ability to identify strengths and weaknesses in one's knowledge, attitudes, and practice. However, the ability to perform an accurate self-assessment has been questioned in a variety of studies about health care workers,[1-5] higher-education professionals,[6,7] and the business community.[8-10] In a systematic review in 2006, Davis et al[11] identified 17 studies of self-assessment involving physicians, of which 13 demonstrated little, no, or an inverse relationship with an externally validated measure of performance. To our knowledge, the role of resident self-assessment in the context of 360 assessments has not been reported.

The purpose of this study was to review the background and literature on self-assessment in the context of resident performance and to determine the correlation between self-assessment across competencies in high- and low-performing residents and assessments performed by raters from a variety of professional roles (peers, nurses, and attending physicians). In addition, this study specifically examined whether the magnitude of differences between self-assessment and the ratings of others was related to the residents' level of performance.

Author Affiliations:
Department of Surgery, Johns Hopkins University Schools of Medicine and Nursing, Baltimore, Maryland (Dr Lipsett); and Department of Medical Education, University of Illinois, Chicago (Drs Harris and Downing).

CONCEPTUAL FRAMEWORK AND BACKGROUND

The reasons for a difference in self-assessment and other performance measures appear to be many and include lack of a gold standard for measurement, a different frame of reference of the assessor vs self, and measurement error. A number of studies found the worst accuracy in self-assessment among the physicians who were the least skilled but were the most confident. Kruger and Dunning[12] have attributed the difficulty in recognizing one's own failures to miscalculation due to deficits in metacognitive skill, a skill that can

be improved with training. Moreover, they demonstrated that high performers slightly underestimate their performance. Krueger and Mueller,[13] on the other hand, attribute their finding of "being unskilled and unaware" to a statistical regression to the mean and relate it to a "better than average" phenomenon. For example, when college professors were asked whether they performed "above-average work," 94% of college professors indicated they were in this category, a figure clearly mathematically impossible[14] but perceptually possible. Our residents have this same overall perception about their performance.

Self-assessment must be placed in context with self-perception and then reconciled with feedback from multiple sources.[15,16] One reported advantage of multisource feedback, especially that from peers, is that individuals are able to calibrate interpretation of the feedback with their own experiences.[17] Trainees may be more willing to accept feedback from peers, because peers evaluate each other having the same frame of reference and experiences and peers can provide specific feedback from observed interactions.[5] In addition, peer assessment has been said to benefit both the assessor, by having experience giving feedback and by conceptually formalizing standards or processes that they use to assess colleagues, and the assessee and may result in deep rather than surface learning for both.[5,18] However, some trainees are suspicious of peer assessment and do not believe their colleagues to be equal to the task of assessing them.[5,17]

Sargeant and colleagues[19] performed a series of quantitative and qualitative studies on practicing family physicians. They demonstrated that practicing physicians agreed with higher ratings more than with lower ratings.[20] Physicians also more frequently disagreed with feedback from medical colleagues than that from patients or coworkers in the same office.[21] Further, they found that physicians responded with negative emotions to feedback that was inconsistent with self-perceptions of performance, questioned its credibility, and were not inclined to use it. Physicians indicated that the credibility of the feedback was related to whether the rater was able to specifically observe the behavior, interaction, or skill and whether the feedback was specific.

Negative emotions interfere with assimilation and acceptance of feedback[22,23] and a period of reflection is required for final acceptance of feedback.[24-26] For trainees with lower levels of skill, it is unknown whether negative emotions are more common or whether feedback from specific groups of raters is more or less likely to evoke negative emotions.[27] Since the intended outcome of feedback is to improve the performance of the resident, the resident must be willing to acknowledge and accept the variance between self-assessment and that from other raters.

METHODS

MULTISOURCE ASSESSMENT PROGRAM: GENERAL ORGANIZATION

Clinical evaluations at Johns Hopkins University School of Medicine Department of Surgery are completed by nurses, peers, and faculty working with surgical residents rotating on a surgical service (P.A.L., unpublished data, 2010). Briefly, residents are assessed on each of the Accreditation Council for Graduate Medical Education competencies using a behaviorally anchored Likert-type scale, with scale points from 1 to 5 (5=outstanding). The scale level of 3, with an associated behavioral anchor, is set to reflect the expected performance of the typical or average postgraduate-level resident (levels 1-5). Each resident was expected to perform a self-assessment similar to the other rating forms, once each 3-month period, for a total of 4 self-assessments per year. In addition, residents were asked to identify 3 areas of strength, 3 areas they would like to improve, and the measures they would take to correct the perceived weaknesses.

