

Forecasting Residents' Performance—Partly Cloudy

Reed G. Williams, PhD, Gary L. Dunnington, MD, and Debra L. Klamen, MD, MHPE

Abstract

The authors offer a practical guide for improving the appraisal of a resident's performance. They identify six major factors that compromise the process of observing, measuring, and characterizing a resident's current performance, forecasting future performance, and making decisions about the resident's progress. Factors that compromise any of these steps lead to individual and collective uncertainty and decrease faculty confidence when making

decisions on a resident's progress. The six factors, addressed in order of importance, are inaccuracies due to (1) incomplete sampling of performance, (2) rater memory constraints, (3) hidden performance deficits of the resident, (4) lack of performance benchmarks, (5) faculty members' hesitancy to act on negative performance information, and (6) systematic rater error. The description of each factor is followed by a number of specific suggestions on what residency

programs can do to eliminate or minimize the impact of these factors. While this article is couched in the context of the performance evaluation of residents, everything included pertains to measuring and appraising medical students' and practicing physicians' clinical performance as well.

Acad Med. 2005; 80:415–422.

We wrote this article to be a practical guide for improving the appraisal of residents' performance. To do justice to this topic, a brief discussion of the inference chain underlying performance appraisal is needed.

Background

The goal of clinical performance appraisal is akin to the goals of weather forecasting and medical diagnosis. All three are designed to forecast the future and to make decisions based on those projections.

Accurate weather forecasting depends on two things: (1) sufficient sampling of the current weather situation to accurately and reliably classify (recognize) the current state of affairs, and (2) knowledge about how similar weather patterns have

behaved historically under similar conditions.

Medical diagnosis depends on similar knowledge. First, physicians collect sufficient information about a patient's current state of affairs to diagnose the patient. Then they use collective knowledge from the field of medicine to establish prognoses for the patient (prognoses if left untreated and after undergoing alternate treatment regimens). Based on this information, a decision is made about the best treatment regimen for this patient.

Making progress decisions about residents depends on the same type of information. First, sufficient sampling of current behavior of a resident must have occurred to allow one to accurately and reliably classify the resident's current performance capabilities. Second, there must be sufficient knowledge about how residents like this one have performed in various practice situations historically to allow a reasoned decision about the prognosis for this resident. Most research on clinical performance appraisal has been focused on the reliability of ratings. Far less attention has been focused on those attributes of the process that compromise the ability to accurately forecast a resident's performance across the range of tasks and situations that make up clinical practice. This forecasting ability determines whether the clinical performance appraisal process serves well its ultimate function of

providing the information necessary to make good resident progress decisions. That is where our attention is directed in this article.

Factors That Compromise Forecasting Performance

Below, we describe six major factors that compromise faculty efforts to systematically observe and characterize the current performance of residents across a range of supervised practice situations, to accurately fill in the gaps about a resident's performance capabilities in situations that the faculty member has not observed, and to use this information to forecast how the resident will perform in the range of clinical practice settings that make up the domain of clinical practice after completion of residency training. This information, in turn, provides the basis for making progress decisions about a resident (e.g., promote without stipulation, promote with remediation requirements, hold back and prescribe a plan of remediation, encourage an alternate career path). Factors that compromise any of these steps lead to individual and collective uncertainty and decrease faculty confidence when making a resident's progress decisions. In this article, we discuss first the factors that most compromise fairness and accuracy of performance appraisal. We also offer a number of suggestions for addressing each. While we focus on the performance

Dr. Williams is professor and vice-chairman for educational affairs, Department of Surgery, Southern Illinois University School of Medicine, Springfield, Illinois.

Dr. Dunnington is professor and chairman, Department of Surgery, Southern Illinois University School of Medicine, Springfield, Illinois.

Dr. Klamen is associate dean for education and curriculum and professor and chair, Department of Medical Education, Southern Illinois University School of Medicine, Springfield, Illinois.

Correspondence should be addressed to Dr. Williams, Department of Surgery Southern Illinois University School of Medicine, 800 N. Rutledge Street, P.O. Box 19638, Springfield, Illinois 62794-9638; telephone: (217) 545-0529; e-mail: (rwilliams@siumed.edu).

appraisal of residents here, everything that is said applies to the appraisal of medical students and practitioners as well.