The ratings for all residents from July 2007 to June 2008 in the General Surgery Program were obtained by us, blinded to the identity of the resident and raters. Residents were identified only by a unique identifier, postgraduate year level, and sex. The study was approved by the Johns Hopkins University School of Medicine and the University of Illinois–Chicago institutional review boards.

DATA ANALYSIS

To assess the results of self-assessment dyads, anonymous data tables were converted to STATA, version 9 (StataCorp, College Station, Texas) for analysis. For each resident, mean values and standard deviations were determined for each competency for each rater group and for self-assessments. For each dyad (self-assessment–peer, self-assessment–nurse, and self-assessment–faculty), the mean values for the performance assessments of residents were divided into quartiles. To account for measurement error, the standard error of measurement was calculated and 95% confidence intervals, determined. Each competency was individually assessed as well as a global assessment of competence. Correlation coefficients were determined using the Pearson product-moment correlation, both overall and by quartile. To determine whether the lowest-performing residents differed in their self-assessments from those of others, self-assessment–other mean differences in the competencies were determined and summed across competencies.

RESULTS

Of the 62 clinically active residents, all residents had at least 3 self-assessments and 3 or more raters from each of the rater groups. Residents represented each of the clinical years (1-5) and both sexes. The results of the self-other dyads are shown in the eTable (http://www.archsurg.com) by competency. **Figure 1** demonstrates the differences between the self-assessment and those of peers, nurses, and faculty.

SELF-FACULTY DYAD

Compared with faculty assessments, residents as a group underestimated their performance in a variety of specific competencies including patient care: clinical judgment and medical knowledge. Residents in the upper quartile of performance underestimated their performance in many of the other specific competencies (Figure 1B). In addition, the self-assessment of residents in the lowest-performance quartile overestimated their own skills in the area of professionalism: compassion (Figure 1A). Further, residents in all quartiles of performance, when asked to provide an overall estimate of their global perfor-