1. Inadequate sampling of performance

Inadequate observation and evaluation of a resident's performance across the range of tasks that constitute the scope of professional practice for physicians is by far the most significant obstacle to developing clear and accurate appraisals of a resident's performance capabilities. It is clear that direct observation of the resident's performance is very limited.¹ Studies on clinical performance evaluation have an average of fewer than ten ratings per resident or medical student. Further, observation is restricted to a narrow range of important clinical tasks. Stillman and her colleagues² found that 30% of internal medicine residents in 14 residency programs reported that they had never been observed performing a complete physical examination and 45% reported they had been observed one time, presumably in the mandated clinical evaluation exercise of the American Board of Internal Medicine. Based on their research, Pulito and colleagues³ concluded that surgery faculty members evaluate students' knowledge and clinical reasoning skills primarily via interactions during conferences and in informal conversations. Professional behavior is inferred based on students' interactions with the surgical team. Rarely are students observed interacting with patients. Data collection and clinical reasoning ability are inferred based on case presentations and patient write-ups. Faculty members also acquire some second-hand information about medical students from comments by residents. These investigators concluded that faculty have little basis for evaluating most other features of clinical performance. Schwind and colleagues⁴ found that surgeons primarily comment on residents' knowledge, clinical performance, work ethic, performance improvement, dependability, surgical and technical skills, and interpersonal and communication skills with other health professionals, suggesting that these are the behaviors most frequently observed. Other aspects of performance such as management and relations with patients were less noted.

What you can do:

Maximize the number of faculty ratings.

Different numbers of observations and evaluations are required to establish

stable estimates of competence for different clinical performance domains. In general, raters rate interpersonal and communication skills less reliably than they do clinical performance.⁵⁻⁷ Since all performance dimensions are measured using the same rating instrument, you must secure enough ratings to reliably rate the least stable performance area. There is also a great deal of variability in the number of ratings needed due to variations in the mix of residents being rated and the mix of attendings observing and rating each resident each year. The results reported by Williams and colleagues⁵ suggest that program directors should strive for 38 ratings per resident per year (approximately three per month) to assure a reproducible estimate of performance for all residents. This strategy serves the dual functions of increasing the number of situations and tasks observed (most important) and balancing the idiosyncrasies of raters by increasing the number of different raters evaluating the resident's performance. While it might be possible to decrease the needed number of ratings by systematically controlling the performances observed by faculty and the dimensions of performance rated, we believe that it will be easier for program directors to direct their attention to increasing the number of ratings obtained. Also since time is always in short supply, it is better to encourage attending physicians to spend less time observing each encounter in favor of observing a broader range of cases.

Include nurses and patients as raters.

Adding residents' ratings by nurses and patients broadens the spectrum of competencies measured. Not only do nurses see aspects of a resident's performance not seen by attending physicians, they also provide a glimpse of how trainees behave when not being observed by physicians. Patients' ratings provide insight regarding perceptions of the professional services residents provide. Obviously, including nurses and patients as raters also increases the absolute number of available ratings for each resident. This is not to say that nurses and patients should use the same rating form as physicians or that their ratings should be averaged with those of attending physicians. The idea in collecting nurses' and patients' evaluations is to broaden the sample of behaviors observed. Nurses and patients

report information about a resident's performance. Attending physicians incorporate this information with other information in making judgments about a resident's progress.

Sample cases broadly and systematically.

Clinical performance has proven to be highly case specific.^{8,9} As a result, trainees who are quite skilled with one type of case may perform poorly with another, even when it is closely related to the first. Different levels of performance are due to differences in the experience, training and personal disposition of the trainee. The implications are that performance needs to be assessed across the entire range of cases. These findings suggest that residency program directors need to pay more attention to arranging observation and evaluation of a resident's performance that more systematically spans the range of patient diagnoses and chief complaints than traditionally has been the case.

Ensure observation and evaluation of all aspects of performance.

There are certain important aspects of a resident's performance that are not commonly observed by attending physicians. For example, attending physicians rarely watch residents take a history and perform physical examinations on patients. Likewise, few attending physicians observe residents as they explain findings to patients, describe the next steps in the diagnostic work-up, or describe alternate approaches to management and obtain informed consent. Attending physicians may observe resident technical and procedural skills but typically do not complete the evaluation form at the time of observation. As a result much of the detail about performance is lost, thus decreasing the educational value and validity of the evaluation process. These problems can be minimized by prescribing the aspects of performance to be evaluated and by having some evaluation forms filled out immediately after observation of the performance.

Supplementing the traditional resident evaluation process with objective structured performance examinations can help achieve these goals. Three such examination procedures used in the general surgery residency at the Southern Illinois University (SIU) School of

Medicine include the Patient Assessment and Management Examination (PAME),¹⁰ the Objective Structured Assessment of Technical Skills (OSATS),¹⁰ and in-depth observations of sentinel cases using tailored rating forms for purposes of evaluating operating room performance on specific cases.¹¹

The PAME examination compresses time and allows the opportunity to observe and evaluate a resident in all stages of an encounter with a patient from initial examination through diagnostic work-up to recommending treatment. The observer is afforded an opportunity to observe the resident's ability to perform the history and physical examination, order diagnostic studies, interpret diagnostic study results to the patient, and to explain management options. In the oral examination portion of the examination, the observer has an opportunity to inquire about the rationale underlying the resident's decisions and actions.

The OSATS examination evaluates the resident's performance of selected technical skills in a systematic, objective manner.

Finally, the use of structured systematic evaluation of a resident's operating room performance for sentinel cases provides a more in-depth assessment of these procedural skills than normally occurs in the operating room. This approach has the added benefit of using rating forms tailored to the case, thus directing attention to important case specific performance attributes including both surgical skills and intraoperative decision making. Focusing on sentinel cases acknowledges the reality that residency programs cannot evaluate all of the residents all of the time. Rather, strategic decisions need to be made to concentrate faculty resources on observation and appraisal of selected activities of residents based on criteria such as frequency of procedures and frequency of associated morbidity and mortality. Use of the increasing numbers of computer simulations and realistic training mannequins¹² will also help increase the opportunities for systematic sampling of residents' performance across the range of cases and tasks that constitute the domain of clinical practice.

2. Inaccuracies of appraisal due to overreliance on memory

It is rare for attending physicians to fill out an evaluation form on a resident's performance at the time of the performance. Rather, these activities are separated by days, if not weeks. Since humans cannot store copious amounts of detailed information in memory, data reduction invariably occurs. As such, the attending physician forms and relies on a more global impression of the resident when it comes time to fill out the evaluation form. Under these conditions physician raters have repeatedly been observed to store a two-dimensional impression of residents' or peers' performance.^{13,14} They form and recall an impression about clinical performance and a separate impression about professional behavior. Any effort to extract a more detailed impression and assessment of performance, say by increasing the number of items on the rating scale, is futile. Specific details about a resident's performance will be forgotten. Second, attending physician recall will be selective. Dramatic events, positive and negative, will be recalled at the expense of a more balanced view. Likewise, observers tend to recall more recent events.

What you can do:

Evaluate performance and give feedback immediately. Attending physicians should be encouraged to evaluate and provide feedback to residents immediately after observing their performance. In this way both parties will recall the details of the situation. The attending physician can remind the resident of the situation, comment on what was done and on the consequences of the act (negative or positive), and suggest alternative behaviors the resident can consider if the observed behavior led to negative outcomes. Done right, these face-to-face dialogues will have more constructive impact on the resident than would written end-of-rotation feedback and will lead to future improvement in the resident's performance. The five-step microskills model of clinical teaching originally proposed by Neher and colleagues¹⁵ provides practical steps clinical faculty can use to improve feedback. This approach has been used extensively in faculty development and its effectiveness has been documented.^{16,17} Separating observation and evaluation for

purposes of supporting residents' learning will allow tuning the end-of-rotation evaluation process to decision making about the resident's progress, where a general impression of overall progress along with general identification of deficiencies is all that is needed.

Encourage immediate recording and evaluation of observations. The attending physician's use of notecards or a diary to record periodic observations of the actions of residents as they occur should result in more accurate recall of these actions. A number of approaches have been used to capture observations and judgments as they are made, including daily performance report cards in anesthesia^{18,19} and obstetrics and gynecology,²⁰ and diaries in business.²¹ These approaches may result in better performance appraisal for three reasons. First, the notecards provide more detail about performance than will be retained in memory. Second, since a record of each observation is jotted down, attending physicians may be more likely to record a negative evaluation, since it can be chalked up to a "bad day." (A pattern of "bad days" may then give a firm basis for remedial action to be taken.) Third, DeNisi and Peters²² demonstrated that those who keep such notes or diaries were better able to recall performance specifics at the time of evaluation, even if they were not allowed to refer to the notes for the rating.

Another way to link the performance observation and evaluation more closely is to introduce a supplementary evaluation process that calls for the faculty member to observe single encounters of the resident with patients and complete the evaluation form immediately afterward. The American Board of Internal Medicine has developed the mini-CEX examination for this purpose.²³ Attending physicians are asked to observe and evaluate selected resident-patient encounters on the spot. At the end of the rotation or year the residency faculty can then review a number of evaluations of these encounters along with the traditional global performance rating results. The SIU General Surgery program use of PAME, originally designed and described by MacRae and colleagues,¹⁰ and the operative procedure rating system¹¹ accomplishes the same purpose as the Mini-CEX by having the faculty member

observe a resident's performance and judge the performance immediately afterward.

Limit the number of items on the clinical performance rating form.