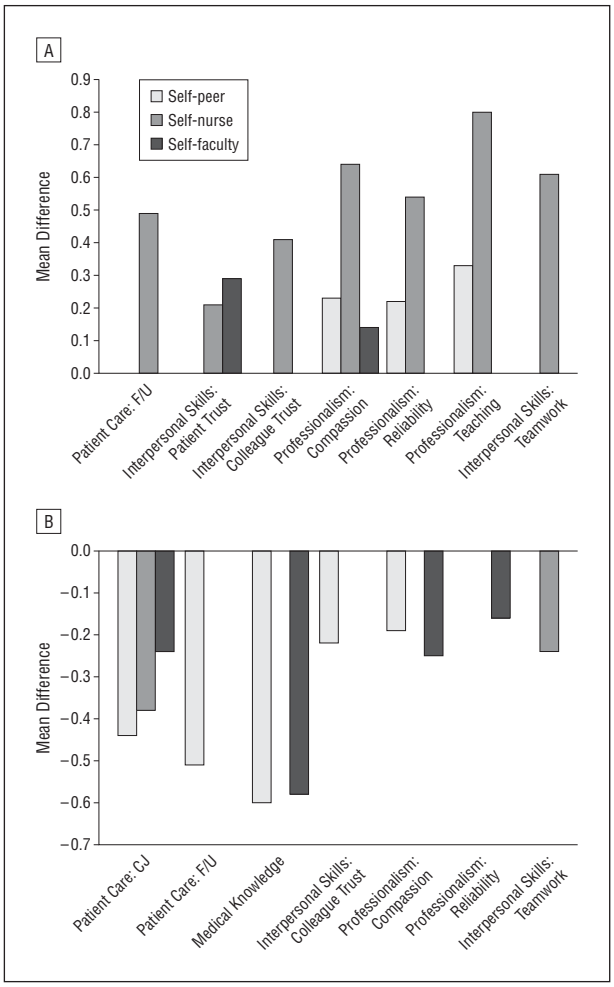

Figure 1. Mean significant difference between the self-assessment and that of different rater groups: self-peer, self-nurse, and self-faculty. A, Lowest-performing quartile. B, Highest-performing quartile. A negative value indicates that self-assessment is lower than mean rater group assessment, while a positive number indicates that self-assessment is higher than mean rater group assessment. A, Note in the lowest-performing quartile that across most competencies self-assessment overestimates performance, especially with the nurse rater group. Faculty are less discriminating and are concordant with resident self-assessment. B, Note in the highest-performing quartile that self-assessment underestimates performance across all rater groups. CJ indicates clinical judgment; F/U, follow-up.

mance, consistently overrated their performance when compared with attending physicians (**Figure 2**).

SELF-PEER DYAD

Peer and self-assessments were similar to the faculty findings. For almost all competencies, the self-assessment of residents in the upper quartile consistently underestimated their peer assessments (Figure 1B). Residents who were in the lowest performance quartiles had a self-assessment that significantly overestimated their performance (Figure 1A). Again, when asked to submit a global assessment, resident self-assessment overestimated performance when compared with peers (Figure 2).

SELF-NURSE DYAD

The resident self-assessment–nurse dyad comparisons accentuated the differences seen in the lowest-performing

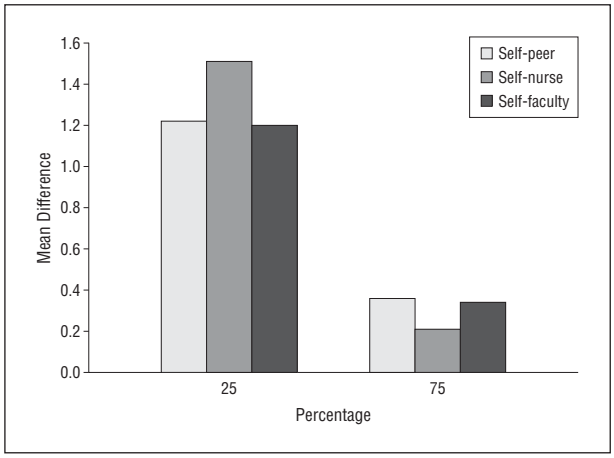

Figure 2. Mean significant difference between the resident self-assessment of global performance and that of rater group (self-peer, self-nurse, and self-faculty) by lowest and highest quartile of performance.

resident quartile. Residents in the lowest-performing quartiles overestimated their performance when compared with nurses (Figure 1A). On the other hand, self-assessments of residents in the upper quartile of performance underestimated nurse assessments in several dimensions (Figure 1B) As was seen with peer and faculty self-other comparisons, the global performance self-assessments of residents in all quartiles overestimated their performance when compared with nurses (Figure 2)

CORRELATION BETWEEN SELF-OTHER

The correlations between self and other professionals are shown in **Table 1** by competency. When compared with nurse evaluators, self-assessments of specific competencies showed significant correlations except in the areas of medical knowledge and professionalism: compassion and reliability and responsibility. Similarly, the self-assessments of residents demonstrated a moderate correlation with those of peer raters, except again in the medical knowledge competency. In contrast, self-assessments and attending physician raters demonstrated significant correlations in all competencies.

MEAN DIFFERENCES IN SELF-ASSESSMENT DYADS

The sum of the mean differences across competencies is shown by performance group and rater in **Table 2**. Residents whose performance was rated to be in the lowest quartile overestimated their performance when the difference between their self-assessment in each competency and each rater group mean was summed. However, the magnitude of the overestimation was concentrated in the competencies that may be considered behavioral, or those related to interpersonal skills, communication, teamwork, and professionalism. In the lowest-performing group, the summed difference was greatest with nurse raters. Residents in the middle and upper quartile underestimated their performance irrespective of rater group, with the greatest underestimation seen in the highest-performance quartile. Unlike the

Table 1. Correlations Between Self-assessments and the Assessments From Raters From Different Professional Roles