Although it is tempting to design a rating form that includes an item for each important attribute of performance, resist! Research results suggest that physician raters have a two-dimensional view of clinical performance.^{13,14} Raters form an impression of clinical performance and professional behavior. This conclusion has been documented consistently in the clinical performance assessment research literature. For more details and a more thorough description of this research see the review article by Williams and colleagues.¹ Recently, Williams and colleagues⁵ demonstrated that adding global rating items to a clinical performance rating form beyond one item per dimension (clinical performance, professional behavior) and one item comparing overall performance of the resident to that of the resident's peers yields virtually no new information. Increasing the number of items beyond this results in a rapid rate of diminishing returns. Faculty also are provided a list of component skills (e.g., technical/procedural skills, intraoperative decision making, dependability, relationship with other medical personnel) and asked to check those where residents have serious problems and those where the resident's performance is consistently outstanding. This approach provides an overall rating of performance and identifies specific strengths and weaknesses.

3. Hidden performance deficits

Clinical settings are busy environments with multiple health care professionals (attending physicians, consultants, fellows, senior residents, junior residents, nurses, medical students) contributing to care for each patient. In this environment the deficiencies of a resident may be masked by the multiple contributions that other team members make. This problem is exacerbated by the limited amount of direct performance observation by attending physicians and the fact that most performance aspects are judged on the basis of case presentations and patient chart entries that reflect the collective knowledge and skill of the health care team. In such a setting, it is difficult to get a clear picture

of a single resident's personal contributions to the care of the patient. Also, residents volunteer information only when they are confident that they know the correct answer. Such selective volunteerism gives a skewed view of a trainee's knowledge and competence level.

What you can do:

Use performance examinations. As stated previously, performance examinations offer many useful advantages when it comes to evaluating a resident's performance. The primary advantage is that the resident's performance reflects exclusively her or his knowledge and skill. There is no opportunity for the resident to solicit assistance from professional colleagues.

Question residents systematically. While not as effective as using performance examinations, attending physicians can partially overcome selective volunteerism and gain a clearer idea of the trainee's knowledge base by directing questions to residents on a systematic basis rather than taking answers from volunteers.

Systematically observe a wider range of activities. Systematic observation of a wider range of a resident's activities will reduce overdependence on those measures that reflect the combined knowledge and ability of the health care team (e.g., estimates based on case presentations and patient chart entries).

4. Lack of meaningful benchmarks

*Merriam-Webster's Collegiate Dictionary*²⁴ defines a benchmark as "a: a point of reference from which measurements may be made; b: something that serves as a standard by which others may be measured or judged; c: a standardized problem or test that serves as a basis for evaluation or comparison." A major problem in judging residents' clinical performance is the absence of useful benchmarks. For example, it is difficult to compare the performance of a resident to his or her peers because residents have all seen different patients with respect to diagnosis, acuity, severity of illness, and complexity of illness. This opportunistic exposure means that direct comparisons of residents who are working up the same case rarely occur in the natural clinical environment. Under these circumstances

it is easy for attending physicians to write off a resident's poor performance by assuming that most peers would perform in a similar manner. Alternatively, attending physicians may be too hard on a resident; assuming that other residents would do a better job with the patient.

Other factors that hamper benchmarking are the absence of well-organized, accessible performance-data norms to allow comparison of a resident's performance to that of other residents, to the resident's own earlier performance, and to the performance of earlier cohorts of residents. As regards data for comparing a resident's performance to that of others, ratings assigned are taken literally based on the descriptors on the scale. A rating of Good is interpreted as good. Decision makers often are not cognizant of the relative frequency of these ratings, and assume a large percentage of negative ratings. The reality is different. Williams and Dunnington²⁵ reported that 80% of all clinical performance ratings of residents were Excellent or Very Good. Only 17% of ratings were Good. Viewed in the light of these findings, a rating of Good is not a very positive statement about a resident's performance. Normative data such as this makes it easier to identify residents whose performance is outside of the normal range of ratings.

Even fewer residency programs maintain and use data that allows comparing a resident's current-year performance to that in prior years. Thus it is impossible to determine systematically whether the clinical performance of a fourth-year resident changed from their third-year performance. Likewise, it is difficult to determine whether the extra reading the resident was asked to do translated into better performance ratings. The work reported by Van der Vleuten and colleagues²⁶ and by Verhoeven and colleagues,²⁷ while not a direct application in a residency program setting, provides a model of how longitudinal performance data can be used to gain a richer understanding of residents' performance (both individuals and cohorts of residents) across years.