	Self vs Nurse		Self vs Peers		Self vs Attending Physician	
	Correlation	*P* Value	Correlation	*P* Value	Correlation	*P* Value
Patient care: clinical judgment	0.4295[a]	.01[a]	0.3482[a]	.03[a]	0.4075[a]	<.01[a]
Patient care: follow-up	0.4109[a]	.02[a]	0.4773[a]	<.01[a]	0.3689[a]	.02[a]
Medical knowledge	0.2931	.23	0.2600	.25	0.3985[a]	<.01[a]
Interpersonal skills: patient trust	0.4386[a]	<.01[a]	0.3890[a]	.01[a]	0.3541[a]	.03[a]
Interpersonal skills: colleague trust	0.4742[a]	<.01[a]	0.4562[a]	<.01[a]	0.4457[a]	<.01[a]
Professionalism: compassion	0.2911	.24	0.2535	.28	0.3678[a]	.01[a]
Professionalism: reliability and responsibility	0.2725	.33	0.3425[a]	.03[a]	0.3935[a]	<.01[a]
Professionalism: teaching	0.3723[a]	.05[a]	0.3543[a]	.02[a]	0.4065[a]	<.01[a]
Interpersonal communication: teamwork	0.2931	.04	0.4915[a]	<.01[a]		

[a] Significant correlation between self-assessment and other rater group.

Table 2. Self-assessment vs Other, Mean Differences, by Quartile

Rater Group	Total Mean Difference	PC and MK	Behavioral[a]	Global
Attending physician				
Lowest 25%	–0.39	0.14	–0.53	–1.2
Top 25%	2.04	0.95	1.09	–0.34
Peer				
Lowest 25%	–0.96	–0.19	–0.98	–1.22
Top 25%	2.57	0.84	1.34	–0.36
Nurse				
Lowest 25%	–5.14	–0.96	–4.18	–1.51
Top 25%	2.35	0.68	1.67	–0.21

Abbreviations: MK, medical knowledge; PC, patient care: clinical judgment and follow-up.
[a] Behavioral = interpersonal skills and communication: patient and colleague trust and teamwork and professionalism: reliability and responsibility, compassion, and teaching.

lower-performing residents, the summed difference was more balanced between cognitive and behavioral competencies.

COMMENT

In this study, we found that when compared with multisource assessment by professional colleagues, some residents are able to self-assess specific competencies when the combined or individual rater group assessment is taken as a gold standard. To account for the problem of interrater reliability and measurement error, each assessment was corrected for reliability obtained from generalizability studies using the standard error of measurement before considering correlations or differences between self-other. While residents were able to make self-assessments about specific competencies that correlated with that of other raters, when asked to rate their global assessment, residents systematically overestimated their overall performance across all rating groups. This study also documented that those surgical trainees who were in the lowest-performing quartile often overestimated their competency-specific performance irrespective of rater group, while the highest-performing residents tended to underestimate their skills. Superficially, these findings may appear to be a regression to the mean performance,[13] but the magnitude of the differences between self-other is greatest in the lowest-performing quartiles and for the behavioral skills, especially with nurse raters. In previous studies of the reliability of each rating group, nurses were the most consistent raters but used the rating scale more widely to differentiate high- and low-performing residents. This suggests that residents who need to acquire knowledge, skills, attitudes, and behaviors that would make them more effective in their roles as residents appear less able to identify their own difficulties.[12,28,29] The uniformity of this finding across rating groups in the multisource assessments makes this finding more generalizeable. The findings of this study support those seen in other physician groups and in other disciplines where the correlation between self-other in student assessment ranged from 0.05 to 0.82, with an average of 0.39.[7,30,31]

In classic studies by Kruger and Dunning,[12] across 4 studies, students in the bottom quartile on tests of humor, grammar, and logic grossly overestimated their test performance and skills. Although their test scores placed them in the 12th percentile, they estimated themselves to be in the 62nd percentile. More recent work by Ehrlinger and colleagues[29] further examined the pattern of overestimation and underestimation of performance described by Kruger and Dunning[12] by extending work from the classroom into more "real-world" situations but nonmedical studies. They further examined whether incentives to enhance accuracy of self-assessments, such as monetary incentives or having to justify their assessment to another

party, would alter how those in the lowest-performing quartile would rate their performance. Somewhat surprisingly, poor performers became more overconfident in the presence of a monetary incentive. Further, when students had to justify their performance to a third party, poor performers once again became more, rather than less, overconfident. Taken together, these data suggest that even with intense focus and effort, those with lower skill levels are not able to accurately self-assess their performance. Finally, Ehrlinger and colleagues[29] examined the origin of the misperceptions in self-assessment. They found that bottom performers had misconceptions of their own performance rather than misconceptions about the performance of others. In contrast, top performers are overly optimistic about the performance of peers, and thus, they exhibit undue modesty about their own performance.