Finally, there is a notable absence of effort to systematically track the careers of graduates after they complete residency training. Historical data such as this would provide a better basis for

determining the prognosis for current residents with similar performance patterns and using this information for making progress decisions. The work by Papadakis and colleagues²⁸ is an example of what can be done in this regard. They found that medical students with a record of professional behavior problems in medical school were more likely to be sanctioned for similar problems in practice. The absence of detailed practice behavior records and outcomes data make this type of benchmarking very difficult today but we anticipate this data will be more available in the future, thus facilitating work that allows better forecasting of residents' performance in practice.

What you can do:

Resist changing the rating forms! When discomfort with the clinical performance rating process is expressed, whether it be complaints about grade inflation, concerns about nonspecificity of diagnoses based on ratings, or something else, the first tendency is to change the form. Invariably the forms get longer with each change. This process virtually never solves the expressed problem and it creates new problems related to benchmarking. Changing the form is like frequently changing the scale on a thermometer. The true meaning of ratings is only acquired over time and with experience. Familiarity with the form and past ratings patterns allows faculty members to recognize outliers. Each time the form is changed, the faculty must gradually reacquire a sense of the new meaning of ratings patterns.

Develop clinical performance rating norms. Attending physicians virtually never assign below average ratings to residents.^{4,25,29} For example, Williams and Dunnington²⁵ reported that five percent or less of all residents' ratings were in the Fair or Poor category. All residents who had 40% or more end-of-rotation ratings below Very Good had stipulations associated with their promotion at the end of the year. Therefore, as with diagnostic test results for patients, clinical performance rating results only acquire meaning with time and experience. A patient's body temperature of 39° centigrade has no meaning to a physician in the absence of systematic data regarding the immediate and long-term outcomes for patients

with these physiologic indicators. The meaning becomes clear only when systematic follow-up occurs and accurate, well-organized records are created and are easily accessible.

Carry out longitudinal analysis of data about residents with performance problems. Longitudinal performance data should be kept and analyzed on a regular basis, looking for predictors of future poor performance. For example, Papadakis et al.²⁸ found that medical students who were cited for professional behavior problems in medical school were twice as likely to be reported to the state board for professional behavior problems in practice. Likewise, Hojat and colleagues³⁰ found that a medical student whose performance was in the bottom quartile of clinical performance ratings during medical school was three times as likely to be in the bottom quartile of performance during residency than in the top quartile. Data similar to that provided by Papadakis et al. and Hojat et al. are more difficult to acquire for graduates of residency programs due to the unavailability or inaccessibility of practice performance data. However, such data are likely to be more easily available in the future and are the key to providing benchmarks for making valid progress decisions about future residents.

Use performance examinations. Performance examinations require each student or resident to work up and otherwise interact with the same patient, providing the opportunity for the direct comparison of their performance on standardized tasks. Anyone who has observed a series of residents work up the same standardized patient case (with real or simulated findings) can attest to the benefits this practice affords in interpreting the competence of a single resident to work up the case. Poor performance stands out. Use of standardized performance examinations provides many of the same benefits for analyzing clinical performance as in-training examinations afford to residency programs for measuring resident knowledge. Additionally, standardized patient encounters make it possible to create cases with gold standards. Evidence such as practice guidelines and data from randomized controlled trials or meta-analyses can be used to decide in advance what information should be elicited from the patient, what diagnoses

should be in the differential, what tests should be ordered, and what management options are optimal. The resident's performance can then be compared to this gold standard of performance.

5. Hesitancy to act

In a survey of medical schools regarding clinical performance evaluation problems, Tonesk and Buchanan³¹ found that two of the main problems were failures to act. First, individual faculty members were often unwilling to report suspected negative performance attributes of medical students. Second, the survey revealed that clerkship directors often didn't act on negative performance reports that were submitted. This failure to act was generally attributed to a perceived absence of sufficient data to make decisions such as dismissing a student or requiring the trainee to repeat a year, or because of fear of retaliation for such action by the trainee in the form of threatened lawsuits etc. Supporting results have been found in surgical residencies. Research by Schwind and colleagues⁴ documented that end-of-rotation ratings by individual faculty members often failed to detect or report general surgery residents' clinical performance deficiencies. In fact, 18% of residents, who at the end of the year were judged to have some performance deficiency that required formal remediation, did not receive a single post-rotation performance rating indicating that deficiency. Further, more post-rotation numeric ratings and written comments contradicted these performance deficits than supported them, especially in the areas of applied knowledge and professional behavior. These findings are consistent with behavioral research findings regarding rater reluctance to transmit bad news. In a review of this research, Tesser and Rosen³² concluded that good news is communicated more frequently, more quickly, and more fully than bad news. Further, they noted that this phenomenon has been observed across a wide range of situations in person-to-person and written communication.