Evans and colleagues[3] found that, on average, peer assessment, especially global rating scales, reflected more accurately those ratings of the trainer rather than self-assessment. They also found that trainee surgeons tended to overestimate their own technical skills, and more so for those with the lowest scores. In contrast, peer-assessment overestimates of ability were not apparent. Evans and colleagues suggest that informal peer assessment may therefore allow for a more open and frank discussion of strengths and weaknesses in resident trainees than attending physician assessment. In contrast to these findings, during standardized assessment of technical skills, obstetrics and gynecology trainees rated task-specific, overall, and global assessments similar to faculty ratings (r=0.32-0.77).[32] Interestingly, in this study, residents tended to rate themselves lower than faculty. Moreover, lower-performing residents were able to self-identify problems, and their areas needing improvement were qualitatively similar to the assessment of faculty.

These 2 studies of trainees in surgical subspecialties differ qualitatively from our study in that they examined self-assessment in context-specific situations, namely a single specific surgical procedure or skill. This finding suggests a possible limitation of our study in that the gold standard for assessments used in our study are those performed by external raters and are thus subject to constraints of the measurement instrument and measurement error. Previous generalizability studies with our residents have documented a high degree of reliability of the assessment within and between rating groups. More specifically, we have demonstrated that the reliability of the instruments both within and between rater groups was high ($G>0.80$) and that the number of raters (21 total) was sufficient to make summative decisions. In addition, in this study, adjustments were made for the reliability of the assessment (P.A.L., unpublished data, 2010). These data also suggest that performance assessments, outside of specific contexts, are greatly influenced by many additional components, such as how both residents and raters collect, process, recall, and communicate their experiences.[33]

The experiences of our residents, nurses, and faculty may not reflect those of other specialties or of other institutions. The particular position of the rater, their personal demographic and cultural characteristics, and the rating scale and instrument, as well as situations in which the resident is assessed, are all likely to influence these findings and are not specifically addressed in this study.[34,35]

How should the findings of this study inform and change our practice? To address the discrepancies between self–other physician assessments, Sargeant et al[26] propose a "directed self-assessment model within a social context." When this model is placed in the context of the findings of our study of surgical residents, the implication is that program directors should pay particular attention to residents performing in the lowest quartile to facilitate reconciliation of differences in their own perceptions of their performance vs those of others. However, provision of negative external feedback can have unintended consequences and cause a further reduction in performance.[16,22,23,25,34] Residents should have a clear understanding of standards and expectations of performance. For those performing at a lower level of technical skill, providing visual examples or practical experiences of an acceptable level of technical skill may enhance learning by providing a shared view of what a good performance "looks like."[35] For residents who need skill building in interpersonal skills, communication, or teamwork, video-taped facilitated review of standardized or actual patient experiences may provide an additional opportunity for learning beyond feedback from others.[25,35,36]

Facilitating trainees' reflections on their external measures of performance and self-assessment is a process that can be enhanced by a skilled facilitator. Program directors, and those providing formal feedback to residents, need to acquire better skills in providing this facilitated reflection. They need to be clear with residents on what and how they are being assessed by their colleagues, nurses, and faculty. In this process, program directors should help residents understand what standards residents are using to assess their own performance and whether these standards are appropriate for their level of development.[37,38] Residents should understand how they judge the quality and specificity of the multisource feedback.

Finally, program directors must recognize and manage emotional reactions to feedback. Program directors must recognize that residents who are performing at the lowest levels may be at the greatest risk for negative emotions (P.A.L., unpublished data, 2010) and that these emotions may inhibit residents from reflecting on and assimilating the feedback. Until residents have reflected on and assimilated the feedback, plans for learning and change are not likely to be realized.

In summary, this study found that surgery residents were able to self-assess specific competencies but overestimated their global performance when compared with raters from any professional group (peers, nurses, or faculty). In addition, residents who were in the lowest-performing quartile overestimated their skills and did so most notably in their behavioral skills. On the other hand, residents in the highest quartile tend to underestimate their skills. Thus, the rater, level of performance, and the competency being assessed all influence the comparison of the resident's self-assessment and those of other raters. These data have important implications for residents, program directors, and the public and suggest that strategies that help the lowest-performing residents recognize areas in need of improvement and further re-

search into the development of effective measures are needed.

Accepted for Publication: November 1, 2010.
Correspondence: Pamela A. Lipsett, MD, MHPE, The Johns Hopkins Hospital, Osler 603, 600 N Wolfe St, Baltimore, MD 21287 (plipsett@jhmi.edu).
Author Contributions: *Study concept and design:* Lipsett and Harris. *Acquisition of data:* Lipsett. *Analysis and interpretation of data:* Lipsett, Harris, and Downing. *Drafting of the manuscript:* Lipsett and Harris. *Critical revision of the manuscript for important intellectual content:* Lipsett, Harris, and Downing. *Statistical analysis:* Lipsett and Downing. *Study supervision:* Harris and Downing.
Financial Disclosure: None reported.
Online-Only Material: The eTable is available at http://www.archsurg.com.