Evidence of failure to act at the residency program level is limited and difficult to interpret due to absence of a "gold standard." Schueneman and colleagues³³ reported that only nine out of 310 surgery residents (3%) were asked to

leave their residency training programs over a period of 15 years. Harris and colleagues,³⁴ in a study of employee performance ratings in the business world, provide the best available evidence that ratings of employees may be inflated when the ratings will be shared with employees and/or may have consequences for them. This study demonstrated that 1% of employees were rated as needing to improve their performance when the ratings were shared with employees and could have consequences for pay raises and/or promotion. The same supervisors, using the same rating instruments, rated 7% of these same employees as needing to improve their performance when the ratings were not to be shared with employees and were guaranteed not to have consequences.

What you can do:

Seek only performance reports at the end of rotations. Do not ask individual faculty members to assign grades or submit resident-progress recommendations at the end of rotations. Rather, simply have them provide quality ratings about performance and provide written performance details to support these ratings. Schwind and colleagues demonstrated that fewer than 1% of post-rotation ratings over a five-year period of time noted a resident's performance deficits. However, during the end-of-the-year meeting, an average of 28% of residents were judged to have a deficiency that required some form of remediation. Martin and colleagues³⁵ also found that faculty ratings of students and residents were almost always overestimates of clinical performance when measured against a gold-standard panel.

Make progress decisions by committee. Work by Schwind and colleagues⁴ and by Williams and colleagues³⁶ suggests that making a resident's progress decisions by committee has several advantages. First progress decisions by committee offer a broader base of information and more perspectives for making a decision (i.e., they provide triangulation on a resident's performance). Second, committee discussion about a resident's performance allows for some calibration of lenient and stringent raters as a group consensus emerges. Third, individuals observing a resident's "bad day" but deciding not to put much stock in it (i.e.,

giving the resident the benefit of the doubt) may find that there has been a pattern of such days once a group discusses the resident's performance. A further benefit of progress decisions by committee is that the evaluation committee meeting provides physician attendings an opportunity to sit down and reflect on the resident's performance. A final benefit is that raising issues of poor performance in a group affords the individual faculty member with some anonymity and protection. The record of observations and judgments and the remediation recommendation comes from the committee rather than from individual faculty members. Some people have expressed concern that making progress decisions by committee will lead to invalid decisions due to the undue influence of single faculty members on the group. Williams and colleagues³⁶ found no evidence to suggest that the group decision making process is swayed by powerful or persuasive individuals in the group.

6. Systematic rater error

Contrary to the impression left by formal and informal comments in the clinical performance appraisal literature and during professional meetings, systematic rater errors are the least serious threat to accurate and reproducible estimates of clinical competence and the easiest to control. These threats are commonly enumerated as the halo effect (the rater's tendency to form a general impression of the resident and to give the same rating to all aspects of performance), central tendency error (the rater's avoidance of the most extreme rating categories on the scale), and stringency bias (a tendency for some raters to be lenient graders and others to be severe graders).

What you can do:

Increase the number of raters. There are two solutions to systematic rater problems. The first solution addresses the issue of stringency. We believe that the best solution to this problem resides in increasing the number of raters who rate each resident's performance. Littlefield and his colleagues³⁷ studied the ratings given by surgery faculty in five medical schools to determine the percentage of lenient and stringent raters. Approximately 13% of 107 raters were significantly more lenient than their colleagues and 14% were significantly

more stringent. However, the effects of lenient and stringent raters cancelled each other out when large numbers of raters (seven or more) rated each resident. It is tempting to "handicap" hawks and doves to normalize their ratings but this is probably not a good idea. Even more extreme ratings may result when the handicapped raters become aware of the process.

Familiarize raters with the evaluation form. There has been surprisingly little research on the effects of rater training when the raters are physicians. The studies that have been done^{38,39} are limited studies suggesting that training has little impact on the accuracy and reliability of physician raters. Much more work on the training of raters has been done in the organizational psychology and management research fields. Woehr and Huffcutt⁴⁰ summarized the results of rater training studies and concluded that training designed to familiarize raters with the dimensions of performance to be rated and with the standards of performance associated with each level on the scale through use of examples proved to be the single most effective training strategy for increasing accuracy of rating and observation. This training also decreased halo errors and leniency errors to some extent. We suspect that one reason why research on training of raters has not been common in the medical and surgical education area is that it is very difficult to get physicians and surgeons to attend these training sessions. For this reason we urge readers to rely more on increasing the number of raters to balance out leniency and stringency. Brief training efforts designed to familiarize raters with the form and to control for halo errors and central tendency errors may be worth trying. These could be incorporated on the form itself.

Summary

In this article we have focused on six factors that compromise the process of observing, evaluating, and characterizing current performance of residents, forecasting future performance, and making residents' progress decisions accordingly.

- First and foremost, the amount of observation that occurs and the range of tasks observed are inadequate as a

basis for judging a resident's ability to handle the range of tasks that constitute clinical practice.

- Second, faculty ratings are normally separated in time from the observations that support them. This separation leads to distortions associated with forgetting and selective recall.
- Third, a resident's performance deficits are often hidden from view due to the collective nature of contributions to patient care by the health care team. Thus, a resident's case presentations and chart entries can, on occasion, lead to false impressions about the resident's clinical competence.
- Fourth, the current system lacks good benchmarks against which to compare a resident's performance. Accessible benchmarks are lacking for comparing a resident's performance to that of peers, and to the resident's own performance in prior years. Further, follow-up data are not available that indicate how past residents with similar residency performance patterns have performed in practice. This is analogous to not knowing how patients, treated with a particular therapy, fared after treatment as a basis for deciding whether to use that therapy in the future.
- Fifth, we noted that there is hesitancy on the part of faculty to deliver bad news about a resident's performance. There are two related reasons for this hesitancy. Faculty members and program directors are aware of the softness of the available performance evaluation data and are acutely aware of the serious consequences for residents of their progress decisions. We believe that making progress decisions by committee improves the scope and quality of the performance information available about residents and helps increase faculty confidence in the information and the decisions they make.
- The sixth factor that compromises the quality of evaluating residents' performance information is systematic rater error (halo effect, central tendency error, and differences in rater stringency). However, these effects are small and are relatively easily controlled with minimal rater training and by increasing the number of ratings to balance rater idiosyncrasies.

The suggestions we offer for controlling these factors should lead to more accurate forecasting of a resident's ability to perform across the range of situations and tasks that constitute clinical practice. This, in turn, should lead to fairer decisions about a resident's progress and increased faculty confidence in making negative decisions about that progress (diagnosing and managing the resident's progress) when merited. Following these guidelines also will aid programs in meeting the legal requirements of due process that pertain to making decisions about a resident's progress.⁴¹

References

- 1 Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical competence ratings. *Teach Learn Med.* 2003;15:270–92.
- 2 Stillman PL, Swanson DB, Smees S, et al. Assessing clinical skills of residents with standardized patients. *Ann Intern Med.* 1986; 105:762–71.
- 3 Pulito AR, Donnelly MB, Plymale M, Mentzer RM. What do faculty observe of medical students' clinical performance? Paper presented at Association for Surgical Education Meetings, Vancouver, BC, Canada, 2003.
- 4 Schwind CJ, Williams RG, Boehler ML, Dunnington GL. Do individual attending post-rotation performance ratings detect resident clinical performance deficiencies? *Acad Med.* 2004;79:453–7.
- 5 Williams RG, Verhulst SJ, Colliver JA, Dunnington GL. Assuring the reliability of resident performance appraisals: more items or more observations? *Surgery.* 2005;137:141–7.
- 6 Viswesvaran C, Ones DS, Schmidt FL. Comparative analysis of the reliability of job performance ratings. *J Appl Psychol.* 1996;81: 557–74.
- 7 Conway JM, Huffcutt AI. Psychometric properties of multisource performance ratings: a meta-analysis of subordinate, supervisor, peer, and self-ratings. *Hum Perform.* 1997;10:331–60.
- 8 Carline JD, Wenrich M, Ramsey PG. Characteristics of ratings of physician competence by professional associates. *Eval Health Prof.* 1989;12:409–23.
- 9 Elstein AS, Shulman LS, Sprafka SA. Medical problem-solving. *J Med Educ.* 1981;56:75–6.
- 10 MacRae H, Regehr G, Leadbetter W, Reznick RK. A comprehensive examination for senior surgical residents. *Am J Surg.* 2000;179:190–3.
- 11 Larson JL, Williams RG, Ketchum J, Boehler MA, Dunnington GL. Feasibility, reliability and validity of an operative performance rating system for evaluating general surgery residents. Paper presented at the 62nd annual meeting of the Central Surgical Association, Tucson, AZ, March 11, 2005.
- 12 Issenberg SB, McGaghie WC, Hart IR, et al. Simulation technology for health care professional skills training and assessment. *JAMA.* 1999;282:861–6.
- 13 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA.* 1993;269:1655–60.
- 14 Verhulst SJ, Colliver JA, Paiva RE, Williams RG. A factor analysis study of performance of first-year residents. *J Med Educ.* 1986;61: 132–4.
- 15 Neher JO, Gordon KC, Meyer B, Stevens N. A five-step "microskills" model of clinical teaching. *J Am Board Fam Pract.* 1992;5:419–24.
- 16 Salerno SM, O'Malley PG, Pangaro LN, Wheeler GA, Moores LK, Jackson JL. Faculty development seminars based on the one-minute preceptor improve feedback in the ambulatory setting. *J Gen Intern Med.* 2002; 17:779–87.
- 17 Furney SL, Orsini AN, Orsetti KE, Stern DT, Gruppen LD, Irby DM. Teaching the one-minute preceptor. A randomized controlled trial. *J Gen Intern Med.* 2001;16:620–4.
- 18 Rhoton MF. A new method to evaluate clinical performance and critical incidents in anaesthesia: quantification of daily comments by teachers. *Med Educ.* 1990;24: 280–9.
- 19 Rhoton MF, Furgerson CL, Cascorbi HF. Evaluating clinical competence in anesthesia: using faculty comments to develop criteria. *Proc Annu Conf Res Med Educ.* 1986;25: 57–62.
- 20 Brennan BG, Norman GR. Use of encounter cards for evaluation of residents in obstetrics. *Acad Med.* 1997;72(10 suppl 1):S43–4.
- 21 Flanagan J, Burns R. The employee performance record: A new appraisal and development tool. *Harv Bus Rev.* 1955;33:95–102.
- 22 DeNisi AS, Peters LH. Organization of information in memory and the performance appraisal process: evidence from the field. *J Appl Psychol.* 1996;81:717–37.
- 23 Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med.* 2003;138: 476–81.
- 24 Mish FC (ed). *Merriam-Webster's Collegiate Dictionary*, 11th ed. Springfield, MA: Merriam-Webster, Inc., 2003.
- 25 Williams RG, Dunnington G. The prognostic value of resident clinical performance ratings. *J Am Coll Surg.* 2004;199:620–7.
- 26 Van der Vleuten C, Verwijnen GM, Wijnen W. Fifteen years of experience with progress testing in a problem-based curriculum. *Med Teach.* 1996;18:103–9.
- 27 Verhoeven BH, Verwijnen GM, Scherpbier AJ, van der Vleuten CP. Growth of medical knowledge. *Med Educ.* 2002;36:711–7.
- 28 Papadakis MA, Hodgson CS, Teherani A, Kohatsu ND. Unprofessional behavior in medical school is associated with subsequent disciplinary action by a state medical board. *Acad Med.* 2004;79:244–9.