REFERENCES

1. Barnsley L, Lyon PM, Ralston SJ, et al. Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. *Med Educ*. 2004;38(4):358-367.
2. Claridge JA, Calland JF, Chandrasekhara V, Young JS, Sanfey H, Schirmer BD. Comparing resident measurements to attending surgeon self-perceptions of surgical educators. *Am J Surg*. 2003;185(4):323-327.
3. Evans AW, McKenna C, Oliver M. Trainees' perspectives on the assessment and self-assessment of surgical skills. *Assess Eval High Educ*. 2005;30(2):163-174. doi:10.1080/0260293042000264253.
4. Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. *Acad Med*. 1991;66(12):762-769.
5. Hodges B, Regehr G, Martin D. Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it. *Acad Med*. 2001;76(10)(suppl):S87-S89.
6. Dunning D, Johnson K, Ehrlinger J, Kruger J. Why people fail to recognize their own competence. *Curr Dir Psychol Sci*. 2003;12:83-87.
7. Falchikov N, Boud D. Student self-assessment in higher education: a meta-analysis. *Rev Educ Res*. 1989;59(4):395-430. doi:10.3102/00346543059004395.
8. Atwater LE, Yammarino FJ. Does self-other agreement on leadership perceptions moderate the validity of leadership and performance predictions? *Person Psychol*. 1992;45(1):141-164. doi:10.1111/j.1744-6570.1992.tb00848.x.
9. Brutus S, Fleenor JW, London M. Does 360-degree feedback work in different industries? a between-industry comparison of the reliability and validity of multisource performance ratings. *J Manage Dev*. 1998;17(3):177-190. doi:10.1108/EUM0000000004487.
10. Sala F, Dwight SA. Predicting executive performance with multirater surveys: whom you ask makes a difference. *Consult Psychol J Pract Res*. 2002;54(3):166-172. doi:10.1037//1061-4087.54.3.166.
11. Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *JAMA*. 2006;296(9):1094-1102.
12. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol*. 1999;77(6):1121-1134.
13. Krueger J, Mueller RA. Unskilled, unaware, or both? the better-than-average heuristic and statistical regression predict errors in estimates of own performance. *J Pers Soc Psychol*. 2002;82(2):180-188.
14. Cross P. Not can but will college teaching be improved? *New Dir Higher Educ*. 1977;(17):1-15. doi:10.1002/he.36919771703.
15. Eva KW, Regehr G. Knowing when to look it up: a new conception of self-assessment ability. *Acad Med*. 2007;82(10)(suppl):S81-S84.
16. DeNisi AS, Kluger AN. Feedback effectiveness: can 360-degree appraisals be improved? *Acad Manage Exec*. 2000;14(1):129-139. http://www.jstor.org/stable/4165614.
17. Evans AW, Leeson RM, Petrie A. Reliability of peer and self-assessment scores compared with trainers' scores following third molar surgery. *Med Educ*. 2007;41(9):866-872.
18. Brown S, Dove P. Opening mouths to change feet: some views on self- and peer assessments. *SCED*. 1991;63:59-65.
19. Sargeant J, Mann K, Sinclair D, Van der Vleuten C, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv Health Sci Educ Theory Pract*. 2008;13(3):275-288.
20. Sargeant JM, Mann KV, Ferrier SN, et al. Responses of rural family physicians and their colleague and coworker raters to a multi-source feedback process: a pilot study. *Acad Med*. 2003;78(10)(suppl):S42-S44.
21. Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: perceptions of credibility and usefulness. *Med Educ*. 2005;39(5):497-504.
22. Brett JF, Atwater LE. 360 degree feedback: accuracy, reactions, and perceptions of usefulness. *J Appl Psychol*. 2001;86(5):930-942.
23. Cron WL, Slocum JW, VandeWalle D, Fu Q. The role of goal orientation on negative emotions and goal setting when initial performance falls short of one's performance goal. *Hum Perform*. 2005;18(1):55-80. doi:10.1207/s15327043hup1801_3.
24. Kluger AN, DeNisi A. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychol Bull*. 1996;119(2):254-284. doi:10.1037//0033-2909.119.2.254.
25. Mann K, Gordon J, MacLeod A. Reflection and reflective practice in health professions education: a systematic review. *Adv Health Sci Educ Theory Pract*. 2009;14(4):595-621.
26. Sargeant JM, Mann KV, van der Vleuten CP, Metsemakers JF. Reflection: a link between receiving and using assessment feedback [published online June 5, 2008]. *Adv Health Sci Educ Theory Pract*. 2009;14(3):399-410.
27. Sargeant J, Mann K, Sinclair D, et al. Learning in practice: experiences and perceptions of high-scoring physicians. *Acad Med*. 2006;81(7):655-660.
28. Kruger J. Lake Wobegon be gone! the "below-average effect" and the egocentric nature of comparative ability judgments. *J Pers Soc Psychol*. 1999;77(2):221-232.
29. Ehrlinger J, Johnson K, Banner M, Dunning D, Kruger J. Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent. *Organ Behav Hum Decis Process*. 2008;105(1):98-121.
30. Risucci DA, Tortolani AJ, Ward RJ. Ratings of surgical residents by self, supervisors and peers. *Surg Gynecol Obstet*. 1989;169(6):519-526.
31. Boud D. Reframing assessment as if learning were important. In: Boud D, Falchikov N, eds. *Rethinking Assessment in Higher Education*. New York, NY: Routledge; 2007:14-25.
32. Mandel LS, Goff BA, Lentz GM. Self-assessment of resident surgical skills: is it feasible? *Am J Obstet Gynecol*. 2005;193(5):1817-1822.
33. Ostroff C, Atwater LE, Feinberg BJ. Understanding self-other agreement: a look at rater and ratee characteristics, context, and outcomes. *Person Psychol*. 2004;57(2):333-375. doi:10.1111/j.1744-6570.2004.tb02494.x.
34. Scullen SE, Mount MK, Goff M. Understanding the latent structure of job performance ratings. *J Appl Psychol*. 2000;85(6):956-970.
35. Hattie J, Timperley H. The power of feedback. *Rev Educ Res*. 2007;77(1):81-112. doi:10.3102/003465430298487.
36. Martin D, Regehr G, Hodges B, McNaughton N. Using videotaped benchmarks to improve the self-assessment ability of family practice residents. *Acad Med*. 1998;73(11):1201-1206.
37. Goodstone MS, Diamante T. Organizational use of therapeutic change strengthening multisource feedback systems through interdisciplinary coaching. *J Consult Clin Psychol*. 1998;50(3):152-163. doi:10.1037//1061-4087.50.3.152.
38. Duffy FD, Holmboe ES. Self-assessment in lifelong learning and improving performance in practice: physician know thyself. *JAMA*. 2006;296(9):1137-1139.