- 29 Thompson WG, Lipkin M Jr, Gilbert DA, Guzzo RA, Roberson L. Evaluating evaluation: assessment of the American Board of Internal Medicine Resident Evaluation Form. *J Gen Intern Med.* 1990;5:214–7.
- 30 Hojat M, Gonnella JS, Veloski JJ, Erdmann JB. Is the glass half full or half empty? A reexamination of the associations between assessment measures during medical school and clinical competence after graduation. *Acad Med.* 1993;68(2 suppl):S69–76.
- 31 Tonesk X, Buchanan RG. An AAMC pilot study by 10 medical schools of clinical evaluation of students. *J Med Educ.* 1987;62:707–18.
- 32 Tesser A, Rosen S. The reluctance to transmit bad news. In: Berkowitz L (ed). *Advances in Experimental Social Psychology*, vol 8. New York: Academic Press, 1975.p. 193–232.
- 33 Scheuneman AL, Carley JP, Baker WH. Residency evaluations. Are they worth the effort? *Arch Surg.* 1994;129:1067–73.
- 34 Harris MM, Smith DE, Champagne D. A field study of performance appraisal purpose: Research- versus administrative-based ratings. *Personnel Psychol.* 1995;48:151–60.
- 35 Martin JA, Reznick RK, Rothman A, Tamblyn RM, Regehr G. Who should rate candidates in an objective structured clinical examination? *Acad Med.* 1996;71:170–5.
- 36 Williams RG, Schwind CJ, Dunnington GL, Fortune J, Rogers DA, Boehler ML. The effects of group dynamics on resident progress committee deliberations. *Teach Learn Med.* 2005;17:96–100.
- 37 Littlefield JH, DaRosa DA, Anderson KD, Bell RM, Nicholas GG, Wolfson PJ. Accuracy of surgery clerkship performance raters. *Acad Med.* 1991;66(9 suppl):S16–8.
- 38 Newble DI, Hoare J, Sheldrake PF. The selection and training of examiners for clinical examinations. *Med Educ.* 1980;14:345–9.
- 39 Noel GL, Herbers JE Jr, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med.* 1992;117:757–65.
- 40 Woehr DJ, Huffcutt AI. Rater training for performance-appraisal: a quantitative review. *J Occup Organ Psychol.* 1994;67:189–205.
- 41 Irby DM, Milam S. The legal context for evaluating and dismissing medical students and residents. *Acad Med.* 1989;64:639–43.